A. N. Gorban
N. Kazantzis
I. G. Kevrekidis
H. C. Öttinger
C. Theodoropoulos
(Eds.)

Springer: COMPLEXITY

# Model Reduction and Coarse-Graining Approaches for Multiscale Phenomena

Springer

Gorban · Kazantzis · Kevrekidis · Öttinger · Theodoropoulos (Eds.)

# Model Reduction and Coarse-Graining Approaches for Multiscale Phenomena

Alexander N.Gorban · Nikolaos K. Kazantzis
Ioannis G. Kevrekidis · Hans Christian Öttinger
Constantinos Theodoropoulos (Eds.)

# Model Reduction and Coarse-Graining Approaches for Multiscale Phenomena

With 50 Figures

Springer

Alexander N.Gorban
University of Leicester
Department of Mathematics
University Road
LE1 7RH Leicester
United Kingdom
*e-mail: ag153@leicester.ac.uk*
and
Institute of Computational Modeling
Russian Academy of Sciences

Ioannis G. Kevrekidis
Princeton University
Department of Chemical Engineering
Engineering Quadrangle A-217
Princeton NJ 08544-5263
USA
*e-mail: yannis@princeton.edu*

Constantinos Theodoropoulos
University of Manchester
School of Chemical Engineering
and Analytical Science
PO Box 88, Sackville St.
Manchester, M60 1QD
United Kingdom
*e-mail: K.Theodoropoulos@manchester.ac.uk*

Nikolaos K. Kazantzis
Worcester Polytechnic Institute
Department of Chemical Engineering
Institute Road 100
Worcester, MA 01609-2280
USA
*e-mail: nikolas@wpi.edu*

Hans Christian Öttinger
ETH Zürich
Institut für Polymere
Wolfgang Pauli-Straße 10
CH-8093 Zürich
Switzerland
*e-mail: hco@mat.ethz.ch*

Library of Congress Control Number: 2006929855

ISBN-10 3-540-35885-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-35885-5 Springer Berlin Heidelberg New York



Typesetting: Data conversion by authors
Production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper 57/3100/YL 5 4 3 2 1 0

# Preface

Even a cursory inspection of the content of the well-known on-line free encyclopedia Wikipedia reveals a simple classification and model typology that is frequently encountered in a wide spectrum of scientific disciplines. In particular, different types of models have traditionally been classified according to the following well-known categorization criteria [1]:

1. Linear vs. nonlinear: If the objective functions and constraints are represented entirely by linear equations, then the model is known as a linear model. If one or more of the objective functions or constraints are represented with a nonlinear equation, then the model is known as a nonlinear model.
2. Deterministic vs. probabilistic (stochastic): A deterministic model performs the same way for a given set of initial conditions, while in a stochastic model, randomness is present, even when given an identical set of initial conditions.
3. Static vs. dynamic: A static model does not account for the element of time, while a dynamic model does. Dynamic models typically are represented with difference equations or differential equations.
4. Lumped parameters vs. distributed parameters: If the model is homogeneous (consistent state throughout the entire system) the parameters are lumped. If the model is heterogeneous (varying state within the system), then the parameters are distributed. Distributed parameters are typically represented with partial differential equations.

The above point of view of model classification and typology, while elementary, remains methodologically important and educationally quite useful. However, these elementary prototype models bear the same relation to modern science and technology, as an elementary wheel-drive and gear-box does to the level of sophistication of a modern car; at a higher level of integration of the various system components, complexity and structural integrity naturally emerge, calling for a new paradigm in systems modeling. Indeed, in order to develop a working model of a realistic and complex physical or engineering

system (or process) we need a "model factory" with a new "technology" for multi-level model construction. The realization of the above ambitious goal will probably follow the methodological path that has served the scientific community rather well over the years: first, elementary fundamental models will continue to be generated with an increasing degree of scientific accuracy and conformity to the fundamental laws of physics, chemistry and biology. On the basis of such elementary models, researchers will continue to build models at the "second level of descriptive power", such as various legacy codes and computational codes of elementary processes and systems. However, confronted with the challenges of complexity inherent in real systems and processes members of the scientific community are constantly motivated to develop models at an even higher level of descriptive power and accuracy, possibly through a smart combination/utilization of the models (building blocks) developed at lower levels of modeling. The procedural ascent to higher levels of modeling accuracy and complexity will continue to be necessitated by the need for the pursuit of scientific and technological breakthroughs, being limited only by the inevitable intellectual and technical capacity constraints. It is envisaged that the above intellectual and research efforts will eventually define and characterize a new scientific discipline that could be named "Model Engineering" [2].

In the present volume, two main thematic directions in the development of this newly emerging discipline are traced, namely Model Reduction and Invariance, as well as Coarse-Graining. For dynamical models describing the behavior of large-scale complex systems, one of the most powerful and rigorous approaches to model reduction is based on the notion of the system's slow invariant manifold. The theory of invariant manifolds was introduced more than a century ago through the work of two legendary figures of mathematics, Lyapunov and Poincaré [3, 4]. It experienced intense development during the 20th century and is currently being vigorously revisited and reexamined as an important and powerful tool in applied mathematics used for mathematical modeling and model reduction purposes. Coarse-Graining is also a one-hundred-year-old idea. Its first appearance in the physics community occurred through the seminal work of the Ehrenfests [5] (but the role of Boltzmann, Gibbs and Einstein was also important) and, moreover, further development of the original ideas in the 20th century led to explicit and transparent connections to all branches of statistical physics, kinetics and thermodynamics. Even though these insightful connections remain quite popular today [6], Coarse-Graining has evolved further, now reaching a much broader field of applications, and becoming an important universal tool for modeling.

It should be pointed out that the problem of multiscale modeling and the physically meaningful "coupling" of models of different levels poses essential difficulties and challenges in model construction. Indeed, non-elementary models are always multiscale ones. Recently, however, a notable scientific breakthrough has occurred in the form of the so-called "equation-free approach," which aims to address some of the above challenges by systematically facili-

tating the development and guiding the integration of models of a higher level of descriptive power with given legacy codes and other computational models at lower levels of modeling [7].

Most of the contribution to the present volume are based on selected talks/presentations given at the workshop entitled "Model Reduction and Coarse-Graining Approaches for Multiscale Phenomena" at the University of Leicester, Leicester, UK, August 24-26, 2005.[1]

The theme of the workshop was deliberately broad in scope and aimed at promoting an informal exchange of new ideas and fresh methodological perspectives in the increasingly important area of Model Reduction and Coarse-Graining for multiscale phenomena.

The main thematic areas of the workshop, which were structured around recently developed theoretical and computational approaches, were:

1. Invariance and model reduction (invariant manifolds for ODEs and PDEs, perturbation theory and applications of new model reduction techniques);
2. Coarse-graining approaches;
3. Accuracy estimation and post-processing algorithms.

Specific areas of study represented at the workshop included dynamical systems, non-equilibrium statistical mechanics, kinetic theory, hydrodynamics and mechanics of continuous media, (bio)chemical kinetics, particulate systems, nonlinear dynamics, nonlinear control and nonlinear estimation.

The goals of this initiative were to assemble a group of people with a wide variety of expertise reflecting the thematically interdisciplinary nature of the workshop, to organize a series of presentations and to encourage discussions in an informal, casual and "interactive" format that fostered and facilitated a fruitful dialogue across disciplines.

It was strongly felt by all participants that the generic nature and power of the pertinent conceptual, analytical and computational frameworks helped eliminate some of the traditional language barriers that, unnecessarily sometimes, impede scientific cooperation, development of a dialogue, as well as interaction among researchers across disciplinary boundaries between physics, chemistry, biology, applied mathematics and engineering.

Motivated by the excellent response, enthusiasm and level of participation, we strongly believe that this book will help not only to disseminate some of the new knowledge and research experience already accumulated in the emerging field of Model Engineering, but most importantly, to encourage other people who would like to study and further develop it in a fruitful dialogue and cooperation.

[1] The workshop was financially supported by EPSRC and LMS, and the authors gratefully acknowledge this support.

## References

1. Wikipedia, `http://en.wikipedia.org/wiki/Mathematical_model`
2. A.N. Gorban, I.V. Karlin: *Invariant Manifolds for Physical and Chemical Kinetics*, Lect. Notes Phys., vol. 660 (Springer, Berlin Heidelberg New York 2005)
3. A.M. Lyapunov: *The General Problem of the Stability of Motion* (Taylor & Francis, London 1992)
4. H. Poincaré: *Les Méthodes Nouvelles de la Mécanique Céleste*, vols. 1–3 (Gauthier–Villars, Paris 1892/1893/1899).
5. P. Ehrenfest, T. Ehrenfest-Afanasyeva: The Conceptual Foundations of the Statistical Approach in Mechanics. In: *Mechanics Enziklopädie der Mathematischen Wissenschaften*, vol. 4. (Leipzig 1911) (Reprinted: P. Ehrenfest, T. Ehrenfest-Afanasyeva: *The Conceptual Foundations of the Statistical Approach in Mechanics* (Dover Phoneix, 2002))
6. H.C. Öttinger. *Beyond Equilibrium Thermodynamics* (Wiley, Hoboken 2005)
7. I.G. Kevrekidis, C.W. Gear, J.M. Hyman, P.G. Kevrekidis, O. Runborg, K. Theodoropoulos: Equation-free coarse-grained multiscale computation: enabling microscopic simulators to perform system-level tasks. *Comm. Math. Sciences* **1**, 715-762 (2003)

Leicester — *Alexander N. Gorban*
Worcester — *Nikolaos Kazantzis*
Princeton — *Ioannis G. Kevrekidis*
Zürich — *Hans Christian Öttinger*
Manchester — *Constantinos Theodoropoulos*
April, 2006

# Contents

# Part I

## Computation of Invariant Manifolds

# A New Model Reduction Method for Nonlinear Dynamical Systems Using Singular PDE Theory

N. Kazantzis[1] and C. Kravaris[2]

[1] Department of Chemical Engineering, Worcester Polytechnic Institute, Worcester, MA 01609, USA, nikolas@wpi.edu
[2] Department of Chemical Engineering, University of Patras, Patras GR-2000, Greece, kravaris@upatras.gr

**Summary.** In the present research study a new approach to the problem of model-reduction for nonlinear dynamical systems is proposed. The formulation of the problem is conveniently realized through a system of singular quasi-linear invariance PDEs, and an explicit set of conditions for solvability is derived. In particular, within the class of real analytic solutions, the aforementioned set of conditions is shown to guarantee the existence and uniqueness of a locally analytic solution, which is then proven to represent the slow invariant manifold of the nonlinear dynamical system under consideration. As a result, an exact reduced-order model for the nonlinear system dynamics is obtained through the restriction of the original system dynamics on the aforementioned slow manifold. The local analyticity property of the solution's graph that corresponds to the system's slow invariant manifold enables the development of a series solution method, which allows the polynomial approximation of the "slow" system dynamics on the slow manifold up to the desired degree of accuracy.

## 1 Introduction

The natural world is dominated by physical and chemical processes that exhibit nonlinear behavior and are typically modeled by systems of nonlinear ordinary (ODEs) or partial differential equations (PDEs) [3, 14, 30]. Despite the fact that the dynamic behavior of linear systems can be mathematically analyzed and insightfully characterized with rigor and elegance [1, 3, 14, 15, 30], it still represents a rather challenging task for nonlinear systems and unboubtedly induces considerable research effort. Among the most notable research objectives in nonlinear systems analysis is certainly the existence of invariant manifolds and the associated problem of finding/computing them [1, 3, 14, 15, 30]. In particular, the problem under consideration has been traditionally motivated by efforts to develop systematic methods for the simplification of the analysis of the behavior of nonlinear dynamical systems through

an effective reduction of the dimensionality of the original problem, and the explicit computation of a reduced-order, yet accurate, description of the system dynamics [2, 5, 7, 8, 12, 13, 14, 17, 18, 19, 22, 24, 25, 26, 27, 28, 29, 31, 32] Two distinct categories of available approaches in the literature rely either on the classical quasi-steady-state (QSS) approximation method and certain variants, or on methods and results from singular perturbation (SP) theory [3, 14, 20, 21, 30, 31]. Notice however, that in both cases, appropriate a priori information is needed for their practical application. Indeed, the QSS method presupposes the explicit physical identification of the system's "fast" state variables, whereas the standard SP approach presupposes the explicit physical identification of a function of the system's parameters which is considered to be "small" (in a certain sense), and its "smallness" is responsible for the underlying time-scale multiplicity or the manifestation of a distinct spectral gap. Please notice that in addition to relying on the above a priori knowledge, both QSS and SP methods are inherently inexact, in the sense that they do not follow exactly the system's slow invariant manifold, thus resulting in long-term inaccuracies in the dynamic response of the reduced-order system/model. On the other hand, a mathematically meaningful and rigorous treatment of the model-reduction problem for nonlinear dynamical systems has to rely on the explicit computation/construction of the system's exact slow invariant manifold, and this is certainly non-trivial [1, 13, 14, 30]. Within the above framework however, the restriction of the system dynamics on the slow invariant manifold results in a reduced-order description of the system dynamics which is exact, in the sense that it generates the actual system trajectory on the slow manifold once the fast transients die out and the system crosses the above manifold (upon which it is bound to evolve for all future times).

The present research study proposes a new systematic approach to the problem of explicitly calculating the system's slow invariant manifold and constructing an exact reduced-order model for the nonlinear system dynamics. The latter represents the restriction of the original system dynamics on the aforementioned slow manifold. From a mathematical standpoint, the above objective is attained by focusing on the study of the invariance PDE and the derivation of a specific set of conditions that ensure the existence and uniqueness of a solution that correspond's exactly to the system's slow manifold.

The present paper is organized as follows: Section 2 contains some mathematical preliminaries that are necessary for the ensuing theoretical developments. The paper's main results are presented in Section 3, accompanied by remarks and comments on the use of the proposed approach and method for model-reduction purposes of nonlinear dynamical systems. Finally, a few concluding remarks are provided in Section 5.

## 2 Mathematical Preliminaries

A nonlinear dynamical system is considered:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x) \tag{1}$$

where $x \in R^n$ is the state vector. It is assumed that $f(x)$ is a real analytic vector function, and without loss of generality, let the origin $x^0 = 0$ be an equilibrium point of (1): $f(0) = 0$. Furthermore, it is assumed that the Jacobian matrix $A = \dfrac{\partial f}{\partial x}(0)$ is Hurwitz (having eigenvalues with negative real parts), and specifically, its eigenspectrum $\sigma(A)$ consists of two distinct subsets of the "fast" eigenvalues $\sigma_f(A)$ and the "slow" eigenvalues $\sigma_s(A)$: $\sigma(A) = \sigma_f(A) \cup \sigma_s(A)$. It is implicitly assumed that the real parts of the "fast" eigenvalues are a few orders of magnitude larger than the real parts (in absolute value) of the "slow" eigenvalues. Under the above assumptions and within the context of model reduction, the primary objective of the present study is the explicit construction of the system's slow manifold and the associated reduced order dynamic system that represents the restriction of the flow of (1) on the aforementioned slow manifold (thus effectively circumventing the effect of the fast dynamic modes).

The following definition is essential for the ensuing theoretical developments.

**Definition 1 [1, 30]**: A set

$$\Omega = \{x \in R^n | \phi(x) = 0\} \tag{2}$$

where $\phi : R^n \to R^m$ is a map with $\phi(0) = 0$, is said to be invariant under the flow of dynamics (1), if for each $\phi(x(0)) \in \Omega$, the integral curve $\{x(t\}$ of (1) satisfying $x(t = 0) = x(0)$, is such that $\phi(x(t)) \in \Omega$ for all $t \in R^+$. An invariant set $\Omega \subset R^n$ passing through the origin $x^0 = 0$ is said to be a real analytic local invariant manifold, if $\phi$ is real analytic and $\Omega$ has the local topological structure of an analytic manifold around the origin.

It follows easily that for $\Omega$ to be rendered invariant under the flow of (1), the map $\phi$ needs to satisfy the following invariance PDE:

$$\frac{\partial \phi}{\partial x}(x) f(x) = 0 \tag{3}$$

Notice, that the above invariance PDE condition is satisfied by all possible invariant manifolds of dynamics (1), and therefore, it admits multiple solutions. The key issue that the present study aims at addressing, is the development of a systematic method that allows the specific construction of the system's slow manifold out of the above multitude of invariant manifolds. Equivalently, a method that allows the explicit mathematical characterization of the system's motion that corresponds to the "slow" eigenmodes, as it evolves on the

slow manifold embedded in state space. Consequently, the restriction of the system dynamics (1) on the above slow manifold represents a reduced-order description of the original nonlinear dynamics (1).

## 3 Main Results

Before embarking on the presentation of the present study's main results, it would be methodologically appropriate to first examine the application of the proposed ideas and methods to linear systems, thus paving the way for the development of the proposed method for nonlinear dynamical systems.

Consider a linear dynamical system:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = Ax \tag{4}$$

where $A$ is a constant matrix of appropriate dimensions whose eigenspectrum satisfies the assumptions stated in the previous section. The invariance condition (3) for a linear manifold:

$$\Omega = \{x \in R^n | \Phi x = 0\} \tag{5}$$

to be rendered invariant under the flow of (4) becomes:

$$\Phi A x = 0 \tag{6}$$

for all $x \in \Omega$, where $\Phi$ is a constant matrix. In order to explicitly compute the particular invariant manifold that corresponds to the system's slow manifold, a standard linear coordinate transformation is employed that can transform the original system (4) into the following block-triangular form [30]:

$$\begin{aligned} \frac{\mathrm{d}x_s}{\mathrm{d}t} &= A_s x_s \\ \frac{\mathrm{d}x_f}{\mathrm{d}t} &= A_{fs} x_s + A_f x_f \end{aligned} \tag{7}$$

where $x_f, x_s$ are the "fast" and "slow" state vectors respectively, with $\sigma(A_s)$ and $\sigma(A_f)$ being exactly the "fast" and "slow" eigenspectra, i.e the sets of "fast" and "slow" eigenmodes of system (4). One can easily show that:

$$\Omega = \{(x_f, x_s) \in R^n | x_f - T x_s = 0\} \tag{8}$$

where matrix $T$ is the unique solution to the Lyapunov-Sylvester equation [9]:

$$T A_s - A_f T = A_{fs} \tag{9}$$

represents the requested slow manifold. Indeed, let: $\Phi = [-T|I]$ and $A = \begin{bmatrix} A_s & 0 \\ A_{fs} & A_f \end{bmatrix}$. Then:

$$\Phi A = [-TA_s + A_{fs}|A_f] = [-A_f T|A_f] = A_f \Phi \tag{10}$$

and $\Phi Ax = A_f \Phi x = 0$ for all $x \in \Omega$. Therefore, $\Omega$ is an invariant manifold for system (7). Furthermore, consider the "off-the-manifold" coordinate: $z = x_f - Tx_s$, which evolves as follows:

$$\begin{aligned} \frac{\mathrm{d}z}{\mathrm{d}t} &= A_{fs}x_s + A_f x_f - TA_s x_s = A_f x_f - A_f T x_s = \\ &= A_f z \end{aligned} \tag{11}$$

The above equation shows that the "off-the-manifold" coordinate $z$ decays according to the system's "fast" eigenvalues, and therefore, $\Omega$ represents the requested slow manifold.

Let us now examine how the above ideas can be generalized to account for nonlinear dynamical systems. First, a special class of nonlinear systems will be considered, namely systems that exhibit the exact triangular structure shown below:

$$\begin{aligned} \frac{\mathrm{d}x_s}{\mathrm{d}t} &= F_s(x_s) \\ \frac{\mathrm{d}x_f}{\mathrm{d}t} &= F_f(x_s, x_f) \end{aligned} \tag{12}$$

where the first dynamic equation describes the "slow" motion and the second the "fast" one. Notice that the second dynamic equation may correspond to a process whose own dynamics is driven by:

(i) either a "slowly" varying input/disturbance dynamics mathematically realized by the first dynamic equation (where input or disturbance changes are modeled and generated as "outputs" of the autonomous nonlinear dynamics of the first equation) [19], or

(ii) a time-varying process parameter vector $x_s(t)$ that follows the "slow" dynamics of the first equation and models phenomena such as catalyst deactivation, enzymatic thermal deactivation, heat-transfer coefficient changes, time-varying (bio)chemical kinetic parameters, etc. [19], or

(iii) by an upstream nonlinear process with slow dynamics modeled through the first dynamic equation in (12). It is useful to remind the reader, that as indicated in the previous section, $F_f(x_s, w_f)$ and $F_s(x_s)$ are assumed to be real analytic vector functions with: $F_f(0, 0) = 0$ and $F_s(0) = 0$.

For system (12), one can easily show that:

$$\Omega = \{(x_f, x_s) \in R^n | x_f - \pi(x_s) = 0\} \tag{13}$$

represents an invariant manifold, if the map $\pi$ satisfies the invariance PDE shown below:

$$\frac{\partial \pi}{\partial x_s} F_s(x_s) = F_f(x_s, \pi(x_s)) \tag{14}$$

Notice that the above system of first-order quasi-linear PDEs has a common principal part [4, 6], which consists of the components of the vector function

$F_s(x_s)$. Moreover, the origin is a **characteristic point** for the above system of PDEs (14), since the principal part vanishes at $(x_s, x_f) = (0, 0)$ (due to the equilibrium condition) [4, 6]. Therefore, the above system of PDEs (14) becomes **singular**. Notice, that in order to solve the above system of PDEs (14) in a neighborhood of the characteristic point $(x_s, x_f) = (0, 0)$, the existence and uniqueness conditions of the Cauchy-Kovalevskaya theorem are not satisfied and the theorem can not be applied [4, 6]. However, for the specific structure of the above system of singular invariance PDEs (14), **Lyapunov's auxiliary theorem** [23] can be employed to guarantee the existence and uniqueness of a locally analytic solution.

**Lyapunov's Auxiliary Theorem [23].** *Consider the following first-order system of quasi-linear partial differential equations:*

$$\frac{\partial w}{\partial x}\phi(x, w) = \psi(x, w) \tag{15}$$

*where:* $w : R^m \longrightarrow R^p$ *is the unknown vector function of (15), and* $\phi(x, w) : R^m \times R^p \longrightarrow R^m$, $\psi(x, w) : R^m \times R^p \longrightarrow R^p$ *are given analytic vector functions which satisfy the following conditions:*

$$\begin{aligned} \phi(0, 0) &= 0 \\ \psi(0, 0) &= 0 \\ \frac{\partial \phi}{\partial w}(0, 0) &= 0 \end{aligned} \tag{16}$$

*It is assumed that the eigenvalues* $k_i, (i = 1, ..., m)$ *of the* $m \times m$ *matrix* $\dfrac{\partial \phi}{\partial x}(0, 0)$ *satisfy the following condition:*

$$0 \notin CH\{k_1, k_2, ..., k_m\} \tag{17}$$

*and are not related to the eigenvalues* $\lambda_i, (i = 1, ..., p)$ *of the* $p \times p$ *matrix* $\dfrac{\partial \psi}{\partial w}(0, 0)$ *through any equation of the type:*

$$\sum_{i=1}^{m} m_i k_i = \lambda_j \tag{18}$$

$(j = 1, ..., p)$*, where all the* $m_i$ *are non-negative integers that satisfy the condition:*

$$\sum_{i=1}^{m} m_i > 0 \tag{19}$$

*Then, the above first-order system of PDEs (15), with initial condition* $w(0) = 0$*, admits a unique analytic solution* $w$ *in a neighborhood of* $x = 0$.

Using Lyapunov's auxiliary theorem one arrives at the following result [19]:

**Theorem 1.** *Consider the nonlinear dynamic system (12) and let all the aforementioned assumptions hold true. Moreover, assume that the eigenvalues $k_i$ of matrix $A_s = \dfrac{\partial F_s}{\partial x_s}(0)$ are not related to the eigenvalues $\lambda_i$ of matrix $A_f = \dfrac{\partial F_f}{\partial x_f}(0,0)$ through any equations of the type (18,19). Then the set $\Omega$ (13) is a real analytic invariant manifold of (12), where $\pi(x_s)$ is the unique solution of the singular invariance PDE (14).*

**Remark 1**: Let us now consider the linear case, where: $\phi(x,w) = A_s x_s$ and $\psi(x,w) = A_{fs}x_s + A_f x_f$, with $A_s, A_f, A_{fs}$ being constant matrices with appropriate dimensions. Then, the unique solution of (14) is: $\pi = \Pi x$, where $\Pi$ is the solution of the following Lyapunov-Sylvester matrix equation [9]:

$$\Pi A_s - A_f \Pi = A_{fs} \tag{20}$$

As proven in [9], the above matrix equation (20), which coincides with (9) in the previously examined linear case, admits a unique solution $\Pi$, as long as the $A_s, A_f$ matrices do not have common eigenvalues, and this is guaranteed by the assumptions of Lyapunov's auxiliary theorem and the spectral gap assumption made earlier. Therefore, the linear result is naturally reproduced.

Furthermore, the following Theorem can be proven as well [19]:

**Theorem 2.** *Let all assumptions of Theorem 1 hold true. Furthermore, let $\Omega$ (13) be an invariant manifold of (12), where $\pi(x_s)$ is the unique locally analytic solution of the invariance PDE (14) and $(x_s(t), x_f(t))$ a solution curve of (12). There exists a neighborhood $U^0$ of the origin and real numbers $M > 0$ and $K > 0$ such that, if $(x_s(0), x_f(0)) \in U^0$, then:*

$$||x_f(t) - \pi(x_s(t))||_2 \leq M \exp(-Kt)||x_f(0) - \pi(x_s(0))||_2 \tag{21}$$

*Furthermore, the rate of decay of the dynamics of the off-the-manifold coordinate $z = x_f - \pi(x_s)$ is governed by the "fast" eigenvalues of matrix $A_f = \dfrac{\partial F_f}{\partial x_f}(0,0)$.*

Theorem 2 states that any trajectory of system (12) starting at a point sufficiently close to the origin, converges to $\Omega$. Therefore, the reduced-order model:

$$\begin{aligned} \frac{\mathrm{d}x_s}{\mathrm{d}t} &= F_s(x_s) \\ x_f(t) &= \pi((x_s(t))) \end{aligned} \tag{22}$$

is a projection of the motion of the original system (12) on the invariant manifold $\Omega$, as a result of neglecting the fast transients of the motion. Equivalently, the invariant manifold $\Omega$ (13) computed through the system of singular invariance PDEs (14) is rendered locally exponentially attractive, and thus, it

represents the system's slow manifold. The latter is the cornerstone of the proposed model-reduction method for nonlinear dynamical systems.

Let us now consider the most generic case where the previously mentioned exact triangularization of the system dynamics is not feasible. However, one can always triangularize the linear part of the system dynamics by transforming the system's Jacobian $A = \dfrac{\partial f}{\partial x}(0)$ into the block triangular form considered earlier in the linear case. In particular, one can always employ a linear coordinate transformation such that the Jacobian $A = \dfrac{\partial f}{\partial x}(0)$ becomes transformed into a block triangular form where the eigenvalues of the diagonal blocks are exactly the "slow" and "fast" eigenvalues of $A$ [30]. As a result, in the new coordinate system the original system dynamics is represented via the following form:

$$\begin{aligned} \frac{\mathrm{d}x_s}{\mathrm{d}t} &= F_s(x_s, x_f) \\ \frac{\mathrm{d}x_f}{\mathrm{d}t} &= F_f(x_s, x_f) \end{aligned} \tag{23}$$

where $F_f(x_s, x_f)$ and $F_s(x_s, x_f)$ are real analytic vector functions with: $F_f(0,0) = 0$, $F_s(0,0) = 0$, $\dfrac{\partial F_s}{\partial x_f}(0,0) = 0$ and $\sigma(A_s) = \sigma(\dfrac{\partial F_s}{\partial x_s}(0,0)$, $\sigma(A_f) = \sigma(\dfrac{\partial F_f}{\partial x_f}(0,0)$ are the set of the "slow" and "fast" eigenvalues of the Jacobian matrix $A$ respectively, as they surface once the block-triangularization of the system's linear part is performed. As in the previous case, one can readily infer that:

$$\Omega = \{(x_f, x_s) \in R^n | x_f - \pi(x_s) = 0\} \tag{24}$$

represents an invariant manifold for system (23), if the map $\pi$ satisfies the quasi-linear invariance PDE below:

$$\frac{\partial \pi}{\partial x_s} F_s(x_s, \pi(x_s)) = F_f(x_s, \pi(x_s)) \tag{25}$$

Notice that the above system of invariance PDEs has a common principal part [4, 6] which consists of the components of the vector function $F_s(x_s, x_f)$, and that the origin represents a **characteristic point** for the above system of invariance PDEs (25) (since the principal part vanishes at $(x_s, x_f) = (0, 0)$ due to the equilibrium condition) [4, 6]. As a consequence, and similarly to the previous case, the above system of PDEs (25) becomes **singular** and the Cauchy-Kovalevskaya theorem can not be applied [4, 6]. However, Lyapunov's auxiliary theorem can be employed to guarantee the existence and uniqueness of a locally analytic solution. Indeed, within a similar framework of analysis as in [19], one can prove the following Theorems:

**Theorem 3.** *Consider the nonlinear dynamical system (23) and let all the aforementioned assumptions hold true. Moreover, assume that the eigenvalues* $k_i$ *of matrix* $A_s = \dfrac{\partial F_s}{\partial x_s}(0,0)$ *are not related to the eigenvalues* $\lambda_i$ *of matrix* $A_f = \dfrac{\partial F_f}{\partial x_f}(0,0)$ *through any equations of the type (18,19). Then the set* $\Omega$ *(24) is a real analytic invariant manifold of (23), where* $\pi(x_s)$ *is the unique solution of the singular invariance PDE (25).*

**Theorem 4.** *Let all assumptions of Theorem 3 hold true. Furthermore, let* $\Omega$ *(24) be an invariant manifold of (23), where* $\pi(x_s)$ *is the unique locally analytic solution of the invariance PDE (25) and* $(x_s(t), x_f(t))$ *a solution curve of (23). There exists a neighborhood* $U^0$ *of the origin and real numbers* $M > 0$ *and* $K > 0$ *such that, if* $(x_s(0), x_f(0)) \in U^0$*, then:*

$$||x_f(t) - \pi(x_s(t))||_2 \leq M \exp(-Kt)||x_f(0) - \pi(x_s(0))||_2 \tag{26}$$

*Furthermore, the rate of decay of the dynamics of the off-the-manifold coordinate* $z = x_f - \pi(x_s)$ *is governed by the "fast" eigenvalues of matrix* $A_f = \dfrac{\partial F_f}{\partial x_f}(0,0)$.

Theorems 3 and 4 imply that $\Omega$ is the system's slow invariant manifold that exponentially attracts all system trajectories once the fast transients die out. Therefore, a reduced-order description of the original system dynamics would be the following one:

$$\begin{aligned} \frac{\mathrm{d}x_s}{\mathrm{d}t} &= F_s(x_s, \pi(x_s)) \\ x_f &= \pi((x_s)) \end{aligned} \tag{27}$$

The above reduced-order model represents exactly the system's actual dynamics on the slow manifold $\Omega$ (the most important stage of the system's life before it reaches the equilibrium state), and can be used in practice since the fast transients are justifiably ignored. Indeed, the proposed reduced-order model implies that almost instantaneously the fast state $x_f$ jumps from its initial condition $x_f(0)$ to $\pi(x_s(0))$ on the manifold $\Omega$ where the system is bound to evolve and the relation $x_f(t) = \pi(x_s(t))$ holds for every $t > 0$.

**Remark 2**: In order to be able to make practical use of the proposed method, one must provide a solution scheme for the associated system of singular invariance PDEs (25). Notice that the method of characteristics is not applicable because the aforementioned system of PDEs (25) is singular [4, 6]. However, since all functions involved are locally analytic around the origin, it is possible to calculate the solution $x_f = \pi(x_s)$ in the form of a multivariate Taylor series around the origin. The method involves expanding all functions involved, as well as the unknown solution $x_f = \pi(x_s)$ in a Taylor series and equating the same order Taylor coefficients of both sides of the

PDEs (25). This procedure leads to linear recursion formulas [23], through which one can calculate the $N$-th order Taylor coefficients of the unknown solution $x_f = \pi(x_s)$, given the Taylor coefficients of $x_f = \pi(x_s)$ up to order $N-1$.

In the derivation of the recursion formulas, it is convenient to use the following tensorial notation:

a) The entries of a matrix $A$ are represented as $a_i^j$, where the subscript $i$ refers to the corresponding row and the superscript $j$ to the corresponding column of the matrix.

b) The partial derivatives of the $\mu$-th component $F_\mu(x_s, x_f)$ of the vector function $F(x_s, x_f)$ with respect to the state variables $x_s$ evaluated at $(x_s, x_f) = (0, 0)$ are denoted as follows:

$$\begin{aligned} F_\mu^i &= \frac{\partial F_\mu}{\partial x_{s,i}}(0,0) \\ F_\mu^{ij} &= \frac{\partial^2 F_\mu}{\partial x_{s,i}\partial x_{s,j}}(0,0) \\ F_\mu^{ijk} &= \frac{\partial^3 F_\mu}{\partial x_{s,i}\partial x_{s,j}\partial x_{s,k}}(0,0) \end{aligned} \tag{28}$$

etc., where $i, j, k, ..=1, ..., n$

c) The standard summation convention where repeated upper and lower tensorial indices are summed up.

Under the above notation the $l$-th component $\pi_l(x_s)$ of the unknown solution $\pi(x_s)$ can be expanded in a multivariate Taylor series as follows:

$$\begin{aligned} \pi_l(x_s) = \frac{1}{1!}\pi_l^{i_1} x_{s,i_1} + \frac{1}{2!}\pi_l^{i_1 i_2} x_{s,i_1} x_{s,i_2} + ... + \\ + \frac{1}{N!}\pi_l^{i_1 i_2 ... i_N} x_{s,i_1} x_{s,i_2} ... x_{s,i_N} + ... \end{aligned} \tag{29}$$

Similarly one expands the components of the vector functions $F_s(x_s, x_f)$, $F_f(x_s, x_f)$ in multivariate Taylor series. Substituting the Taylor expansions of $\pi(x_s)$ and $F_s(x_s, x_f)$, $F_f(x_s, x_f)$ into the system of PDEs (25) and matching the Taylor coefficients of the same order, the following relation for the $N$-th order terms can be obtained:

$$\sum_{L=0}^{N-1} \sum_{\binom{N}{L}} \pi_l^{\mu i_1 ... i_L} F_{s,\mu}^{i_{L+1} ... i_N} = F_{f,l}^{\mu} \pi_\mu^{i_1 ... i_N} + f_l^{i_1 ... i_N}(\pi^{i_1 ... i_{N-1}}) \tag{30}$$

where $f_l^{i_1 ... i_N}(\pi^{i_1 ... i_{N-1}})$ is a function of Taylor coefficients of the unknown solution $\pi(x_s)$ calculated in the previous recursive steps. Note that the second summation symbol in (30) should be regarded as summing up the relevant quantities over the $\binom{N}{L}$ possible combinations of the indices $(i_1, ..., i_N)$. Furthermore, equations (30) represent a set of linear algebraic equations in

the unknown coefficients $\pi_{\mu}^{i_1,\ldots,i_N}$, and this is precisely the underlying mathematical reason that allows the series solution method to be accomplished in an automated fashion by exploiting the computational capabilities and commands of a symbolic software package such as MAPLE. Finally, it should be also pointed out, that occasionally the Taylor series solution method for the invariance PDEs (25) exhibits slow convergence. In these cases, significant improvement of the convergence properties of the PDE solution scheme can be achieved if direct Newton-type methods as described in [10] are employed, or relaxation methods such as the ones reported in [7, 16].

## 4 Conclusions

A new approach to the problem of model-reduction for nonlinear dynamical systems was proposed in the present study. The formulation of the problem was conveniently realized through a system of singular quasi-linear invariance PDEs, and a set of conditions for solvability was derived. In particular, within the class of real analytic solutions, the aforementioned set of conditions was shown to guarantee the existence and uniqueness of a locally analytic solution. The solution of the system of singular invariance PDEs was then proven to be the slow invariant manifold of the nonlinear dynamical system under consideration, and an exact reduced-order model for the nonlinear system dynamics was obtained through the restriction of the original system dynamics on the aforementioned slow manifold. The local analyticity property of the above solution (whose graph corresponds to the system's slow invariant manifold) enabled the development of a series solution method, which allows the polynomial approximation of the "slow" system dynamics on the slow manifold up to the desired degree of accuracy, and can be easily implemented with the aid of a symbolic software package such as MAPLE.

*Acknowledgement.* The authors would like to thank Professor A. Gorban and Professor B. Leimkuhler, Department of Mathematics, Centre for Mathematical Modelling, University of Leicester for the support and wonderful hospitality during the workshop "Model-Reduction and Coarse-Graining Approaches for Multiscale Phenomena," August 24-26, 2005, Leicester, UK. Financial support from the National Science Foundation through grant CTS-9403432 is gratefully acknowledged by the first author.

## References

1. V.I. Arnold: *Geometrical Methods in the Theory of Ordinary Differential Equations* (Springer, Berlin Heidelberg New York 1983)
2. A. Astolfi, R. Ortega: Immersion and invariance: A new tool for stabilization and adaptive control of nonlinear systems. *IEEE Trans. Autom. Contr.* **48**, 590–606 (2003)

3. P.D. Christofides: *Nonlinear and Robust Control of PDE Systems* (Birkhauser, Boston, MA 2001)
4. R. Courant, D. Hilbert: *Methods of Mathematical Physics, Volume II* (John Wiley & Sons, New York 1962)
5. S.M. Cox, A.J. Roberts: Initial conditions for models of dynamical systems. *Physica D* **85**, 126 (1995)
6. L.C. Evans: *Partial Differential Equations* (American Mathematical Society, Providence, RI 1998)
7. C. Foias, M.S. Jolly, I.G. Kevrekidis, G.R. Sell, E.S. Titi: On the computation of inertial manifolds. *Phys. Lett. A* **131**, 433 (1989)
8. C. Foias, R. Sell, E.S. Titi: Exponential tracking and approximation of inertial manifolds for dissipative equations. *J. Dynam. Diff. Equat* **1**, 199 (1989)
9. F.R. Gantmacher: *The Theory of Matrices* (Chelsea Publishing Company, New York, 1960)
10. A.N. Gorban, I.V. Karlin: Methods of invariant manifolds and regularization of acoustic spectra. *Transp. Theor. Stat. Phys.* **23**, 559 (1994)
11. A.N. Gorban, I.V. Karlin, V.B. Zmievskii, S.V. Dymova: Reduced description in the reaction kinetics. *Physica A* **275**, 361 (2000)
12. A.N. Gorban, I.V. Karlin: Method of invariant manifold for chemical kinetics. *Chem. Engn. Sci.* **58**, 4751 (2003)
13. A.N. Gorban, I.V. Karlin, A.Y. Zinovyev: Constructive methods of invariant manifolds for kinetic problems. *Phys. Reports* **396**, 197 (2004)
14. A.N. Gorban, I.V. Karlin: *Invariant Manifolds for Physical and Chemical Kinetics*, Lecture Notes in Physics, **660** (Springer, Berlin Heidelberg New York 2005)
15. J. Guckenheimer, P.J. Holmes: *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (Springer, Berlin Heidelberg New York 1983)
16. S Jin, M. Slemrod: Regularization of the Burnett equations for rapid granular flows via relaxation. *Physica D* **150**, 207 (2001)
17. T. Kaper, N. Koppel, C.K.R.T. Jones: Tracking invariant manifolds up to exponentially small errors. *SIAM J. Math. Anal.* **27**, 558 (1996)
18. N. Kazantzis: Singular PDEs and the problem of finding invariant manifolds for nonlinear dynamical systems. *Phys. Lett. A* **272**, 257 (2000)
19. N. Kazantzis, T. Good: Invariant manifolds and the calculation of the long-term asymptotic response of nonlinear processes using singular PDEs *Comp. Chem. Eng* **26**, 999 (2002)
20. P.V. Kokotovic, H.K. Khaliland, J.O. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design* (Academic Press 1986)
21. A. Kumar, P.D. Christofides, P. Daoutidis: Singular perturbation modeling of nonlinear processes with nonexplicit time-scale multiplicity *Chem. Engn. Sci.* **53**, 1491 (1998)
22. G. Li, H. Rabitz: A general analysis of exact nonlinear lumping in chemical kinetics. *Chem. Engn. Sci.* **49**, 343 (1994)
23. A.M. Lyapunov: *The General Problem of the Stability of Motion*, (Taylor & Francis Ltd, London 1992).
24. G. Moore: Geometric methods for computing invariant manifolds. *Appl. Numer. Math.* **17**, 319 (1995)
25. A.J. Roberts: Low-dimensional modelling of dynamics via computer algebra. *Comp. Phys. Commun.* **100**, 215 (1997)

26. M.R. Roussel: Forced-convergence iterative schemes for the approximation of invariant manifolds. *J. Math. Chem.* **21**, 385 (1997)
27. M.R. Roussel, S.J. Fraser: Invariant manifold methods for metabolic model reduction. *Chaos* **11**, 196 (2001)
28. S.Y. Shvartsman, I.G. Kevrekidis: Nonlinear model reduction for control of distributed systems: A computer-assisted study. *AICHE J.* **44**, 1579 (1998)
29. S.Y. Shvartsman, C. Theodoropoulos, R. Rico-Martinez, I.G. Kevrekidis, E.S. Titi, T.J. Mountziaris: Order reduction for nonlinear dynamic models of distributed reacting systems. *J. Proc. Contr.* **10**, 177 (2000)
30. S. Wiggins: *Introduction to Applied Nonlinear Dynamical Systems a nd Chaos* (Springer, Berlin Heidelberg New York 1990)
31. A. Zagaris, H.G. Kaper, T.J. Kaper: Analysis of the computational singular perturbation reduction method for chemical kinetics. *J. Nonl. Sci.* **14**, 59 (2004)
32. V.B. Zmievski, I.V. Karlin, M. Deville: The universal limit in dynamics of dilute polymeric solutions. *Physica A* **275**, 152 (2000)

# A Versatile Algorithm for Computing Invariant Manifolds

H. W. Broer[1], A. Hagen[2], and G. Vegter[1]

[1] Department of Mathematics and Computing Science, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands
[2] Department of Mathematics, University of Texas at Arlington, 411 S. Nedderman Dr., Arlington TX 76019, USA, hagen@nethere.com

**Summary.** This paper deals with the numerical computation of invariant manifolds using a method of discretizing global manifolds. It provides a geometrically natural algorithm that converges regardless of the restricted dynamics. Common examples of such manifolds include limit sets, co-dimension 1 manifolds separating basins of attraction (separatrices), stable/unstable/center manifolds, nested hierarchies of attracting manifolds in dissipative systems and manifolds appearing in bifurcations. The approach is based on the general principle of normal hyperbolicity, where the graph transform leads to the numerical algorithms. This gives a highly multiple purpose method. The algorithm fits into a continuation context, where the graph transform computes the perturbed manifold. Similarly, the linear graph transform computes the perturbed hyperbolic splitting. To discretize the graph transform, a discrete tubular neighborhood and discrete sections of the associated vector bundle are constructed. To discretize the linear graph transform, a discrete (un)stable bundle is constructed. Convergence and contractivity of these discrete graph transforms are discussed, along with numerical issues. A specific numerical implementation is proposed. An application to the computation of the 'slow–transient' surface of an enzyme reaction is demonstrated.

## 1 Introduction

Invariant manifolds of dynamical systems typically determine the skeleton of the dynamics, around which a further analysis may be in order. This is true whether the system is dissipative or conservative. For dissipative systems, the phase space often contains a nested hierarchy of attracting manifolds $V_i \subset V_{i+1}$, $i = 0, \ldots, n$. The manifold $V_i$ is composed of initial data which evolves slowly compared to initial data in the rest of $V_{i+1}$. The manifold $V_0$ contains the global attractor, which may be an equilibrium point or more complicated set. The long-time (medium-time) dynamics is described by the system restricted to $V_0$ ($V_1$). By restricting the system to a lower dimensional

manifold, fast transients are removed from consideration. Thus, the dimension of the model is reduced while retaining the essential features of the dynamics.

Analytical formulae for the lower dimensional manifolds and the corresponding reduced systems are only obtainable in special cases. Hence, methods of approximating these manifolds are desirable. For example, in applied bifurcation theory, the center manifold of an equilibrium is approximated locally by polynomials, using a recursive algebraic procedure [23]. This allows the local approximation of the system restricted to the center manifold, up to sufficiently high-order terms. An analysis of the bifurcation is then performed on the approximate center manifold.

In the present paper, we focus on a numerical algorithm which computes global invariant manifolds. This allows a global approximation of the system restricted to the invariant manifold, in principle to arbitrary accuracy. This may aid further analysis of long-time non-local dynamics.

The algorithmic approach is based on the principle of normal hyperbolicity. According to the Invariant Manifold Theorem, normally hyperbolic invariant manifolds persist smoothly under small perturbations of the system. To be specific, the Invariant Manifold Theorem is concerned with the following setup. Given a diffeomorphism $F$ and an $F$–invariant submanifold $V$, the invariant manifold $\tilde{V}$ for a nearby diffeomorphism $\tilde{F}$ is constructed. Based on this, an invariant manifold $\tilde{V}$ for the system of interest, $\tilde{F}$, may be computed given an analytically known initial manifold $V$ for a nearby system $F$. It turns out that a rough estimate of an initial manifold $V$ is often enough. In addition, the algorithm may be repeated with computed initial data, allowing the potential to compute invariant manifolds of systems not necessarily near a system with a known manifold.

The algorithm is adapted from one of the classical approaches to the proof of the Invariant Manifold Theorem, the graph transform. The theory of invariant manifolds using the graph transform is well developed [21]. In particular the convergence properties of the graph transform are inherited by the algorithm. This complete theory of convergence is one thing that distinguishes this approach from many other approaches to computing invariant manifolds in the literature.

The implementation of methods for computing (non-local) manifolds of dimension $\geq 2$ is fairly recent. Some of the related work in this category concerns quasiperiodic (for example [17]) or attracting (for example [10]) tori, parts of global attractors [9] or global (un)stable manifolds [22]. The computations of tori use global parametrizations of the tori where simplicial complexes are used in the present paper. The computations of parts of global attractors use successive subdivisions of a covering of part of the global attractor. This approach computes global attractors which are smooth or non–smooth. The computations of global (un)stable manifolds are concerned with extending a given piece of the manifold, to fill out the global (un)stable manifold. The present paper has the antecedents [2, 3, 5, 27]. In [5, 27] a method to compute saddle–type manifolds is presented. The graph transform and simplicial com-

plexes are used to approximate manifolds. The present paper, starting with a simplicial complex, uses a piecewise polynomial approximation. To do this, a discrete tubular neighborhood is constructed. An approximation of arbitrary order for any manifold is obtained. A tubular neighborhood of $V$ is the geometrical setting of the graph transform. Thus, a discrete tubular neighborhood is a natural approach which allows an analogous development of a discrete graph transform. In addition, the construction of a discrete (un)stable bundle allows a natural derivation of the discrete linear graph transform.

Compared to related work, the present approach gives a general purpose algorithm. It applies to manifolds of arbitrary topological type, attracting or saddle–type, regardless of the restricted dynamics. There is a satisfactory theory of convergence in this general setting. If the manifold is not normally hyperbolic, however, a different approach should be used, see for example [17]. Other novel features of the present paper include the following. In Section 5, a practical approach to solving the global equations associated with the discrete graph transform is proposed. In Section 6, the graph transform approach is used to compute a part of the 'slow–transient' surface of an enzyme reaction model. This is the first time this approach has been used to compute this type of surface. For numerical methods designed specifically for this type of problem, see [15, 16, 30].

To repeatedly apply the algorithm, both the perturbed manifold $\tilde{V}$ and its hyperbolic splitting must be approximated. This is done by first using the graph transform $\mathit{\Gamma}$ to obtain $\tilde{V}$ and then the linear graph transform $\mathcal{L}$ to compute the hyperbolic splitting of $\tilde{V}$. Thus, in Section 2, $\mathit{\Gamma}$ and $\mathcal{L}$ are formulated. This includes a discussion of normal hyperbolicity, the Invariant Manifold Theorem, tubular neighborhoods and hyperbolic splittings. In Section 3, the discretizations of the domains of $\mathit{\Gamma}$ and $\mathcal{L}$ are formulated. To do this, a discrete tubular neighborhood along with a space of discrete sections of the associated vector bundle are constructed. In Section 4, discrete versions $\mathit{\Gamma}_D$ of $\mathit{\Gamma}$ and $\mathcal{L}_D$ of $\mathcal{L}$ are formulated, based on the discrete approximating sections of Section 3. Analyses of the convergence and contractivity of $\mathit{\Gamma}_D$ and $\mathcal{L}_D$ are given. In Section 5, an outline of a computer implementation of the algorithm is given. Some auxiliary numerical techniques, along with numerical conditioning and error, are also discussed. Section 6 contains an application to an enzyme reaction model. For more examples, see [2, 3] or the DISC project website, http://home.nethere.net/hagen.

## 2 Invariant Manifolds

In this section, the basic theory of normally hyperbolic invariant manifolds is introduced. An overview of some definitions and results from [21] is given. For locating a perturbed manifold, the graph transform is formulated. The linear graph transform is formulated to locate the hyperbolic splitting of this perturbed manifold. In later sections, discrete versions of these graph trans-

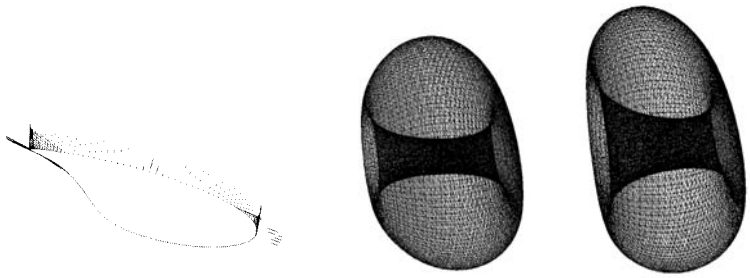


Fig. 1: Lorenz system orbit and hyperbolic splitting; two tori in the Lorenz-84 system, moving away from a Hopf saddle–node bifurcation [23].

forms, suitable for a numerical implementation, will be given. This will be done by replacing the basic elements, like tubular neighborhoods and sections of vector bundles, with discrete constructions.

### 2.1 Normal Hyperbolicity

The starting point is a $C^r$ diffeomorphism $F$ on a $C^\infty$ Riemannian manifold $M$, with an invariant submanifold $V \subset M$. Here, $V$ is a compact, $C^r$, $r$–normally hyperbolic submanifold of $M$, $r \geq 1$. The submanifold $V$ is $r$–normally hyperbolic for $F$ if there is a $DF$–invariant splitting

$$T_V(M) = N^u(V) \oplus T(V) \oplus N^s(V), \tag{1}$$

and a Riemann structure on the tangent bundle $T_V(M)$, such that, for $y \in V$, $i \geq 0$, and $0 \leq k \leq r$:

$$\begin{aligned} &\|DF^i \,|\, N^s_y(V)\| \cdot \|(DF^i \,|\, T_y(V))^{-1}\|^k \leq c\mu^i, \\ &\|(DF^i \,|\, N^u_y(V))^{-1}\| \cdot \|DF^i \,|\, T_y(V)\|^k \leq c\,(1/\lambda)^i\,, \end{aligned} \tag{2}$$

for some $0 < \mu < 1 < \lambda < \infty$ and $0 < c < \infty$. Here the operator norms are associated with the Riemann structure on $T_V(M)$. For example, consider the attracting case, $N^u_y(V) = \{0\}$, $y \in V$ and $r = 1$. Condition (2) concerns the linearization of $F$ at $V$, in other words $DF$ on $T_V(M)$. It states that under the action of the linearization, vectors normal to $V$ are asymptotically contracted more than vectors tangent to $V$. This means that under the action of the dynamical system $F$, a neighborhood of a point in $V$ is flattened in the direction of the manifold.

The *Invariant Manifold Theorem* [21, Theorem 4.1] states that a $C^r$ diffeomorphism $\tilde{F}$, that is $C^r$–near $F$, has an $r$–normally hyperbolic invariant manifold $\tilde{V}$, that is $C^r$ and $C^r$–near $V$. This theorem and its proof suggests

that it may be possible to compute an approximation to $\tilde{V}$ from a given $V$. To implement this idea, we look more closely at a proof of the invariant manifold theorem.

First, we focus on a tubular neighborhood of $V$ [20, 24]. A tubular neighborhood of $V$ in $M$ is a vector bundle $E$ with base space $V$, an open neighborhood $U$ of $V$ in $M$, an open neighborhood $Z$ of the zero section in $E$ and a homeomorphism $\phi : Z \to U$. Here, $\phi$ must satisfy $\phi \circ \sigma_0 = i$, where $\sigma_0 : V \to E$ is the zero section and $i : V \to M$ is the inclusion. For example, the normal bundle $E = \bigcup_{p \in V} T_p(V)^\perp$ of $V$ in $M$ gives a tubular neighborhood of $V$, at least if $r \geq 2$. In fact, any Lipschitz vector bundle $N(V)$, transverse to $T(V)$ in $T_V(M)$, gives a tubular neighborhood of $V$ in $M$. In the following, $\tilde{V}$ is constructed in the neighborhood $U$ in $M$, or equivalently in the neighborhood $Z$ in $N(V)$. A slight technical adjustment is made here. Namely, below, $Z$ is the closure of a neighborhood, $Z = Z(\epsilon) = \{(p, v) \in N(V) : |v|_p \leq \epsilon\}$.

For any Lipschitz transverse vector bundle $N(V)$, the invariant splitting (1) induces a splitting $N(V) = N^u(V) \oplus N^s(V)$ into stable and unstable parts. The hyperbolic splitting $T_V(M) = N^u(V) \oplus T(V) \oplus N^s(V)$ has the same growth properties (2) as the invariant splitting. Sections of $Z$ may now be written $\sigma(p) = (p, v^s(p), v^u(p))$, where $v^s(p) \in Z^s_p = N^s_p(V) \cap Z$, $v^u(p) \in Z^u_p = N^u_p(V) \cap Z$.

### 2.2 The Graph Transform

The graph transform uses the $\tilde{F}$–dynamics near $V$ to locate $\tilde{V}$. The domain of the graph transform is a certain space of sections of the vector bundle $Z = Z(\epsilon)$. The graphs of the sections in the domain are the Lipschitz manifolds near $V$ in Lipschitz norm. In fact, the graph transform is a contraction on a space of Lipschitz sections $\sigma : V \to Z$. To define the Lipschitz constant of a section, a $C^0$ connection in $T_V(M)$ is used [25]. A connection gives a way to compare points in different fibers of $T_V(M)$. It does this using a continuous family of horizontal subspaces $H(y)$, $y \in T_V(M)$, which extend the tangent spaces of $V$. More precisely, a $C^0$ connection in the vector bundle $\pi : T_V(M) \to V$ is a $C^0$ distribution $H : T_V(M) \to T(T_V(M))$ with $T_y(T_V(M)) = H(y) \oplus V(y)$, $y \in T_V(M)$, where $V(y)$ is the kernel of $D\pi$. Here, it is also required that the horizontal subspace of the associated frame bundle corresponding to $H(y)$ be invariant under the structure group. This implies, in particular, that if $\sigma_0 : V \to T_V(M)$ is the zero section, then $H(\sigma_0(p)) = D\sigma_0(T_p(V))$.

To define the slope of a section $\sigma : V \to T_V(M)$ at $p \in V$, let $\theta : V \to T_V(M)$ be a $C^1$ section with $\theta(p) = \sigma(p)$ and $D\theta(T_p(V)) = H(\sigma(p))$. Then the slope of $\sigma$ at $p$ is

$$\text{slope}_p(\sigma) = \limsup_{x \to p} \frac{|\sigma(x) - \theta(x)|_x}{d_V(x, p)},$$

[21]. Since $Z^s$ and $Z^u$ are subbundles of $T_V(M)$, this also gives a natural definition of the slope of sections $\sigma^s : V \to Z^s$ and $\sigma^u : V \to Z^u$. From this,

the Lipschitz constant of $\sigma^s$ is $\mathrm{Lip}(\sigma^s) = \sup_{p\in V} \mathrm{slope}_p(\sigma^s)$, and similarly for $\sigma^u$. Now, the Lipschitz constant of a section $\sigma(p) = (p, v^s(p), v^u(p))$ of $Z$ is $\mathrm{Lip}(\sigma) = \max\{\mathrm{Lip}(\sigma^s), \mathrm{Lip}(\sigma^u)\}$, where $\sigma^s(p) = (p, v^s(p))$ and $\sigma^u(p) = (p, v^u(p))$. The domain of the graph transform is $\mathcal{S}_{\epsilon,\delta} = \{\sigma : V \to Z : \mathrm{Lip}(\sigma) \leq \delta\}$. The norm on $\mathcal{S}_{\epsilon,\delta}$ is $\|\sigma\| = \max\{\,|\sigma^s|_s, |\sigma^u|_u\}$, where $|\cdot|_s$ and $|\cdot|_u$ are the natural $C^0$ norms on sections of $Z^s$ and $Z^u$, respectively. With this norm, $\mathcal{S}_{\epsilon,\delta}$ is complete.

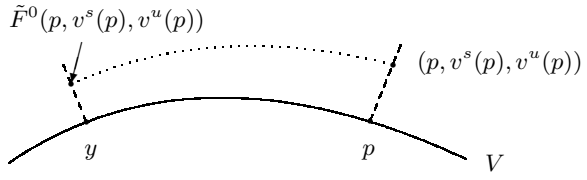



Fig. 2: Invariance condition (3).

To formulate the graph transform, the starting point is the $\tilde{F}$–invariance condition $\phi \circ \sigma(V) = \tilde{F} \circ \phi \circ \sigma(V)$. This is split into two coupled equations, a part on $V$ and a part normal to $V$. We put $\tilde{F}^0 = \phi^{-1} \circ \tilde{F} \circ \phi$ and work in $N(V)$. The image of $\phi \circ \sigma$ is $\tilde{F}$–invariant if and only if

$$\begin{aligned} (y, v^s(y), v^u(y)) &= \tilde{F}^0(p, v^s(p), v^u(p)), \\ y &= \pi \circ \tilde{F}^0(p, v^s(p), v^u(p)), \end{aligned} \tag{3}$$

for $p \in V$, where $\pi : N(V) \to V$ is the vector bundle projection. See Figure 2. Under our hypotheses, $y = \pi \circ \tilde{F}^0(p, v^s(p), v^u(p))$ may be solved for a unique $p \in V$ given $y \in V$ and $\sigma \in \mathcal{S}_{\epsilon,\delta}$ for small $\epsilon, \delta$ and $\theta = \|F - \tilde{F}\|_{C^1}$. Denote this solution by $p = p(y, v^s, v^u)$. Now, given $\sigma \in \mathcal{S}_{\epsilon,\delta}$, $\sigma(p) = (p, v^s(p), v^u(p))$, the *graph transform* of $\sigma$ is the section $\Gamma(\sigma)(p) = (p, w^s(p), w^u(p))$. Here, $w^s$ is defined by

$$w^s(y) = P^s_y \circ \tilde{F}^0(p, v^s(p), v^u(p)),\ \ p = p(y, v^s, v^u), \tag{4}$$

for $y \in V$, where $P^s_y : N_y(V) \to N_y(V)$ is the linear projection with range $N^s_y(V)$ and nullspace $N^u_y(V)$. The unstable part $w^u$ is defined implicitly by

$$\begin{aligned} v^u(y) &= P^u_y \circ \tilde{F}^0(p, v^s(p), w^u(p)), \\ y &= \pi \circ \tilde{F}^0(p, v^s(p), w^u(p)), \end{aligned} \tag{5}$$

for $p \in V$, where $P^u_y : N_y(V) \to N_y(V)$ is the linear projection with range $N^u_y(V)$ and nullspace $N^s_y(V)$. In (5), there is a unique solution for $w^u(p)$ for small $\theta, \epsilon$, and $\delta$.

If $\sigma = \Gamma(\sigma)$, then (4) and (5) imply (3). Hence $\sigma$ is a fixed point of $\Gamma$ if and only if the graph of $\sigma$ is $\tilde{F}$–invariant. By replacing $\tilde{F}$ with $\tilde{F}^N$ above, for some large integer $N$, $\Gamma$ becomes a contraction on $\mathcal{S}_{\epsilon,\delta}$ whose fixed point $\sigma^*$ satisfies $\phi \circ \sigma^*(V) = \tilde{V}$.

### 2.3 The Linear Graph Transform

Two linear graph transforms $\mathcal{L}^s$ and $\mathcal{L}^u$ are used to determine the hyperbolic splitting $N^u(\tilde{V}) \oplus T(\tilde{V}) \oplus N^s(\tilde{V})$ of $\tilde{V}$. Here, $\mathcal{L}^s$ determines $N^s(\tilde{V})$ and $\mathcal{L}^u$ determines $N^u(\tilde{V})$. These two linear graph transforms are contractions on certain spaces of sections. These spaces of sections are determined by the initial data for $\mathcal{L}^s$ and $\mathcal{L}^u$.

To illustrate the details, here $\mathcal{L}^u$ is formulated. Given a transverse bundle $N(\tilde{V})$, first the initial data for $\mathcal{L}^u$ in $N(\tilde{V})$ is determined. Let $Q : T_{\tilde{V}}(M) \to T_{\tilde{V}}(M)$, be, on each fiber $T_y(M)$, the linear projection with range $N_y(\tilde{V})$ and nullspace $T_y(\tilde{V})$. Initial data $N(\tilde{V}) = N^{u,0}(\tilde{V}) \oplus N^{s,0}(\tilde{V})$ are then

$$N^{u,0}(\tilde{V}) = Q(N^{u,1}(\tilde{V})),\ N^{s,0}(\tilde{V}) = Q(N^{s,1}(\tilde{V})),$$

where $N^{u,1}_y(\tilde{V})$, $N^{s,1}_y(\tilde{V})$ are obtained from $N^u_p(V)$, $N^s_p(V)$, $y = \phi \circ \sigma^*(p)$, by parallel translation $T_p(M) \to T_y(M)$ along $\phi$–images of fibers of $N(V)$, [1, 25]. There exists $\alpha > 0$, where $\alpha \to 0$ as $\epsilon + \delta + \theta \to 0$, such that, if $\{\lhd N(V), T(V)\}$, $\{\lhd N(\tilde{V}), T(\tilde{V})\} \geq \alpha > 0$, then this procedure produces non-degenerate initial data $N^{u,0}(\tilde{V}), N^{s,0}(\tilde{V})$.

The domain of $\mathcal{L}^u$ is a space of sections whose graphs are the $j$–plane bundles near $N^{u,0}(\tilde{V})$ in $N(\tilde{V})$, where $j$ is the dimension of $N^{u,0}(\tilde{V})$. These are sections of the bundle $L(\tilde{V})$ whose fiber at $y \in \tilde{V}$ is the space of linear transformations $N^{u,0}_y(\tilde{V}) \to N^{s,0}_y(\tilde{V})$, $L(N^{u,0}_y(\tilde{V}), N^{s,0}_y(\tilde{V}))$, [21]. The domain of $\mathcal{L}^u$ is $\mathcal{S}_\eta = \{\sigma : \tilde{V} \to L(\tilde{V}) : \sup_y \|\sigma(y)\| \leq \eta\}$, where the operator norm $\|\cdot\|$ is associated with the Riemann structure on $T_{\tilde{V}}(M)$. The space $\mathcal{S}_\eta$ is complete with respect to the norm $|\sigma| = \sup_y \|\sigma(y)\|$.

To formulate $\mathcal{L}^u$, the starting point is the invariance condition. The linear mapping induced by $D\tilde{F} : T_{\tilde{V}}(M) \to T_{\tilde{V}}(M)$ on $N(\tilde{V}) \subset T_{\tilde{V}}(M)$ is $\Phi = Q \circ D\tilde{F}|_{N(\tilde{V})} : N(\tilde{V}) \to N(\tilde{V})$. The graph of $\sigma \in \mathcal{S}_\eta$ is $\Phi$–invariant if and only if $\Phi(\text{graph}\{\sigma(x)\}) = \text{graph}\{\sigma(y)\}$, $y = \tilde{F}(x)$, $x \in \tilde{V}$. This condition is split into a part in $N^{u,0}(\tilde{V})$ and a part in $N^{s,0}(\tilde{V})$. Let $P^u_y : N_y(\tilde{V}) \to N_y(\tilde{V})$ be the linear projection with range $N^{u,0}_y(\tilde{V})$ and nullspace $N^{s,0}_y(\tilde{V})$. Define $P^s_y$ analogously. Then the graph of $\sigma \in \mathcal{S}_\eta$ is $\Phi$–invariant if and only if

$$\begin{aligned} \sigma(y)(\tilde{\rho}) &= P^s_y \circ \Phi(\rho, \sigma(x)(\rho)), \\ \tilde{\rho} &= P^u_y \circ \Phi(\rho, \sigma(x)(\rho)), \end{aligned} \tag{6}$$

for $\rho \in N^{u,0}_x(\tilde{V})$, $x \in \tilde{V}$, where $y = \tilde{F}(x)$. The second equation in (6) is a linear mapping $N^{u,0}_x(\tilde{V}) \to N^{u,0}_y(\tilde{V})$, $\rho \to \tilde{\rho}$, which is invertible for small

$\epsilon$, $\delta$, $\theta$ and $\eta$. Denote the inverse $B_y(\tilde{\rho}) = \rho$. Then, the graph transform of $\sigma$ is the section $\mathcal{L}^u(\sigma)(y) = P^s_y \circ \Phi \circ (\mathrm{id}, \sigma(x)) \circ B_y$ for $y \in \tilde{V}$. Here, $(\mathrm{id}, \sigma(x)) : N^{u,0}_x(\tilde{V}) \to N_x(\tilde{V})$ is $(\mathrm{id}, \sigma(x))(\rho) = (\rho, \sigma(x)(\rho))$.

The graph of $\sigma$ is $\Phi$–invariant if and only if $\sigma$ is a fixed point of $\mathcal{L}^u$. By replacing $\Phi$ with $\Phi^N$ above, for some large integer $N$, and for $\epsilon$, $\delta$, $\theta$ and $\eta$ small, $\mathcal{L}^u$ is a contraction on $\mathcal{S}_\eta$ whose fixed point $\sigma^*$ gives the $\Phi$–invariant bundle $N^u(\tilde{V})$. The formulation of $\mathcal{L}^s$ is analogous.

To summarize, one step of the proposed continuation algorithm has two parts. The initial data is an $F$–invariant manifold $V$ with hyperbolic splitting $N^u(V) \oplus T(V) \oplus N^s(V)$. The first step uses the graph transform $\Gamma$ on $V$ with $N^u(V) \oplus T(V) \oplus N^s(V)$ to determine the $\tilde{F}$–invariant manifold $\tilde{V}$. That is, starting with the zero section $\sigma_0$, $\Gamma$ is iterated, $\Gamma^i(\sigma_0) \to \sigma^*$ in $C^0$ norm as $i \to \infty$. The second step uses linear graph transforms $\mathcal{L}^s$ and $\mathcal{L}^u$ together with initial data determined by $\tilde{V}$ and $N^u(V) \oplus T(V) \oplus N^s(V)$ to determine the hyperbolic splitting $N^u(\tilde{V}) \oplus T(\tilde{V}) \oplus N^s(\tilde{V})$ of $\tilde{V}$. Now the first and second steps are repeated with initial data $\tilde{V}$, $N^u(\tilde{V}) \oplus T(\tilde{V}) \oplus N^s(\tilde{V})$.

## 3 Discrete Sections

In this section, discrete versions of $V$, its hyperbolic splitting, transverse bundle and sections of the transverse bundle are constructed. From this, the discrete version of the graph transform in Section 4 follows. Here, the manifold $M = \mathbb{R}^n$ with the constant Riemann metric induced by the usual inner product. This is not, in principle, a reduction of the generality of the method, since $V$ may be embedded in $\mathbb{R}^n$ and the property of normal hyperbolicity (2) is independent of the Riemann structure.

The initial manifold $V$ is approximated by a geometric simplicial complex $\mathcal{C} \subset \mathbb{R}^n$ supporting $V \subset \mathbb{R}^n$, [6, 26]. Recall that the polyhedron $P \subset \mathbb{R}^n$ of $\mathcal{C}$ is the set of all points in the simplices of $\mathcal{C}$ with the subspace topology. A simplicial complex $\mathcal{C}$ supports $V$ if the vertices of all simplices are in $V$ and $P$ is homeomorphic to $V$. If $H$ is the maximal diameter of the simplices of $\mathcal{C}$ then $P$ converges to $V$ in Lipschitz norm as $H \to 0$. Denote by $\mathcal{C}_1 \ldots \mathcal{C}_N$ the $d$–simplices of $\mathcal{C}$, $d = \dim V$. For the uniformity of the polynomial approximations on each $\mathcal{C}_i$ as $H \to 0$, it is required that $\{\mathcal{C}_i\}_{i=1}^N$ be a regular family. This means that, if $h_i$ is the diameter of $\mathcal{C}_i$ and $\rho_i$ the supremum of the diameters of the inscribed spheres of $\mathcal{C}_i$, then $h_i/\rho_i$ is bounded uniformly for all $i$ and $H \to 0$, [8].

Next, discrete approximations to the transverse bundle and hyperbolic splitting of $V$ are described. The approximation to the hyperbolic splitting will be given by vector bundles $N^s(P)$ and $N^u(P)$, where $N(P) = N^s(P) \oplus N^u(P)$ is the transverse bundle associated with a tubular neighborhood of $P$.

To be specific, a tubular neighborhood of $P$ is induced by a transverse field of $k_0$–planes $\mu : P \to \mathcal{G}_{n,k_0}$ = the Grassmann manifold of $k_0$–planes of $\mathbb{R}^n$, $k_0 = \mathrm{codim}\, V$, provided $\mu$ is locally Lipschitz with respect to Riemannian

metrics [20, 32]. Note that the approximation to the hyperbolic splitting satisfies $N_x(P) = N_x^s(P) \oplus N_x^u(P) \subset T_x(\mathbb{R}^n)$, $x \in P$. Here, $T_x(\mathbb{R}^n)$, $x \in P$, are as usual identified with the ambient space $\mathbb{R}^n$ containing $V$ and also the underlying space $\mathbb{R}^n$ of the Grassmann manifold via the standard basis. By this identification, the field $\mu$ gives a transverse bundle $N(P)$. In fact, the field $\mu$ is made up of two parts, $\mu(x) = \mu_1(x) \oplus \mu_2(x)$, $x \in P$, where $\mu_i : P \to \mathcal{G}_{n,k_i}$ for $i = 1, 2$, $k_1 = \dim N^s(V)$ and $k_2 = \dim N^u(V)$. Here, $\mu_1$ gives $N^s(P)$ and $\mu_2$ gives $N^u(P)$.

The bundle $N^s(V)$ is approximated by $N^s(P)$ as follows. The given $N(V)$ induces a homeomorphism $\psi : P \to V$. Let $N^{s,0}(P)$ be the vector bundle over $P$ whose fiber at $y \in P$ is $N^s_{\psi(y)}(V)$. To approximate $N^s(V)$, the Lipschitz field $\vartheta : P \to \mathcal{G}_{n,k_1}$, $\vartheta(y) = N_y^{s,0}(P)$, is approximated by a field $\mu_1 : P \to \mathcal{G}_{n,k_1}$.

The field $\mu_1$ is constructed by interpolating a given finite set of data points in $\mathcal{G}_{n,k_1}$. These data points are the $k_1$–planes $\{N_y^s(V) : y \in \mathcal{C}^0\}$, where $\mathcal{C}^0$ is the set of vertices of $\mathcal{C}$. The interpolation is performed in the space of frames for the $k_1$–planes of $\mathcal{G}_{n,k_1}$. Since the same procedure is used for $\mu_2$, in the following we will use $k$ to denote a variable which may be $k_1$ or $k_2$. Recall that $\mathcal{F}_{n,k}$, the space of $k$–frames in $\mathbb{R}^n$, $k \leq n$, is given the structure of a smooth manifold by its natural identification with the space of $n \times k$ matrices of rank $k$. The space of $n \times k$ matrices of rank $k$ is a smooth manifold due to its identification with an open subset of $\mathbb{R}^{nk}$, [1].

In the case $k = 1$, the following method may be used to interpolate the $k$–plane fibers at the vertices of a $d$–simplex $\mathcal{C}_i$. Given $d+1$ nearby 1–plane fibers at the vertices of $\mathcal{C}_i$, choose $d + 1$ unit vector bases $b_1 \dots b_{d+1}$ for the fibers, all contained in a small neighborhood in the frame manifold. Then a basis for the interpolating 1–plane fiber at the barycentric coordinates $(t_1, \dots, t_{d+1})$ [7] is obtained by normalizing the vector $v = t_1 \cdot b_1 + \dots + t_{d+1} \cdot b_{d+1}$. This is numerically practical since the nearness of the bases $b_1 \dots b_{d+1}$ implies that $|v|$ is near one.

For the construction of discrete $k$–plane bundles in the case $k > 1$, see [4]. Here, plane rotation matrices are used to interpolate special orthonormal bases for the $k$–plane fibers at the vertices of a $d$–simplex.

Next, a discrete approximation of a section in $\mathcal{S}_{\epsilon,\delta}$ is constructed. The field of $k_0$–planes $\mu : P \to \mathcal{G}_{n,k_0}$ induces a vector bundle $N(P)$ with base space $P$, whose fiber at $x \in P$ is the $k_0$–plane $\mu(x)$. This $N(P)$ gives a tubular neighborhood of $P$. Analogous to the approach in Section 2, we work in a neighborhood of the zero section in $N(P)$, which is equivalent to a neighborhood of $P$ in $\mathbb{R}^n$. Any $C^r$, $r \geq 1$, manifold $\tilde{V}$ Lipschitz–near $V$ corresponds to the graph of a section $\sigma$ of $N(P)$, for small $H$. The section $\sigma$ is $C^r$ on each $\mathcal{C}_i$. A candidate manifold $\tilde{V}$ is approximated by a section $\sigma_D$ of $N(P)$ which is polynomial on each $\mathcal{C}_i$ in appropriate coordinates. On each $\mathcal{C}_i$, $\sigma_D$ is a polynomial map into the fibers of $N(P)$. In fact, $N(P) = N^s(P) \oplus N^u(P)$, where the fiber of $N^s(P)$ at $x \in P$ is the $k_1$–plane $\mu_1(x)$ and the fiber of $N^u(P)$ at $x \in P$ is the $k_2$–plane $\mu_2(x)$. The approximating section is $\sigma_D(x) = (x, v^s(x), v^u(x))$,

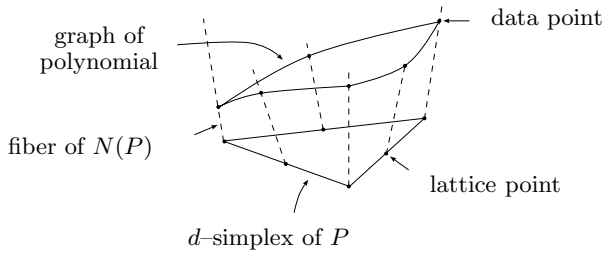



Fig. 3: Approximation to $\tilde{V}$, attracting case, $p = 2$.

where $v^s(x) \in N^s_x(P)$, $v^u(x) \in N^u_x(P)$. In appropriate coordinates, on each $\mathcal{C}_i$, $v^s$ and $v^u$ are Lagrange polynomials of order $p \geq 1$, [7].

The section $\sigma$ on $\mathcal{C}_i$ is approximated by interpolating a discrete data set consisting of the values of $\sigma$ at certain points of $\mathcal{C}_i$. The discrete data set for $\sigma_D$ on $\mathcal{C}_i$ consists of the points of intersection of the graph of $\sigma$ in $N_{\mathcal{C}_i}(P)$ with the fibers $N_x(P)$, for points $x$ in the principal lattice of order $p$ of $\mathcal{C}_i$. See Figure 3. The principal lattice of order $p$ of $\mathcal{C}_i$, denoted $\Sigma_i$, is the set of points in $\mathcal{C}_i$ with barycentric coordinates $b_1 \dots b_{d+1} \in \{0, 1/p \dots (p-1)/p, 1\}$, [7]. Denote the points of $\Sigma_i$ by $x_{i,j} \in \mathcal{C}_i \subset P$, $j = 1 \dots m$. Then the points of intersection of the graph of $\sigma$ in $N_{\mathcal{C}_i}(P)$ with the fibers $N_x(P)$, $x \in \Sigma_i$, are

$$(x_{i,j}, v^s_{i,j}, v^u_{i,j}) \in N_{\mathcal{C}_i}(P), \text{ for some } v^s_{i,j} \in N^s_{x_{i,j}}(P), v^u_{i,j} \in N^u_{x_{i,j}}(P),$$

$j = 1 \dots m$. The discrete section $\sigma_D$ is composed of stable and unstable parts, $v^s(x)$ and $v^u(x)$. Here, $v^s(x)$, $x \in \mathcal{C}_i$, is fitted to $v^s_{i,j}$, $j = 1 \dots m$, and $v^u(x)$, $x \in \mathcal{C}_i$, is fitted to $v^u_{i,j}$, $j = 1 \dots m$.

Coordinates on $N^s_{\mathcal{C}_i}(P)$, $i = 1 \dots N$, are induced by smooth orthonormal moving frames. Namely, an orthonormal basis of $N^s_x(P)$ is given by the columns of an $n \times k_1$ matrix $E_i(x)$ which depends smoothly on $x \in \mathcal{C}_i$. For each $x \in \mathcal{C}_i$, this matrix induces an invertible linear transformation $\xi_i(x) : \mathbb{R}^{k_1} \to N^s_x(P)$, $\xi_i(x)(\rho) = E_i(x)\rho$. There is a unique Lagrange polynomial $\eta^s_i : \mathcal{C}_i \to \mathbb{R}^{k_1}$ of total degree $p$ fitting the data

$$\eta^s_i(x_{i,j}) = \xi_i(x_{i,j})^{-1}(v^s_{i,j}), \ j = 1 \dots m,$$

[7, 8]. Now put $v^s(x) = \xi_i(x) \circ \eta^s_i(x)$ for $x \in \mathcal{C}_i$.

The construction of $v^u$ is analogous to the construction of $v^s$. The resulting approximating section $\sigma_D(x) = (x, v^s(x), v^u(x))$ of $N(P)$ is continuous. If $\tilde{V}$ is of smoothness class $C^{p+1}$, $\sigma_D$ is an approximation to $\sigma$ of order $p$. That is, $\sup\{ |v(x) - v_D(x)|_x : x \in P\} = O(H^{p+1})$ as $H \to 0$, where $\sigma(x) = (x, v(x))$ and $\sigma_D(x) = (x, v_D(x))$.

## 4 The Discrete Graph Transform

In this section the discrete graph transform $\Gamma_D$, used to approximate $\tilde{V}$, is formulated. This is done in Sections 4.1 and 4.2 by replacing the components of the graph transform described in Section 2 with the discrete counterparts of Section 3. Namely, $N(V) = N^u(V) \oplus N^s(V)$ is replaced by $N(P) = N^u(P) \oplus N^s(P)$ in Section 4.1 and the sections $\sigma$ of $N(V)$ are replaced by discrete sections $\sigma_D$ of $N(P)$ in Section 4.2.

In addition, the discrete linear graph transforms $\mathcal{L}_D^u$ and $\mathcal{L}_D^s$, used to approximate the hyperbolic splitting of $\tilde{V}$, are formulated. The approximations of the stable and unstable bundles, $N^s(P)$ and $N^u(P)$, lead to $\mathcal{L}_D^s$ and $\mathcal{L}_D^u$ in Section 4.3.

### 4.1 The Graph Transform of Sections of $N(P)$

In this section, the graph transform is formulated as in Section 2.2, replacing $N(V) = N^u(V) \oplus N^s(V)$ by $N(P) = N^u(P) \oplus N^s(P)$. The difference between this section and Section 2.2 is that here $N(P)$ is Lipschitz rather than smooth.

The Lipschitz constant of a section $\sigma^s$ of $N^s(P)$ is defined as follows. First, $N(P)$ induces a homeomorphism $\psi : V \to P$. Suppose $N^s(V)$ is the vector bundle over $V$ whose fiber at $p \in V$ is $N^s_{\psi(p)}(P)$. Since $N^s(V)$ is a subbundle of $T_V(M)$, the Lipschitz constant of the section $\sigma^s \circ \psi$ of $N^s(V)$ is defined in Section 2. Hence, $\mathrm{Lip}\{\sigma^s\} = \mathrm{Lip}\{\sigma^s \circ \psi\}$, and similarly for $\sigma^u$. Now, $\mathrm{Lip}\{\sigma\}$ for a section $\sigma$ of $N(P)$ is defined as in Section 2.2. Suppose $Z = Z(\epsilon) = \{(x, v) \in N(P) : |v|_x \leq \epsilon\}$ and $\mathcal{S}_{\epsilon,\delta} = \{\sigma : P \to Z : \mathrm{Lip}(\sigma) \leq \delta\}$. The space $\mathcal{S}_{\epsilon,\delta}$ with the $C^0$ norm $\|\cdot\|$ described in Section 2.2 is complete.

Given $\sigma \in \mathcal{S}_{\epsilon,\delta}$, $\sigma(x) = (x, v^s(x), v^u(x))$, the graph transform of $\sigma$ is a section $\Gamma(\sigma)(x) = (x, w^s(x), w^u(x))$ of $N(P)$. Here, $w^s(x)$ is the stable part of the intersection of the $\tilde{F}^0$–image of the graph of $\sigma$ with the fiber $N_x(P)$. Thus, to define $w^s(x)$ for a given $x \in P$, first solve

$$x = \pi \circ \tilde{F}^0(p, v^s(p), v^u(p)), \tag{7}$$

for $p \in P$, where $\pi : N(P) \to P$ is the vector bundle projection. In (7) we are solving for the unique $p \in P$ such that $\tilde{F}^0 \circ \sigma(p)$ is contained in the fiber $Z_x(P)$. Equation (7) has a unique solution for $p \in P$, provided $\epsilon, \delta$, $\theta$ and $H$ are small. Denote this solution by $p = p(x, v^s, v^u)$. Now, $w^s(x)$ is given by the formula

$$w^s(x) = P^s_x \circ \tilde{F}^0(p, v^s(p), v^u(p)), \tag{8}$$

for $x \in P$, where $P^s_x : N_x(P) \to N_x(P)$ is the linear projection with range $N^s_x(P)$ and nullspace $N^u_x(P)$.

The unstable part $w^u$ is defined implicitly by eliminating $x$ in

$$v^u(x) = P^u_x \circ \tilde{F}^0(p, v^s(p), w^u(p)),\ x = \pi \circ \tilde{F}^0(p, v^s(p), w^u(p)), \tag{9}$$

for $p \in P$, where $P_x^u : N_x(P) \to N_x(P)$ is the linear projection with range $N_x^u(P)$ and nullspace $N_x^s(P)$. In (9) we are solving for the vector $w = w^u(p) \in Z_p^u(P)$ such that the $\tilde{F}^0$–image of $(p, v^s(p), w)$ has unstable component in the graph of $v^u$. There is a unique solution for $w^u(p)$ in (9) for small $\epsilon$, $\delta$, $\theta$ and $H$. The proof that there are unique solutions in (7) and (9) follows from the Lipschitz implicit function theorem [12, page 207]. As in Section 2.2, by replacing $\tilde{F}$ with $\tilde{F}^N$ if necessary, $\Gamma$ becomes a contraction on $\mathcal{S}_{\epsilon,\delta}$ whose fixed point gives the $\tilde{F}$–invariant manifold $\tilde{V}$.

### 4.2 The Discrete Graph Transform

In this section, the formulation of $\Gamma_D$ started in Section 4.1 is finished. The domain of $\Gamma$ from Section 4.1 is restricted to the subset of $\mathcal{S}_{\epsilon,\delta}$ consisting of discrete sections. For $\sigma_D \in \mathcal{S}_{\epsilon,\delta}$, where $\sigma_D$ is a discrete section of the form constructed in Section 3, $\Gamma(\sigma_D)$ is not a discrete section. Thus, define $\Gamma_D(\sigma_D) = \mathcal{I} \circ \Gamma(\sigma_D)$, where $\mathcal{I} \circ \sigma$ is the discrete section approximating $\sigma$ described in Section 3. Whether $\Gamma_D$ leaves $\mathcal{S}_{\epsilon,\delta}$ invariant depends on the effect $\mathcal{I}$ has on both the $C^0$ norm and the Lipschitz constant of sections in $\mathcal{S}_{\epsilon,\delta}$.

To be precise, a formula for $\mathcal{I}(\sigma)$ is obtained. A section $\sigma \in \mathcal{S}_{\epsilon,\delta}$ is

$$\sigma(x) = (x,\, \xi_i^s(x) \circ f_i^s(x),\, \xi_i^u(x) \circ f_i^u(x)), \quad x \in \mathcal{C}_i \tag{10}$$

for some $f_i^s : \mathcal{C}_i \to \mathbb{R}^{k_1}$ and $f_i^u : \mathcal{C}_i \to \mathbb{R}^{k_2}$. Here, $\xi_i^s$ and $\xi_i^u$ are defined in Section 3. Recall that $\xi_i^s(x) : \mathbb{R}^{k_1} \to N_x^s(P)$, $\xi_i^s(x)(\rho) = E_i^s(x)\rho$, where the columns of the $n \times k_1$ matrix $E_i^s(x)$ form an orthonormal basis for $N_x^s(P)$, $x \in \mathcal{C}_i$. The description of $\xi_i^u(x)$ is analogous. Recall that $\Sigma_i$, defined in Section 3, is the principal lattice of order $p \geq 1$ of the $d$–simplex $\mathcal{C}_i$. Then $\mathcal{I}(\sigma)$ is the discrete section $\sigma_D$ of $N(P)$ whose data on $\mathcal{C}_i$ consists of the points of intersection of the graph of $\sigma$ in $N_{\mathcal{C}_i}(P)$ with the fibers $N_x(P)$, $x \in \Sigma_i$. To be specific,

$$\mathcal{I}(\sigma)(x) = (x,\, \xi_i^s(x) \circ L_i^s \circ f_i^s(x),\, \xi_i^u(x) \circ L_i^u \circ f_i^u(x))$$

for $x \in \mathcal{C}_i$, where $L_i^s$ and $L_i^u$ are the standard Lagrange interpolation operators on functions on $\mathcal{C}_i$. Here, the Lagrange interpolation operators are defined as follows. Given $f : \mathcal{C}_i \to \mathbb{R}^{k_1}$, $L_i^s \circ f : \mathcal{C}_i \to \mathbb{R}^{k_1}$ is the unique polynomial of total degree $p$ with $L_i^s \circ f(x) = f(x)$ for $x \in \Sigma_i$. The definition of $L_i^u$ is analogous.

The maximum factor of growth of the $C^0$ norm of a section under $\mathcal{I}$ is $C_p = \sup\{\, \|\mathcal{I}(\sigma)\|/\|\sigma\| : \sigma \in \mathcal{S}_{\epsilon,\delta}\}$. The maximum factor of growth of the Lipschitz constant of a section under $\mathcal{I}$ is $C_p' = \sup\{\, \mathrm{Lip}\{\mathcal{I}(\sigma)\}/\mathrm{Lip}\{\sigma\} : \sigma \in \mathcal{S}_{\epsilon,\delta}\}$. Here, $C_p$ and $C_p'$ are bounded as $H \to 0$. The Lipschitz constant of $\mathcal{I}$ is also bounded by $C_p$ for $p \geq 1$. If $C_p = C_p' = 1$, $\mathcal{I}$ has no deleterious effect on $\Gamma$, and $\Gamma_D$ is a contraction on $\mathcal{S}_{\epsilon,\delta}$ with no adjustments to any parameters. In general, however, $C_p, C_p' > 1$. Note that $C_p$ and $C_p'$ are smaller for smaller $p \geq 1$. Even for $p = 1$, though, $C_p' > 1$.

To deal with $C_p > 1$ or $C'_p > 1$, one of the parameters of $\Gamma$ is modified. For simplicity, consider the attracting case. Suppose that $0 < \alpha < 1$ is the factor of (weakest) normal contraction toward $V$ under $F$. Also, $0 < \mu < 1$ from (2) is a bound on $\alpha/\{$the factor of (strongest) tangential contraction under $F\}$. Given $\sigma \in \mathcal{S}_{\epsilon,\delta}$, the $C^0$ norm and Lipschitz constant of $\Gamma(\sigma)$ are multiplied by factors $c\alpha^N + o(1)$ and $c\mu^N + o(1)$, respectively, as $\epsilon + \delta + \theta + H \to 0$. The $C^0$ norm and Lipschitz constant of $\Gamma_D(\sigma)$ are multiplied by factors $C_p\, c\alpha^N + o(1)$ and $C'_p\, c\mu^N + o(1)$, respectively. Thus, by choosing $N$ large enough, we obtain $\Gamma_D : \mathcal{S}_{\epsilon,\delta} \to \mathcal{S}_{\epsilon,\delta}$. Also, $\Gamma_D$ is a contraction since

$$\mathrm{Lip}\{\Gamma_D\} \le \mathrm{Lip}\{\mathcal{I}\}\mathrm{Lip}\{\Gamma\} = C_p\, c\alpha^N + o(1)$$

as $\epsilon + \delta + \theta + H \to 0$.

Alternatively, it is possible to estimate $\mathrm{Lip}\{\mathcal{I}(\sigma)\}$ using the constant $C''_p = H \sup\{\, \mathrm{Lip}\{\mathcal{I}(\sigma)\}/\|\sigma\| : \sigma \in \mathcal{S}_{\epsilon,\delta}\}$, which is bounded as $H \to 0$. In this case, there exists a constant $c > 0$ and a positive function $\omega(H) \to 0$ as $H \to 0$, such that the following holds. If $\epsilon = cH\delta$, $\omega(H) < c\delta$, $\theta < c\epsilon$, $\delta$ is sufficiently small and $N$ sufficiently large, then $\Gamma_D : \mathcal{S}_{\epsilon,\delta} \to \mathcal{S}_{\epsilon,\delta}$ is a contraction [2]. This result does not use the full hypothesis of normal hyperbolicity, but only the existence of a $C^1$, 0–normally hyperbolic manifold $\tilde{V}$, [21]. This explains why $\Gamma_D$ is a contraction, in practice, for some dynamical systems even in the absence of normal hyperbolicity.

In either of the scenarios in the preceding two paragraphs, $\Gamma_D$ has a fixed point $\sigma^*_D \in \mathcal{S}_{\epsilon,\delta}$, where $\phi \circ \sigma^*_D(P) \to \tilde{V}$ in $C^0$ norm as $H \to 0$. In fact, $\phi \circ \sigma^*_D(P) \to \tilde{V}$ in Lipschitz norm as $H \to 0$ if $p = 1$ or $r \ge 2$. In addition, if $\tilde{V}$ is of smoothness class $C^{p+1}$, then $\phi \circ \sigma^*_D(P)$ is a $C^0$ approximation to $\tilde{V}$ of order $p$.

### 4.3 The Discrete Linear Graph Transform

This section deals with the computation of the approximate hyperbolic splitting of $\tilde{V}$. In Section 4.2, an approximation $\phi \circ \sigma^*_D(P)$ to $\tilde{V}$ was obtained for $H \to 0$. The simplicial complex $\tilde{\mathcal{C}}$ with vertices $\phi \circ \sigma^*_D(\mathcal{C}^0)$, where $\mathcal{C}^0$ is the set of vertices of $P$, supports the manifold $\phi \circ \sigma^*_D(P)$. Suppose $\tilde{P} \subset \mathbb{R}^n$ is the polyhedron of $\tilde{\mathcal{C}}$ and $N(\tilde{P})$ is a given transverse bundle. Given such an $N(\tilde{P})$, the approximate hyperbolic splitting of $\tilde{V}$ is given by a splitting $N(\tilde{P}) = N^u(\tilde{P}) \oplus N^s(\tilde{P})$.

In this section, the discrete linear graph transforms $\mathcal{L}^u_D$ and $\mathcal{L}^s_D$ are used to determine $N^u(\tilde{P})$ and $N^s(\tilde{P})$. Here it is assumed that $N(\tilde{P})$ and $N(P)$ are approximately normal in the following sense. Each $d$–simplex subspace $P_i$, $i = 1 \ldots N$, of $P$ is a manifold with boundary with tangent bundle $T(P_i)$. Then

$$\inf\{\sphericalangle N_x(P), T_x(P_i) :\ \text{all } P_i \text{ containing } x,\ x \in P\} \to \pi/2$$

as $H \to 0$. Next, $\mathcal{L}^u_D$ is formulated. The formulation of $\mathcal{L}^s_D$ is analogous.

The initial data for $\mathcal{L}_D^u$ is a splitting $N(\tilde{P}) = N^{u,0}(\tilde{P}) \oplus N^{s,0}(\tilde{P})$. This splitting is obtained from $N(P) = N^u(P) \oplus N^s(P)$ by parallel translation followed by projection onto the fibers of $N(\tilde{P}) \subset T_{\tilde{P}}(\mathbb{R}^n)$ using $Q$, as in Section 2.3. To be specific, suppose $\pi$ is the vector bundle projection of $N(P)$. Then $N_y^{u,1}(\tilde{P})$, $N_y^{s,1}(\tilde{P})$ are obtained from $N_p^u(P)$, $N_p^s(P)$, $p = \pi \circ \phi^{-1}(y)$, by parallel translation $T_p(\mathbb{R}^n) \to T_y(\mathbb{R}^n)$ along $\phi$–images of fibers of $N(P)$. In the present case, parallel translation is trivially defined by the identification of $T_x(\mathbb{R}^n)$, $x \in \mathbb{R}^n$, with the ambient space $\mathbb{R}^n$. In the present setting,

$$Q : T_{\tilde{P}}(\mathbb{R}^n) \to N(\tilde{P}) \subset T_{\tilde{P}}(\mathbb{R}^n),$$

is, on each fiber $T_x(\mathbb{R}^n)$, the linear orthogonal projection with range $N_x(\tilde{P})$. The initial data are then

$$N^{u,0}(\tilde{P}) = Q(N^{u,1}(\tilde{P})),\ N^{s,0}(\tilde{P}) = Q(N^{s,1}(\tilde{P})).$$

This procedure produces non-degenerate initial data for $\epsilon + \delta + \theta + H \to 0$.

As in Section 2.3, $L(\tilde{P})$ is the bundle whose fiber at $y \in \tilde{P}$ is the space of linear transformations $N_y^{u,0}(\tilde{P}) \to N_y^{s,0}(\tilde{P})$. The domain of $\mathcal{L}_D^u$ is a subset of the space of sections $\mathcal{S}_\eta = \{\sigma : \tilde{P} \to L(\tilde{P}) : \sup_y \|\sigma(y)\| \leq \eta\}$, where the operator norm $\|\cdot\|$ is associated with the Riemann structure on $T_{\tilde{P}}(\mathbb{R}^n)$. The space $\mathcal{S}_\eta$ is complete with respect to the norm $|\sigma| = \sup_y \|\sigma(y)\|$.

The domain of $\mathcal{L}_D^u$ is the subset of $\mathcal{S}_\eta$ consisting of discrete sections. A discrete section in $\mathcal{S}_\eta$ is constructed using the construction of a discrete field of $k_2$–planes $\mu : \tilde{P} \to \mathcal{G}_{n,k_2}$ in Section 3. A discrete section $\sigma_D$ of $L(\tilde{P})$ is constructed from given data $\{\sigma_D(x) \in L_x(\tilde{P}) : x \in \tilde{\mathcal{C}}^0\}$, where $\tilde{\mathcal{C}}^0$ is the set of vertices of $\tilde{P}$, as follows. Using the method of Section 3, construct the field $\mu : \tilde{P} \to \mathcal{G}_{n,k_2}$ of $k_2$–planes determined by the set of $k_2$–plane data points

$$\left\{\text{graph}\{\sigma_D(x)\} \subset N(\tilde{P}) \subset T_{\tilde{P}}(\mathbb{R}^n) : x \in \tilde{\mathcal{C}}^0\right\}.$$

The discrete section $\sigma_D$ is then uniquely characterized by $\text{graph}\{\sigma_D(x)\} = \mu(x)$, $x \in \tilde{P}$.

To construct $\mathcal{L}_D^u$, first the linear graph transform $\mathcal{L}^u$ is formulated in the present setting, replacing $N(\tilde{V})$ by $N(\tilde{P})$. Thus, instead of a smooth manifold and transverse bundle, here they are only Lipschitz. In addition, the formulation of $\mathcal{L}^u$ in this section is slightly different from the formulation of $\mathcal{L}^u$ in Section 2.3 because $\tilde{P}$ is not $\tilde{F}$–invariant. Second, the domain of $\mathcal{L}^u$ is restricted to discrete sections, $\mathcal{L}_D^u(\sigma_D) = \mathcal{I} \circ \mathcal{L}^u(\sigma_D)$, $\sigma_D \in \mathcal{S}_\eta$. Here, for $\sigma \in \mathcal{S}_\eta$, $\mathcal{I}(\sigma)$ is the discrete section of $L(\tilde{P})$ defined by the data $\{\sigma(x) : x \in \tilde{\mathcal{C}}^0\}$.

To formulate $\mathcal{L}^u$, the invariance condition is derived. To define the mapping $\Phi$ induced by $D\tilde{F}$ on $N(\tilde{P})$, suppose $\pi$ is the vector bundle projection of $N(\tilde{P})$ and $\phi : Z \to U$ is the homeomorphism, defined in Section 2.1, associated with the tubular neighborhood of $\tilde{P}$ induced by $N(\tilde{P})$. Then the linear mapping induced by $D\tilde{F}_x : T_x(\mathbb{R}^n) \to T_y(\mathbb{R}^n)$, $y = \tilde{F}(x)$, $x \in \tilde{P}$, on $N(\tilde{P})$ is

$$\Phi = Q \circ \gamma \circ D\tilde{F}|_{N(\tilde{P})} : N(\tilde{P}) \to N(\tilde{P}).$$

Here $\gamma : T_y(\mathbb{R}^n) \to T_p(\mathbb{R}^n)$, $p = \pi \circ \phi^{-1}(y)$, $y \in U$, is parallel translation. Note that $y \in U$ for small $H$ because $\tilde{P} \to \tilde{V}$ in $C^0$ norm as $H \to 0$.

Given a section $\sigma \in \mathcal{S}_\eta$, the linear graph transform $\mathcal{L}^u(\sigma)$ is characterized by the condition $\Phi(\text{graph}\{\sigma(x)\}) = \text{graph}\{\mathcal{L}^u(\sigma)(y)\}$ where $y = \pi \circ \phi^{-1} \circ \tilde{F}(x)$. To calculate $\mathcal{L}^u(\sigma)(y)$ for a given $y \in \tilde{P}$, first solve $y = \pi \circ \phi^{-1} \circ \tilde{F}(x)$ for $x \in \tilde{P}$. Next, given an orthonormal basis $e_1 \dots e_{k_2}$ for $N_y^{u,0}(\tilde{P})$, solve $e_i = P_y^u \circ \Phi(\rho_i, \sigma(x)(\rho_i))$ for $\rho_i \in N_x^{u,0}(\tilde{P})$, $i = 1 \dots k_2$. Then $\mathcal{L}^u(\sigma)(y)$ is given by the formula

$$\mathcal{L}^u(\sigma)(y)(e_i) = P_y^s \circ \Phi(\rho_i, \sigma(x)(\rho_i)),$$

$i = 1 \dots k_2$. If $\Phi$ is replaced by $\Phi^N$, then $\mathcal{L}^u : \mathcal{S}_\eta \to \mathcal{S}_\eta$ is a contraction for $\epsilon + \delta + \theta + \eta + H$ small and $N$ large.

Next, conditions are determined which guarantee $\mathcal{L}_D^u(\sigma_D) \in \mathcal{S}_\eta$ for $\sigma_D \in \mathcal{S}_\eta$ and that $\mathcal{L}_D^u : \mathcal{S}_\eta \to \mathcal{S}_\eta$ is a contraction. Recall $\mathcal{L}_D^u(\sigma_D) = \mathcal{I} \circ \mathcal{L}^u(\sigma_D)$ for $\sigma_D \in \mathcal{S}_\eta$. Thus, the norm of $\mathcal{I}(\sigma)$, $\sigma \in \mathcal{S}_\eta$ and the Lipschitz constant of $\mathcal{I}$ on $\mathcal{S}_\eta$ must be estimated. For $\sigma \in \mathcal{S}_\eta$, $|\mathcal{I}(\sigma)| \leq \eta + o(1)$ and $\text{Lip}\{\mathcal{I}\} = 1 + o(1)$ as $H \to 0$. Thus, $\mathcal{L}_D^u : \mathcal{S}_\eta \to \mathcal{S}_\eta$ is a contraction for $\epsilon + \delta + \theta + \eta + H$ small and $N$ large.

The fixed point $\sigma_D^* \in \mathcal{S}_\eta$ of $\mathcal{L}_D^u$ gives an approximation to $N^u(\tilde{V})$ in the following sense. Suppose $\gamma : N_x(\tilde{V}) \to N_y(\tilde{P})$, $y = \pi \circ \phi^{-1}(x)$, is parallel translation and $\sigma$ is a section of $L(\tilde{P})$ satisfying $\text{graph}\{\sigma(y)\} = \gamma(N_x^u(\tilde{V}))$, $y = \pi \circ \phi^{-1}(x)$, $y \in \tilde{P}$. Then $|\sigma - \sigma_D^*| \to 0$ as $H \to 0$.

## 5 Numerical Implementation

In this section, a specific computer implementation of the discrete graph transform is outlined. In Section 5.1, a practical numerical approach for solving equations (7), (8) and (9) is proposed. The main part is solving (7), as well as the second equation in (9), for a point $p \in V$. Note that this is a global problem. In Section 5.2, numerical conditioning and error for these problems is discussed. Also, some important smoothing techniques are mentioned. These are useful for stabilizing a computation in which non-smooth data appears.

The discrete graph transform/linear graph transform algorithm takes as input an approximation to $V$ and its hyperbolic splitting. It returns as output an approximation to $\tilde{V}$ and its hyperbolic splitting. Then, the algorithm may be repeated taking as input the newly computed data. In practice, the input/output to the algorithm are the following: $(i)$ A polyhedron $P$ Lipschitz–near a $C^r$ $F$–invariant submanifold $V \subset \mathbb{R}^n$, $r \geq 1$. $(ii)$ Approximately normal fibers $N_x(P)$, $x \in \mathcal{C}^0$ = the vertices of $P$, and a splitting $N_x(P) = N_x^u(P) \oplus N_x^s(P)$, $x \in \mathcal{C}^0$, which is near the hyperbolic splitting.

The graph transform algorithm, which returns as output an approximation to $\tilde{V}$, is the subject of Section 5.1. The linear graph transform algorithm, which returns as output an approximation to the hyperbolic splitting of $\tilde{V}$, will not be discussed further here. It is less complicated than the graph transform algorithm since it presents no additional nonlinear equations to solve.

### 5.1 The Discrete Graph Transform Algorithm

The graph transform algorithm starts with the zero section $\sigma_D^0$ of $Z(P)$ and for $i \geq 0$ repeats (graph transform step) until the convergence criteria are met. The graph transform step takes as input a discrete section $\sigma_D^i$ of $Z(P)$ and returns as output a discrete section $\sigma_D^{i+1} = \Gamma_D \circ \sigma_D^i$ of $Z(P)$. Here, $Z(P) = \{(x, v) \in N(P) : |v|_x \leq \epsilon\}$ is from Section 4.1 and $\Gamma_D$ is from Section 4.2. The convergence criteria for the graph transform are the following. The iteration of (graph transform step) is stopped when $|\sigma_D^{i+1} - \sigma_D^i| < error$ and the contraction factor $|\sigma_D^{j+2} - \sigma_D^{j+1}|/|\sigma_D^{j+1} - \sigma_D^j| < 1$ is approximately constant for all $j < i$ sufficiently large [5].

The graph transform step consists of the following. Recall that $\Sigma_i$, defined in Section 3, is the principal lattice of order $p \geq 1$ of the $d$–simplex $\mathcal{C}_i$. A discrete section of $Z(P)$ is determined by a discrete set of data points, one in each fiber $Z_x(P)$, $x \in G = \bigcup\{\Sigma_i : i = 1 \ldots N\} \subset P$. Thus for the graph transform step, the *input* is the set of data points $\sigma_D^i(x)$, $x \in G$, and the *output* is the set of data points $\sigma_D^{i+1}(x) = (\Gamma_D \circ \sigma_D^i)(x)$, $x \in G$. The sections have stable and unstable parts, $\sigma_D^i(x) = (x, v^{s,i}(x), v^{u,i}(x))$ and $\sigma_D^{i+1}(x) = (x, v^{s,i+1}(x), v^{u,i+1}(x))$. Hence, the graph transform step has two independent stages, one for determining the stable part $v^{s,i+1}(x)$, $x \in G$ and one for determining the unstable part $v^{u,i+1}(x)$, $x \in G$.

Some notation used below is $\phi$, defined in Section 2.1 and $\tilde{F}^0 = \phi^{-1} \circ \tilde{F} \circ \phi$, defined in Section 2.2.

**Graph transform step:** *Stable part*
For $x \in G$:
1. Put $v^s = v^{s,i}, v^u = v^{u,i}$ in (7) and (8).
2. Solve (7) for $p \in P$.
   2.1 Determine a neighborhood containing $p \in P$.
   $A_j \equiv \cup\{\mathcal{C}_k : \mathcal{C}_k \cap \mathcal{C}_j \neq \emptyset\}$ for $j = 1 \ldots N$.
   Find $j^* \in \{1 \ldots N\}$ with $\tilde{F}^0 \circ \sigma_D^i(A_{j^*}) \cap Z_x(P) \neq \emptyset$.
   (a) $\mathcal{C}_j^0 \equiv$ vertices of $\mathcal{C}_j$, $j = 1 \ldots N$.
   (b) $B_j \equiv d$–simplex with vertices $\phi \circ \tilde{F}^0 \circ \sigma_D^i(\mathcal{C}_j^0)$, $j = 1 \ldots N$.
   (c) For $j = 1 \ldots N$: Test $B_j \cap \phi(Z_x(P)) \neq \emptyset$. If true, return $j = j^*$.
   2.2 Locate $p \in A_{j^*}$ to a desired tolerance.
   (a) Search for $p$ in each $\mathcal{C}_k \subset A_{j^*}$ using a standard root finding method [14].
   (b) If no root found in (a), search $\mathcal{C}_k$ in successively larger regions around $A_{j^*}$.

3. Evaluate (8) at $p$ to obtain $v^{s,i+1}(x) = w^s(x)$.

In 2.1, a simple geometrical test is used to find $A_{j^*}$. This step is typically only necessary for $i = 0$, the same $j^*$ may be used for $i > 0$, since the location of $p \in P$ may not change much as $i$ increases. The approach in 2.1 is justified by the fact that $\sigma_D^i$ is kept approximately flat over $\mathcal{C}_j$ and $\tilde{F}^0$ is well approximated by its linearization over the set $\sigma_D^i(\mathcal{C}_j)$ as $H \to 0$.

**Graph transform step:** *Unstable part*
For $p \in G$:
1. Put $v^s = v^{s,i}, v^u = v^{u,i}$ in (9).
2. Solve (9) for $w = w^u(p) \in Z_p^u(P)$.
   Comment: Use a standard root finding method [14] with initial guess $w = 0$. Function evaluations in the root finding method require a call to the following subroutine.
   2.1 Given $w \in Z_p^u(P)$, solve the second equation in (9) for $x = x(w) \in P$.
      (a) $y \equiv \phi \circ \tilde{F}^0(p, v^{s,i}(p), w)$.
      (b) $x \in P$ is the point near $y$ with $y - x$ parallel to $\phi(Z_x(P))$. There are two stages to solving for $x$, similar to *Stable part* step 2.
3. Put $v^{u,i+1}(p) = w$.

### 5.2 Numerical Conditioning and Smoothing Techniques

The global equations (7), (8) and (9) associated with the graph transform pose a numerically well-conditioned problem. To be specific, solving (7) for $p \in P$ is numerically optimally conditioned for $N(P)$ chosen perpendicular to $V$, as is evaluation of the second equation of (9). In practice, $N(P)$ is an approximate normal bundle in the sense of Section 4.3. In the evaluation of (8) at $p$, hyperbolicity damps the numerical discretization and rounding error. Solving (9) for $w^u$ is a well-conditioned problem. This is because the normal hyperbolicity of $V$ implies that small errors in $w^u$ produce large deviations in the right hand side of the first equation of (9).

As discussed in Section 4.2, it may be necessary to control the Lipschitz constant of discrete sections $\sigma_D(x) = (x, v^s(x), v^u(x))$, $x \in P$. The Lipschitz constant of sections is effectively controlled in practice using two techniques. The first is even redistribution of the grid points $G$. This replaces $P$ with a nearby polyhedron $P'$ with each $\mathcal{C}_i \subset P'$ close to the shape of the standard $d$–simplex. The second technique is local fairing [11] of the data $v^s(x) \in N_x^s(P)$ and $v^u(x) \in N_x^u(P)$, $x \in \Sigma_i$, which smooths out graph$\{\sigma_D\}$. Consider for example the attracting case. Here, the data $\sigma_D^i(x) \in Z_x(P)$, $x \in \Sigma_i$, is tested for large deviations. If an undesirable data point $\sigma_D^i(x^*)$ is detected, it is replaced by the average of $\sigma_D^i(x)$, $x \neq x^*$, $x \in \Sigma_i$. To be precise, the average $y \in \mathbb{R}^n$ of $\phi \circ \sigma_D^i(x) \in \mathbb{R}^n$, $x \neq x^*$, $x \in \Sigma_i$, is obtained. Then, $y$ is projected onto the affine $k_1$–plane $\phi \circ Z_x^*(P)$ to obtain $z \in \phi \circ Z_x^*(P) \subset \mathbb{R}^n$. The data point $\sigma_D^i(x^*)$ is replaced by $\phi^{-1}(z)$. Prior to these steps, it is important to

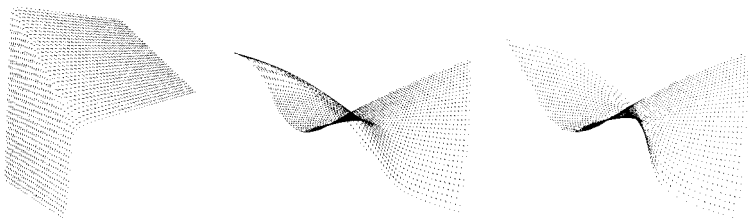


Fig. 4: Enzyme reaction surfaces: left $k_p = 0.1, k_1 = 10^3$; middle $k_p = 0.1, k_1 = 1.0$; right $k_p = 1.0, k_1 = 1.0$.

use local averaging of the fibers of $N(P)$, to make $N_x(P)$, $x \in \mathcal{C}_i$, more nearly parallel. For each $x \in \mathcal{C}^0$, $N_x(P)$ is replaced by the average of the $N_y(P)$ for $y \in \mathcal{C}^0$ near $x$. This is sometimes necessary because, in practice, small bumps in $P$ can introduce degeneracies in its approximate normal bundle $N(P)$.

## 6 An Application

This section deals with a problem of chemical kinetics. The 'slow–transient' surface of an enzyme reaction is computed for a variety of parameter values. This application requires a modification to the algorithm of Section 5. This modification allows the computation of just a part of an invariant manifold. This is a necessary adaptation in cases where the invariant manifold is so large that its data cannot be held in computer memory.

The 'slow–transient' surface, in the phase space of chemical species concentration variables, is useful in chemical kinetics for model reduction. After a short time interval, the $n$–tuple of chemical species concentrations is restricted to the surface, at least for experimentally measurable tolerances. The dynamics of the reaction after this short time interval is described by the dynamics on the surface. In principle, once this surface is known, the system may be reduced to a 2D system on the surface. In chemical kinetics, the steady state and equilibrium approximations, as well as variations on these, have been used to approximate the slow–transient surface [13]. These approximations are typically valid in limiting cases.

In the enzyme reaction model

$$\begin{aligned} \dot{s} &= -k_1(e_0 - c - q)\, s + k_{-1}\, c \\ \dot{c} &= k_1(e_0 - c - q)\, s - (k_{-1} + k_2)\, c + k_{-2}\, q\,, \quad (s, c, q) \in \mathbb{R}^3, \\ \dot{q} &= k_2\, c - (k_{-2} + k_p)\, q \end{aligned} \tag{11}$$

the variables $s, c$ and $q$ are the concentrations of different chemical species undergoing chemical reaction [30]. Here, $k_1, k_{-1}, k_2, k_{-2}, k_p > 0$ are the rate constants and $e_0 > 0$ is the concentration of the enzyme, taken to be constant. The attracting equilibrium is 0 in the physical region $\{0 \leq s < \infty,\, c + q \leq e_0,\, 0 \leq c, q\} \subset \mathbb{R}^3$. In Figure 4, the part of the slow–transient surface in the physical region restricted to $\{0 \leq s \leq 2\}$ is computed for three parameter choices. In every case, $e_0 = 1.0$, $k_{-1} = 1.0$, $k_2 = 1.0$ and $k_{-2} = 1.0$. The middle surface is computed by alternate means in [30].

In the present example, the dynamics are described by a nested hierarchy of attracting invariant manifolds in 3D. This is an equilibrium point contained in a curve contained in a surface, the slow–transient surface, which separates the physical region of phase space. The rate of attraction toward the surface is faster than toward the curve in the surface. The rate of attraction toward the curve in the surface is faster than toward the point in the curve. The part of the slow–transient surface in the physical region restricted to $\{0 \leq s \leq 2\}$ is a manifold with boundary $S$. A technical obstacle here is that $S$ is only part of an invariant surface and is not overflowing invariant. For a diffeomorphism $F$, a compact manifold with boundary $S$ is overflowing invariant under $F$ if $S \subset F(S^0)$, where $S^0 = S \setminus \partial S$ is the interior of $S$. For such manifolds, the graph transform works in principle with no modification [12]. For the present example, a modification to the general purpose algorithm presented in Section 5 is required. Namely, local extrapolation of $S$ at its boundary is used after each graph transform step. This means the following. In the present case, the order of approximation is $p = 1$. Thus, the output data of a graph transform step is $\sigma^i_D$ where $\text{graph}\{\sigma^i_D\} = P$ is a polyhedral manifold with boundary. The $d$–simplices of $P$ whose points are on the boundary of $P$ are flatly extended to form a slightly larger polyhedron $P' \supset P$. This $P'$ is used as input to the next graph transform step. For other approaches to computing the slow–transient surface in chemical kinetics, see [15, 16, 30].

*Acknowledgement.* This work is partially supported by the Netherlands Organisation for Scientific Research (NWO), project nr. 613-02-201.

## References

1. W.M. Boothby: *An Introduction to Differentiable Manifolds and Riemannian Geometry* (Academic Press, New York 1975)
2. H.W. Broer, A. Hagen, G. Vegter: Multiple purpose algorithms for invariant manifolds. *Dynam. Contin. Discrete Implus. Systems B* **10**, 331–44 (2003)
3. H.W. Broer, A. Hagen, G. Vegter: Numerical approximation of normally hyperbolic invariant manifolds. In: *Proceedings of the 4th AIMS meeting 2002 at Wilmington, DCDS 2003,* supplement volume, ed. by S. Hu (AIMS Press, Springfield MO 2003)
4. H.W. Broer, A. Hagen, G. Vegter: *Numerical continuation of invariant manifolds.* Preprint (2006)

5. H.W. Broer, H.M. Osinga, G. Vegter: Algorithms for computing normally hyperbolic invariant manifolds. *Z. angew. Math. Phys.* **48**, 480–524 (1997)
6. S. Cairns: A simple triangulation method for smooth manifolds. *Bull. Amer. Math. Soc.*, **67**, 389–90 (1961)
7. G. Carey, J. Oden: *Finite Elements,* vol 3 (Prentice-Hall, New Jersey 1984)
8. P.G. Ciarlet, P. Raviart: General Lagrange and Hermite interpolation in $\mathbb{R}^n$ with applications to finite element methods. *Arch. Rational Mech. Anal.* **46**, 177–99 (1972)
9. M. Dellnitz, G. Froyland, O. Junge: The algorithms behind GAIO-set oriented numerical methods for dynamical systems. In: *Ergodic Theory, Analysis and Efficient Simulation of Dynamical Systems,* ed. by B. Fiedler (Springer, Berlin 2001)
10. L. Dieci, J. Lorenz: Computation of invariant tori by the method of characteristics. *SIAM J. Numer. Anal.* **32**, 1436–74 (1995)
11. G. Farin: *Curves and Surfaces for Computer-Aided Geometric Design: a practical guide* (Academic Press, New York 1997)
12. N. Fenichel: Persistence and smoothness of invariant manifolds for flows. *Indiana Univ. Math. J.* **21**, 193–226 (1971)
13. S.J. Fraser: The steady state and equilibrium approximations: a geometrical picture. *J. Chem. Phys.* **88**, 4732–8 (1988)
14. G. Golub, J.M. Ortega: *Scientific Computing: an introduction with parallel computing* (Academic Press, San Diego 1993)
15. A.N. Gorban, I.V. Karlin, A.Yu. Zinovyev: Constructive methods of invariant manifolds for kinetic problems. *Physics Reports* **396**, 197–403 (2004)
16. A.N. Gorban, I.V. Karlin, A.Yu. Zinovyev: Invariant grids for reaction kinetics. *Physica A* **333**, 106–54 (2004)
17. A. Haro, R. De La Llave: *A parametrization method for the computation of invariant tori and their whiskers in quasi-periodic maps: numerical implementation and examples.* Preprint (2005)
18. A. Hagen: Hyperbolic Structures of Time Discretizations and the Dependence on the Time Step. Ph.D. Thesis, University of Minnesota, Minnesota (1996)
19. A. Hagen: Hyperbolic trajectories of time discretizations. *Nonlinear Anal.* **59**, 121–32 (2004)
20. M.W. Hirsch: *Differential Topology* (Springer, Berlin Heidelberg New York 1994)
21. M.W. Hirsch, C.C. Pugh, M. Shub: *Invariant Manifolds* (Springer, Berlin Heidelberg New York 1977)
22. B. Krauskopf, H.M. Osinga: Computing geodesic level sets on global (un)stable manifolds of vector fields. *SIAM J. Appl. Dyn. Sys.* **4**, 546–69 (2003)
23. Y. Kuznetsov: *Elements of Applied Bifurcation Theory* (Springer, Berlin Heidelberg New York 1998)
24. S. Lang: *Introduction to Differentiable Manifolds* (Springer, Berlin Heidelberg New York 2002)
25. D. Martin: *Manifold Theory: an introduction for mathematical physicists.* (Ellis Horwood Limited, England 2002)
26. C. Maunder: *Algebraic Topology* (Van Nostrand Reinhold, London 1970)
27. H.M. Osinga: Computing Invariant Manifolds. Ph.D. Thesis, University of Groningen, The Netherlands (1996)
28. J. Palis, F. Takens: *Hyperbolicity & Sensitive Chaotic Dynamics at Homoclinic Bifurcations* (Cambridge University Press, Cambridge 1993)

29. M. Phillips, S. Levy, T. Munzner: Geomview: an interactive geometry viewer. *Notices of the Amer. Math. Soc.* **40**, 985–8 (1993)
30. M.R. Roussel, S.J. Fraser: On the geometry of transient relaxation. *J. Chem. Phys.* **94**, 7106–13 (1991)
31. D. Ruelle: *Elements of Differentiable Dynamics and Bifurcation Theory* (Academic Press, Boston 1989)
32. Y. Wong: Differential geometry of grassmann manifolds. *Proc. NAS* **57** 589–94 (1967)

# Covering an Invariant Manifold with Fat Trajectories

M. E. Henderson

IBM Research Division, T. J. Watson Research Center, Yorktown Heights, NY 10598, USA, mhender@watson.ibm.com,
www.research.ibm.com/people/h/henderson/

**Summary.** Invariant manifolds are important objects in the study of dynamical systems, as well as several applications. They are challenging to compute because even in simple systems they can be very complicated surfaces, demanding adaptive schemes to deal with large curvatures.

This paper describes a method that represents the invariant manifold as a set of circular disks in the tangent space, projected onto the manifold which overlap and cover the manifold. These disks are found by integrating fat trajectories, which add tangent and curvature information to the usual point in phase space, and integrates these quantities along a trajectory.

Using a covering eliminates the usual problems with advancing front approaches, and the dual of the covering is a triangulation, should one be needed.

## 1 Introduction

One of the more important concepts in dynamical systems is that of an invariant manifold. By dynamical system we mean a flow in a phase space $\mathbb{R}^n$

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x), \qquad x \in \mathbb{R}^n \tag{1}$$

An *invariant set* of points in phase space is such that points on the trajectory $t \in (-\infty, \infty)$ passing through any point in the set are also in the set. So a collection of trajectories through a discrete set of initial points is an invariant set. If the initial points lie on a smooth curve and the flow $f(x)$ is smooth, the invariant set will be a smooth surface. This is a consequence of the smooth dependance of trajectories on initial conditions (Figure 1.)

Any point in the initial set can be moved forward or backward along a trajectory without changing the invariant manifold, so the curve of initial points defining an invariant manifold is not unique, and a smooth invariant manifold need not be defined by a smooth curve. If a smooth curve of initial points $M_0$ can be found for an invariant manifold $M$, $M_0$ is called a *global transversal* of $M$.

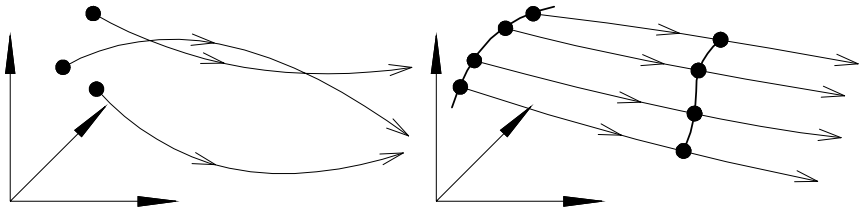


Fig. 1: Invariant sets. (left) defined by a set of points, (right) defined by points on a smooth curve.

For a large class of invariant manifolds $M_0$ is part of the definition of the manifold. For example, the unstable invariant manifold of a hyperbolic fixed point is the image of a small ball in the unstable eigenspace of the fixed point. However, there are interesting invariant manifolds for which finding a global transversal is part of the problem. A periodic orbit can be found by finding a fixed point of a Poincar'e return map. The fixed point is a global transversal. Not all invariant tori have a closed global transversal, but a torus which contains a quasiperiodic motion does, and a global transversal can be found which is an invariant circle of a return map that is similar to the Poincaré return map [18], [20]. There are better ways to compute periodic orbits [3] and quasiperiodic tori [17], [19], which find $M_0$ and $M$ together by solving a larger nonlinear system. The literature on all these problems is extensive, and the citations above are not meant to be exhaustive.

Some commonly computed invariant manifolds are summarized in Table 1.

| Motion | Geometry | $M_0$ |
|---|---|---|
| Fixed Pt. | Point | Point |
| Periodic Motion | Closed Curve | Point |
| Heteroclinic Motion | Curve connecting two Fixed Pts. | Point |
| Quasiperiodic Motion | Torus | Closed curve |
| Unstable manifold of hyperbolic equilibrium | $\mathbb{R} \times (k-1)$ Sphere | $(k-1)$ Sphere |
| Inertial manifold | Attracting $k$d manifold | $(k-1)$d manifold |

Table 1: Some common invariant manifolds and the manifolds of starting points which define them.

Certain complex behaviors in dynamical systems are associated with particular configurations of invariant manifolds. However, they can also be useful

by themselves. In orbital mechanics, for instance, it has been proposed [5], [10] to use the unstable manifold of periodic motions (orbits) to design trajectories for spacecraft "missions". These trajectories start in an unstable orbit about a planet or Lagrange point. To stay in the unstable orbit small thrusts are needed. By choosing an appropriate point and direction on the unstable manifold and burning a small amount of fuel, the vehicle can coast along a trajectory on the invariant manifold, and reach certain destinations with no further expense of fuel. The destination might be another unstable periodic orbit, which would allow the spacecraft to return home.

Invariant manifolds are also commonly used in fluid flow visualization, where they are called stream surfaces. In experiments, smoke or dye is introduced into a steady flow along a wire or a tube, and swept downstream. Flow visualization software simulates the experiment by computing the image of the wire under the flow.

In this paper we describe an algorithm for computing a well distributed set of points on a two dimensional invariant manifold when a global transversal is given (a curve in $\mathbb{R}^n$). The algorithm is described in detail for invariant manifolds of dimension two and greater in [9]. The points are spaced along a set of trajectories, and the trajectories are spaced by "fattening" the trajectories. That is, trajectories are not allowed to pass into an interval around the other trajectories.

## 2 Basic Definitions

The "forward" part of an invariant manifold $M^+(M_0)$ consists of all trajectories which start at a point on a smooth curve $M_0 \subset \mathbb{R}^n$. There is also a "backward" part $M^-(M_0)$, found by integrating trajectories backward in time from $M_0$. This is simply a change of the sign of $f(x)$, so in what follows we drop the superscripts $\pm$, and consider only the forward image of $M_0$.

The "natural" parameterization of $M$ uses the coordinate $\sigma$ of a parameterization of $M_0$, and the time $t$. Any point on an invariant manifold $M(M_0)$ can be written as $x(\sigma, t) \in \mathbb{R}^n$ where

$$M(M_0) = \left\{ x(\sigma,t) \;\middle|\; x(\sigma,0) = M_0(\sigma), \quad \frac{\mathrm{d}}{\mathrm{d}t}x(\sigma,t) = f(x(\sigma,t)) \right\}.$$

In many interesting cases the natural parameterization is poor (Figure 2). The tangent vectors of the coordinate lines of the natural parameterization are $x^i_{,\sigma}(\sigma,t)$, and $x^i_{,t}(\sigma,t) = f^i(x(\sigma,t))$. A poor parameterization is one where these become nearly linearly dependant, and/or become large or small in norm.

We use the usual tensor notation [16], where the superscript refers to the coordinates of a vector $x \in \mathbb{R}^n$. The subscript with a comma refers to the derivative with respect to the subscript. We will also use the Einstein

summation convention, where the appearance of an index twice in a product indicates a sum over that index. The inner product of two vectors is written $x^p y^p$, to mean $\sum_p x^p y^p$. We will try to use $p, q, r, ...$ for indices which are summed over, and $i, j, ...$ for other indices.

The metric $g$ at a point on $M(M_0)$ is the $2 \times 2$ matrix

$$g = \begin{pmatrix} x^p_{,\sigma} x^p_{,\sigma} & x^p_{,\sigma} x^p_{,t} \\ x^p_{,t} x^p_{,\sigma} & x^p_{,t} x^p_{,t} \end{pmatrix}.$$

A "good" parameterization is one for which $g$ is everywhere close to the identity. That is $x_{,\sigma}$ and $x_{,t}$ are unit vectors, and $x_{,t}$ is orthogonal to $x_{,\sigma}$.

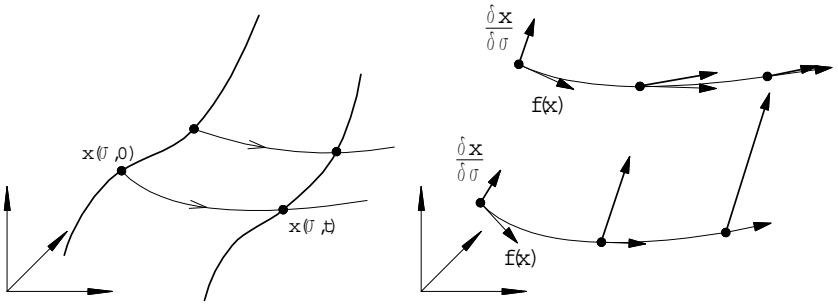



Fig. 2: The natural parameterization. (left) A "good" parameterization, (right) two types of "poor" parameterizations. The lower trajectory has tangent vectors that are roughly orthogonal (that is, $g$ is diagonal), but not unit vectors. The upper trajectory suffers from shear, where the tangent vectors are far from orthogonal ($g$ has large off diagonal elements.

One way to understand the literature on computing invariant manifolds is to consider how the natural parameterization $M$ is improved. (A recent survey [14] describes and constrasts the various approaches.) [11] and [13] use a diagonal scaling, while [6], [7] and [15] use an upper triangular scaling. These scalings are done indirectly, by adapting a mesh, and if the coordinate curves no longer align with the trajectory, some sort of interpolation must be done. The approach described here uses a parameterization that is locally Euclidean near a trajectory (i.e. $g$ in the new parameterization is the identity), and advances points, tangents and curvature along trajectories (a *fat* trajectory).

## 3 Fat Trajectories

In order to build an interval about a trajectory we need the tangent space and curvature of the invariant manifold $M$. The interval trajectory will be the set of points which when projected to the tangent space of the nearest point on the fat trajectory, lie inside a disk about the origin. The radius of the disk will be allowed to vary along the trajectory according to the curvature of the trajectory. A good choice is $R = \sqrt{\epsilon/2 \parallel x'' \parallel}$, where $x''$ is the curvature and $\epsilon$ controls the distance between the tangent space and the manifold (see [8] for details).

In the natrual parameterization the $t$ tangent vector is $x_{,t} = f(x)$. The $\sigma$ tangent must be integrated along a trajectory starting at $M_0(\sigma)$ –

$$\frac{\mathrm{d}}{\mathrm{d}t} x^i_{,\sigma} = f^i_{,p} x^p_{,\sigma}. \tag{2}$$

With a little differential geometry it can be shown [9] that an orthonormal basis $x^i_0$, $x^i_1$ for the tangent space which changes as little as possible along a trajectory (Figure 3) evolves according to

$$\frac{\mathrm{d}}{\mathrm{d}t} x^i_{,j} = f^i_{,p} x^p_{,j} - \left( x^p_{,r} f^p_{,q} x^q_{,j} \right) x^i_{,r} \tag{3}$$

Equation 3 is similar in form to Equation 2, but a linear combination of the tangent vectors has been subtracted, and this maintains the orthonormality of the basis. Evolution equations can also be found for the curvature (or more precisely the derivatives of the tangent vectors). In the natural parameters

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} x^i_{,t,t} &= f^i_{,p} x^p_{,t} \\ \frac{\mathrm{d}}{\mathrm{d}t} x^i_{,\sigma,t} &= f^i_{,p} x^p_{,\sigma} \\ \frac{\mathrm{d}}{\mathrm{d}t} x^i_{,\sigma,\sigma} &= f^i_{,p} x^p_{,\sigma,\sigma} + f^i_{,p,q} x^p_{,\sigma} x^q_{,\sigma} \end{aligned} \tag{4}$$

and in the orthonormal basis

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} x^i_{,j,k} = {} & f^i_{,p} x^p_{,j,k} + f^i_{,p,q} x^p_{,j} x^q_{,k} \\ & -(x^p_{,r} f^p_{,q} x^q_{,j}) x^i_{,r,k} - (x^p_{,r} f^p_{,q} x^q_{,k}) x^i_{,r,j} \\ & -(x^p_{,w} f^p_{,q} x^q_{,j,k} + x^p_{,w} f^p_{,q,r} x^q_{,j} x^r_{,k} + x^p_{,j,k} f^p_{,q} x^q_{,w}) x^i_{,w} \end{aligned} \tag{5}$$

Though the expressions are of course more complicated, the form of Equation 5 is the same as Equation 4 except that this time linear combinations of both the second derivative vectors and the tangents have been subtracted.

Initial conditions for the basis at points on $M_0$ can be found using Gram–Schmidt orthogonalization starting with the natural parameterization. Since the second derivatives in the natural parameterization are easily found they can be transformed into second derivatives in the new basis [9].

This coordinate system is analogous to the Riemannian Normal Coordinates (RNC) used in general relativity to find an inertial frame along geodesics. Here the trajectory plays the role of the geodesic. The coordinate system is also a parallel transport.

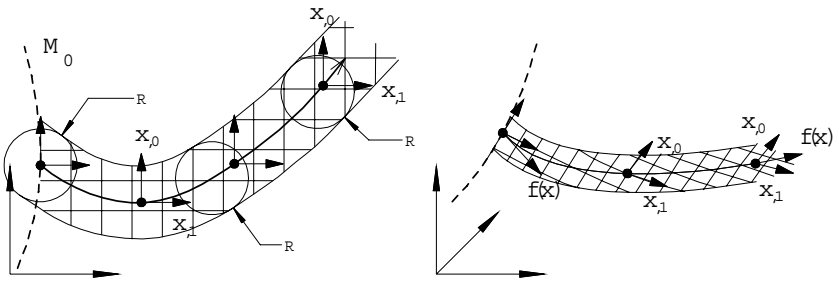



Fig. 3: A sketch of the new coordinate system near a trajectory. (left) Looking "down" on $M$. (right) the same in space. Note that the flow direction $f(x)$ is not one of the two basis vectors. The lines paralleling the trajectory are a neighborhood on $M$ of the trajectory with width $R$.

*Fat trajectories* are neighborhoods of width $R(x)$ about a trajectory, with $R(x)$ varying along the trajectory (Figure 3). To cover $M$, a set of points is distributed on $M_0$ using $R(x)$, and fat trajectories are integrated forward from these points. The integration is stopped if the trajectory enters a previously integrated fat trajectory. This may leave uncovered parts of $M$ if the flow expands (which is common). To cover the rest of $M$ we must locate points and construct initial conditions for starting more fat trajectories. To do this we use circular disks in the tangent space of the fat trajectory at points spaced on the trajectory according to $R(x)$. This allows us to use a representation of the boundary of the covered part of $M$ to locate an interpolation point.

## 4 Flying Disks

In [8] the author developed a method of representing manifolds as the union of overlapping spherical balls of different radii. The representation was used to compute implicitly defined manifolds (i.e. solutions of $F(x) = 0$ with $F : \mathbb{R}^n \to \mathbb{R}^{n-k}$), and has been used for computing other types of manifolds as well. The approach computes an approximate "restricted Laguerre–Voronoi"

tessellation of $M$ based on the spherical balls. This is instead of the more usual grid, or triangulation, though the dual Delaunay triangulation can be used if a triangulation is required (Figure 4.) Voronoi and Delaunay diagrams are described in [2]. The restricted Laguerre–Voronoi diagram, or restricted power diagram, is decribed in [1], [4] and [12]. Using the Voronoi tessellation avoids the well known problems with advancing triangulations, and the radius of the spherical ball provides a way of equi-distributing points on $M$. Roughly, points are no closer than $R$, or further apart than $2R$.

Below we describe the two dimensional case, but the same approach works in higher dimensions, with a polyhedral Voronoi tessellation instead of polygonal tesselation.

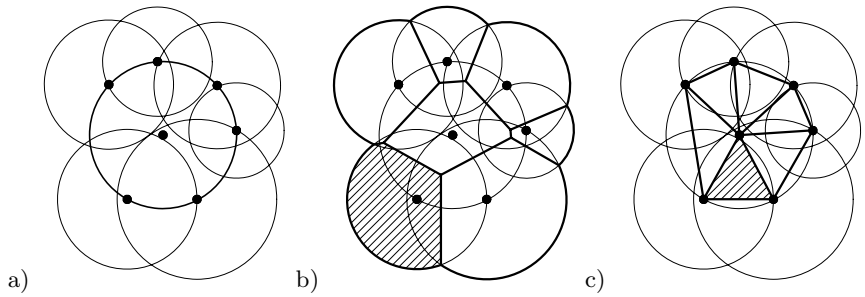



Fig. 4: A set of circular neighborhoods (a), the corresponding restricted Laguerre-Voronoi (b) and dual Delaunay diagrams (c).

A triangulation would probably be the first choice to represent a manifold. To iteratively find a set of points on a manifold a point on the boundary would be identified (easily done for a triangulation) and advanced some distance normal to the boundary. The new point would then be used to define a triangle (or simplex in higher dimensions) which is added to the mesh. This keeps the triangulation moving "outward" from the initial point, but there are many cases in which the new triangle is incompatible with the existing triangulation. That is, the new triangle overlaps the existing triangulation.

A covering is a set of neighborhoods centered at points on the manifold. The neighboroods are allowed to overlap, as long as every point on $M$ lies in some neigborhood. A covering does not have the difficulty with compatibility of new neighborhoods as triangulations (they are meant to overlap). However, it is not obvious how to find a point near the boundary of a union of neighborhoods. The polygonal Voronoi tiles provide a way to find a point on the boundary.

Finding the Voronoi tiles is simple. The points are found at the same time, so this is not the usual incremental computation of a Voronoi diagram, where the points are givem. If we have one circular disk, and a square which contains

it, then the boundary of the disk is the part of the circle inside the square. When a second disk is added, the intersection of the circles bounding the two disks lies on a line (in higher dimensions a plane) orthogonal to the line between the centers of the two disks, and the part of the circle on the boundary of the union is the part on the appropriate side of this line. If complementary halfplanes are removed from the squares surrounding the two disks, the part of the boundary of each disk is the part inside the resulting polygon (figure 5). By identifying neighboring disks when a new disk is added, the Voronoi tiles (the part of the disk inside the polygon) are updated by removing complementary halfplanes from the new disk and each disk which it overlaps.

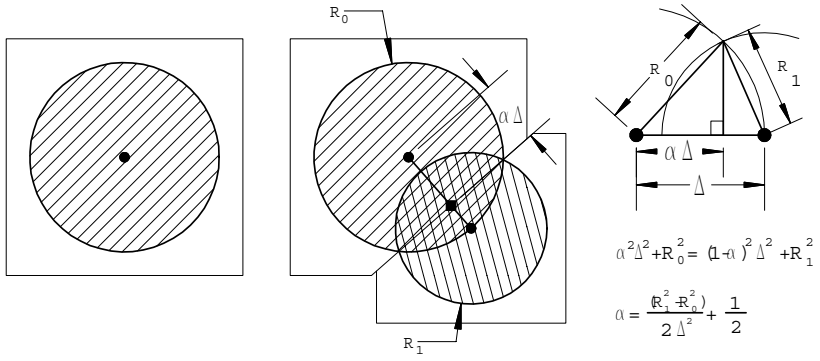



Fig. 5: Updating Voronoi tiles. Each tiles starts as a square, then for each disk which overlaps the disk the polygon is clipped against the line by the intersection of the circles of the two disks.

When disks are in different tangent spaces, they must be projected to a common tangent space before updating the polygons. If the radius of the disk is small relative to the curvature of $M$, the projection of one disk to the tangent space of an overlapping disk will almost be a circular disk, and the previous procedure can be used to update the polygons. There is an error committed, but the effect is that points that are identified as boundary points may be slightly inside the boundary (Figure 6).

The invariant manifold $M$ is represented as the union of the projections of a set of circular disks onto $M$. This is a list, or "atlas" of "charts", which consist of a point on $M$, tangent vectors of $M$ at that point, and a radius (these represent the circular disk), together with a polygon (the Voronoi tile). As points are added to the list, the polygons are updated by clipping the polygon against a line.

To approximate a fat trajectory we start with a point $x_0 \in M_0$, or an interpolated point, and compute the initial orthonormal basis and second derivatives. This forms the first chart on the fat trajectory. The trajectory,

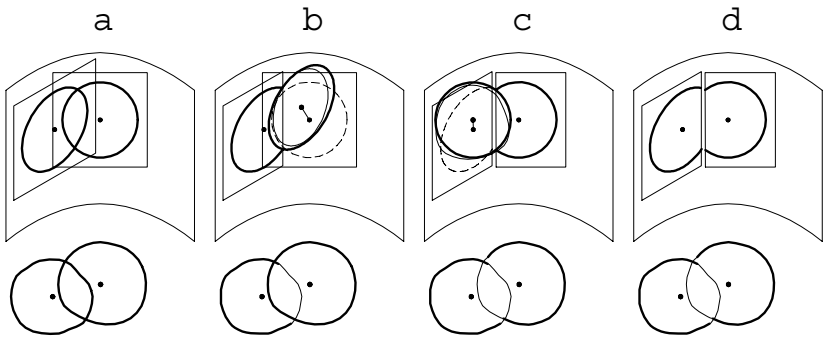



Fig. 6: On a curved surface (a) the update of a disk's polygon is done in the disk's tangent space. The center of each overlapping disk is projected into the tangent space (b), and a circle with the radius of the overlapping disk is used to update the polygon. There is an error involved, since the neighborhood is actually on the manifold (sketched below the surface), and the circle is distorted by the projection onto $M$ and then the projection into the tangent space. This process is then repeated (c) to update the polygon of the neighboring disk. The result (d) is still part of the boundary of the projection onto $M$, but we may think that a point is on the boundary when it is actually a little inside. The size of this error is of the order of the distance between the tangent space and $M$ on the circle.

tangent space and curvature are integrated a distance $R$, and another chart is added. This process is repeated until a maximum time is reached, or the trajectory enters an existing chart. A sketch of a fat trajectory that has been covered this way is shown in figure 7.

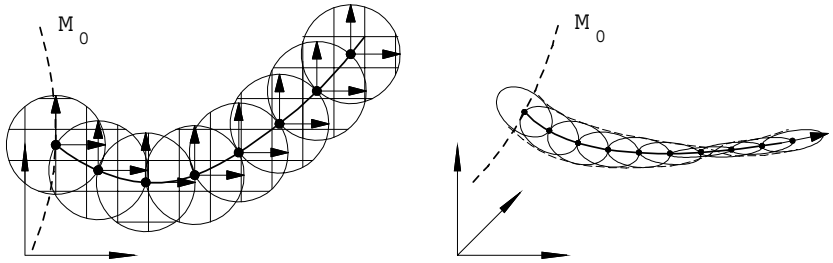



Fig. 7: Circular neighborhoods (charts) along a trajectory. (left) Looking "down" on $M$. (right) the same in space, showing how the disk "rolls" and "pitches", but does not "yaw".

## 5 Interpolation

When there are no more points on $M_0$ which are outside the union of charts (Figure 8), a starting point must be found that is not on $M_0$. A point in the interior of the union will be inside a fat trajectory, so we use the polygons to find a point near the boundary, where the new trajectory will leave the interior of the union.

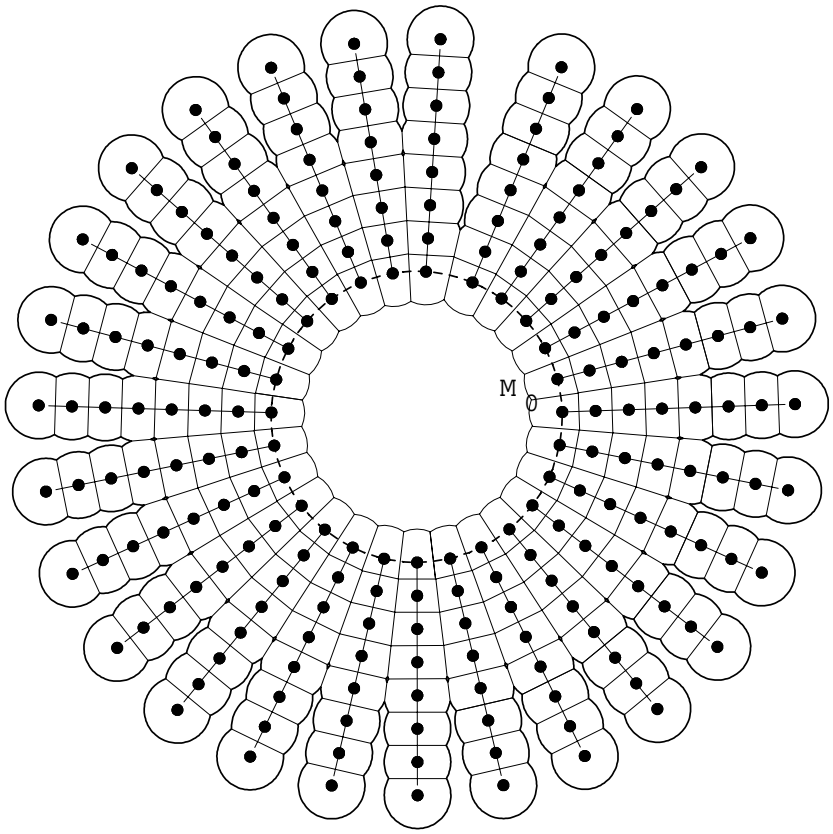



Fig. 8: When the manifold of starting points $M_0$ is covered, some other point must be found that can be used to start a new trajectory.

In [9] the author used an argument based on a modified nonlinear optimization problem to show that such a point exists. There is a technical requirement that is satisfied for 2d surfaces once $M_0$ is covered, and the variation in $f(x)$ over the disk must be small relative to the radius of the disk. The optimization

problem looks for a point on the boundary of the union which is furthest back (locally) toward $M_0$ along trajectories on $M$. While there is no objective in this optimization problem, the usual optimality conditions must hold. These conditions are that an optimum be a stable fixed point of a modified flow (on a boundary the component of $f(x)$ normal to the boundary is projected out). The problem is posed on the part of $M$ which is *outside* the disks, and not further than a maximum time $T_{max}$ from $M_0$ (measured along trajectories).

There can be no fixed points on the extrior of the union, since it takes an infinite amount of time to reach the fixed point, and a maximum time has been imposed. The point furthest back toward $M_0$ must therefore lie on the boundary of the disks. For a point on the boundary to be a fixed point, the flow vector must lie in the positive cone of normal vectors. This is just another way of introducing Lagrange multipliers. There are only two types of point on the boundary, those which lie on a single circle, and those at the intersection of two circles. For a fixed point on a single circle $f(x)$ must be parallel to the normal of the circle, and point away from the center of the circle. That is, *the extension of the flow vector on the boundary backward in time passes through the center of the circle* (Figure 9). However, such a point cannot be a minimum, since it lies on a circle which curves backward in time.

At fixed points of the modified flow lying at the intersection of two circles must have $f(x)$ in the positive cone formed by the normals of the two circles. The normals are parallel to lines starting at the center and passing through the intersection point (Figure 10). That is, *the extension the flow vector at the boundary point backward in time crosses the interior of the edge between the two centers*. This point is the minimum, and if we use the point on the edge between the centers to start a new trajectory, the initial values can be interpolated from the values at the two centers, and if $f(x)$ does not vary much over a disk the new trajectory will leave the union near the intersection point.

Intersection points can be easily found from the polygons associated with the disks (Figure 10). They are points where an edge of the polygon crosses the circular boundary of the disk. A list of the disks on the boundary can be maintained (boundary disks have polygons with at least one exterior vertex). To find an interpolation point this list is transversed, and the edges of the polygon are tested for crossing. If one endpoint is inside and the other outside this is trivial. If both endpoints are outside, the distance between the edge and the center being less than the radius indicates a crossing.

## 6 Example

As an illustration, we consider a periodically forced pendulum with damping (this example is from [21]). When the forcing is zero, the phase space consists of a set of hyperbolic fixed points at $x_1 = (2n+1)\pi$, $(x_1)_{,t} = 0$, where the pendulum points straight up, and centers at $x_1 = 2n\pi$, $(x_1)_{,t} = 0$ with the

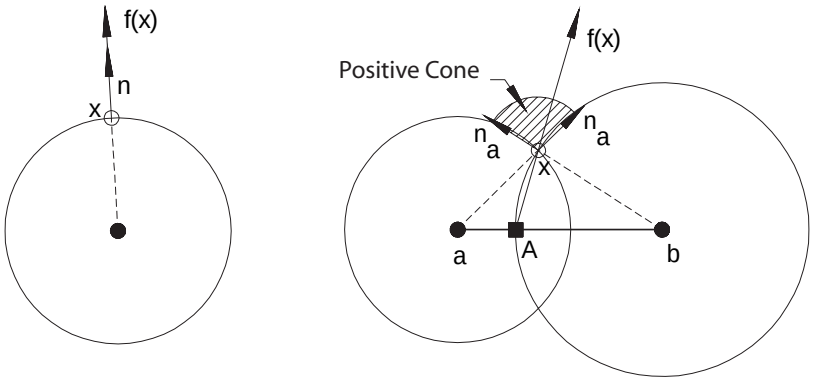



Fig. 9: (left) A point $x$ on the boundary of a single disk is a fixed point of the modified flow if $f(x)$ is parallel to the normal and points in the "outward" direction. That is, the extension of $f(x)$ backward in time passes through the center of the disk. Moving $x$ a little in either direction on the circle moves $x$ further back towards $M_0$, so it is not the "minimum". (right) A point $x$ at the intersection of two circles is a fixed point of the modifed flow if $f(x)$ lies in the positive cone formed from the two normals. That is, the extension of $f(x)$ backward in time crosses the line between centers $a$ and $b$ (point $A$). This is a local "minimum", and a trajectory started at $A$ – if $f$ does not change too rapidly over the radius of a disk – will pass out of the interior of the disks near $x$. Initial values for the tangents and curvature can be interpolated from centers $a$ and $b$.

pendulum pointing down. With periodicity the phase space can be reduced to $x_1 \in [-\pi, \pi]$. Figure 11 shows the unperturbed nonlinear single pendulum. Gravity acts on the pendulum bob and the equations are

$$\frac{\mathrm{d}}{\mathrm{d}t} x_1 = x_2$$

$$\frac{\mathrm{d}}{\mathrm{d}t} x_2 = -\sin x_1$$

Without the forcing and dmping there is an energy $E = x_2^2/2 - \cos x_1$ that is conserved on trajectories. For initial energies $E > 1$ the pendulum "runs", that is $x_1$ continually increasing or decreasing depending on the initial velocity. For $E < 1$ the pendulum oscillates about the downward pointing fixed point. If the pendulum is started with $E = 1$ it will swing to the top and stop. This last is a heteroclinic orbits shown as dark curve in Figure 11 center.

The perturbated equations, analyzed in [21] are

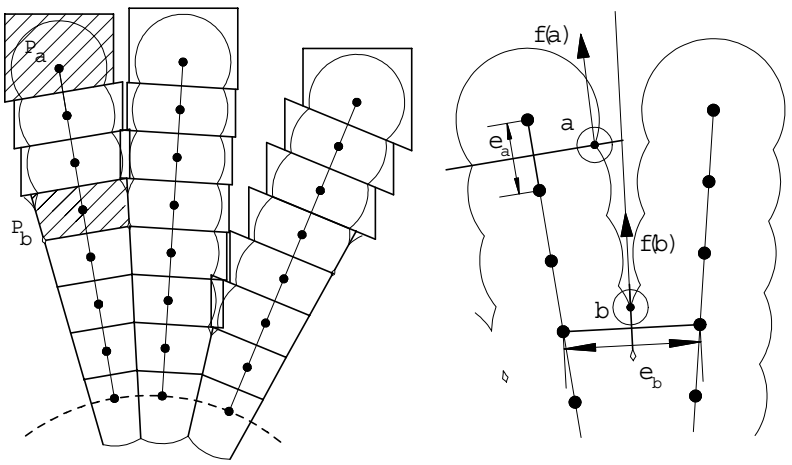



Fig. 10: Interpolating: (left) Some of the disks and polygons from Fig. 8. Polygons with a vertex outside the disk indicate that the disk is on the boundary. We highlight two polygons, $P_a$ and $P_b$. (right) The edges of $P_a$ and $P_b$ which cross the boundary of the disks, and two of the crossing points, $a$ and $b$. All of the other edges of the polygons have been removed. The centers that these two edges separate form an edge. At point $a$ the flow vector $f(a)$ extended backward does not intersect $e_a$, while at the point $b$ $f(b)$ extends backward to cross $e_b$. A trajectory started at the point where the two cross will leave the union of the disks, and initial values can be interpolated between the centers at the ends of $e_b$.

$$\frac{\mathrm{d}}{\mathrm{d}t}x_1 = x_2$$

$$\frac{\mathrm{d}}{\mathrm{d}t}x_2 = -\sin x_1 + \epsilon\left(\gamma \sin(\Omega x_3)\sin x_1 - \delta x_2\right))$$

$$\frac{\mathrm{d}}{\mathrm{d}t}x_3 = 1$$

The perturbation is time dependant, so time is introduced as a phase space coordinate to make the flow autonomous (a standard trick called *suspending* the flow).

For small perturbations ($\epsilon << 1$) there is a critical forcing amplitude

$$\gamma^* = \frac{4\delta}{\pi\Omega}\sinh\frac{\pi\Omega}{2}$$

For $\gamma < \gamma^*$ the damping removes more "energy" than the periodic forcing puts into the system, and the pendulum eventually will spiral into the fixed

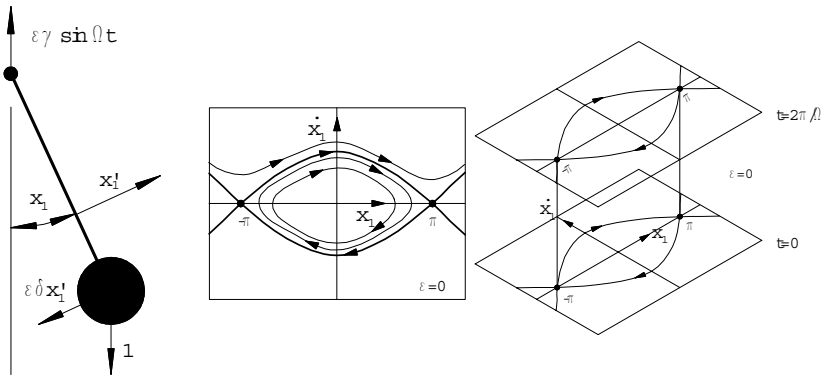



Fig. 11: The periodically forced pendulum from [21]. (left) A periodic vertical force is applied to the pivot, as well as a damping. (center) the behavior of the pendulum when $\epsilon = 0$. If the energy $x_1^2/2 - \cos x_1$ is greater than one the pendulum swings around and around. If the energy is less than one the pendulum oscillates (there is no damping at $\epsilon = 0$. When the energy is exactly one, the pendulum comes to rest with the bob above the pivot. (right) with $\epsilon > 0$ time becomes a variable, periodic over $2\pi/\Omega$. Following [21] we will use this box to display the image of a line of initial points.

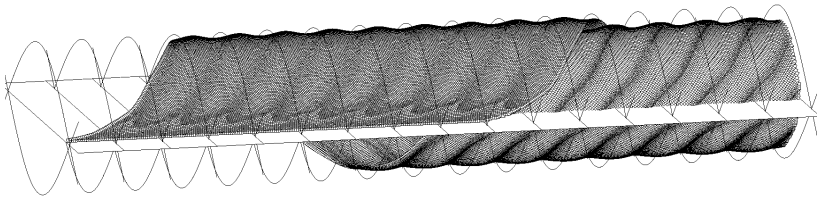


Fig. 12: The periodically forced pendulum. $\epsilon = .2$, $\gamma = 1.5$, $\delta = .2$, and $\Omega = 5$. The calculation used 176 points on seven replicas of the fundemental region $[0, \pi/\Omega)$, and those 176 start points created 84,943 disks. In addition, 48 interpolations were needed, for a total of 91,240 disks.

point at zero (which is now the straight line $(0, 0, t)$). For $\gamma \geq \gamma^*$ heteroclinic tangles appear (a type of chaotic motion).

For illustration we chose $M_0$ to be the line $(x_1, x_2, x_3) = (3.0, -0.1, x_3)$, which is near one of the unstable fixed points for $\epsilon = 0$. Figure 12 shows the surface that was computed. The time coordinate $x_3$ is periodic with period $2\pi/\Omega$, and Figure 12 shows sixteen periods of $x_3$. If we use the same compu-

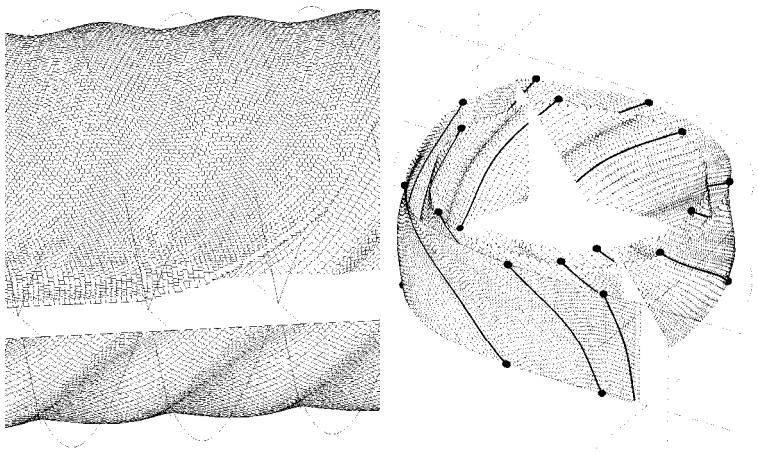


Fig. 13: (left) A closer view of the invariant manifold in Figure 12. (right) The same invariant manifold but brought back to a single period in the forcing. The dark black line is a trajectory starting near the fixed point at $x_1 = \pi$. The mapping from the plane $t = 0$ to $t = 2\pi/\Omega$ is used in the analysis, and the black dots are the orbit of one point under that map. For these parameters the motion decays to the fixed point of the map at $(0, 0)$

tational results and collapse it to two periods of $x_3$ we can see some of the structure that leads to a heteroclinic tangle and chaotic motion.

## References


1. F. Aurenhammer, Power diagrams: Properties, algorithms and applications. *SIAM Journal of Computing* **16**, 78–96 (1987)
2. F. Aurenhammer, Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM Computing Surveys* **23**, 345–405 (1991)
3. E.J. Doedel, Nonlinear numerics. *International Journal of Bifurcation and Chaos* **7**, 2127–2143 (1997)
4. H. Edelsbrunner, The union of balls and its dual shape. *Discrete and Computational Geometry* **13**, 415–440 (1995)
5. G. Gomez, W.S. Koon, M.W. Lo, J.E. Marsden, J. Masdemont, S.D. Ross, Invariant manifolds, the spatial three-body problem and space mission design. In: *International Conference on Differential Equations*, 1167–1181 (Berlin 1999)
6. J. Guckenheimer, A. Vladimirsky, A fast method for approximating invariant manifolds. *SIAM J. Appl. Dyn. Systems* **3**, 232–260 (2004)
7. J. Guckenheimer, P. Worfolk, Dynamical systems: Some computational problems. In: *Bifurcations and Periodic Orbits of Vector Fields*, ed. by D. Schlomiuk, 241–277 (Kluwer Academic Publishers 1993)

8. M.E. Henderson, Multiple parameter continuation: Computing implicitly defined $k$–manifolds. *International Journal of Bifurcation and Chaos* **12**, 451–476 (2002)
9. M.E. Henderson, Computing invariant manifolds by integrating fat trajectories. *SIAM Journal on Applied Dynamical Systems* **4**, 832–882 (2005)
10. K.C. Howell, B.T. Barden, M.W. Lo, Application of dynamical systems theory to trajectory design for a libration point mission. *The Journal of the Astronautical Sciences* **45**, 161–178 (1997)
11. J.P.M. Hultquist, Constructing stream surfaces in steady 3D vector fields. In: *IEEE Proceedings Visualization '92,* pp. 171–178 (Boston 1992) .
12. H. Imai, M. Iri, K. Murota, Voronoi diagram in the Laguerre geometry and its applications. *SIAM Journal on Computing* **14**, 93–105 (1985)
13. M. E. Johnson, M. S. Jolly, I. G. Kevrekidis, Two-dimensional invariant manifolds and global bifurcations: some approximation and visualization studies. *Numerical Algorithms* **14**, 125–140 (1997)
14. B. Krauskopf, H. Osinga, E. Doedel, M. Henderson, J. Guckenheimer, A. Vladimirsky, M. Dellnitz, O. Junge, A survey of methods for computing (un)stable manifolds of vector fields. *International Journal of Bifurcation and Chaos* **15**, 2127–2143 (2005)
15. B. Krauskopf, H.M. Osinga, Computing geodesic level sets on global (un)stable manifolds of vector fields. *SIAM Journal on Applied Dynamical Systems* **2**, 546–569 (2003)
16. D. Lovelock, H. Rund, *Tensors, Differential Forms, and Variational Principles* (John Wiley & Sons, New York 1975)
17. B. Rassmussen, Numerical Methods for the Continuation of Invariant Tori, PhD thesis, Georgia Institute of Technology, School of Mathematics, December 2003
18. V. Reichelt, Computing invariant tori and circles in dynamical systems of fixed points. In: *The IMA Volumes in Mathematics and its Applications,* Vol. 119, ed. by E. Doedel, L. S. Tuckerman, 407–437 (Springer, Berlin Heidelberg New York 2000)
19. F. Schilder, H.M. Osinga, W. Vogt, Continuation of quasi-periodic invariant tori *SIAM J. Applied Dynamical Systems* **4**, pp. 459–488 (2005)
20. M. van Veldhuizen, A new algorithm for the numerical approximation of an invariant curve. *SIAM J. Sci. Stat. Comput.* **8**, 951–962 (1987)
21. S. Wiggins, *Global Bifurcations and Chaos*, Applied Mathematical Sciences, 73 (Springer, Berlin Heidelberg New York 1988)

# "Ghost" ILDM-Manifolds and Their Identification

S. Borok[1], I. Goldfarb[2], V. Gol'dshtein[3], and U. Maas[4]

[1] Ben-Gurion University of the Negev, P.O. Box 653, Beer-Sheva, Israel, borok@bgu.ac.il
[2] Ben-Gurion University of the Negev, P.O. Box 653, Beer-Sheva, Israel, goldfarb@cs.bgu.ac.il
[3] Ben-Gurion University of the Negev, P.O. Box 653, Beer-Sheva, Israel, vladimir@bgu.ac.il
[4] Institute for Technical Thermodynamics, Karlsruhe University (TH), Karlsruhe 84105, Germany, maas@ftt.karlsruhe-uni.de

**Summary.** One of the popular methods (Intrinsic Low-Dimensional Manifolds – ILDM)) of decomposition of multiscale systems into fast and slow sub-systems for reduction of their complexity is considered in the present paper. The method successfully locates a position of slow manifolds of considered system and as any other numerical approach has its own disadvantages. In particular, an application of the ILDM-method produces so-called "ghost"-manifolds that do not have any connection to the true dynamics of the system. It is shown analytically that for two-dimensional singularly perturbed system (for which the fast-slow decomposition has been already done in analytical way) the "ghost"-manifolds appear. The problem of discrimination/identification of the "ghost"-manifolds is under consideration and two numerical criteria for their identification are proposed. A number of analyzed examples demonstrate efficiency of the suggested approach.

## 1 Introduction

In this paper, following Maas and Pope [16] we consider Intrinsic Low- Dimensional Manifolds Method (ILDM) for systems of ordinary differential equations. The main aim of this paper is to demonstrate that application of the conventional ILDM machinery can produce additional artificial objects ("ghost" manifolds) that do not have any connection to the true slow invariant manifold of considered system. Two various approaches for the "ghost" manifolds identification/discrimination are suggested and their application is demonstrated.

The paper is organized as follows. In Sect. 2, we give a review of several reduction methods, which are used in combustion and chemical kinetics problems. In Sect. 3, we give the examples of the "ghost" manifolds phenom-

enon. In Sect. 4, we suggest two criteria for identification/discrimination of the "ghost" objects. In Sect. 5, we conclude the results.

## 2 Theoretical Background

In this section the method of invariant manifolds, iterative method of Fraser, Inflector-method, Intrinsic Low-Dimensional Manifolds (ILDM) method and its modification (TILDM) will be briefly described.

### 2.1 Method of Invariant Manifolds (MIM)

Consider a singularly perturbed system of ordinary differential equations

$$\epsilon \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, \mathbf{y}, \epsilon) \tag{1}$$

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{g}(\mathbf{x}, \mathbf{y}, \epsilon) \tag{2}$$

Here $\mathbf{x} \in \Re^m,\ \mathbf{y} \in \Re^n$ are vectors in Euclidean space, $t \in (t_0, +\infty)$ is a time-like variable, $0 < \epsilon < \epsilon_0 << 1$, functions $\mathbf{f}: \Re^m \times \Re^n \to \Re^m,\ \mathbf{g}: \Re^m \times \Re^n \to \Re^n$ are supposed to be sufficiently smooth for all $\mathbf{x} \in \Re^m,\ \mathbf{y} \in \Re^n,\ 0 < \epsilon < \epsilon_0$. The values $|\mathbf{f_i}(\mathbf{x}, \mathbf{y}, \epsilon)|,\ |\mathbf{g_i}(\mathbf{x}, \mathbf{y}, \epsilon)|,\ (i = 1, ...m;\ j = 1, ...n)$ are assumed to be comparable with the unity as $\epsilon \to 0$.

**Definition 1.** *A smooth surface in the phase space $M \in \Re^m \times \Re^n \times \Re$ is called an invariant manifold of the system (1)-(2), if any phase trajectory $(\mathbf{x}(t, \epsilon), \mathbf{y}(t, \epsilon))$ such that $(\mathbf{x}(t_1, \epsilon), \mathbf{y}(t_1, \epsilon)) \in M$ belongs to $M$ for any $t > t_1$. If the last condition holds only for $t \in [t_1, T]$, then $M$ is called a local invariant manifold.*

The simplest examples of invariant manifolds are phase trajectory and phase space.

The manifold's existence leads to the fact that the analysis of the system's behaviour can be considerably simplified by reducing a dimension of the system. We are interested in the invariant manifolds of dimension $m$ (the dimension of the slow variable) that can be represented as a graph of the vector-valued function:

$$\mathbf{x} = \mathbf{h}(\mathbf{y}, \epsilon). \tag{3}$$

The invariant manifolds mentioned above are called manifolds of slow motions (this term was adopted from the nonlinear mechanics). The system's dynamics on this manifold is described by the equation

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{g}(\mathbf{h}(\mathbf{y}, \epsilon), \mathbf{y}, \epsilon). \tag{4}$$

If $\mathbf{y}(t,\epsilon)$ is a solution of the Eq.(4), then the pair $\mathbf{x}(t,\epsilon),\mathbf{y}(t,\epsilon)$ where $\mathbf{x}(t,\epsilon) = \mathbf{h}(\mathbf{y}(t,\epsilon),\epsilon)$ is a solution of the original system (1)-(2), since it determines a trajectory on the invariant manifold.

A usual approach in the qualitative study of (1)-(2) is to consider first the degenerate system, which is obtained by substituting $\epsilon = 0$ into the system

$$0 = \mathbf{f}(\mathbf{x},\mathbf{y},0) \tag{5}$$

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{g}(\mathbf{x},\mathbf{y},0), \tag{6}$$

and then to draw conclusions for the qualitative behaviour of the full system for sufficiently small $\epsilon$. The Eq.(5) determines the slow surface. The slow surface is the zeroth approximation of the slow invariant manifold. It is assumed that the Eq.(5) has an isolated smooth solution $\mathbf{x} = \mathbf{h_0}(\mathbf{y})$. Moreover, the next relation should take place

$$\lim_{\epsilon\to 0} \mathbf{h}(\mathbf{y},\epsilon) = \mathbf{h_0}(\mathbf{y})$$

In addition, only these manifolds are important here that are stable (attractive). By the famous Tikhonov's theorem, the question of stability of an invariant manifold can be reduced to study of its zeroth approximation stability.

Invariant manifold $\mathbf{x} = \mathbf{h}(\mathbf{y},\epsilon)$ of the system (1)-(2) is stable, if the real parts of all eigenvalues of the matrix $D_{\mathbf{x}}\mathbf{f}(\mathbf{h_0(y)},\mathbf{y},0)$ are negative.

Points of the slow surface determined by (5) are sub-divided into two types: *standard* points and *turning* points. A point $(\mathbf{x},\mathbf{y})$ is a standard point of the slow surface if in some neighborhood of this point the surface can be represented as a graph of a function $\mathbf{x} = \mathbf{h_0}(\mathbf{y})$ such that $\mathbf{f}(\mathbf{h_0}(\mathbf{y}),\mathbf{y},0) = 0$. It means that the condition of the Implicit Function Theorem $D_{\mathbf{x}}\mathbf{f}(\mathbf{h_0(y)},\mathbf{y},0) \neq 0$ holds and the slow surface has the dimension of slow variable. Points where this condition does not hold are turning points of the slow surface. In other words, turning points are defined as solutions of the following system

$$\mathbf{f}(\mathbf{x},\mathbf{y},0) = 0$$

$$\mathbf{f}_x(\mathbf{x},\mathbf{y},0) = 0$$

The asymptotic method described below can not be applied there.

Problems of existence, uniqueness and stability of invariant manifolds have been studied by many authors. The main results of these studies can be summarized in the following theorems.

**Theorem 1.** *(Mitropolsky and Lykova, 1973) Let the system (1)-(2) satisfies the following conditions:*

*(i) The equation* $\mathbf{f}(\mathbf{x},\mathbf{y},0) = 0$ *has an isolate solution* $\mathbf{x} = \mathbf{h_0}(\mathbf{y})$ *in some domain* $G = \{(\mathbf{x},\mathbf{y},\epsilon) :\ \mathbf{y} \in \Re^n, 0 < \epsilon < \epsilon_0, ||\mathbf{x} - \mathbf{h_0}(\mathbf{y})|| \leq \rho\}$.

*(ii) The functions* $\mathbf{f},\mathbf{g},\mathbf{h_0}$ *and their first and second partial derivatives are uniformly continues and bounded in* $G$.

*(iii) The eigenvalues* $\lambda_i(\mathbf{y}),\ i = 1, 2, ..., n$ *of the matrix* $D_{\mathbf{x}}\mathbf{f}(\mathbf{h_0(y)}, \mathbf{y}, 0)$ *satisfy the condition* $Re[\lambda_i(\mathbf{y})] \leq -\beta,\ i = 1, 2, ..., n,\ \mathbf{y} \in \Re^n$ *for some* $\beta > 0$.

*Then there exists an* $\epsilon_1 :\ 0 < \epsilon_1 < \epsilon_0$, *such that for every* $\epsilon :\ 0 < \epsilon < \epsilon_1$ *the system (1)-(2) has a unique invariant manifold* $\mathbf{x} = \mathbf{h}(\mathbf{y}, \epsilon)$, *where the function* $\mathbf{h}$ *satisfies the equality* $\mathbf{h}(\mathbf{y}, 0) = \mathbf{h_0}(\mathbf{y})$.

**Theorem 2.** *(Strygin and Sobolev, 1988) Let the assumptions (i)-(iii) of the previous theorem hold. Then there exists an* $\epsilon_1 :\ 0 < \epsilon_1 < \epsilon_0$, *such that for every* $\epsilon :\ 0 < \epsilon < \epsilon_1$ *the invariant manifold* $\mathbf{x} = \mathbf{h}(\mathbf{y}, \epsilon)$ *is stable.*

In general situations the determination of the exact form and location of the slow invariant manifold is impossible. Therefore, methods of approximation are necessary. One of them finds the slow invariant manifold as a power series with respect to the small parameter $\epsilon$:

$$\mathbf{h}(\mathbf{y}, \epsilon) = \mathbf{h_0}(\mathbf{y}) + \sum \epsilon^i \mathbf{h_i}(\mathbf{y})$$

**Theorem 3.** *(Strygin and Sobolev, 1988) Let the assumptions of the previous theorem hold. Then the invariant manifold* $\mathbf{x} = \mathbf{h}(\mathbf{y}, \epsilon)$ *can be represented as*

$$\mathbf{h}(\mathbf{y}, \epsilon) = \mathbf{h_0}(\mathbf{y}) + \sum_{i=1}^{k} \epsilon^i \mathbf{h_i}(\mathbf{y}) + \mathbf{h^*}(\mathbf{y}, \epsilon) \tag{7}$$

*for some* $k$, *where* $\mathbf{h^*}(\mathbf{y}, \epsilon)$ *is a smooth function with a bounded norm, such that* $|\mathbf{h^*}(\mathbf{y}, \epsilon)| = O(\epsilon^{k+1})$ *for all* $\mathbf{y} \in \Re^n$.

It is not hard to see from the Eq. 7 that the slow surface $\mathbf{x} = \mathbf{h_0}(\mathbf{y})$ is $O(\epsilon)$ approximation of the slow invariant manifold, except the turning points. Thus, the general scheme of application of this technique for singularly perturbed system can be subdivided to analysis of the fast and slow motions. The analysis can be considerably simplified by this decomposition and reducing the dimension of the system to the dimension of the slow variable $\mathbf{y}$ and to the dimension of the fast variable $\mathbf{x}$. It means that in $O(\epsilon)$ approximation of the slow invariant manifold, the analysis of the original system can be reduced to the analysis of system's dynamics on the slow surface. On the slow surface the changes of the slow and fast variables are comparable (i.e. the fast and the slow processes are balanced). Beyond the slow surface the slow variables are fixed (quasi-stationary). Hence, each system's trajectory can be approximated by fast motions (which are beyond the slow manifold) that are described by the fast sub-system

$$\epsilon \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, \mathbf{y^0}, \epsilon);\ \mathbf{y} = \mathbf{y^0} = const,$$

and slow motions (which are on the slow manifold) that are given by the slow sub-system (4) with $\mathbf{h}(\mathbf{y}, \epsilon) = \mathbf{h_0}(\mathbf{y})$.

The method of invariant manifolds has been used for study of singularly perturbed systems of ordinary differential equations by many authors (see, for example [4], [12], [32]). The asymptotics of the slow invariant manifold are given explicitly, for example, in [17], [20].

### 2.2 Iterative Method of Fraser

The method of functional iteration for finding slow manifold was supposed in [5], further developed and applied to enzyme kinetics in [3], [18], [26], [27], [25]. In [13] there was done the asymptotic analysis of the method and comparison with the ILDM-method.

The method was inspired by the phase space geometry of an enzyme kinetics model involving a fast and a slow species, where the slow manifold is a curve in the phase plane, and extended naturally to multidimensional systems with higher-dimensional slow manifolds.

The idea of the method is as follows. Consider the planar dynamical system

$$\dot{x} = f(x, y)$$
$$\dot{y} = g(x, y),$$

where $x$ can be considered as a slow variable and $y$ as a fast one. Taking $g = 0$ as a zeroth iteration the procedure matches the slope of the slow manifold. From the trajectory equation

$$y'(x)f(x, y) = g(x, y)$$

there is obtained functional equation

$$y = \varphi(x, y')$$

and from here iterative scheme

$$y_{n+1} = \varphi(x, y'_n).$$

The procedure is explicit if the vector field is linear in the fast variable and implicit otherwise. In [13] there was considered general singularly perturbed system of ordinary differential equations, which is linear for the fast variable

$$\dot{\mathbf{y}} = \mathbf{f_1}(\mathbf{y}, \epsilon)\mathbf{z} + \mathbf{f_2}(\mathbf{y}, \epsilon),$$
$$\epsilon\dot{\mathbf{z}} = \mathbf{g_1}(\mathbf{y}, \epsilon)\mathbf{z} + \mathbf{g_2}(\mathbf{y}, \epsilon)$$

It was shown that for such system the iterative method generates, term by term, the asymptotic expansion of the slow invariant manifold. Starting from the slow surface, the $i$-th iteration of the algorithm yields the correct expansion coefficient at $O(\epsilon^i)$. Thus, after $l$ applications, the expansion is accurate up to and including the terms of $O(\epsilon^l)$.

### 2.3 Inflector Method

In this sub-section we describe very briefly the definition of inflector and some its properties. This object was introduced by Japanese mathematician Masami Okuda in the early eighties [21], [22], [23]. This investigation is interest for us because the Inflector can be considered as some prediction of Intrinsic Low-Dimensional Manifolds. The study deals with two-dimensional dynamical systems, but can be naturally generalized for higher dimensional problems.

### Definitions of Inflector, A-inflector and R-inflector

Here we remind the definitions of inflector, A-inflector and R-inflector. Consider two-dimensional dynamical system of the type

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}), \tag{8}$$

where

$$\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \ \mathbf{F}(\mathbf{x}) = \begin{pmatrix} f(x,y) \\ g(x,y) \end{pmatrix}.$$

Let $A = A(\mathbf{x})$ be a Jacobian matrix of $\mathbf{F} = \mathbf{F}(\mathbf{x})$:

$$A = \frac{\partial \mathbf{F}}{\partial \mathbf{x}} = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}.$$

Let $\lambda_i = \lambda_i(x)$ $(i = 1, 2)$ be the eigenvalues of $A$, and assume $|\lambda_1| \leq |\lambda_2|$. For the dynamical system (8) the author defined three sets: $C$ (inflector), $C_a$ (A-inflector), $C_r$ (R-inflector) by

$$C = \{\mathbf{x} \mid (A - \lambda_i \mathbf{I})\mathbf{F} = 0, \ i = 1 \ or \ 2\}, \tag{9}$$

$$C_a = \{\mathbf{x} \mid \lambda_2 < 0, \mid \lambda_1/\lambda_2 \mid < 1, \ (A - \lambda_1 \mathbf{I})\mathbf{F} = 0\}, \tag{10}$$

$$C_r = \{\mathbf{x} \mid \lambda_2 > 0, \mid \lambda_1/\lambda_2 \mid < 1, \ (A - \lambda_1 I)\mathbf{F} = 0\}, \tag{11}$$

where $I$ is the unit matrix. The definition (10) means that the A-inflector is found as all the points in the phase plane where the vector field is parallel to the slow eigenvector. Notice here that A-inflector (R-inflector) was called attractor (repellor) in the previous author's study (1976).

Eliminating $\lambda_i$ from Eq. (9), one can obtain another expression for $C$

$$C = \{\mathbf{x} \mid f(g_x f + g_y g) - g(f_x f + f_y g) = 0\} \tag{12}$$

It is obvious that

$$C_a \subset C, \ \ C_r \subset C,$$

but $C_a \cup C_r$ is not always $C$.

### The relation between the A-inflector (R-inflector) and the attracting (repelling) naive trajectory

In [21], [23] and [22] the author investigated properties of the inflector. Let us remind here very briefly one of them concerning asymptotics of the inflector.

Consider according [23] the system

$$\dot{x} = u(x, y, \epsilon) \tag{13}$$

$$\epsilon \dot{y} = v(x, y, \epsilon), \tag{14}$$

where $\epsilon > 0$ and the functions $u$ and $v$ have the power-series expansions in powers of $\epsilon$. It is assumed that the trajectory equation

$$\epsilon u(x, y, \epsilon)\mathrm{d}y = v(x, y, \epsilon)\mathrm{d}x \tag{15}$$

has solution $y = Y^0(x, \epsilon)$ in the neighborhood of $\epsilon = 0$ in some region $\Omega$ in the phase plane. The author denoted this trajectory as

$$T^0(\epsilon) = \{\mathbf{x} \mid y = Y^0(x, \epsilon)\} \tag{16}$$

and called $T^0(\epsilon)$ *a naive trajectory* (NT) in the neighborhood of $\epsilon = 0$.

From the definition of the function $Y^0(x, \epsilon)$ follows that it has the power-series expansion

$$Y^0(x, \epsilon) = \sum_{i=0}^{\infty} \psi_j(x)\epsilon^j \tag{17}$$

which converges in the neighborhood of $\epsilon = 0$ uniformly in $x$ in $\cap_{|\epsilon|<\epsilon_0}\{x \mid \mathbf{x} \in T^0(\epsilon)\}$ with some $\epsilon_0 > 0$. For calculation of the functions $\psi_j(x)$ in [23] the standard procedure was used:substitution of Eq. (17) into Eq. (15) and equating coefficients of like powers of $\epsilon$. Then we have

$$v_0(x, \psi_0(x)) = 0, \tag{18}$$

$$\psi_1(x) = -(\bar{u}_0\bar{v}_{0x} + \bar{v}_1\bar{v_{0y}})/\bar{v}_{0y}^2, \tag{19}$$

where $\bar{u}_0 = u_0(x, \psi_0(x))$, $\bar{v}_{0y} = v_{0y}(x, \psi_0(x))$, etc.

An attracting part and a repelling part of the naive trajectory was defined as follows:

$$T_a^0(\epsilon) = \{\mathbf{x} \mid y = Y^0(x, \epsilon),\ D(\mathbf{x}) < 0\} \tag{20}$$

$$T_r^0(\epsilon) = \{\mathbf{x} \mid y = Y^0(x, \epsilon),\ D(\mathbf{x}) > 0\}, \tag{21}$$

where $D(\mathbf{x})$ is so-called *a repulsion rate*. This object was introduced by the author in [22] for stability analysis in transient states. In that article he gave the mathematical expression for the repulsion rate and found two properties of the inflector with relation to it. The repulsion rate $D(\mathbf{x})$ has the following properties for the dynamical system (8) [22]: (i) Let $T(\mathbf{x})$ be a section of the trajectory passing through a point $\mathbf{x}$. If $D(\mathbf{x}) < 0$, then any state point in the neighborhood of $\mathbf{x}$ will approach $T(\mathbf{x})$ at that point of time, and if $D(\mathbf{x}) > 0$ it will go away from $T(\mathbf{x})$. (ii) If $\mathbf{x}$ is a regular point belonging to the A-inflector $C_a$ (R-inflector $C_r$), then $D(\mathbf{x}) < 0$ ($D(\mathbf{x}) > 0$).

Let $\Omega^0(\chi)$ be the region $\{\mathbf{x} \mid \mid v_y^0 \mid \geq \chi\}$, where $\chi$ is an arbitrary positive constant independent of $\epsilon$ and $v^0 \equiv v(x, y, 0) = v_0(x, y)$. Then the repulsive rate can be written as

$$D(\mathbf{x}) = \bar{v}_{0y}\epsilon^{-1} + O(1) \tag{22}$$

for a regular point $\mathbf{x} \in T^0(\epsilon)$ as $\epsilon \to 0$ in the region $\Omega^0(\chi)$.

The following important property was proved in [23] (Property 1.1): The A-inflector (R-inflector) is a first order approximation to $T_a^0(\epsilon)$ ($T_r^0(\epsilon)$) for sufficiently small $\epsilon$ in $\Omega^0(\chi)$ except singular points.

### 2.4 Intrinsic Low-Dimensional Manifold Method (ILDM)

Let us describe here very briefly the essential steps of the ILDM method. Consider differential system

$$\frac{\mathrm{d}\mathbf{Z}}{\mathrm{d}t} = \mathbf{F}(\mathbf{Z}) \tag{23}$$

Assume that this system can be represented locally as a multi-scale system for a corresponding choice of a local basis. The last depends on the choice of an arbitrary point $\mathbf{Z}$ in the $n$-dimensional Euclidean space $\mathfrak{R}^{\mathbf{n}}$. It means that in this local basis a separation of variables in accordance with their rates of changes is possible (i.e. the considered system can be rewritten in this local basis for some neighborhood of the point $\mathbf{Z}$ as singularly perturbed system). According to the assumption, the system can be subdivided locally into fast relaxing and slow or non-relaxing subsystems. Suppose that the fast sub-system has the same dimension $n_f$ ($n_f < n$) at any point $\mathbf{Z} \subseteq \mathfrak{R}^{\mathbf{n}}$.

For typical situations a set of all steady states of the fast subsystem represents an $n_s$-dimensional slow manifold ($n_s = n - n_f$) and our aim is to determine its location. The authors of ILDM suggested that the dynamics of the overall system from arbitrary initial condition should decay very quickly onto this $n_s$-dimensional manifold. The ILDM allows to identify approximately (as a set of separate points) the slow invariant manifolds (so-called intrinsic low-dimensional manifolds – ILDM-manifolds). These manifolds can be found in the following manner [16]. Suppose a local basis of the original phase space is formed by the invariant subspaces of the Jacobi matrix $M_j$ of the vector field $\mathbf{F}$ at an arbitrary point $\mathbf{Z_0}$. If the set of eigenvalues $\lambda_i$ can be sub-divided into two groups

$$max\{Re[\lambda_i], i = 1, ..., n_f\} << \tau < min\{Re[\lambda_i], i = n_f + 1, ..., n\} \tag{24}$$

(where $\tau < 0$) one can introduce invariant sub-spaces $T_f$ and $T_s$. The sub-space $T_f$ is spanned by the eigenvectors corresponding to eigenvalues with large negative (fast) real parts. In turn, the sub-space $T_s$ is spanned by the eigenvectors corresponding to eigenvalues with small negative or positive (slow) real parts. Therefore, the new basis $Q(\mathbf{Z})$, which is constructed from the eigenvectors of the Jacobi matrix and transition matrix from the standard basis to this local basis $Q^{-1}(\mathbf{Z})$ can be written like two block matrices

$$Q = \left( Q_f \; Q_s \right); \; Q^{-1} = \begin{pmatrix} \tilde{Q}_f \\ \tilde{Q}_s \end{pmatrix} \tag{25}$$

where matrices $Q_f$ and $Q_s$ correspond to the fast and slow subspaces ($Q_f$ is $n \times n_f$ matrix of the fast eigenvectors, $Q_s$ is $n \times n_s$ matrix of the slow eigenvectors, $\tilde{Q}_f$ is $n_f \times n$ matrix and $\tilde{Q}_s$ is $n_s \times n$ matrix). The parameter $\tau$ is a time scale splitting parameter. This splitting parameter determines the dimensions of the slow ($n_s$) and fast ($n_f$) sub-spaces.

Using a standard lineaization of the RHS of (23) at the point $\mathbf{Z_0}$ we get

$$\frac{\mathrm{d}\mathbf{Z}}{\mathrm{d}t} = \mathbf{F}(\mathbf{Z}) \approx \mathbf{F}(\mathbf{Z_0}) + \frac{\partial \mathbf{F}}{\partial \mathbf{Z}} \mid_{\mathbf{Z}=\mathbf{Z_0}} (\mathbf{Z} - \mathbf{Z_0}) \tag{26}$$

The Jacobian at the point $\mathbf{Z_0}$ can be represented as a product of three matrices: the transition matrix $Q$, a two-blocks representation $J_{M_J}$ of the Jacobian in the eigenvectors basis and inverse of the transition matrix $Q^{-1}$

$$\frac{\partial \mathbf{F}}{\partial \mathbf{Z}} \mid_{\mathbf{Z}=\mathbf{Z_0}} = M_J(\mathbf{Z_0}) = QJ_{M_J}Q^{-1} = M_J = \begin{pmatrix} Q_f & Q_s \end{pmatrix} \begin{pmatrix} J_{M_f} & 0 \\ 0 & J_{M_s} \end{pmatrix} \begin{pmatrix} \tilde{Q}_f \\ \tilde{Q}_s \end{pmatrix} \tag{27}$$

The square $(n \times n)$ matrix $J_{M_J}$ is decomposed into a two-block matrix. The blocks $J_{M_f}$, $J_{M_s}$ correspond to fast and slow invariant sub-spaces. The matrix $J_{M_f}$ is $n_f \times n_f$ and the matrix $J_{M_s}$ is $n_s \times n_s$.

Introduce the intermediate variable $\boldsymbol{\phi} = \mathbf{Z} - \mathbf{Z_0}$ and rewrite the expression (26) in the form

$$\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} = \mathbf{F}(\mathbf{Z_0}) + M_J(\mathbf{Z_0})\boldsymbol{\phi} = \mathbf{F}(\mathbf{Z_0}) + Q(\mathbf{Z_0})J_{M_J}(\mathbf{Z_0})Q^{-1}(\mathbf{Z_0})\boldsymbol{\phi} \tag{28}$$

Multiply both sides of (28) by the inverse matrix $Q^{-1}(\mathbf{Z_0})$

$$Q^{-1}(\mathbf{Z_0})\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} = Q^{-1}(\mathbf{Z_0})\mathbf{F}(\mathbf{Z_0}) + J_{M_J}(\mathbf{Z_0})Q^{-1}(\mathbf{Z_0})\boldsymbol{\phi}$$

and introduce new variable (this is a point of the transition from the original basis to the new one, which allows the decomposition into fast and slow motions)

$$\boldsymbol{\Psi} = Q^{-1}(\mathbf{Z_0})\boldsymbol{\phi}$$

With respect to the new variable the equation can be written in the form

$$\frac{\mathrm{d}\boldsymbol{\Psi}}{\mathrm{d}t} = \boldsymbol{\phi}\frac{\mathrm{d}Q^{-1}}{\mathrm{d}t}(\mathbf{Z_0}) + Q^{-1}(\mathbf{Z_0})\mathbf{F}(\mathbf{Z_0}) + J_{M_J}(\mathbf{Z_0})\boldsymbol{\Psi}$$

One can show that the first term in the RHS of the last equation is negligible under certain special conditions [16]. The equation is reduced to the simple equation in the form

$$\frac{\mathrm{d}\boldsymbol{\Psi}}{\mathrm{d}t} \approx Q^{-1}(\mathbf{Z_0})\mathbf{F}(\mathbf{Z_0}) + J_{M_J}(\mathbf{Z_0})\boldsymbol{\Psi}$$

According to the original algorithm of Maas and Pope (1992) the Intrinsic Low-Dimensional Manifold (ILDM) is determined by the following system of equations

$$\tilde{Q}_f(\mathbf{Z})\mathbf{F}(\mathbf{Z}) = \mathbf{0} \tag{29}$$

This definition means that the fast component of the original vector field $\mathbf{F}(\mathbf{Z})$, that corresponds to the ("big") fast block $\mathbf{J_{M_f}}$ of the Jacoby matrix representation, is vanished.

**ILDM-algorithm for singularly perturbed system**

Suppose that we initially have differential system in singularly perturbed form and we are interested in asymptotic expansion of ILDM equation in order to compare it with the invariant manifold. In this case the transition matrix $Q$, its inverse $Q^{-1}$ and the vector field have the following representation

$$Q = \begin{pmatrix} Q_{ff} & Q_{sf} \\ Q_{fs} & Q_{ss} \end{pmatrix}, \; \tilde{Q} = \begin{pmatrix} \tilde{Q}_{ff} & \tilde{Q}_{fs} \\ \tilde{Q}_{sf} & \tilde{Q}_{ss} \end{pmatrix}, \; \mathbf{F} = \begin{pmatrix} \epsilon^{-1}\mathbf{f} \\ \mathbf{g} \end{pmatrix}, \tag{30}$$

where $\tilde{Q}_{ff}$ is $n_f \times n_f$ matrix, $\tilde{Q}_{fs}$ is $n_f \times n_s$ matrix, $\tilde{Q}_{sf}$ is $n_s \times n_f$ matrix and $\tilde{Q}_{ss}$ is $n_s \times n_s$ matrix.

The ILDM-equation gets the form

$$\tilde{Q}_{ff}\mathbf{f} + \epsilon\tilde{Q}_{fs}\mathbf{g} = \mathbf{0}$$

In the zero approximation $\epsilon \to 0$ the equation is

$$\tilde{Q}_{ff}\mathbf{f} = \mathbf{0}.$$

If $\det \tilde{Q}_{ff} = 0$, then the last equation gets additional solutions ("ghost" manifolds) except the slow manifold $\mathbf{f} = \mathbf{0}$. This is one of reasons for "ghost" manifolds appearance. The others will be considered in the future works of the authors.

**Connection of the ILDM and the $C_a$-inflector**

Consider a two-dimensional system (8):

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}), \; \mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \; \mathbf{F}(\mathbf{x}) = \begin{pmatrix} f(x,y) \\ g(x,y) \end{pmatrix}.$$

Assume that $|\lambda_1| < |\lambda_2|$ and $\lambda_2 < 0$ hold in some domain $D$ of the phase plane. According to the definition of the $C_a$-inflector (10) its equation can be written as

$$fg_x + g(g_y - \lambda_1) = 0 \tag{31}$$

According to the ILDM-method, $\lambda_1$ is a slow eigenvalue and $\lambda_2$ is a fast eigenvalue in $D$. The equation for the ILDM in this case looks as

$$\frac{1}{\det(Q)}(fg_x + g(g_y - \lambda_1)) = 0, \tag{32}$$

where $Q$ is the new basis matrix, which is built from the eigenvectors of the Jacobi matrix of the system.

From the direct calculation we get

$$\det(Q) = g_x(\lambda_2 - \lambda_1) \tag{33}$$

Equations (31) and (32) show that in $D$ ILDM coincides with $C_a$-inflector up to the expression $g_x$.

**Remarks about possible non-coincidence of ILDM and slow invariant manifold**

The interesting fact is that the concept of the Intrinsic Low-Dimensional Manifolds is well known and widely used in the reduction methods [21]-[23], [25]. The following hints of "ghost"-manifolds existence are given in these studies:

(*i*) Consider the definition of the A-inflector (Eq.10) Sect. 2.3. We see that the points where the eigenvalues of the Jacobian are equal are out of that definition. In our study (see, for example, [6]) we show that the original algorithm cannot treat these points (curves, surfaces). It can be easily shown by formula (33). Namely, the expression $1/\det(Q)$ is always involved into an ILDM-equation and $\lambda_1 - \lambda_2 = 0$ is one of possibilities to the determinant to vanish. Therefore numerical application of the ILDM-algorithm yields "ghost"-objects in the points $\lambda_1 - \lambda_2 = 0$.

(*ii*) Note, that the asymptotic comparison between the A-inflector (R-inflector) and $T_a^0(\epsilon)$ ($T_r^0(\epsilon)$) in [23] was performed in the region $\Omega^0(\chi) = \{\mathbf{x} \mid \mid v_y^0 \mid \geq \chi\}$. One can show that the original algorithm does not work in the zones where $v_y^0 = 0$ and their neighborhoods.

(*iii*) In some cases it can be shown for a two-dimensional singularly perturbed system that in turning zones eigenvalues of Jacobi matrix are complex. It means that their real parts are identical. By the definition, $v_y^0 = 0$ in turning points. From the above we can conclude that turning zones are problematic for ILDM-method. The analysis of the algorithm shows that existence of complex eigenvalues is one of the main problems of the method.

(*iiii*) It should be noticed that the ILDM-method was used in [25] for analysis of fast-slow planar dynamical systems. In this study the Intrinsic Low-Dimensional Manifold was called a *slow tangent manifold*. It was defined as the curve on which the slow eigenvector is parallel to the velocity field (this definition coincides with the ILDM definition, see [16], [13]). It was shown that the slow tangent manifold lies close to the slow invariant manifold. In our study we demonstrate that in some situations the ILDM does not coincide with the slow invariant manifolds, and different disruptions of the original algorithm are reasoned by different types of non-linearity of a vector field of the considered ODE system.

### 2.5 TILDM

The remarks (*i*)-(*iiii*) show that the ILDM-algorithm has several disadvantages and some improved version is needed.

TILDM-method [2] is a modified version of the original ILDM approach of Maas and Pope. The additional letter "T" comes from the word "Transpose". The basic difference between the algorithms is that the TILDM uses the symmetric matrix $T = J \cdot J^t$ instead of the Jacobi matrix $J$. It is known that any symmetric matrix has real eigenvalues and orthogonal eigenvectors.

This solves one of the main problems of the ILDM-approach (complex eigenvalues with a large negative real part of the Jacobi matrix) and also problems connected to non-orthogonality of the eigenvectors. Note, that idea to exploit properties of a symmetrized (in some special sense) matrix was suggested in [9, 10].

Consider the differential system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, \mathbf{z}, \epsilon)$$

$$\epsilon\dot{\mathbf{z}} = \mathbf{g}(\mathbf{y}, \mathbf{z}, \epsilon),$$

where $\mathbf{y} \in K_1 \subset \Re^m$, $\mathbf{z} \in K_2 \subset \Re^n$, $0 < \epsilon \leq \epsilon_0$, functions $\mathbf{f}$, $\mathbf{g}$ derivatives are proportional to the unity when $\epsilon \to 0$.

Fix an arbitrary point $(\mathbf{y}, \mathbf{z})$. The Jacoby matrix is

$$J = \begin{pmatrix} D_y\mathbf{f} & D_y\mathbf{g} \\ \epsilon^{-1}D_z\mathbf{f} & \epsilon^{-1}D_z\mathbf{g} \end{pmatrix}$$

The corresponding symmetric matrix $T$ is

$$T = J \cdot J^t = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}$$

$T_{11}$ is $m \times m$ matrix with the elements proportional to $O(\epsilon^0)$, $T_{12}$ is $m \times n$ matrix with the elements proportional to $O(\epsilon^{-1})$, $T_{21}$ is $n \times m$ matrix with the elements proportional to $O(\epsilon^{-1})$, $T_{22}$ is $n \times n$ matrix with the elements proportional to $O(\epsilon^{-2})$. For arbitrary point $(\mathbf{y}, \mathbf{z})$ the matrix $T$ has positive eigenvalues and orthogonal eigenvectors. The eigenvalues of $T$ fall into two distinct groups: $n$ fast eigenvalues ( proportional to $O(\epsilon^{-2})$ and $m$ slow ones (proportional to $O(\epsilon^0)$).

From linear algebra we know that in some orthonormal basis $Q$ the matrix $T$ has a diagonal form with its eigenvalues in the diagonal. The eigenvalues can appear along the diagonal in any desirable order.

$$T = QT_dQ^t,$$

where

$$Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}, \quad Q^t = \begin{pmatrix} Q_{11}^t & Q_{21}^t \\ Q_{12}^t & Q_{22}^t \end{pmatrix}$$

Here $\begin{pmatrix} Q_{11} \\ Q_{21} \end{pmatrix}$ is the orthonormal basis of the fast sub-space, $\begin{pmatrix} Q_{12} \\ Q_{22} \end{pmatrix}$ is the orthonormal basis of the slow sub-space. $T_d$ is a following diagonal matrix

$$T_d = \begin{pmatrix} \Lambda_f & 0 \\ 0 & \Lambda_s \end{pmatrix}$$

Here $\Lambda_f$ is a fast block ($n \times n$ block of the fast eigenvalues), Here $\Lambda_s$ is a slow block ($m \times m$ block of the slow eigenvalues). By the definition, the equation

for TILDM manifold is multiplication of the fast part of the matrix $Q^t$ by the vector field $\mathbf{F} = (\mathbf{f},\ \epsilon^{-1}\mathbf{g})^t$:

$$Q_{11}^t \mathbf{f} + \epsilon^{-1} Q_{21}^t \mathbf{g} = \mathbf{0}. \tag{34}$$

Asymptotic analysis with respect to the small parameter $\epsilon$ shows that zeroth approximation of the TILDM coincides with the zeroth approximation of the slow invariant manifold (slow surface $\mathbf{g} = \mathbf{0}$). It must be noticed that the turning points are not problematic for the TILDM-algorithm and this fact is one of the most important advantages of the method.

## 3 "Ghost" ILDM-Manifolds Examples

In this section the examples of "ghost"-manifolds appearance will be demonstrated. The examples 1-3 are theoretical ones; the example 4 is practical one. All presented systems are written in the singularly perturbed form. Nevertheless the "ghost" objects appear when we apply the ILDM method.

*Example 1.* This example will demonstrate appearance of a large number of "ghost" manifolds because of non-correct fast direction defined by the ILDM method. It should be noticed that the slow manifold of this system does not have turning points and also it is stable. Consequently according to the conjecture [24] the ILDM manifold should coincide with the invariant manifold, but this statement is not true for this example. In other words, the present example can be considered as a counterexample for the conjecture suggested in [24].

Consider the following system of differential equations with small parameter $\epsilon$ :

$$\epsilon \dot{x} = -x - \sin(x) - \sin(y)$$

$$\dot{y} = -y$$

The slow manifold (the manifold of critical points) is given by the equation

$$-x - \sin(x) - \sin(y) = 0 \tag{35}$$

The slow manifold is shown as the central object on Fig.1(below). Application of the ILDM method for this example provides us with two equations for domains with different hierarchy of the eigenvalues $\lambda_{1,2}$:

$$-x - \sin(x) - \sin(y) + \frac{\epsilon y \cos(y)}{-1 + \epsilon - \cos(x)} = 0,\ |\lambda_1| > |\lambda_2|$$

$$y = 0,\ |\lambda_2| > |\lambda_1|$$

Fig.1(upper) demonstrates two ILDM manifolds (solid lines)and a system's trajectory (thick dashed line). Fig.1(below) demonstrates the slow curve (solid

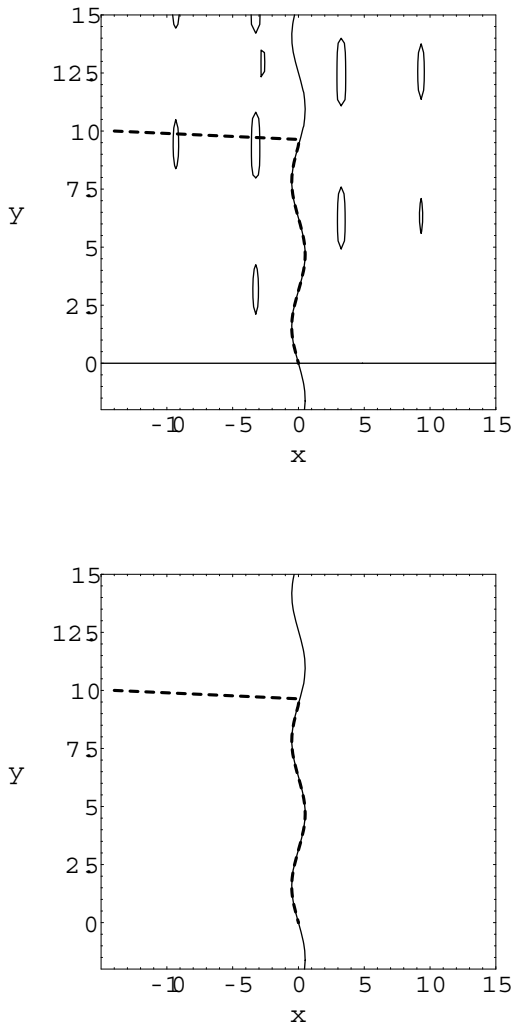



Fig. 1: Example 1. Upper graph – ILDM and trajectory, lower graph – slow curve and trajectory

line)and a system's trajectory (thick dashed line). On the figure we can see that the trajectory with arbitrary initial conditions approaches the ILDM curve passing through "ghost" manifolds (fast motion, almost parallel to the x axe). it must be noticed that one of the ILDM-manifolds (the central part of Fig.1(upper)) is very close to the slow manifold.

*Example 2.* This example will demonstrate the essential perturbations produced by the ILDM algorithm on unique slow manifold. Consider the following

system of differential equations with small parameter $\epsilon$

$$\epsilon\dot{x} = -x - \sin(x) - \sin(y) + 10$$

$$\dot{y} = -2y - \sin(y)$$

The method of invariant manifolds provides us with the slow manifold as follows (dashed line on Fig.2).

$$-x - \sin(x) - \sin(y) + 10 = 0 \tag{36}$$

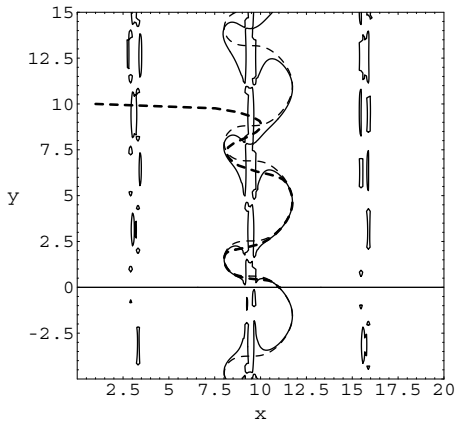



Fig. 2: Application of ILDM algorithm for Theoretical Example 2

As in the previous example we get two ILDM-equations (it depends on which of the eigenvalues is "fast" in the considered domain) applying the algorithm. Fig.2 demonstrates two ILDM manifolds (solid lines), the slow curve (central dashed line) and a system's trajectory (thick dashed line).

*Example 3.* Consider the following system of differential equations with small parameter $\epsilon$

$$\epsilon\dot{x} = -x/2 - \sin(x) - \sin(xy)$$

$$\dot{y} = -y$$

We will show that this example is pathological for the ILDM algorithm in some sense. The method of invariant manifolds provides us with the slow manifold

$$-x/2 - \sin(x) - \sin(xy) = 0$$

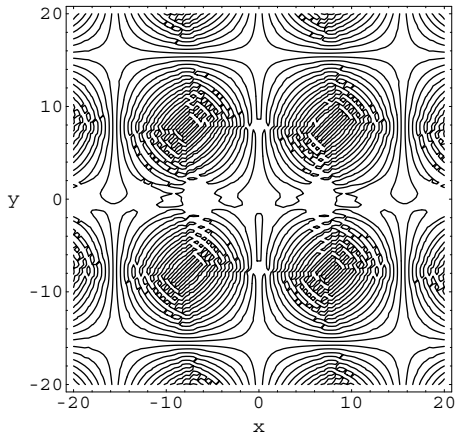



Fig. 3: The curve on which the eigenvalues of the Jacobian in Example 3 are equal one to another

The analysis of the eigenvalues shows that either $|\lambda_1| >> |\lambda_2|$ or $|\lambda_1| = O(|\lambda_2|)$ and the last relation holds in almost all points of the phase plane. On Fig.3 the curve is depicted, on which $\lambda_1 = \lambda_2$. Then, in some small vicinity of this curve the eigenvalues are comparable. We see that the curve fills up the whole plane and has a very interesting form. Let us remark that the system is written in the singularly perturbed form with explicit small parameter.

Fig. 4 shows that the ILDM method approximates the slow manifold very well.

*Example 4.* Consider classical model of thermal explosion in a gas. The dimensionless model reads as

$$\epsilon \frac{\mathrm{d}\theta}{\mathrm{d}t} = \eta \exp\left(\frac{\theta}{1+\beta\theta}\right) - \alpha\theta = f(\theta, \eta) \tag{37}$$

$$\frac{\mathrm{d}\eta}{\mathrm{d}t} = -\eta \exp\left(\frac{\theta}{1+\beta\theta}\right) = g(\theta, \eta) \tag{38}$$

$$\theta(0) = 0,\ \eta(0) = 1 \tag{39}$$

Here $\theta$ is a dimensionless temperature, $\eta$ is a dimensionless concentration, $\alpha$ is a dimensionless heat loss parameter, $\epsilon$ is a reciprocal of the dimensionless adiabatic temperature rise, $\beta$ is a dimensionless ambient temperature. For realistic combustible gas mixtures typical values of $\epsilon$ lie in the interval $(0.01,\ 0.1)$ and the following relation is satisfied: $\beta^2 < \epsilon < \beta$. Therefore this system can be considered as a singularly perturbed system with small parameter $\epsilon$, where $\theta$ is a fast variable, $\eta$ is a slow variable.

Note that the dynamics of the system is known very well, see, for example, [2], [6], [7], [8]. In particular, in [7], [8] the dynamics of the system was

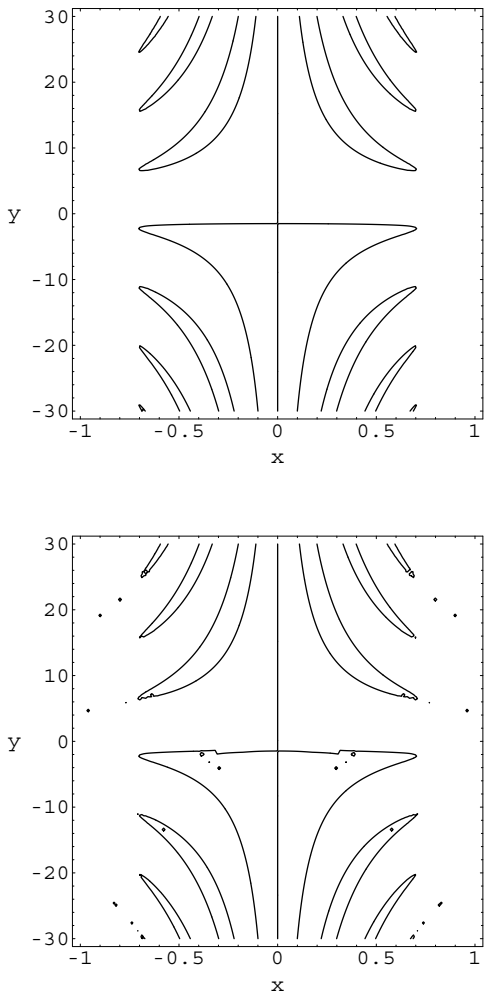



Fig. 4: Example 3. Upper graph – the slow manifold, lower graph - ILDM

analyzed in framework of the method of invariant manifolds (MIM); in [6] the detailed analysis of the ILDM algorithm application to the system (37)-(39) was performed; in [2] the modification of ILDM (TILDM) was applied. Here we remind only basic results of the ILDM-method application.

According to the ILDM-method, the Jacobian of the system is

$$J = \begin{pmatrix} \epsilon^{-1} f_\theta & \epsilon^{-1} f_\eta \\ g_\theta & g_\eta \end{pmatrix}$$

The eigenvalues are

$$\lambda_{1,2} = 1/2(\epsilon^{-1} f_\theta + g_\eta \pm \sqrt{D(\theta,\eta)}),$$

where

$$D(\theta,\eta) = (\epsilon^{-1} f_\theta + g_\eta)^2 - 4\epsilon^{-1}(f_\theta g_\eta - f_\eta g_\theta)$$

There are three possibilities depending on the sign of the discriminant $D(\theta,\eta)$:
*a)* $D(\theta,\eta) > 0$. The Jacobi matrix provides us with two real different eigenvalues. Depending on order of magnitude of the eigenvalues two ILDM equations are obtained for different domains of the phase space.
*b)* $D(\theta,\eta) = 0$. The Jacobian provides us with two identical eigenvalues. In this case one of the main assumptions of the ILDM approach does not hold, namely, the eigenvalues can not be sub-divided into two different groups (24). It means that there is no splitting on fast and slow eigenvalues and the ILDM-method can not be applied.
*c)* $D(\theta,\eta) < 0$. The Jacobian provides us with two complex eigenvalues. It means that their real parts are identical. We can repeat the previous argument to conclude that the original technique does not work in this case. The region in the phase plane corresponding to this case is the domain between the curves $Y_+$ and $Y_-$ (see Fig. 5).

Fig. 5 shows all the curves $(M1, M2, Y_\pm)$ obtained by the ILDM-algorithm (thick solid lines) and the slow manifold (dashed line).

The functions $Y_\pm(\theta)$ are the solutions of the equation $D(\theta, Y_\pm(\theta)) = 0$ and they have their own sense. These functions serve as the separating lines on the phase plane between domains of real and complex eigenvalues.

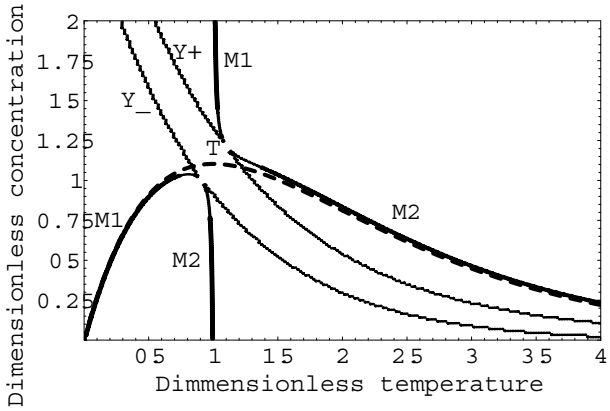



Fig. 5: ILDM and the slow curve for the Semenov's model

Let us now illustrate briefly the basic steps of the system's analysis by the method of invariant manifolds.

In accordance with Sect. 2.1, the slow curve of the system (37)-(39) is given by

$$f(\theta,\eta) \equiv \eta \exp\left(\frac{\theta}{1+\beta\theta}\right) - \alpha\theta = 0 \tag{40}$$

Eq.(40) has a unique isolated solution $\theta(\eta)$ for all $\eta$, except at the turning points, at which $f = 0,\ f_\theta = 0$. The slow curve has two turning points. On Fig. 5 we can see one of them $T$. The second point has a very big $\theta$-coordinate for reasonable values of the system's parameters. On the slow curve the relative rates of the processes are comparable, and the system's dynamics is governed by the reduced system on the slow curve:

$$\frac{\mathrm{d}\eta}{\mathrm{d}t} = -\eta \exp\left(\frac{\theta(\eta)}{1+\beta\theta(\eta)}\right)$$

where $\theta(\eta)$ is given by (40).

The first approximation of the slow invariant manifold reads

$$\eta = \alpha\theta \exp\left(-\frac{\theta}{1+\beta\theta}\right) + \epsilon\frac{\theta(1+\beta\theta)}{\theta-(1+\beta\theta)^2}$$

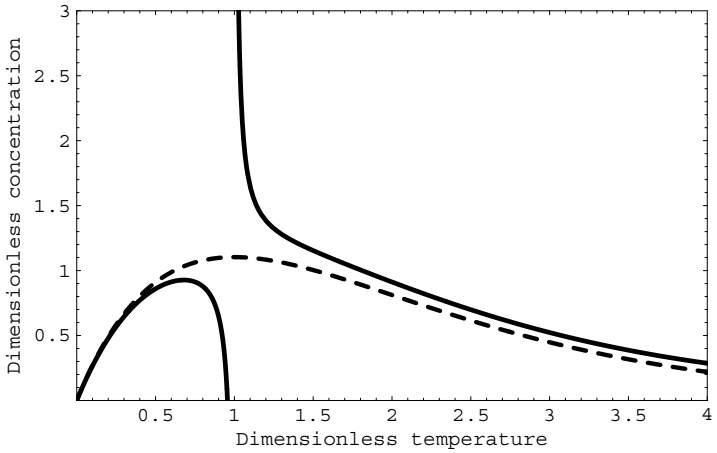



Fig. 6: The zeroth and the first approximations of the slow invariant manifold for the Semenov's model

Fig. 6 represents both the slow curve (dashed line) and the first approximation (solid line) of the exact manifold.

If we compare Fig. 5 and Fig. 6 we see that the first approximation of the invariant manifold and the intrinsic low-dimensional manifolds are identical,

except the lines $Y_{\pm}(\theta)$. These lines separate in the phase plane domains of real and complex eigenvalues. Between $Y_-$ and $Y_+$ the "transition zone" is located (the "gray zone" in [6]), which is large in this example, because of the existence of complex eigenvalues far from the "transition point". In this zone the ILDM method does not work and it confirms [7] that there is no division into fast and slow processes in "transition zone".

### 3.1 Conclusions

We can see that there exist "ghost" manifolds as a result of ILDM-method application. The first example demonstrated that even for two-dimensional singularly perturbed system the slow manifold of which is stable and does not have turning points the ILDM does not coincide with the invariant manifold. The second example demonstrated appearance of "ghost"- manifolds in neighborhoods of the turning points. It is known (see, for example[1], [2], [6]) that in these zones the original ILDM method doesn't work. Let us remind that the $C_a$-inflector (see Sect. 2.3) is not defined in zones containing turning points. The third example is pathological for the ILDM-algorithm in some sense. In spite of the original system of equations is done in the standard singularly perturbed form the processes involved are comparable in almost all phase plane. Nevertheless, the algorithm locates the slow manifold well. The application of the algorithm gives "ghost" manifolds. The forth example demonstrated one of the main ILDM-method's problems: existence of complex eigenvalues of the Jacobi matrix.

## 4 Criteria for "Ghost"-Manifolds Identification

In Sect. 3 we demonstrated that application of the original Maas and Pope algorithm produces so-called "ghost"-manifolds. In this section we suggest two criteria that allow to distinguish the "ghost"-manifolds from the correct ones.

### 4.1 Criterion 1: "Normal Vector"

The idea of the criterion "Normal vector" can be described as follows. Fix an arbitrary point that belongs to the invariant manifold of the system (1)-(2). In this point the vector field $\mathbf{F} = (\epsilon^{-1}f,\ g)^T$ and vector normal to the invariant manifold $\mathbf{n}$ are $\epsilon$-close to orthogonal pair, i.e. the value of $(\mathbf{F},\ \mathbf{n})$ is comparable with $\epsilon$. If a point is far from the invariant manifold then the vector field and the normal have some angle $\mid \alpha - \pi/2 \mid \sim O(1)$ and $(\mathbf{F},\ \mathbf{n})$ cannot be small.

Apply the suggested criterion for discrimination of "ghost"-manifolds in theoretical example 1. The slow manifold for this system is exactly known (Eq. (35), Fig.1). Fig.7 demonstrates result of application of the criterion.

The horizontal axe is $x$-coordinate of the checked point, the vertical axe shows values of $\log(\mathbf{F}, \mathbf{n})$ for different $x$. For $x \in (-1,\ 1)$ we have $\log(\mathbf{F}, \mathbf{n}) = O(1)$. It means that $(\mathbf{F}, \mathbf{n}) = O(\epsilon)$. According to the suggested criterion the point belongs to the correct ILDM-branch. For $x$ from any other zone we have $(\mathbf{F}, \mathbf{n}) = O(1)$. That is, the point belongs to "ghost"-manifold.

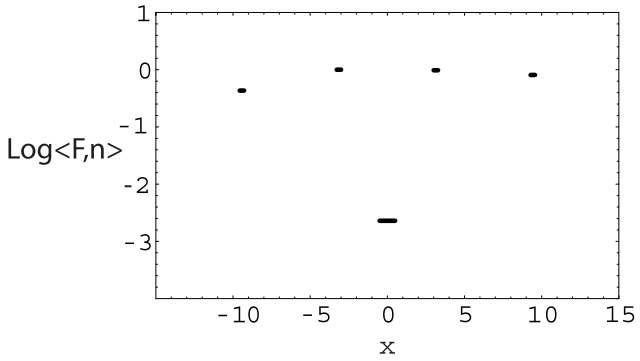



Fig. 7: Application of criterion "Normal vector" for Theoretical Example 1

The obtained results are confirmed by Fig.1. For $x \in (-1,\ 1)$ the ILDM coincides with the slow manifold (we do not see "ghost" manifolds in this zone); for $x$ out of this interval there is only "ghost" ILDM.

Apply the suggested criterion for discrimination of "ghost"-manifolds in theoretical example 2. The slow manifold for this system is exactly known (Eq. (36), Fig.2).

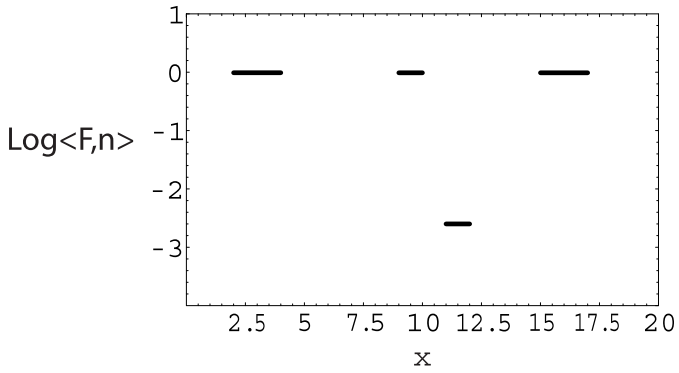



Fig. 8: Application of criterion "Normal vector" for Theoretical Example 2

Application of the criterion is shown on Fig.8. Difference of values $\log(\mathbf{F}, \mathbf{n})$ is easily seen. According to Fig.8 points from $x \in (11,\ 12)$ belong to the real ILDM, because $\log(\mathbf{F}, \mathbf{n}) = O(1)$ and so $(\mathbf{F}, \mathbf{n}) = O(\epsilon)$. Points from other intervals belong to "artificial" ILDM-branches. This result is confirmed by Fig.2. We can see that for $x \in (11,\ 12)$ the ILDM coincides with the slow manifold. For $x$ out of this interval there is only "ghost" ILDM.

### 4.2 Criterion 2: "Slow Matrix"

Consider the system of ordinary differential equations (23). Suppose that this system can be represented as a singularly perturbed system (1)-(2) in some coordinate system. In this sub-section we find an invariant that does not depend on a choice of coordinate system and can distinguish between ILDM-manifolds that correspond to true invariant manifold and ILDM-manifolds that are far from any invariant manifold. Our analysis is asymptotic one.

We know from Sect. 2.3 that that ILDM-manifolds are solutions of Eq. (29). Denote the intrinsic low-dimensional manifold by $S$. Let us analyze values of $\tilde{Q}_s(\mathbf{Z})\mathbf{F}(\mathbf{Z}) = \mathbf{0}$ for different points $(\mathbf{x},\ \mathbf{y})$. The matrix $\tilde{Q}_s$ can be represented as $\tilde{Q}_s = (\tilde{Q}_{sf}\ \ \tilde{Q}_{ss})$ see Eq. (30). That is, we have

$$\tilde{Q}_s\mathbf{F} = \frac{1}{\epsilon}\tilde{Q}_{sf}\mathbf{f} + \tilde{Q}_{ss}\mathbf{g} \tag{41}$$

Let us remind that zeroth approximation of the slow invariant manifold is defined by $\mathbf{f} = \mathbf{0}$. If the ILDM-manifold $S$ belongs to $\epsilon$-neighborhood of the slow invariant manifold, then the term $\epsilon^{-1}\mathbf{f}$ has the order $O(1)$ on $S$. If the ILDM-manifold is far from the slow invariant manifold, then the term $\epsilon^{-1}\mathbf{f}$ is comparable with the value $O(\epsilon^{-1})$ on $S$.

From (41) we can conclude that

($i$) If ILDM-manifold $S$ belongs to $\epsilon$-neighborhood of slow invariant manifold, then $\tilde{Q}_s(\mathbf{Z})\mathbf{F}(\mathbf{Z})$ has the order $O(\mid \mathbf{g} \mid)$ on $S$.

($ii$) If ILDM-manifold is far from any slow invariant manifold then $\tilde{Q}_s(\mathbf{Z})\mathbf{F}(\mathbf{Z}) >> \mid \mathbf{g} \mid$ on $S$.

Then, the described criterion suggests to use values of $\tilde{Q}_s(\mathbf{Z})\mathbf{F}(\mathbf{Z})$ for discrimination of "ghost"-manifolds.

Apply the suggested criterion for discrimination of "ghost" manifolds in the theoretical Example 1. The eigenvalues of the Jacobi matrix are $\lambda_1 = (-1 - cos(x))/\epsilon$, $\lambda_2 = -1$. Consider any artificial branch, for example, $x \in (2,\ 4)$. The eigenvalues analysis shows that in this region $|\lambda_2| > |\lambda_1|$. Then, $\tilde{Q}_f\mathbf{F} = -y$ and $\tilde{Q}_s\mathbf{F} = -x - \sin(x) - \sin(y) + \frac{\epsilon y \cos(y)}{-1+\epsilon-\cos(x)}$. The result of application of the criterion is given in Table 1.

Table 1 shows that the values of $\tilde{Q}_s\mathbf{F}$ are much bigger than $g$. Therefore according to the suggested criterion the points from this region belong to "ghost" manifold.

Check the points of the ILDM that belong to $\epsilon$-neighborhood of the slow invariant manifold, $x \in (-1,\ 1)$. Then, $\tilde{Q}_f\mathbf{F} = -x - \sin(x) - \sin(y) + \frac{\epsilon y \cos(y)}{-1+\epsilon-\cos(x)}$

Table 1: Application of criterion "Slow matrix" for Theoretical Example 1

| $x$ | $y$ | $\tilde{Q}_s\mathbf{F}$ |
|---|---|---|
| 2.588 | 0 | -311.375 |
| 3.439 | 0 | -314.597 |
| 2.89 | 0 | -313.92 |

and $\tilde{Q}_s\mathbf{F} = -y \equiv g$. Therefore for these points $|\tilde{Q}_s\mathbf{F}| = O(|g|)$. According to the criterion this means that all the points from the considered interval belong to correct ILDM manifold.

Results of the suggested criterion are conformed by the method of invariant manifolds, criterion 1 and Fig.1.

## 5 Conclusions

The present paper represents a natural continuation of the authors work on a comparative analysis of the two powerful asymptotic methods ILDM and MIM.

As any other algorithm, ILDM has its own restrictions, which were partly demonstrated in the present paper on a number of examples. It was shown, that ILDM can not treat the regions of the phase space, where the leading eigenvalues of the Jacobi matrix are equal. In particular, it means, that the ILDM approach may face problems in the vicinity of the turning surfaces, where the leading eigenvalues are normally complex (their real values are equal and there is no splitting in rates of change of the processes involved). As a result of the ILDM application in these regions of the phase space, so called ghost manifolds can appear. It is illustrated by a number of examples.

The problem of the determination and elimination of the ghost manifolds is of high importance. A numerical criterion allowing distinguishing the ghost manifolds from the true ones is suggested in the present paper. The criterion is based on the unique properties of the true invariant manifolds. The efficiency of the suggested criterion is demonstrated on the number of the examples introduced earlier.

## References

1. S. Borok, I. Goldfarb, V. Gol'dshtein: "Ghost" ILDM – Manifolds and their Discrimination. In: 20th Annual Symposium of the Israel Section of the Combustion Institute, 55–57 (Beer-Sheva, Israel 2004)
2. V. Bykov, I. Goldfarb, V. Gol'dshtein, U. Maas: On a Modified Version of ILDM Approach: Asymptotic Analysis Based on Integral Manifolds. *IMA J. Appl. Math.*, submitted

3. M. Davis, R. Skodje: Geometric Investigation of Low-Dimensional Manifolds in Systems Approaching Equilibrium. *J. Chem. Phys.* **111** (3), 859–874 (1999)
4. N. Fenichel: Geometric Singular Perturbation Theory for Ordinary Differential Equations. *J. Diff. Eq.* **31**, 53–98 (1979)
5. S.J. Fraser: The Steady State and Equilibrium Approximations: A Geometrical Picture. *J. Chem. Phys.* **88** (8), 4732–4738 (1988)
6. I. Goldfarb, V. Gol'dshtein, U. Maas: Comparative Analysis of Two Asymptotic Approaches Based on Integral Manifolds. *IMA J. Appl. Math.* **69**, 353–374 (2004)
7. V. Gol'dshtein, V. Sobolev: *Qualitative Analysis of Singularly Perturbed Systems (in Russian)* (Institute of Mathematics, Siberian Branch of USSR Academy of Science, Novosibirsk 1988)
8. V. Gol'dshtein, V. Sobolev: Singularity Theory and Some Problems of Functional Analysis. *AMS Translations, Series 2*, **153**, 73–92 (1992)
9. A.N. Gorban, I.V. Karlin: Methods of invarinat manifolds for chemical kinetics. *Chemical Engineering Science* **58** (21), 4751–4768 (2003).
10. A.N. Gorban, I.V. Karlin, A.Yu. Zinoviev: Constructive methods of invarinat manifolds for kinetic problems. *Physics Reports* **396**, 197–403 (2004).
11. M. Hadjinicolaou, D.M. Goussis, Asymptotic Solutions of Stiff PDEs with the CSP Method: the Reaction Diffusion Equation. *SIAM Journal of Scientific Computing* **20**, 781–910 (1999)
12. J.K. Hale: *Ordinary Differential Equations* (Wiley–Interscience, New-York 1969)
13. H.G. Kaper, T.J. Kaper: Asymptotic Analysis of Two Reduction Methods for Systems of Chemical Reactions. *Phys. D* **165** (1-2), 66–93 (2002)
14. S.H. Lam, D.M. Goussis: Understanding Complex Chemical Kinetics with Computational Singular Perturbation. *Proc. Comb. Inst.* **22**, 931–941 (1988)
15. S.H. Lam, D.M. Goussis: The CSP Method for Simplifying Kinetics. *J. Chem. Kinetics* **26**, 461–486 (1994)
16. U. Maas, S. B. Pope: Simplifying Chemical Kinetics: Intrinsic Low-Dimensional Manifolds in Composition Space. *Combustion and Flame* **88**, 239–264 (1992)
17. E.F. Mishchenko, N.Kh. Rozov: *Differential Equations with Small Parameter and Relaxation Oscillations* (Plenum Press, New York 1980)
18. A.H. Nguyen, S.J. Fraser: Geometrical Picture of Reaction in Enzyme Kinetics. *J. Chem. Phys.* **91** (1), 186–193 (1989)
19. K. Nipp: Invariant Manifolds of Singularly Perturbed Ordinary Differential Equations. *ZAMP* **36**, 309–320 (1985)
20. K. Nipp: Numerical Integration of Stiff ODEs of Singular Perturbation Type. *ZAMP* **42**, 53–79 (1991)
21. M. Okuda: A New Method of Nonlinear Analysis for Threshold and Shaping Actions in Transient States. *Prog. Theoret. Phys.* **66** (1), 90–100 (1981)
22. M. Okuda: A Phase–Plane Analysis of Stability in Transient States. *Prog. Theoret. Phys.* **68** (1), 37–48 (1982)
23. M. Okuda: Inflector Analysis of the second stage of the transient phase for an enzymatic one-substrate reaction. *Prog. Theoret. Phys.* **68** (6), 1827–1840 (1982)
24. C. Rhodes, M. Morari, S. Wiggins: Identification of the Low Order Manifolds: Validating the Algorithm of Maas and Pope. *Chaos* **9** (1), 108–123 (1999)
25. M.R. Roussel: A rigorous approach to steady-state kinetics applied to simple enzyme mechanisms. Ph.D. Thesis, U. Toronto (1994)

26. M.R. Roussel, S.J. Fraser: Geometry of the Steady-State Approximation: Perturbation and Accelerated Convergence Methods. *J. Chem. Phys.* **93** (2), 1072–1081 (1990)
27. M.R. Roussel, S.J. Fraser: On the Geometry of Transient Relaxation. *J. Chem. Phys.* **94** (11), 7106–7113 (1991)
28. M.R. Roussel: Forced–Convergence Iterative Schemes for the Approximation of Invariant Manifolds *J. Chem. Math.* **21**, 385–393 (1997)
29. M.R. Roussel, S.J. Fraser: Invariant Manifold Methods for Metabolic Model Reduction. *Chaos* **11** (1), 196–206 (2001)
30. N.N. Semenov: *Z. Phys. Chem.* **48**, 571–581 (1928)
31. V. A. Sobolev: *Geometrical Decomposition of Multi-scale Systems*, (Boole Centre for Research in Informatics, University College Cork, Preprint)(2003)
32. B.B. Strygin, V.A. Sobolev: *Decomposition of motions by the Integral Manifolds Method (in Russian)* (Nauka, Moscow 1988)
33. M. Valorani, D.M. Goussis: Explicit Time–Scale Splitting Algorithm for Stiff Problems: Auto–Ignition of Gaseous Mixtures behind a Steady Shock. *J. Comput. Phys.* **169**, 44–79 (2001)
34. A. Zagaris, H.G. Kaper, T.J. Kaper: Analysis of the Computational Singular Perturbation Reduction Method for Chemical kinetics. *J. Nonlinear Sci.* **14**, 59–91 (2004)

# Dynamic Decomposition of ODE Systems: Application to Modelling of Diesel Fuel Sprays

V. Bykov[1], I. Goldfarb[2], V. Gol'dshtein[3], S. Sazhin[4] and E. Sazhina[5]

[1] Institute for Technical Thermodynamics, Karlsruhe University, Kaiserstrasse 12, 76128 Karlsruhe, Germany, bykov@itt.uni-karlsruhe.de
[2] Department of Mathematics, Ben-Gurion University of the Negev, P.O.B. 653, Beer-Sheva 84105, Israel, goldfarb@cs.bgu.ac.il
[3] Department of Mathematics, Ben-Gurion University of the Negev, P.O.B. 653, Beer-Sheva 84105, Israel, vladimir@bgu.ac.il
[4] School of Engineering, Faculty of Science and Engineering, University of Brighton, Cockcroft Building, Brighton BN2 4GJ, UK, S.Sazhin@brighton.ac.uk
[5] School of Engineering, Faculty of Science and Engineering, University of Brighton, Cockcroft Building, Brighton BN2 4GJ, UK, e.m.sazhina@brighton.ac.uk

**Summary.** A new method of decomposition of multiscale systems of ordinary differential equations is suggested. The suggested approach is based on the comparative analysis of the magnitudes of the eigenvalues of the matrix $\mathbf{JJ}^*$, where $\mathbf{J}$ is the local Jacobi matrix of the system under consideration. The proposed approach provides with the separation of the variables into fast and slow ones. The hierarchy of the decomposition is subject of variation with time, therefore, this decomposition is called dynamic. Equations for fast variables are solved by a stiff ODE system solver with the slow variables taken at the beginning of the time step. This is considered as a zeroth order solution for these variables. The solution of equations for slow variables is presented in a simplified form, assuming linearised variations of these variables during the time evolution of the fast variables. This is considered as the first order approximation for the solution for these variables or the first approximation for the fast manifold. The new approach is applied to numerical simulation of diesel fuel spray heating, evaporation and the ignition of fuel vapour/ air mixture. The results show advantages of the new approach when compared with the one proposed by the authors earlier and the conventional CFD approach used in computational fluid dynamics codes, both from the point of view of accuracy and CPU efficiency.

## 1 Introduction

The decomposition of complex systems into simpler subsystems is almost universally used in engineering and physics applications. It allows the numerical

simulation to focus on the subsystems, thus avoiding substantial difficulties and instabilities related to numerical simulation of the original systems.

As an example of such decomposition we can mention the solutions of ordinary and partial differential equations (ODEs and PDEs) describing spray dynamics in computational fluid dynamics (CFD) codes. Numerical spray modelling is traditionally based on the Lagrangian approach coupled with the Eulerian representation of the gas phase. This permits the decomposition of complicated and highly nonlinear systems of PDEs, describing interactions between computational cells, and the systems of ODEs that govern processes in individual computational cells, including liquid/gas phase exchange and chemical kinetics. The systems of ODEs are usually integrated using much shorter time steps $\delta t$ (typically $10^{-6}$ s) than the global time steps used for calculating the gas phase $\Delta t$ (typically $10^{-4}$ s). Thus the decomposition of ODEs and PDEs is *de facto* used although its basis has not been rigorously investigated [1] - [2].

Further decomposition of the system of ODEs, describing droplet dynamics inside individual computational cells, is widely used as well. The simplest decomposition of this system is based on the sequential solution of individual subsystems comprising this system. In this approach, the solution of each individual subsystem for a given subset of variables is based on the assumption that all the other variables are fixed. The sequence of solutions of individual subsystems is often chosen rather arbitrarily and the results sometimes vary substantially depending on the order in which these subsystems are solved. Undoubtedly, an arbitrary choice of decomposition and sequential integration of subsystems might lead to substantial and uncontrollable errors. As a result, in the case of a multiscale system, the reliability of this approach becomes questionable altogether.

A similar system decomposition into lower dimension subsystems has been used for modelling $CO_2$ lasers [3] and analyses of the Shell model equations [4], in constructing reduced chemical mechanisms based on Intrinsic Low-Dimensional Manifolds (ILDM) [5] - [7] and its modification TILDM [8], in development of constructive methods for invariant manifolds for problems of chemical kinetics [9] - [10], in the method of Computational Singular Perturbation (CSP) [11] - [15]. There are many similarities between these methods. They are based on a rigorous separation of timescales, such that 'fast' and 'slow' subspaces of the chemical source term are defined, and mechanisms of much reduced stiffness are constructed.

A useful analytical tool for the analysis of stiff systems of ODEs, used for the modelling of spray heating, evaporation and ignition of fuel vapour/air mixture, could be based on the geometrical asymptotic approach to singularly perturbed systems (Method of Integral Manifolds - MIM) as developed by Gol'dshtein and Sobolev [16] - [17] for combustion applications. This approach is based on the general theory of the integral manifolds [18] - [19].

These approaches to decomposing systems of ODEs were developed and investigated with a view to their application to rather special problems (e.g.

complex chemical kinetics), and were based on a number of assumptions, the justification of which is not at first obvious in most engineering applications. The underlying philosophy of these approaches, however, seems to be attractive for application to the analysis in a wide range of physical and engineering problems including spray modelling in CFD codes.

A general method of decomposition of stiff systems of ODEs into fast and slow parts was suggested in [20, 21]. This decomposition was based on the comparison of the rates of change of variables. In contrast to most previous approaches, where this decomposition was fixed in time (fixed decomposition), in the model developed in [20, 21] it was allowed to change with time (dynamic decomposition). The efficiency of this approach was demonstrated in the example of its application to the numerical modelling of heating, evaporation, and ignition of diesel fuel spray.

The model described in this paper is based on the further development of the model described in [20, 21]. Its main idea is described in the Section 2. The mathematical background of the model and the assumptions on which it is based are discussed in Sections 3 and 4. The application of the model to the problem of numerical modelling of heating, evaporation, and ignition of diesel fuel spray (a problem similar to the one discussed in [20, 21]) is presented in Section 5. The main results of the paper are summarised in Section 6.

## 2 Dynamic Fast-Slow Decomposition: Underlying Philosophy

As in the case of the model described in [20, 21], the focus of the new model will be on stiff systems of ODEs. The stiffness of these systems is known to create problems in their numerical solution. In our approach, however, this stiffness can play a positive role and make the numerical solution of the systems of ODEs easier. Referring to the terminology of asymptotic analysis, stiffness means that the system is multiscale and there can be at least two essentially distinct time scales. This allows us to subdivide all variables into fast and slow ones.

Our approach is ultimately based on the Method of Integral Manifolds (MIM) mentioned in the Introduction. This method is essentially focused on the analysis of the systems of ODEs written in the form:

$$\varepsilon \frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} = \mathbf{F}(\mathbf{X},\ \mathbf{Y}, \varepsilon) \tag{1}$$

$$\frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}t} = \mathbf{G}(\mathbf{X},\ \mathbf{Y}, \varepsilon), \tag{2}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are $n$ and $m$-dimensional vector variables, and $0 < \varepsilon \ll 1$ is a small positive parameter. The rate of change of vector $\mathbf{X}$ tends to infinity when $\varepsilon \longrightarrow 0$ if $\mathbf{F}(\mathbf{X},\mathbf{Y}, \varepsilon) \neq 0$). Hence, Equation (1) describes the so called

fast sub-system, while Equation (2) describes the so called slow sub-system. In practical implementations of the integral manifold method a number of simplifying assumptions have been made. These include the assumption that the slow variable is constant during the fast processes. When $\varepsilon \to 0$ and the functions in the RHS of Equations (1)-(2) are of the same order of magnitude (at least, in some bounded domain), the system (1)-(2) shows a dynamical behaviour characterised by the presence of two sufficiently different time scales. The difference between the rates of change of two vectors $(\mathbf{F}, \mathbf{G})$ is determined by $\varepsilon$.

Although this method has been widely used for the qualitative analyses of thermal explosion of combustible mixture of fuel droplets and air [22] - [24], its direct application to quantitative modelling of realistic physical systems is rather restrictive. This is due to a number of factors. Firstly, these systems are usually described by many equations, with different characteristic time scales, and their division into slow and fast subsystems is not at first obvious. Also, even if this division is possible, then the value $\varepsilon$ is expected to be small but not infinitely small. This implies that we can no longer assume that fast variable changes infinitely fast and slow variable is constant. Finally, in the original version of this method, the subdivision of the variables into fast and slow ones was fixed and not allowed to change with time (fixed decomposition). This certainly does not reflect the physical reality where the characteristic time scales of all variables change with time.

Some of these restrictions of the original Method of Integral Manifolds were overcome in the method described in our earlier presentation [20, 21]. In this method the characteristic time scales of variables in the system of ODEs were organised in ascending order $(\tau_0 \leq \tau_1 \leq ..... \leq \tau_i \leq .... \leq \tau_N)$. Then the authors looked for a possible gap in these timescales such that $\tau_i/\tau_{i+1} < \epsilon$ where $i = 1, 2, ..., N-1$, $\epsilon$ is an *a priori* chosen small parameter. If this gap was found then the first $i$ equations formed the fast system and the remaining $N - i$ equations formed the slow system. In this case the system of ODEs was rewritten in the form (1)-(2). In contrast to the original Method of Integral Manifolds, however, this decomposition of variables was not fixed, but was allowed to change at each time step. Hence, it was suggested to call this decomposition dynamic. When analysing this decomposed system, the system of equations for fast variables was solved numerically using an ODE system solver, while the variations of the slow variables were assumed to be linear in time. This approach to System (1)-(2) is more realistic that the original MIM, and opens the way to implementing this approach into CFD codes. Some preliminary results of this implementation were reported in [20, 21, 25].

Although this approach was shown to be effective for implementation into CFD codes in some cases, it still had a number of limitations. The required threshold $\tau_i/\tau_{i+1} < \epsilon$ to enable the application of this method could be found in exceptional rather than typical cases. Also, this method was most effective when $i$ was close to 1. When $i$ was close to $N$, then any advantages of this method became questionable.

In contrast to the method described in [20, 21], the new method suggested in this paper does not rely on the existence of the gap in the characteristic time scales of the change of variables. That means that potentially it can be applied to both stiff and non-stiff system of equations, although in the latter case it is expected to be less efficient than in the first case. This new method is focused on finding a "global" (possibly "nonlinear") transformation of the original coordinate axes such that the original system of ODEs becomes the singularly perturbed system (SPS) with the required gap between fast and slow variables. In other words, our task is to find the direction of "fast" motion described by System (1)-(2) for a fixed point (**X,Y**) of the phase space. This direction may or may not exist. If this direction exists for the original system of ODEs then the method described in [20, 21] can be applied. If it does not exist then it can be potentially found using the new method. Hence, the new method can be considered as a straightforward generalisation of the method described in [20, 21].

The procedure of finding fast and slow variables can be iterative and result in a hierarchical division of the original system. For example the 'slow' subsystem can, in its turn, be subdivided into 'slow' and 'very slow' subsystems.

## 3 Decomposition of the System of Equations

Let us consider the system, the state of which is characterized by $n$ dimensionless variables $Z_n$ ($n = 1,\ 2,\ .....n$). The value of each of these variables for a given place in space depends on time $t$, i.e. $Z_n = Z_n(t)$. This system can be described by $n$ equations, which can be presented in the vector form:

$$\frac{\mathrm{d}\mathbf{Z}}{\mathrm{d}t} = \mathbf{\Phi}(\mathbf{Z}), \tag{3}$$

where:

$$\mathbf{Z} = (Z_1, Z_2, ....., Z_n), \qquad \mathbf{\Phi} = (\Phi_1, \Phi_2, ....., \Phi_n).$$

Although a rigorous coupled numerical solution of this system could be found, using one of ODEs solvers, this may not always be practical, when too many equations are involved. An alternative approach to this solution could be based on the decomposition of this system, as discussed in Sections 1 and 2.

The solution of System (3) depends on whether it is not stiff (not multi-scale) or stiff (multi-scale). In the first case, we cannot offer any alternative to the conventional numerical solution of this system. In the second case, the system can be potentially decomposed. This decomposition can take place either without changing the scales of variables (for example, equations can merely be reordered according to scales of variables as in [20, 21]) or with the change of these scales. The second approach can be focused on the systems which are multi-scale, but the hierarchy of these scales is 'hidden' inside the

system. In both cases, the analysis of the systems starts with finding local transformation matrices $\mathbf{Q}$ such that the original system (3) is transformed to the system which can be presented as System (1)-(2). The latter is known as a Singularly Perturbed System or SPS.

There are a number of possible ways to specify matrix $\mathbf{Q}$. Some of these will be discussed in the next section. If this matrix is found, it can be split into two distinct parts $\mathbf{Q_f}$ and $\mathbf{Q_s}$ responsible for the transformation of the initial vector variable $\mathbf{Z}$ into the ordered combination of the new fast and slow variables, respectively [26]:

$$\mathbf{Q} = (\mathbf{Q_f Q_s})\,; \mathbf{Q^{-1}} = \begin{pmatrix} \mathbf{Q_f^{-1}} \\ \mathbf{Q_s^{-1}} \end{pmatrix}. \tag{4}$$

Hence, we can write:

$$\mathbf{Z} = \mathbf{Q} \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}, \tag{5}$$

where $\mathbf{U}$ and $\mathbf{V}$ are new vector variables: $\mathbf{U}$ contains all fast scalar variables, $\mathbf{V}$ contains all slow scalar variables. The splitting of matrix $\mathbf{Q}$ implies that a gap in scales exists between these variables. This splitting is fixed over a specified period (time step), but can change beyond this period (dynamic decomposition).

The equations for $\mathbf{U}$ and $\mathbf{V}$ can be presented in vector forms as:

$$\frac{\mathrm{d}\mathbf{U}}{\mathrm{d}t} = \mathbf{Q_f^{-1}F}\left(\mathbf{Q}\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}\right) = \mathbf{\Phi}_f \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}, \tag{6}$$

$$\frac{\mathrm{d}\mathbf{V}}{\mathrm{d}t} = \mathbf{Q_s^{-1}F}\left(\mathbf{Q}\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}\right) = \mathbf{\Phi}_s \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}. \tag{7}$$

Having introduced a new small positive parameter $\varepsilon \ll 1$ and remembering that:

$$||\mathbf{\Phi}_s|| \ll ||\mathbf{\Phi}_f||, \tag{8}$$

we can rewrite Equations (6) and (7) in the form similar to the one used in the Method of Integral Manifolds:

$$\varepsilon\frac{\mathrm{d}\mathbf{U}}{\mathrm{d}t} = \varepsilon\mathbf{Q}_f^{-1}(\mathbf{Z}_0)\mathbf{\Phi}\left(\mathbf{Q}(\mathbf{Z}_0)\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}\right) \equiv \mathbf{\Phi}_{f\varepsilon}\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}, \tag{9}$$

$$\frac{\mathrm{d}\mathbf{V}}{\mathrm{d}t} = \mathbf{Q}_s^{-1}(\mathbf{Z}_0)\mathbf{\Phi}\left(\mathbf{Q}(\mathbf{Z}_0)\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}\right) \equiv \mathbf{\Phi}_s\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}, \tag{10}$$

where $\mathbf{\Phi}_{f\varepsilon} = \varepsilon\mathbf{\Phi}_f$. In this presentation the right hand sides of Equations (9) and (10) are expected to be of the same order of magnitude over the same period during which the original decomposition of matrix $\mathbf{Q}$ is valid.

Equations (9) and (10) need to be integrated over the time period $\Delta t$ : $t_k \rightarrow t_{k+1}$. The zeroth order solution of Equation (10) (slow subsystem) is just a constant value of the slow variable: $\mathbf{V}_{k+1}^0 = \mathbf{V}_k = \mathbf{V}(t_k)$, where the

superscript $^0$ indicates the zeroth approximation, while the subscripts $_k$ and $_{k+1}$ indicate the points in time. The zeroth order for the fast variable is found from Equation (9) with $\mathbf{V} = \mathbf{V}_k$. The latter condition can be interpreted as the equation for the slow variable on the fast manifold. Hence, Equation (9) can be approximated by the following system:

$$\frac{\mathrm{d}\mathbf{U}}{\mathrm{d}t} = \mathbf{\Phi}_f \begin{pmatrix} \mathbf{U} \\ \mathbf{V_k} \end{pmatrix}. \tag{11}$$

The solution of Equation (11) at $t = t_{k+1}$ ($\mathbf{U}^0_{k+1}$) is the zeroth order approximation of the fast motion on the fast manifold determined at $t = t_k$. Note that the system described by Equation (11) can be stiff in the general case, but with a reduced level of stiffness, compared with the original system (3). Hence, the suggested method is expected to reduce the level of stiffness of the system and not eliminate the stiffness altogether.

Under the same zeroth order approximation the slow variables would remain constant over the same time period. This assumption was used in the original formulation of the Method of Integral Manifolds [19, 18, 17]. This, however, leads to an unphysical result in many applications when slow variables would remain constant for any time $t > t_0$. Hence, the need to calculate slow variables using at least the first order approximation. In the case when $\varepsilon$ is not asymptotically small, higher order approximations need to be considered. In this case we introduce the new time scale $\tau = t/\varepsilon$, and formally present the slow and fast variables as:

$$\left.\begin{array}{l} V(\tau) = V^{(0)} + \varepsilon V^{(1)}(\tau) + \varepsilon^2 V^{(2)}(\tau) + \ \ldots\ldots\ldots\ldots \\ U(\tau) = U^{(0)}(\tau) + \varepsilon U^{(1)}(\tau) + \varepsilon^2 U^{(2)}(\tau) + \ \ldots\ldots\ldots \end{array}\right\}. \tag{12}$$

Having substituted Expressions (12) into Equation (10) we obtain:

$\frac{\mathrm{d}(V^{(0)} + \varepsilon V^{(1)}(\tau) + \varepsilon^2 V^{(2)}(\tau) + \ldots\ldots)}{\mathrm{d}\tau} =$

$$\varepsilon \mathbf{\Phi}_s \begin{pmatrix} U^{(0)}(\tau) + \varepsilon U^{(1)}(\tau) + \varepsilon^2 U^{(2)}(\tau) + \ \ldots\ldots \\ V^{(0)} + \varepsilon V^{(1)}(\tau) + \varepsilon^2 V^{(2)}(\tau) + \ \ldots\ldots \end{pmatrix}. \tag{13}$$

Equation (13) allows us to obtain the first order solution for the slow variable in the form:

$$V_{k+1} = V^{(0)}_{k+1} + \varepsilon V^{(1)}_{k+1} = V^{(0)}_k + \varepsilon \mathbf{\Phi}_s \begin{pmatrix} U^{(0)}_{k+1} \\ V^{(0)}_k \end{pmatrix} \Delta\tau. \tag{14}$$

Returning to the original variables we can write the expression for $\mathbf{V}(t_{k+1}) \equiv \mathbf{V}_{k+1}$ in the form:

$$\mathbf{V}_{k+1} = \mathbf{V}^{(0)}_k + \mathbf{\Phi}_s \begin{pmatrix} \mathbf{U}^{(0)}_{k+1} \\ \mathbf{V}^{(0)}_k \end{pmatrix} \Delta t. \tag{15}$$

To increase accuracy of calculations one could continue the process to take into account the first order solution for the fast motion (similar to (11)).

Then the second order solution for the slow motion (similar to (15)) could be obtained etc.

## 4 Choice of Decomposition

The focus of this section is on the determination of the transformation matrix **Q**. This matrix is expected to provide us with the required subdivision of the original system of ODEs into two smaller subsystems - fast and slow. There are a number of ways to find **Q**. These can be based on (a) invariant subspaces (eigenspaces) of Jacobi matrix of the original system **J** (ILDM, see [5]), (b) so called principal subspaces (eigenspaces) of a matrix $\mathbf{JJ}^*$ (TILDM - further development of ILDM - see [8]), (c) invariant subspaces of the symmetrised Jacobi matrix of the original system (this was based on the system entropy) ([9] - [10]), (d) reordering of the original system of equations according to the values of RHSs of the system suggested in [20, 21].

ILDM was originally suggested for the numerical solution of the systems containing a large number of ODEs. The method has proven to be an efficient tool for the simplification of equations describing detailed chemical kinetics [5] - [7], [27]. This approach uses the local Jacobian eigenspaces (more precisely Shur basis for the local Jacobian matrix) for the transformation of coordinates. Nevertheless, even for singularly perturbed systems, identification of fast and slow eigenvectors is not always possible and a more delicate analysis is required. It can be shown that the Jacobian matrix, does not always show the hierarchy of the scales even if this hierarchy exists [27, 28]. Hence, the information about eigenspaces may be insufficient, even not relevant for establishing the hierarchical structure of the scales of the system. This can be illustrated for the following system of ODEs:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -x - \frac{1}{\varepsilon_1}y, \tag{16}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -y, \tag{17}$$

where $\varepsilon_1$ is a small positive parameter.

The Jacobian of this system is matrix **A** of the RHS of the system (16) - (17). Both eigenvalues of this matrix are equal to $-1$. According to the traditional interpretation of the ILDM approach, there is no internal hierarchy in System (16) - (17). Nevertheless, as one can readily see, this system has essentially different rates of change of variables (at least in the region, where the functions $x(t)$ and $y(t)$ are of the same order: $x(t)$ is the fast variable, while $y(t)$ is the slow one. Multiplying both parts of (16) - (17) by $\varepsilon_1$, we can re-write this system in the conventional SPS form:

$$\varepsilon_1 \frac{\mathrm{d}x}{\mathrm{d}t} = \Theta_f(x, y), \quad \Theta_f(x, y) = -\varepsilon_1 x - y, \tag{18}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \Theta_s(x,y), \quad \Theta_s(x,y) = -y, \tag{19}$$

Thus even in the simplest linear case with a strongly determined hierarchy, the eigenvalues of the Jacobian do not always provide us with correct information regarding possible reduction of the system to a singularly perturbed one. Other examples illustrating this idea can be presented. Using geometrical language, we can conclude that the traditional ILDM procedure fails to decompose the vector field ($\mathbf{Z}$) into fast and slow components. In other words, the information about its eigenspaces may be insufficient to evaluate the 'hidden' hierarchical structure of the system. Essentially the same limitation can be attributed to TILDM (see [27, 28] for the details).

In contrast to ILDM, we suggest not choosing Jacobian but $\mathbf{J}\mathbf{J}^*$ (following TILDM) for construction of the transformation matrix. The latter matrix provides us with information regarding decomposition in the wider domain than the Jacobian $\mathbf{J}$ does ([8]). Presenting the image of the unit sphere (on which the transformation is performed) as a hyper ellipsoid, the eigenvalues of the matrix $\mathbf{J}\mathbf{J}^*$ represent the lengths of its semi axes. The corresponding eigenvectors coincide with the directions of the semi axes. Note that all eigenvalues of matrix $\mathbf{J}\mathbf{J}^*$ are real and positive and corresponding eigenvectors are orthogonal.

The proposed algorithm for finding 'hidden' fast and slow variables of the original system 3 (construction of matrix $\mathbf{Q}$ at an arbitrary point $\mathbf{Z}$) contains the following steps:

1. Build matrix $\mathbf{T}(\mathbf{Z}) \equiv \mathbf{J}(\mathbf{Z})\mathbf{J}^*(\mathbf{Z})$ and determine its eigenvalues $\lambda_i$ $(i = 1, 2, , n)$.

2. Check the scales of $\lambda_i$ and establish whether it is possible to find $\tau > 0$ such that $\lambda_i >> \tau > \lambda_j,\ i = 1, ..., n_f,\ j = n_f + 1, n_f + 2, .., n,\ f + s = n$, where the eigenvalues have been reordered in descending order. Large and small eigenvalues determine 'principal' eigenspaces (invariant subspaces) of $\mathbf{T}$.

3. Build the transformation matrix $\mathbf{Q}$ in such a way that the eigenvectors $\mathbf{v}_i, i = 1, ..., f, f + 1, ..., f + s$ of the matrix $\mathbf{T}$ are presented in the same order as the eigenvalues determined at Step 2:

$$\mathbf{Q} = \left( \mathbf{v}_1 \vdots \mathbf{v}_{n_f}\ \mathbf{v}_{n_f+1} \vdots \mathbf{v}_{n_f+n_s} \right),\ \mathbf{T} = \mathbf{Q} \cdot \begin{pmatrix} \mathbf{\Lambda}_f & 0 \\ 0 & \mathbf{\Lambda}_s \end{pmatrix} \cdot \mathbf{Q}^{-1}, \tag{20}$$

$$\mathbf{\Lambda}_f = \begin{pmatrix} \boldsymbol{\lambda}_1 \ldots & 0 \\ \vdots \ \ddots & \vdots \\ 0 \ \ldots & \boldsymbol{\lambda}_{n_f} \end{pmatrix},\ \mathbf{\Lambda}_s = \begin{pmatrix} \boldsymbol{\lambda}_{n_f+1} \ldots & 0 \\ \vdots \ \ddots & \vdots \\ 0 \ \ldots & \boldsymbol{\lambda}_{n_f+n_s} \end{pmatrix}. \tag{21}$$

Once matrix $\mathbf{Q}$ has been obtained, further analysis of the system is performed following the guidelines discussed in Section 3.

# 5 Application

## 5.1 Equations for Spray Combustion

In this subsection basic equations used for modelling droplets heating, evaporation and combustion are summarised. A number of processes, including droplet dynamics, break-up and coalescence, and the effects of the temperature gradient inside droplets [29]-[32] are ignored at this stage. This can be justified by the fact that the main emphasis of this paper is on the investigation of the new method of the solution of the systems of ODEs relevant to spray combustion modelling rather than providing a detailed analysis of the processes involved. The system under consideration contains equation for the droplet mass in the $i$th parcel $m_{di}$, droplet temperature in the $i$th parcel $T_{di}$, fuel vapour density $\rho_{fv}$, density of oxygen $\rho_{O_2}$ and temperature of the gas $T_g$ [33]:

$$\dot{m}_{di} = 4\pi \frac{\overline{k}_g R_{di}}{c_{pF}} \ln(1 + B_M), \tag{22}$$

$$m_{di} c_l \frac{\mathrm{d}T_{di}}{\mathrm{d}t} = 4\pi R_{di}^2 h (T_g - T_{di}) - \dot{m}_{di} L + 4\pi R_{di}^2 \sigma \overline{Q}_a \theta_R^4, \tag{23}$$

$$\alpha_g \frac{\mathrm{d}\rho_{fv}}{\mathrm{d}t} = -\alpha_g \mathrm{CT} + \left[ \sum_i \dot{m}_{di} \Big/ \mathrm{V} \right], \tag{24}$$

$$\frac{\mathrm{d}\rho_{O_2}}{\mathrm{d}t} = -18.5 \frac{M_{O_2}}{M_f} \mathrm{CT} = -3.48235 \ \mathrm{CT}, \tag{25}$$

$$c_{\mathrm{mix}} \rho_{\mathrm{mix}} \frac{\mathrm{d}T_g}{\mathrm{d}t} = \alpha_g Q_f \mathrm{CT} - \left[ \sum_i m_{di} c_l \frac{\mathrm{d}T_{di}}{\mathrm{d}t} + \sum_i \dot{m}_{di} L + \sum_i \dot{m}_{di} c_{pF} (T_g - T_{di}) \right] \Big/ \mathrm{V}, \tag{26}$$

where $B_M$ is the Spalding mass number, $L$ is the specific latent heat of vaporization, $Q_L$ is the heat spent on droplet heating, $R_d$ is droplet radius, $c_{pF}$ is specific heat capacity of fuel vapour, $c_{\mathrm{mix}}$ is specific heat capacity of the mixture of fuel vapour and air, $\theta_R$ is the radiative temperature, $\overline{Q}_a$ is the average absorption efficiency factor, $\rho_{fv}$ is the fuel vapour density, $\rho_{O_2}$ is the density of oxygen, $\rho_{\mathrm{mix}}$ is the density of the mixture of fuel vapour and air. $\alpha_g$ is the volume fraction of gas assumed equal to 1 in our calculations, the summation is assumed over all droplets in volume V, $Q_f$ is the heat released per unit mass of burnt fuel vapour (in J/kg). CT is the chemical term (in kg/(m$^3$s)) presented as [20, 21]:

$$\mathrm{CT} = A \, M_{O_2}^{-1.5} \, M_f^{0.75} \rho_{fv}^{0.25} \, \rho_{O_2}^{1.5} \exp\left[-E/(BT)\right], \tag{27}$$

where $M_{O_2} = 32$ kg/kmol, and $M_f = 170$ kg/kmol are molar masses of oxygen and fuel respectively in kg/kmol,

$$A = 3.8 \times 10^{11} \ \frac{1}{\mathrm{s}} \left(\frac{\mathrm{mole}}{\mathrm{cm}^3}\right)^{-0.75} = 2.137 \times 10^9 \ \frac{1}{\mathrm{s}} \left(\frac{\mathrm{kmole}}{\mathrm{m}^3}\right)^{-0.75};$$

$$E = 30 \ \frac{\mathrm{kcal}}{\mathrm{mole}} = 1.255 \times 10^8 \ \frac{\mathrm{J}}{\mathrm{kmole}}.$$

This model is similar to the one used in [20, 21].

### 5.2 Values of Parameters and Solution Procedure

The method described in Sections 3 and 4 is applied to simulate polydisperse spray heating, evaporation and ignition, based on equations given in the previous subsection. The model on which the analysis is based is chosen to be rather simple, but capable nevertheless of capturing the essential features of the process. We consider spray consisting of 3 groups of droplets with initial radii 5 $\mu$m, 9 $\mu$m and 13 $\mu$m respectively. The initial temperatures of all droplets is taken to be equal to 400 K. The gas temperature is taken to be equal to 880 K [23]. The gas volume is chosen such that if the droplets are fully evaporated without combustion, then the equivalence ratio of fuel vapour/air mixture is equal to 4. This is the situation typical for diesel engines in the vicinity of the nozzle. The initial density of oxygen is taken equal to 2.73 kg/m$^3$ (this corresponds to air pressure equal to 3 MPa). The initial mass fraction of fuel is taken equal to zero. These values of the parameters can be considered as an approximation of the actual conditions in diesel engines [23].

Since Equations (22) and (23) are solved separately for each of 3 groups of droplets, the maximal number of equations to be solved is equal to 10. Note that the density of the fuel vapour/air mixture could be derived algebraically from mass conservation. It was preferred, however, to solve the ODE for it to enable us to monitor the mass conservation in the system as a validity check.

Firstly, these equations were solved simultaneously in a coupled way using DLSODAR stiff solver from the ODEPACK developed in LLNL laboratory (software is available from [35]). The second approach is based on the decomposing of the original system following the procedure described in [20, 21]. Finally, the third approach is based on the algorithm, presented in the Section 4.

### 5.3 Results and Discussion

The total number of equations solved, and the number of equations for fast variables can change with time as expected. The corresponding plots of the numbers of these equations are shown in Fig.1. The left graph is based on the approach suggested in [20, 21], the right graph is based on the methods described in Sections 3 and 4. Initially all 10 equations were solved, when the

approach based on the coupled solution of the full system of ODEs was used. About 0.6 ms after the beginning of the process this number was reduced to 8 when the smallest droplet evaporated. Then approximately about 1.8 ms after the beginning of the process it further reduced 6 when the smallest droplets evaporated.

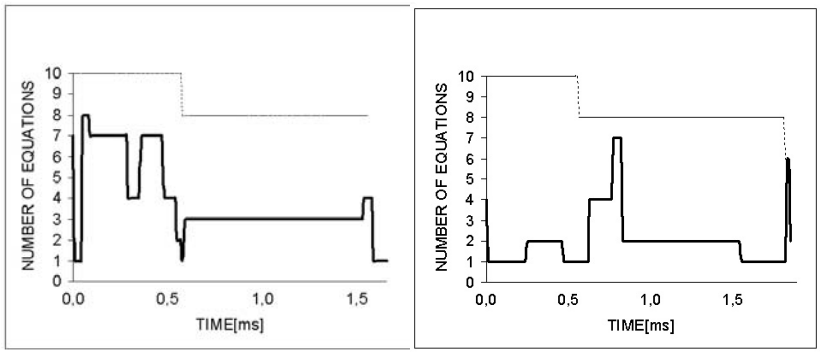



Fig. 1: Plots of the total number of equations solved (dashed) and the number of equations for fast variables (solid) for the values of parameters as specified in the text. Left - dynamic decomposition performed based on the comparison of the RHSs of the system (22) - (26) [21]; right - dynamic decomposition performed based on the algorithm suggested in the present paper.

In the case when the decomposition based on the approach suggested in [20, 21] was applied, the number of equations for fast variables to be solved was always less than the total number of equations. Initially the number of equations for fast variables was equal to seven, then it dropped to just one equation describing fuel vapour density. After that, the number of equations jumped to eight, then most part of the time between 0.1 ms and 0.5 ms the number of equations for fast variables was equal to seven. Between 0.6 and 1.5 ms the number of equations for fast variables was equal to three. After a short jump to four equations, again only the equation for fuel density was solved. To summarise these results, up to about 0.6 ms the difference between these numbers was relatively small to justify the application of decomposition techniques. This could be done between approximately 0.5 ms and 1.5 ms when the number of fast equations (three) was noticeably less than the total number of equations (eight).

The situation became rather different when the transformation matrix described in Section 4 was used. In this case the number of fast equation was typically much less than the total number of equations. Initially the number of equations for fast variables was equal to four, then it dropped to just one equation describing fuel vapour density. Between about 0.25 ms and 0.5 ms the number of equations for fast variables was equal to two (equations for fuel

vapour density and the radius of the smallest droplet). Then again just the equation for fuel density was solved. Between about 0.6 ms and 0.8 ms the number of fast equations was comparable with the total number of equations. During this period the decomposition of the system is not expected to be useful. After about 0.8 ms and until about 1.8 ms, only one equation (fuel density) or two equations (fuel density and the radius of the second droplet) were solved. In this case the decomposition technique described in Section 4 is expected to be particularly important.

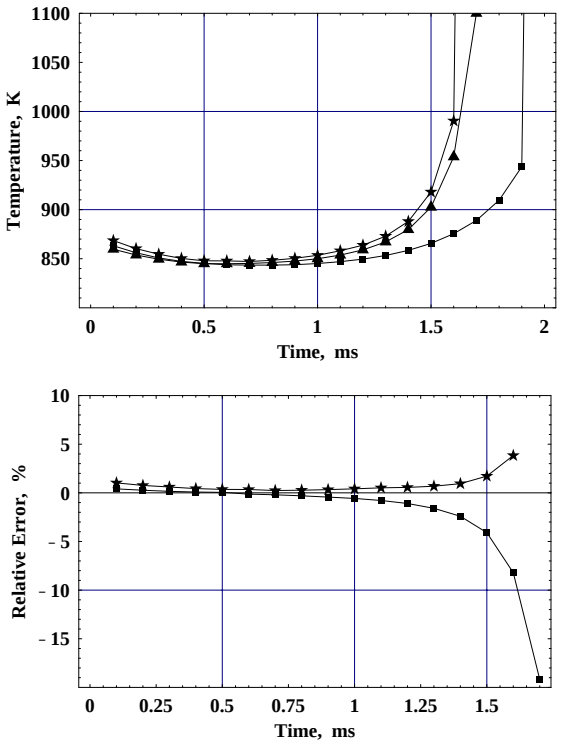



Fig. 2: Plots of gas temperature versus time (upper) and relative errors of calculation (lower). Gas temperature was calculated using dynamic decomposition based on the comparison of the RHSs of the system (22) - (26) [20, 21] (cubes), dynamic decomposition based on the algorithm suggested in the present paper (stars), coupled solution of the full system of equations (triangles). The errors of the dynamic decomposition approaches were calculated relative to the results of the coupled solution of full system of equations. Time step was taken equal $10^{-4}$ s.

The time evolution of gas temperature, calculated using the aforementioned approaches, and the relative errors of calculations based on the dynamic decomposition for two time steps are shown in Figs 2, 3. As follows from these

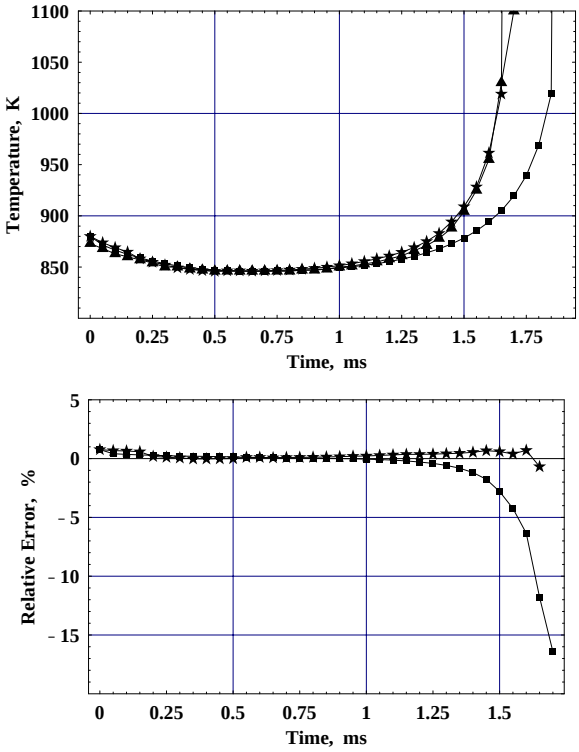



Fig. 3: The same as Fig. 2 but for time step $5 \cdot 10^{-5}$ s.

figures for both time steps, the new approach to dynamic decomposition leads to significantly smaller relative errors that the approach suggested in [20, 21].

To compare the CPU efficiency of both approaches to dynamic decomposition (the one suggested in [20, 21] and the new one) and the conventional approach used in computational fluid dynamics CFD codes (fixed decomposition), a series of runs for various time steps were performed for the problem solved above. For fixed time steps the CPU requirements of all three approaches were about the same. As shown above, however, the accuracy of the new approach is almost always higher than that of the one used in [20, 21]. The latter approach, in its turn, is more accurate than the one based on the fixed decomposition approach used in CFD codes [20, 21]. For example, relative errors less than 1.5% were achieved for the time step of $1.3 \times 10^{-5}$ s for the conventional CFD approach, and for the time step of $2.4 \times 10^{-5}$ s for the new approach described in Sections 2 and 3. When this effect was taken into account then in all cases under consideration, the CPU time for the new method was always smaller than that for the fixed decomposition approach. In some cases, the CPU reduction for the new approach was as high as a factor of

3. The CPU requirements for the approach described in [20, 21] were between those for the new and the fixed decomposition approach. The CPU time was estimated based on the customised function DATE_AND_TIME [34].

## 6 Conclusions

A new approach to dynamic decomposition of the systems of ODEs to fast and slow subsystems is suggested. The suggested approach is based on the comparative analysis of the magnitudes of the eigenvalues of the matrix $\mathbf{JJ}^*$, where $\mathbf{J}$ is the local Jacobi matrix of the system under consideration. The eigenvectors of this matrix are used for construction of the transformation matrix $\mathbf{Q}$. The hierarchy of the decomposition is allowed to vary with time. Hence, this decomposition is called dynamic (it depends on the specific computation cell and on the time layer). As in our earlier approach [20, 21], equations for fast variables are solved by a stiff ODE system solver with the slow variables taken at the beginning of the time step. This is considered as a zeroth order solution for these variables. The solution of equations for slow variables is presented in a simplified form, assuming linear variations of these variables during the time evolution of the fast variables. This is the first order approximation for the solution for these variables or the first approximation for the fast manifold.

The new approach is applied to numerical simulation of diesel fuel spray heating, evaporation and the ignition of fuel vapour/ air mixture. The results show advantages of the new approach when compared with the one discussed in our previous paper [20, 21] and the conventional CFD approach used in computational fluid dynamics codes, both from the point of view of accuracy and CPU efficiency.

*Acknowledgement.* The authors are grateful to the European Regional Development Fund Franco-British INTERREG IIIa (Project Ref 162/025/247), to the German-Israeli Foundation (Grant G-695-15.10/2001) and to the Minerva Fellowship Program of the Max Planck Society for the partial financial support of this project.

## References

1. E.M. Sazhina, S.S. Sazhin, M.R. Heikal, V.I. Babushok, R.A. Johns: A detailed modelling of the spray ignition process in Diesel engines. *Combustion Science and Technology* **160**, 317-344 (2000)
2. S.V. Utyuzhnikov: Numerical modeling of combustion of fuel-droplet-vapour releases in the atmosphere. *Flow, Turbulence and Combustion* **68**, 137-152 (2002)
3. S.S. Sazhin, P. Wild, C. Leys, D. Toebaert,E.M. Sazhina: The three temperature model for the fast-axial-flow $CO_2$ laser. *J Physics D: Applied Physics* **26**, 1872-1883 (1993)

4. S.S. Sazhin, E.M. Sazhina, M.R. Heikal, C. Marooney, S.V. Mikhalovsky: The Shell autoignition model: a new mathematical formulation. *Combustion and Flame* **117**, 529-540 (1999)
5. U. Maas, S.B. Pope: Simplifying Chemical Kinetics: Intrinsic Low-Dimensional Manifolds in Composition Space. *Combustion and Flame* **117**, 99-116 (1992)
6. C. Rhodes, M. Morari, S. Wiggins: Identification of the Low Order Manifolds: Validating the Algorithm of Maas and Pope. *Chaos* **9**, 108-123 (1999)
7. H.G. Kaper, T.J. Kaper: *Asymptotic Analysis of Two Reduction Methods for Systems of Chemical Reactions,* Argonne National Lab, preprint ANL/MCS-P912-1001 (2001)
8. V. Bykov, I. Goldfarb, V. Gol'dshtein, U. Maas: *IMA J. of Applied Mathematics.* **69**, 353-374 (2004)
9. A. Gorban, I. Karlin: Methods of invarinat manifolds for kinetic problems. *Chem. Eng. Sci.* **396**, 197-403 (2002)
10. A. Gorban, I. Karlin, A. Zinovyev: Constructive methods of invarinat manifolds for kinetic problems. *Physics Reports* **396**, 197-403 (2004)
11. S.H. Lam, D.M. Goussis: The GSP method for simplifying kinetics. *International Journal of Chemical Kinetics* **26**, 461-486 (1994)
12. M. Hadjinicolaou, D.M. Goussis: Asymptotic Solutions of Stiff PDEs with the CSP Method: the Reaction Diffusion Equation, *SIAM journal of Scientific Computing.* **20**, 781-910 (1999)
13. A. Masias, D. Diamantis, E. Mastorakos, D.A. Goussis: An algorithm for the construction of global reduced mechanisms with CSP data. *Combustion and Flame* **117**, 685-708 (1999)
14. M. Valorani, D.M. Goussis: Explicit Time-Scale Splitting Algorithm for Stiff Problems: Auto-Ignition of Gaseous Mixtures behind a Steady Shock. *Journal of Computational Physics* **169**, 44-79 (2001)
15. M.K. Neophytou, D.A. Goussis, M. van Loon, E. Mastorakos: Reduced chemical mechanism for atmospheric pollution using Computational Singular Perturbation analysis. *Atmospheric Environment* **38**, 3661-3673 (2004)
16. V. Gol'dshtein, V. Sobolev: *Qualitative Analysis of Singularly Perturbed Systems* (in Russian) (Institute of Mathematics, Siberian Branch of USSR Academy of Science, Novosibirsk 1988)
17. V. Gol'dshtein, V. Sobolev: Integral manifolds in chemical kinetics and combustion. In: *Singularity theory and some problems of functional analysis*, 73–92 (American Mathematical Society 1992)
18. N. Fenichel: Geometric singular perturbation theory for ordinary differential equations. *J Differential Equations* **31**, 53-98 (1979)
19. Yu.A. Mitropolskiy, O.B. Lykova, *Lectures on the methods of integral manifolds* (in Russian) (Institute of Mathematics Ukrainen Akademy of Science, Kiev 1968)
20. V. Bykov, I. Goldfarb, V. Goldshtein, S.S. Sazhin, E.M. Sazhina: System decomposition technique: application to spray modelling in CFD codes. In: *20th Annual Symposium of the Israeli Section of the Combustion Institute. Book of Abstracts*, p. 16 (Ben-Gurion University, Beer-Sheva, Israel 2004)
21. V.Bykov, I.Goldfarb, V.Goldshtein, S.S.Sazhin, E.M.Sazhina: An Asymptotic Approach to Numerical Modelling of Spray Autoignition in Hot Gas Proceedings of the Fourth Mediterranean Combustion Symposium, MCS4, Paper vol. 6, Lisbon (Portugal), October 6-10, (2005)

22. I. Goldfarb, V. Gol'dshtein, G. Kuzmenko, S.S. Sazhin: Thermal radiation effect on thermal explosion in gas containing fuel droplets, *Combustion Theory and Modelling* **3**, 769-787 (1999).
23. S.S. Sazhin, G. Feng, M.R. Heikal, I. Goldfarb, V. Goldshtein, G. Kuzmenko: Thermal ignition analysis of a monodisperse spray with radiation, *Combustion and Flame* **124** 684-701, (2001).
24. V. Bykov, I. Goldfarb, V. Gol'dshtein, J.B. Greenberg: Thermal explosion in a hot gas mixture with fuel droplets: a two reactants model, *Combustion Theory and Modelling* **6**, 1-21 (2002).
25. E.M. Sazhina, V. Bykov, I. Goldfarb, V. Goldshtein, S.S. Sazhin, M.R. Heikal: Modelling of spray autoignition by the ODE system decomposition technique. In: *Proceedings of HEFAT2005 (4th International Conference on Heat Transfer, Fluid Mechanics and Thermodynamics)*, Paper number: SE2 (Cairo, Egypt 2005)
26. V. Bykov, I. Goldfarb, V. Gol'dshtein: Novel numerical decomposition approaches for multiscale combustion and kinetic models. *Journal of Physics: Conference Series*, **22**, 1-29 (2005)
27. I. Goldfarb, V. Gol'dshtein, U. Maas: Comparative analysis of two asymptotic approaches based on integral manifolds. *IMA journal of Applied Mathematics*, **69**, 353-374 (2004)
28. V. Bykov, I. Goldfarb, V. Gol'dshtein, U. Maas: On a Modified Version of ILDM Approach: Asymptotic Analysis Based on Integral Manifolds. *IMA Journal of Applied Mathematics*, to appear (2005)
29. B. Abramzon, S.S. Sazhin: Droplet vaporization model in the presence of thermal radiation, *Int J Heat Mass Transfer* **48**, 1868-1873 (2005)
30. S.S. Sazhin, W.A. Abdelghaffar, E.M. Sazhina, M.R. Heikal: Models for droplet transient heating: effects on droplet evaporation, ignition, and break-up. *Int J Thermal Science* **44**, 610–622 (2005)
31. S.S. Sazhin, W.A. Abdelghaffar, P.A. Krutitskii, E.M. Sazhina, M.R. Heikal: New approaches to numerical modelling of droplet transient heating and evaporation. *Int J Heat Mass Transfer* **48** (19-20), 4215-4228 (2005)
32. B. Abramzon, S. Sazhin: Convective vaporization of fuel droplets with thermal radiation absorption: *Fuel* **85**, 32-46 (2006)
33. S.S. Sazhin: Advanced models of fuel droplet heating and evaporation. *Progress in Energy and Combustion Science* (in press) (2006)
34. Yu. Shramkova: Implementation of simple singularly perturbed model in numerical simulations of Diesel engines, M.Sc. Thesis (2006)
35. LLNL public domain webpage: `www.llnl.gov/CASC/odepack/`

# Model Reduction of Multiple Time Scale Processes in Non-standard Singularly Perturbed Form

N. P. Vora[1], M.-N. Contou-Carrere[2], and P. Daoutidis[3]

[1] GE Water & Process Technologies, nishith.vora@ge.com
[2] Innovene USA LLC, marie.contou-carrere@innovene.com
[3] Department of Chemical Engineering and Materials Science, University of Minnesota, currently at Aristotle University of Thessaloniki, Greece, Corresponding author: daoutidi@cems.umn.edu

**Summary.** This work considers multiple time scale models in non standard singularly perturbed form. These systems naturally arise as descriptions of detailed rate-based process models of fast-rate chemical processes with several large parameters of different orders of magnitude. We propose a systematic framework to derive representations of the dynamics in individual time scales. A nonlinear coordinate transformation is presented to yield a standard singularly perturbed form of the original system. This approach is applied to a representative multiple time scale chemical process with reactions whose reaction rates span different orders of magnitude.

## 1 Introduction

Chemical processes typically exhibit multiple time scale dynamics owing to the presence of fast heat/mass transfer [10], multiple fast and slow reactions [10, 15], fast flow of gases and liquids [11, 14], etc. This feature is observed in a broad range of processes such as catalytic reactors [2], fluidized catalytic crackers [3], multi-phase reactors [10], chemical reaction networks [15], biochemical processes [1], distillation and reactive distillation processes [14]. Dynamic models capturing both the fast and slow phenomena of these multiple time scale processes are *stiff*, as they contain parameters of different orders of magnitude; as a result they are difficult and costly to simulate. This motivates the need to obtain reduced order models capturing the dynamics only in the time scale of interest, while approximating the dynamics in other (faster) time scales in a systematic fashion.

Singular perturbation theory has proven to be a natural framework for the systematic decomposition of the system dynamics in different time scales. There exists an extensive literature on the application of singular perturbation

theory for the model reduction, analysis and control of systems with two time scales (see e.g. [8, 9, 6]).

The vast majority of the existing research has focused on two time scale systems modeled in the so-called "standard" singularly perturbed form, where the fast and slow variables are explicitly separated due to the presence of a small parameter $\epsilon$ (the singular perturbation parameter) that multiplies the time derivative of the vector of "fast" state variables (see e.g. [8, 9]). However, modeling a two time scale process in the standard singularly perturbed form is in itself a nontrivial task. In some processes, e.g. catalytic reactors [2] and fluidized catalytic cracking [3], there is an *a priori* knowledge of the variables with slow and fast dynamics. This allows modeling of such processes directly in the standard singularly perturbed form, through an appropriate definition of $\epsilon$. However, for most chemical processes with fast heat/mass transfer, fast reactions, fast flow of gases and liquids, etc. the fast and slow dynamics can not be associated with distinct state variables, and the corresponding dynamic models are not in the standard singularly perturbed form. Only recently, a class of nonlinear systems in the nonstandard singularly perturbed form that arise naturally as rate-based models of fast-rate chemical processes were studied, and nonlinear coordinate changes were developed that allow transforming such systems in a standard singularly perturbed form [10].

Many real processes are modeled by dynamic models containing several small/large parameters that arise due to the presence of more than one large reaction rate constants, heat/mass transfer coefficients, time constants and other physical constants. The presence of several large parameters in the dynamic model does not necessarily imply the existence of dynamics over several distinct time scales. If these parameters are of the same order of magnitude, then the system is multi-parameter but not multi-time-scale [7], and it is usually approached as a single parameter (two time scale) singular perturbation problem. This is achieved by expressing the small parameters (inverses of the large parameters) as multiples of a single parameter. On the other hand, if the large parameters are of different order of magnitudes, then the system may exhibit multiple time scale dynamics [7]. The modeling, analysis and control of multiple time scale systems has received very little attention; most research efforts have focused on systems in standard singularly perturbed form.

First, we focus on multiple time scale systems in standard singularly perturbed form and briefly review the derivation of reduced models in each individual time scale. Then, we consider a broad class of multiple time scale models in nonstandard singularly perturbed form that arise naturally in detailed rate-based process models of fast-rate chemical processes with several large parameters of different orders of magnitude. We present a systematic approach to decompose this system into dynamics in individual time scales. Finally, we propose a nonlinear coordinate transformation to yield a standard singularly perturbed representation of the original system. Such a non linear diffeomorphism is derived for a representative multiple time scale chemical process in non standard singularly perturbed form.

## 2 Standard Singularly Perturbed Form

A standard singularly perturbed form of multiple time scale systems can be expressed as follows:

$$\begin{aligned}\dot{\zeta} &= F(\zeta, \eta_1, \ldots, \eta_M, u, \epsilon)\\ \epsilon_j \dot{\eta_j} &= G_j(\zeta, \eta_1, \ldots, \eta_M, u, \epsilon), \qquad j = 1, \ldots, M\end{aligned}$$

where $\zeta \in \mathbb{R}^n$ and $\eta_j \in \mathbb{R}^{m_j}$ are the state variables, $u \in \mathbb{R}^q$ is the vector of manipulated inputs, $F$, $G_j \in \mathbb{R}^{m_j}$ are smooth vector fields of dimensions $n$ and $m_j$ respectively, and $\epsilon = [\epsilon_1, \ldots, \epsilon_M]^T$ is a vector of small positive parameters $\epsilon_1, \ldots, \epsilon_M$, known as the singular perturbation parameters, which satisfy:

$$\frac{\epsilon_{j+1}}{\epsilon_j} \to 0 \text{ as } \epsilon_1 \to 0, \qquad j = 1, \ldots, M \tag{1}$$

Conditions for regular degeneration for multiple time scale systems have been derived, in analogy with two time scale systems, in terms of the properties of the Jacobian matrices in the individual time scales. Specifically, they require that the matrix $(\partial G_j(\zeta, \eta_1, \ldots, \eta_M, u, 0)/\partial \eta_j)$ for $j = 1, \ldots, M$ is non-singular, and additionally this condition is satisfied with $\epsilon$ replaced by $\epsilon_j$ for $j = 1, \ldots, M$ [5]. Under these conditions, the system (1) with small parameters satisfying (1) exhibits $M$ distinct fast time scales and 1 slow time scale [5]. Note that $\epsilon_M$ denotes the smallest parameter (responsible for the fastest fast time scale) and $\epsilon_1$ represents the largest of the small parameters (responsible for the slowest fast time scale). This essentially implies that the variable $\eta_{j+1}$ is faster than the variable $\eta_j$, for $j = 1, \ldots, M-1$. Note that such a hierarchy of fast subsystems is a characteristic feature that distinguishes multi-time-scale systems from two time scale ones.

In the limiting case when $\epsilon \to 0$ the dynamic order of the system of (1) degenerates from $(n + \sum_j m_j)$ to $n$, and the slow subsystem is obtained as:

$$\begin{aligned}\dot{\zeta} &= F(\zeta, \eta_1, \ldots, \eta_M, u, 0)\\ 0 &= G_j(\zeta, \eta_1, \ldots, \eta_M, u, 0), \qquad j = 1, \ldots, M\end{aligned}$$

The quasi-steady-state solutions $\eta_j = \sigma_j(\zeta, u)$ for $j = 1, \ldots, M$ are readily obtained by (locally) solving the set of algebraic equations $0 = G_j(\zeta, \eta_1, \ldots, \eta_M, u, 0)$. Then the slow subsystem can be expressed as:

$$\dot{\zeta} = F(\zeta, \sigma_1(\zeta, u), \ldots, \sigma_M(\zeta, u), u, 0) \tag{2}$$

Introducing a "stretched" fastest fast time scale $\tau_M = t/\epsilon_M$, the system in (1) takes the form:

$$\frac{\mathrm{d}\zeta}{\mathrm{d}\tau_M} = \epsilon_M F(\zeta, \eta_1, \ldots, \eta_M, u, \epsilon)$$

$$\frac{\mathrm{d}\eta_j}{\mathrm{d}\tau_M} = \frac{\epsilon_M}{\epsilon_j} G_j(\zeta, \eta_1, \ldots, \eta_M, u, \epsilon), \qquad j = 1, \ldots, M-1$$

$$\frac{\mathrm{d}\eta_M}{\mathrm{d}\tau_M} = G_M(\zeta, \eta_1, \ldots, \eta_M, u, \epsilon)$$

In the limit $\epsilon_M \to 0$, the dynamics of the slow variables $\zeta$ and $\eta_j$ for $j = 1, \ldots, M-1$ become negligible, and the representation of the dynamics corresponding to the fastest fast time scale $\tau_M$ is obtained as:

$$\frac{\mathrm{d}\eta_M}{\mathrm{d}\tau_M} = G_M(\zeta, \eta_1, \ldots, \eta_M, u, 0) \tag{3}$$

where the slow variables $\zeta$ and the slower fast variables (i.e., $\eta_j$, $j = 1, \ldots, M-1$) are "frozen" at their initial conditions and treated as constant parameters. The fast subsystem in (3) represents the *fastest boundary layer* subsystem.

In general, the introduction of the "stretched" $l$th fast time scale, where $1 \leq l \leq M$, $\tau_l = t/\epsilon_l$ results in the following description of the system in (1):

$$\frac{\mathrm{d}\zeta}{\mathrm{d}\tau_l} = \epsilon_l F(\zeta, \eta_1, \ldots, \eta_M, u, \epsilon)$$

$$\frac{\mathrm{d}\eta_j}{\mathrm{d}\tau_l} = \tfrac{\epsilon_l}{\epsilon_j} G_j(\zeta, \eta_1, \ldots, \eta_M, u, \epsilon), \qquad j = 1, \ldots, l-1$$

$$\frac{\mathrm{d}\eta_l}{\mathrm{d}\tau_l} = G_l(\zeta, \eta_1, \ldots, \eta_M, u, \epsilon)$$

$$\frac{\epsilon_j}{\epsilon_l}\frac{\mathrm{d}\eta_j}{\mathrm{d}\tau_l} = G_j(\zeta, \eta_1, \ldots, \eta_M, u, \epsilon), \qquad j = l+1, \ldots, M$$

In the limit $\epsilon_l \to 0$, the dynamics of the slow variables $\zeta$ become negligible, and since $\dfrac{\epsilon_j}{\epsilon_l} \to 0$ for $j = l+1, \ldots, M$, and $\dfrac{\epsilon_l}{\epsilon_j} \to 0$ for $j = 1, \ldots, l-1$, we obtain $\eta_j = 0$ for $j = 1, \ldots, l-1$, and the differential equations for $\eta_j$ for $j = l+1, \ldots, M$, are replaced by a set of algebraic equations $0 = G_j(\zeta, \eta_1, \ldots, \eta_M, u, 0)$, $j = l+1, \ldots, M$. The representation of the $l$th boundary layer subsystem corresponding to the fast variables $\eta_l$ is then obtained as:

$$\frac{\mathrm{d}\eta_l}{\mathrm{d}\tau_l} = G_l(\zeta, \eta_1, \ldots, \eta_M, u, 0)$$

$$0 = G_j(\zeta, \eta_1, \ldots, \eta_M, u, 0), \qquad j = l+1, \ldots, M$$

where the slow variables $\zeta$ and $\eta_j$ for $j = 1, \ldots, l-1$, are "frozen" at their initial conditions $\zeta(0), \eta_j(0)$ and treated as constant parameters, and the variables $\eta_j$, $j = l+1, \ldots, M$ are obtained as quasi-steady-state solutions of the algebraic equations $G_j(\zeta, \eta_1, \ldots, \eta_M, u, 0) = 0$, $j = l+1, \ldots, M$.

The nested application of single parameter singular perturbation (two time scale) theory to multi-time-scale systems illustrated above (for more details, see e.g.[5, 6]), has been used for stability analysis of linear [12] and nonlinear systems (see e.g.[4]). In a different vein, a class of multi-parameter systems with several small parameters of the same order of magnitude, but with unknown relations between them has been studied using the concept of D-stability (see e.g.[7]).

## 3 Nonstandard Singularly Perturbed Form

In this section, we generalize the previous approach to multiple time scale systems of the following general form

$$\dot{x} = f(x) + g(x)u + \sum_{j=1}^{M} \frac{1}{\epsilon_j} b_j(x) k_j(x) \tag{4}$$

where $x \in X \subset \mathbb{R}^n$ is the vector of state variables, $f(x)$ is smooth vector field of dimension $n$, $g(x)$ represents a matrix of dimension $n \times q$, $k_j(x)$ denote smooth vector fields of dimensions $p_j$ for $j = 1, \ldots, M$, $b_j(x)$ denote matrices of dimensions $n \times p_j$, and $\sum_j p_j < n$. We assume that the matrices $b_j(x)$ and the Jacobian matrices $(\partial k_j(x)/\partial x)$ have full column rank and full row rank, respectively.

Such systems arise in the modeling of chemical processes with reactions, heat/mass transfer, etc. occurring in multiple time scales (e.g. [13]). The $1/\epsilon_j$ terms in (4) represent parameters in the dynamic model corresponding to large heat/mass transfer coefficients, large reaction rate constants, etc. We assume that the small parameters $\epsilon_j$ satisfy the relationship of (1), and thus the system (4) is a multiple time scale one.

Let us proceed with the derivation of representations of the system dynamics in the different time scales. A representation of the fastest dynamics is obtained by introducing the "stretched" fastest time scale $\tau_M = t/\epsilon_M$, considering the limit $\epsilon_M \to 0$, and observing that $\lim_{\epsilon_M \to 0} \frac{\epsilon_M}{\epsilon_j} = 0$ for $j < M$, and has the form:

$$\frac{\mathrm{d}x}{\mathrm{d}\tau_M} = b_M(x)\, k_M(x) \tag{5}$$

The system in (5) represents the *fastest boundary layer* subsystem. Similarly, in the slow time scale, $t$, multiplying (4) by $\epsilon_M$ and considering the limit $\epsilon_M \to 0$, the following (linearly independent, as $\left[\frac{\partial k_M(x)}{\partial x}\right]$ has full rank) constraints are obtained:

$$k_{M_i}(x) = 0, \qquad i = 1, \ldots, p_M \tag{6}$$

where $k_{M_i}(x)$ denotes the $i$th component of $k_M(x)$. These constraints must be satisfied in the slow subsystem. Defining $\lim_{\epsilon_M \to 0} \frac{k_{M_i}(x)}{\epsilon_M} = z_{M_i}$ and taking the limit $\epsilon_M \to 0$ in the system of (4), the following system is obtained:

$$\begin{aligned} \dot{x} &= f(x) + g(x)u + \sum_{j=1}^{M-1} \frac{1}{\epsilon_j} b_j(x) k_j(x) + b_M(x) z_M \\ 0 &= k_M(x) \end{aligned}$$

which describes the slow dynamics (after the *fastest boundary layer*) of (4), where $z_M$ denotes the $p_M$-dimensional vector comprising of the variables $z_{M_i}$. Note that the system in (7) is still stiff, as it contains several parameters ($\epsilon_j$, $j = 1, \ldots, M-1$) of different orders of magnitude. Also, the system (7) is a DAE system of nontrivial index, as we do not have algebraic equations to evaluate $z_M$. For most practical cases, the matrix $(L_{b_M} k_M(x))$ is nonsingular, and hence the variables $z_M$ can be obtained after one differentiation of the constraints $k(x)$. This also fixes the index of the DAE system in (7) as two, and the number of slow and fast variables in this fastest time scale as $(n - p_M)$ and $p_M$, respectively. In this case, a solution for the variables $z_M$ can be readily obtained as:

$$z_M = -\left(L_{b_M} k_M(x)\right)^{-1} \left\{ L_f k_M(x) + L_g k_M(x) u + \sum_{j=1}^{M-1} \frac{1}{\epsilon_j} (L_{b_j} k_M(x)) k_j(x) \right\} \tag{7}$$

Observe that the solution for $z_M$ in (7) contains terms $(\frac{1}{\epsilon_j}(L_{b_j} k_M(x)) k_j(x))$, for $j = 1, \ldots, M-1$ that are indeterminate in the limit as $\epsilon_j \to 0$. These terms are implicitly determined by the additional constraints that will be obtained in the subsequent time scales. A state-space realization of the DAE system of (7) can be readily obtained as:

$$\begin{aligned} \dot{x} = f(x) + g(x)u + \sum_{j=1}^{M-1} \frac{1}{\epsilon_j} b_j(x) k_j(x) \\ - b_M(x) \left(L_{b_M} k_M(x)\right)^{-1} \left\{ L_f k_M(x) + L_g k_M(x) u + \sum_{j=1}^{M-1} \frac{1}{\epsilon_j} (L_{b_j} k_M(x)) k_j(x) \right\} \\ 0 = k_M(x) \end{aligned}$$

We can now proceed to obtain a description of the next fastest dynamics (i.e. the dynamics in the $(M-1)$th fast time scale). To this end, we initially rearrange the system in (8) by collecting together terms containing the parameter $\epsilon_{M-1}$ as:

$$\begin{aligned}
\dot{x} = &\left(f(x) - b_M(x)\left(L_{b_M}k_M(x)\right)^{-1} L_f k_M(x)\right) \\
&+ \left(g(x) - b_M(x)\left(L_{b_M}k_M(x)\right)^{-1} L_g k_M(x)\right) u \\
&+ \left\{\sum_{j=1}^{M-2} \frac{1}{\epsilon_j} b_j(x)k_j(x) - b_M(x)\left(L_{b_M}k_M(x)\right)^{-1} \sum_{j=1}^{M-2} \frac{1}{\epsilon_j}(L_{b_j}k_M(x))k_j(x)\right\} \\
&+ \frac{1}{\epsilon_{M-1}}\left\{b_{M-1}(x)k_{M-1}(x) - b_M(x)\left(L_{b_M}k_M(x)\right)^{-1}\ (L_{b_{M-1}}k_M(x))k_{M-1}(x)\right\} \\
0 = &\ k_M(x)
\end{aligned}$$

Furthermore, introducing the "stretched" $(M-1)$th fast time scale $\tau_{M-1} = \dfrac{t}{\epsilon_{M-1}}$ and considering the limit $\epsilon_{M-1} \to 0$, we obtain the following description of the $(M-1)$th fast dynamics of the system in (4):

$$\begin{aligned}
\frac{\mathrm{d}x}{\mathrm{d}\tau_{M-1}} &= \left[\, b_{M-1}(x) \mid b_M(x) \,\right] \begin{bmatrix} k_{M-1}(x) \\ -\left(L_{b_M}k_M(x)\right)^{-1} (L_{b_{M-1}}k_M(x))k_{M-1}(x) \end{bmatrix} \\
0 &= k_M(x)
\end{aligned}$$

The system in (8) represents the $(M-1)$*th boundary layer* subsystem. Assuming that the matrix $\left[\, b_{M-1}(x) \mid b_M(x) \,\right]$ has full column rank, the constraints obtained, in addition to $k_M(x) = 0$, after the $(M-1)$*th boundary layer* are $k_{M-1}(x) = 0$.

Moreover, considering the limit $\epsilon_{M-1} \to 0$ in (8) results in the following description of the slow dynamics after the $(M-1)$th boundary layer:

$$\begin{aligned}
\dot{x} = &\left(f(x) - b_M(x)\left(L_{b_M}k_M(x)\right)^{-1} L_f k_M(x)\right) \\
&+ \left(g(x) - b_M(x)\left(L_{b_M}k_M(x)\right)^{-1} L_g k_M(x)\right) u \\
&+ \left\{\sum_{j=1}^{M-2} \frac{1}{\epsilon_j} b_j(x)k_j(x) - b_M(x)\left(L_{b_M}k_M(x)\right)^{-1} \sum_{j=1}^{M-2} \frac{1}{\epsilon_j}(L_{b_j}k_M(x))k_j(x)\right\} \\
&+ \left[b_{M-1}(x) - b_M(x)\ (L_{b_M}k_M(x))^{-1}\ (L_{b_{M-1}}k_M(x))\right]\ z_{M-1} \\
0 = &\ k_{M-1}(x) \\
0 = &\ k_M(x)
\end{aligned}$$

where $z_{M-1}$ denotes the $p_{M-1}$-dimensional vector comprising of the variables $z_{M-1_i}$ defined as, $z_{M-1_i} = \lim\limits_{\epsilon_{M-1}\to 0} \dfrac{k_{M-1_i}(x)}{\epsilon_{M-1}}$, $i = 1, \ldots, p_{M-1}$. Note that the variables $z_{M-1}$ are implicitly fixed by the constraints $k_{M-1}(x) = 0$.

*Remark 1.* Note that the additional constraints $k_{M-1}(x) = 0$ obtained after the $(M-1)$th boundary layer are the same as the ones that would be obtained in the limit $\epsilon_{M-1} \to 0$ from (7). This implies that the term $\frac{1}{\epsilon_{M-1}} (L_{b_{M-1}} k_M(x))\, k_{M-1}(x)$ in (7) does not introduce additional constraints in the subsequent slow time scales. This indeed is the case as in the limit as $\epsilon_{M-1} \to 0$, we obtain $(L_{b_{M-1}} k_M(x))\, k_{M-1}(x) = 0$ from (7), which is automatically satisfied for $k_{M-1}(x) = 0$.

Proceeding in a similar fashion as above, in the slow time scale after the *$l$th boundary layer*, and assuming that the $(n \times \sum_{j=l}^{M} p_j)$ matrix $[b_l(x) \mid \cdots \mid b_M(x)]$ has full column rank, in the limit $\epsilon_l \to 0$, we obtain the additional constraints $k_l(x) = 0$. These constraints along with the earlier constraints corresponding to faster time scales must be satisfied in the slow subsystem, i.e., $k_j(x) = 0$ for $j = l, \ldots, M$. Defining $\lim_{\epsilon_l \to 0} \frac{k_{l_i}(x)}{\epsilon_l} = z_{l_i}$ and taking the limit $\epsilon_l \to 0$ in the system of (4), the following system is obtained:

$$\begin{aligned} \dot{x} &= f(x) + g(x)u + \sum_{j=1}^{l-1} \frac{1}{\epsilon_j} b_j(x)\, k_j(x) + \sum_{j=l}^{M} b_j(x)\, z_j \\ 0 &= k_j(x) \qquad j = l, \ldots, M \end{aligned}$$

which describes the slow dynamics (after the *$l$th boundary layer*) of (4). Note that the system in (8) is "less stiff" in comparison with the system in (7), as it contains fewer parameters ($\epsilon_j$, $j = 1, \ldots, l-1$) of different orders of magnitude. Also, the system (8) is again a DAE system of nontrivial index, as we do not have algebraic equations to evaluate the algebraic variables $z_j$, $j = l, \ldots, M$. To this end, we assume that (also see (7),(7) and the related discussion) the $\sum_{j=l}^{M} p_j \times \sum_{j=l}^{M} p_j$ matrix $(L_b k(x))_l$ defined as:

$$(L_b k(x))_l := \begin{bmatrix} L_{b_M}\, k_M & \ldots & L_{b_M}\, k_j & \ldots & L_{b_M}\, k_l \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ L_{b_j}\, k_M & \ldots & L_{b_j}\, k_j & \ldots & L_{b_j}\, k_l \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ L_{b_l}\, k_M & \ldots & L_{b_l}\, k_j & \ldots & L_{b_l}\, k_l \end{bmatrix} \tag{8}$$

is nonsingular for $\forall\, l \in [1, M]$. This fixes the index of the DAE system in (8) as two, and the number of slow and fast variables as $(n - \sum_{j=l}^{M} p_j)$ and $\sum_{j=l}^{M} p_j$, respectively. Observe that the nonsingularity of the matrix $(L_b k(x))_l$ for $l = 1, \ldots, M$ implies that all principal minors of $(L_b k(x))_l$ are nonzero, which ensures the solution for the variables $z_l$ in the individual time scales, after just one differentiation of the corresponding constraints.

Proceeding in an analogous manner, a description of the slow dynamics is obtained after the slowest boundary layer corresponding to $l = 1$. Specifically, in the limit as $\epsilon_1 \to 0$, we obtain the system:

$$\begin{aligned} \dot{x} &= f(x) + g(x)u + \sum_{j=1}^{M} b_j(x)\, z_j \\ 0 &= k_j(x) \qquad j = 1, \ldots, M \end{aligned}$$

which describes the slow dynamics (after the *slowest boundary layer*) of (4). Note the *non-stiff* character of the system in (9).

Note that the above approach to derive representations of dynamics in individual time scales does not identify slow and fast variables associated with the individual time scales. Let us now address the derivation of nonlinear changes of coordinates which allow the transformation of system in (4) into a standard singularly perturbed form, thus identifying explicitly the slow and fast variables. The fact that $k_l(x)$, $l = 1, \ldots, M$ are identically equal to zero in the slow systems after the $l$th boundary layer (see (7), (8), (9)), and are non-zero in the $l$th boundary layer, indicates that they should be used in the definition of the fast variables in such a coordinate change. A nonlinear coordinate change can be obtained along these lines yielding a standard singularly perturbed representation of the system in (4), and is given in the following theorem.

**Theorem 1.** *Consider the system in (4), and assume that,*

*(i) the* $\left( \sum_{j=l}^{M} p_j \times \sum_{j=l}^{M} p_j \right)$ *matrix* $(L_b k(x))_l$ *defined in (8) is nonsingular*

*(ii) the* $\left( \sum_{j=l}^{M} p_j \times n \right)$ *matrix* $\left[ \left( \frac{\partial k_l(x)}{\partial x} \right)^T \mid \cdots \mid \left( \frac{\partial k_M(x)}{\partial x} \right)^T \right]^T$ *has full row rank*

$\forall\, l \in [1, M]$. *Then there exists a coordinate change of the form:*

$$\begin{bmatrix} \zeta \\ \eta_1 \\ \vdots \\ \eta_j \\ \vdots \\ \eta_M \end{bmatrix} = T(x, \epsilon) = \begin{bmatrix} \phi(x) \\ \dfrac{k_1(x)}{\epsilon_1} \\ \vdots \\ \dfrac{k_j(x)}{\epsilon_j} \\ \vdots \\ \dfrac{k_M(x)}{\epsilon_M} \end{bmatrix} \tag{9}$$

*where* $\zeta \in \mathbb{R}^{n - \sum_j p_j}$, $\eta_j \in \mathbb{R}^{p_j}$, $j = 1, \ldots, M$, *under which the multiple time scale system of (4) takes the following standard singularly perturbed form:*

$$\begin{aligned}
\dot{\zeta} &= \tilde{f}(\zeta,\epsilon\eta) + \tilde{g}(\zeta,\epsilon\eta)\, u + \sum_{i=1}^{M} \left\{ \tilde{b}_i(\zeta,\epsilon\eta)\eta_i \right\} \\
\epsilon_1 \dot{\eta_1} &= \bar{f}^1(\zeta,\epsilon\eta) + \bar{g}^1(\zeta,\epsilon\eta)\, u + \sum_{i=1}^{M} \left\{ \bar{b}_i^1(\zeta,\epsilon\eta)\eta_i \right\} \\
&\vdots \\
\epsilon_j \dot{\eta_j} &= \bar{f}^j(\zeta,\epsilon\eta) + \bar{g}^j(\zeta,\epsilon\eta)\, u + \sum_{i=1}^{M} \left\{ \bar{b}_i^j(\zeta,\epsilon\eta)\eta_i \right\} \\
&\vdots \\
\epsilon_M \dot{\eta_M} &= \bar{f}^M(\zeta,\epsilon\eta) + \bar{g}^M(\zeta,\epsilon\eta)\, u + \sum_{i=1}^{M} \left\{ \bar{b}_i^M(\zeta,\epsilon\eta)\eta_i \right\}
\end{aligned}$$

*where* $\tilde{f} = L_f\phi(x), \tilde{g} = L_g\phi(x), \tilde{b} = L_b\phi(x), \bar{f}^j = L_f k_j(x), \bar{g}^j = L_g k_j(x), \bar{b}_i^j = L_{b_i} k_j(x)$ *are evaluated at* $x = T^{-1}(\zeta,\epsilon\eta)\ \forall\, i,j$, *and the* $\sum_{j=l}^{M} p_j \times \sum_{j=l}^{M} p_j$ *dimensional matrix* $Q_l(\zeta,0) = (L_b k(x))_l$ *evaluated at* $x = T^{-1}(\zeta,0)$ *is non-singular uniformly in* $\zeta \in \mathbb{R}^{n-\sum_j p_j}$, $\forall\, l \in [1,M]$.

*Proof.* Note that given the linear independence of the $\sum_{j=1}^{M} p_j$ scalar functions in $k_j(x)$, $j = 1,\ldots,M$ (condition (ii) for $l = 1$), there exist $(n - \sum_{j=1}^{M} p_j)$ scalar functions $\phi(x)$ such that (9) is a local diffeomorphism, thus qualifying as a valid coordinate change. It can then be directly verified that the system of (4) in the new coordinates takes the form of (10). In order to prove that (10) is in the standard singularly perturbed form, we must show that the variables $\eta_j$, for $j = l,\ldots,M$ can be solved for their quasi-steady-state solution from the resulting algebraic equations in the limit $\epsilon_l \to 0$, for any arbitrary $l \in [1,M]$. The condition (i) in the theorem guarantees the existence of such a quasi-steady-state solution for the variables $\eta_j$, as the matrix $Q_l(\zeta,0)$ is uniformly nonsingular in the limit $\epsilon_l \to 0$, $\forall\, l \in [1,M]$. This proves the theorem.

Note that the system in (10) allows obtaining the entire hierarchy of the boundary layer subsystems and the corresponding slow subsystems for any arbitrary $l \in [1,M]$.

## 4 Application

In this section, we derive a coordinate change for a representative multiple time scale chemical system in non standard singularly perturbed form.

We consider an isothermal CSTR of volume $V$ where reactants $A$ and $E$ are fed at a flow rate $F^{in}$ at concentrations $C_A^{in}$ and $C_E^{in}$, respectively. The following reactions occur:

$$\begin{gathered}
A \rightleftharpoons B \to C \rightleftharpoons D \\
E \to F
\end{gathered}$$

The reaction rates $R_1$ and $R_3$ for the reversible reactions $A \rightleftharpoons B$ and $C \rightleftharpoons D$, respectively, are given by

$$R_1 = k_1 \left( C_A - \frac{C_B}{\kappa_1} \right)$$

and

$$R_3 = k_3 \left( C_C - \frac{C_D}{\kappa_3} \right)$$

where $C_A$, $C_B$, $C_C$ and $C_D$ denote the concentrations of species $A$,$B$,$C$ and $D$, $k_1$ and $k_3$ are the reaction rate constants whereas $\kappa_1$ and $\kappa_3$ are equilibrium constants. The reaction rates $R_2$ and $R_4$ for the irreversible reactions $B \rightarrow C$ and $E \rightarrow F$ are given by

$$R_2 = k_2 C_B$$

and

$$R_4 = k_4 C_E$$

where $C_E$ denotes the concentration of species $E$.

The dynamic model of this process takes the following form

$$\begin{aligned}
\dot{V} &= F^{in} - F^{out} \\
\dot{C_A} &= \frac{F^{in}}{V} \left( C_A^{in} - C_A \right) - k_1 \left( C_A - \frac{C_B}{\kappa_1} \right) \\
\dot{C_B} &= -\frac{F^{in}}{V} C_B + k_1 \left( C_A - \frac{C_B}{\kappa_1} \right) - k_2 C_B \\
\dot{C_C} &= -\frac{F^{in}}{V} C_C + k_2 C_B - k_3 \left( C_C - \frac{C_D}{\kappa_3} \right) \\
\dot{C_D} &= -\frac{F^{in}}{V} C_D + k_3 \left( C_C - \frac{C_D}{\kappa_3} \right) \\
\dot{C_E} &= \frac{F^{in}}{V} \left( C_E^{in} - C_E \right) - k_4 C_E \\
\dot{C_F} &= -\frac{F^{in}}{V} C_F + k_4 C_E
\end{aligned} \tag{10}$$

It is assumed that the reaction rates are such that the following inequalities hold $k_2 \ll k_4 \ll k_3 \ll k_1$. Moreover, the reaction equilibrium constants $\kappa_1$ and $\kappa_3$ are assumed to be different. Owing to the presence of reaction rates of widely spread orders of magnitude, the process exhibit multiple time scale dynamics.

The system in (10) is in the nonstandard singularly perturbed form of (4) where the singular perturbation parameters are defined as

$$\epsilon_1 = \frac{1}{k_1} \quad \epsilon_2 = \frac{1}{k_3} \quad \epsilon_3 = \frac{1}{k_4}$$

In this description, the vector of state variables $x$ and the vector $f(x)+g(x)u$ take the form

$$x = \begin{bmatrix} V \\ C_A \\ C_B \\ C_C \\ C_D \\ C_E \\ C_F \end{bmatrix} \qquad f(x)+g(x)u = \begin{bmatrix} F^{in} - F^{out} \\ \frac{F^{in}}{V}\left(C_A^{in} - C_A\right) \\ -\frac{F^{in}}{V}C_B - k_2 C_B \\ -\frac{F^{in}}{V}C_C + k_2 C_B \\ -\frac{F^{in}}{V}C_D \\ \frac{F^{in}}{V}\left(C_E^{in} - C_E\right) \\ -\frac{F^{in}}{V}C_F \end{bmatrix} \tag{11}$$

The remaining vectors are

$$\begin{aligned} b_1(x) &= \begin{bmatrix} 0 & -1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}^T & k_1(x) &= C_A - C_B/\kappa_1 \\ b_2(x) &= \begin{bmatrix} 0 & 0 & 0 & -1 & 1 & 0 & 0 \end{bmatrix}^T & k_2(x) &= C_C - C_D/\kappa_3 \\ b_3(x) &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}^T & k_3(x) &= C_E \end{aligned} \tag{12}$$

Assumptions $(i)$ and $(ii)$ in the theorem are fulfilled for this system. Specifically, the matrices $(L_b k(x))_l$ for $l \in [1,2,3]$ defined in (8) are nonsingular since $\kappa_1 \neq \kappa_3$ is assumed

$$(L_b k(x))_1 = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 - \frac{1}{\kappa_3} & 0 \\ 0 & 0 & -1 - \frac{1}{\kappa_1} \end{bmatrix} \tag{13}$$

$$(L_b k(x))_2 = \begin{bmatrix} -1 & 0 \\ 0 & -1 - \frac{1}{\kappa_3} \end{bmatrix} \qquad (L_b k(x))_3 = \begin{bmatrix} -1 \end{bmatrix}$$

It can also be easily verified that the Jacobian matrices

$$\left[ \left( \frac{\partial k_l(x)}{\partial x} \right)^T \mid \cdots \mid \left( \frac{\partial k_M(x)}{\partial x} \right)^T \right]^T$$

for $l \in [1,2,3]$ have full row rank, given that the Jacobians of the scalars $k_1(x)$, $k_2(x)$ and $k_3(x)$ are

$$\begin{aligned}
\frac{\partial k_1}{\partial x} &= \begin{bmatrix} 0\ 1\ -\frac{1}{\kappa_1}\ 0\ 0\ 0\ 0 \end{bmatrix} \\
\frac{\partial k_2}{\partial x} &= \begin{bmatrix} 0\ 0\ 0\ 1\ -\frac{1}{\kappa_3}\ 0\ 0 \end{bmatrix} \\
\frac{\partial k_3}{\partial x} &= \begin{bmatrix} 0\ 0\ 0\ 0\ 0\ 1\ 0 \end{bmatrix}
\end{aligned} \tag{14}$$

Since assumptions $(i)$ and $(ii)$ hold, the theorem guarantees the existence of a coordinate change leading to a standard singularly perturbed form of the system in (10). In particular, we consider the following coordinate change

$$T\left(x, \epsilon_1, \epsilon_2, \epsilon_3\right) = \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \zeta_4 \\ \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} V \\ C_A + C_B \\ C_C + C_D \\ C_E + C_F \\ k_1\left(C_A - \dfrac{C_B}{\kappa_1}\right) \\ k_3\left(C_C - \dfrac{C_D}{\kappa_3}\right) \\ k_4 C_E \end{bmatrix} \tag{15}$$

Under this diffeomorphism, the multiple time scale system of (10) yields the following standard singularly perturbed system

$$
\begin{aligned}
\dot{\zeta_1} &= F^{in} - F^{out} \\
\dot{\zeta_2} &= \frac{F^{in}}{V} C_A^{in} - \frac{F^{in}}{V}\zeta_2 - \frac{k_2}{1+\frac{1}{\kappa_1}}\left(\zeta_2 - \epsilon_1\eta_1\right) \\
\dot{\zeta_3} &= -\frac{F^{in}}{V}\zeta_3 + \frac{k_2}{1+\frac{1}{\kappa_1}}\left(\zeta_2 - \epsilon_1\eta_1\right) \\
\dot{\zeta_4} &= \frac{F^{in}}{V} C_E^{in} - \frac{F^{in}}{V}\zeta_4 \\
\epsilon_1\dot{\eta_1} &= \frac{F^{in}}{V}\left(C_A^{in} - \epsilon_1\eta_1\right) - \left(1+\frac{1}{\kappa_1}\right)\eta_1 + \frac{k_2}{1+\frac{1}{\kappa_1}}\left(\zeta_2 - \epsilon_1\eta_1\right) \\
\epsilon_2\dot{\eta_2} &= -\frac{F^{in}}{V}\epsilon_2\eta_2 - \left(1+\frac{1}{\kappa_3}\right)\eta_2 + \frac{k_2}{1+\frac{1}{\kappa_1}}\left(\zeta_2 - \epsilon_1\eta_1\right) \\
\epsilon_3\dot{\eta_3} &= \frac{F^{in}}{V}\left(C_E^{in} - \epsilon_3\eta_3\right) - \eta_3
\end{aligned}
\tag{16}
$$

where fast and slow variables are clearly identified.

## 5 Conclusion

In this work, we considered a nonstandard singularly perturbed form of multiple time scale systems arising as models of chemical processes. For this class of systems, we derived representations of the subsystems describing the dynamics in individual time scales following a nested application of singular perturbation arguments. We also proposed a nonlinear coordinate transformation that yields a standard singularly perturbed form. The results were exemplified through a chemical reactor example with several reactions having rates of different orders of magnitude.

## References


1. G.E. Bailey, D.F. Ellis: *Biochemical engineering fundamentals.* McGraw Hill chemical engineering series (1977)
2. H.C. Chang, M. Aluko: Multi-scale analysis of exotic dynamics in surface catalyzed reaction-I. *Chem. Eng. Sci.* **39**, 37–50 (1984)
3. M.M. Denn: *Process modeling* (Longam Inc. 1986)
4. C.A. Desoer, S.M. Shahruz: Stability of nonlinear systems with three time scales. *Circt. Syst. Sig. Pro.* **5**, 449–464 (1986)
5. F. Hoppensteadt: Properties of solutions of ordinary differential equations with small parameters. *Comm. Pure and Appl. Math.* **24**, 807–840 (1971)
6. R.E. O'Malley Jr: *Singular perturbation methods for ordinary differential equations* (Springer, Berlin Heidelberg New York 1991)

7. H.K. Khalil, P.V. Kokotovic: Control of linear systems with multiparameter singular perturbations. *Automatica* **15**, 197–207 (1979)
8. P.V. Kokotovic, H.K. Khalil, J. O'Reilly: *Singular perturbations in control: analysis and design* (Academic press, London 1986)
9. P.V. Kokotovic, Jr. R.E. O'Malley, P. Sannuti: Singular perturbations and order reduction in control theory - an overview. *Automatica* **12**, 123–132 (1976)
10. A. Kumar, P.D. Christofides, P. Daoutidis: Singular perturbation modeling of nonlinear processes with non-explicit time-scale separation. *Chem. Eng. Sci.* **53** 1491–1504 (1998)
11. A. Kumar, P. Daoutidis: *Control of non-linear differential algebraic equation systems* (Chapman and Hall/CRC 1999)
12. G.S. Ladde, D.D. Siljak: Multiparameter singular perturbaton of linear systems with multiple time scales. *Automatica* **19**, 385–394 (1983)
13. N. Vora: *Nonlinear model reduction and control of multiple time scale chemical processes: complex reaction systems and reactive distillation columns.* PhD thesis, Univ. of Minnesota, Minneapolis (2000)
14. N. Vora, P. Daoutidis: Nonlinear model reduction of chemical reaction systems. In: *Proc. of 1999 ACC*, 1583–1587 (San Diego, CA 1999)
15. N. Vora, P. Daoutidis: Dynamics and control of an ethyl acetate reactive distillation column. *Ind. Eng. Chem. Res.* **40**, 833–849 (2001)

# Part II

## Coarse-Graining and Ideas of Statistical Physics

# Basic Types of Coarse-Graining

A. N. Gorban

University of Leicester, Leicester LE1 7RH, UK, ag153@le.ac.uk

**Summary.** We consider two basic types of coarse-graining: the Ehrenfests' coarse-graining and its extension to a general principle of non-equilibrium thermodynamics, and the coarse-graining based on uncertainty of dynamical models and $\varepsilon$-motions (orbits). Non-technical discussion of basic notions and main coarse-graining theorems are presented: the theorem about entropy overproduction for the Ehrenfests' coarse-graining and its generalizations, both for conservative and for dissipative systems, and the theorems about stable properties and the Smale order for $\varepsilon$-motions of general dynamical systems including structurally unstable systems. Computational kinetic models of macroscopic dynamics are considered. We construct a theoretical basis for these kinetic models using generalizations of the Ehrenfests' coarse-graining. General theory of reversible regularization and filtering semigroups in kinetics is presented, both for linear and non-linear filters. We obtain explicit expressions and entropic stability conditions for filtered equations. A brief discussion of coarse-graining by rounding and by small noise is also presented.

## 1 Introduction

Almost a century ago, Paul and Tanya Ehrenfest in their paper for scientific Encyclopedia [1] introduced a special operation, the coarse-graining. This operation transforms a probability density in phase space into a "coarse-grained" density, that is a piece-wise constant function, a result of density averaging in cells. The size of cells is assumed to be small, but finite, and does not tend to zero. The coarse-graining models uncontrollable impact of surrounding (of a thermostat, for example) onto ensemble of mechanical systems.

To understand reasons for introduction of this new notion, let us take a phase drop, that is, an ensemble of mechanical systems with constant probability density localized in a small domain of phase space. Let us watch evolution of this drop in time according to the Liouville equation. After a long time, the shape of the drop may be very complicated, but the density value remains the same, and this drop remains "oil in water." The ensemble can tend to the equilibrium in the weak sense only: average value of any continuous function

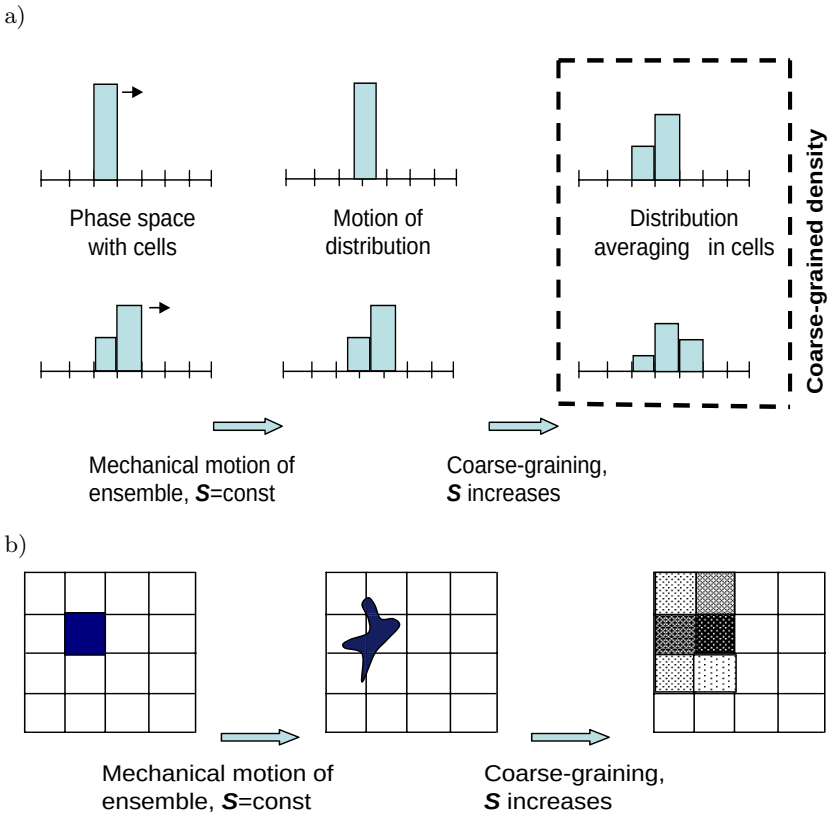



Fig. 1: The Ehrenfests' coarse-graining: two "motion – coarse-graining" cycles in 1D (a, values of probability density are presented by the height of the columns) and one such cycle in 2D (b, values of probability density are presented by hatching density).

tends to its equilibrium value, but the entropy of the distribution remains constant. Nevertheless, if we divide the phase space into cells and supplement the mechanical motion by the periodical averaging in cells (this is the Ehrenfests' idea of coarse-graining), then the entropy increases, and the distribution density tends uniformly to the equilibrium. This periodical coarse-graining is illustrated by Fig. 1 for one-dimensional (1D)[1] and two-dimensional (2D) phase spaces.

Recently, we can find the idea of coarse-graining everywhere in statistical physics (both equilibrium and non-equilibrium). For example, it is the central idea of the Kadanoff transformation, and can be considered as a background

[1] Of course, there is no mechanical system with one-dimensional phase space, but dynamics with conservation of volume is possible in 1D case too: it is a motion with constant velocity.

of the Wilson renormalization group [6] and modern renormalisation group approach to dissipative systems [7, 8]. [2] It gave a simplest realization of the projection operators technique [2] long before this technic was developed. In the method of invariant manifold [3, 4] the generalized Ehrenfests' coarse-graining allows to find slow dynamics without a slow manifold construction. It is also present in the background of the so-called equation-free methods [9]. Applications of the Ehrenfests' coarse-graining outside statistical physics include simple, but effective filtering [10]. The Gaussian filtering of hydrodynamic equations that leads to the Smagorinsky equations [14] is, in its essence, again a version of the Ehrenfests' coarse-graining. In the first part of this paper we elaborate in details the Ehrenfests' coarse-graining for dynamical systems.

The central idea of the Ehrenfests' coarse-graining remains the same in most generalizations: we combine the genuine motion with the periodic *partial equlibration*. The result is the Ehrenfests' chain. After that, we can find the macroscopic equation that does not depend on an initial distribution and describes the Ehrenfests' chains as results of continuous autonomous motion [5, 11]. Alternatively, we can just create a computational procedure without explicit equations [9]. In the sense of entropy production, the resulting macroscopic motion is "more dissipative" than initial (microscopic) one. It is the theorem about entropy overproduction. In its general form it was proven in [12].

Kinetic models of fluid dynamics become very popular during the last decade. Usual way of model simplification leads from kinetics to fluid dynamics, it is a sort of dimension reduction. But kinetic models go back, and it is the simplification also. Some of kinetic equations are very simple and even exactly solvable. The simplest and most popular example is the free flight kinetics, $\partial f(\boldsymbol{x},\boldsymbol{v},t)/\partial t = -\sum_i v_i \partial f(\boldsymbol{x},\boldsymbol{v},t)/\partial x_i$, where $f(\boldsymbol{x},\boldsymbol{v},t)$ is one-particle distribution function, $\boldsymbol{x}$ is space vector, $\boldsymbol{v}$ is velocity. We can "lift" a continuum equation to a kinetic model, and than approximate the solution by a chain, each link of which is a kinetic curve with a jump from the end of this curve to the beginning of the next link. In this paper, we describe how to construct these curves, chains, links and jumps on the base of Ehrenfests' idea. Kinetic model has more variables than continuum equation. Sometimes simplification in modeling can be reached by dimension increase, and it is not a miracle.

In practice, kinetic models in the form of lattice Boltzmann models are in use [19]. The Ehrenfests' coarse-graining provides theoretical basis for kinetic models. First of all, it is possible to replace projecting (partial equilibration) by involution (i.e. reflection with respect to the partial equilibrium). This *entropic involution* was developed for the lattice Boltzmann methods in [89]. In the original Ehrenfests' chains, "motion–partial equilibration–motion–...," dissipation is coupled with time step, but the chains "motion–involution–motion–..." are conservative. The family of chains between conservative (with

[2] See also the paper of A. Degenhard and J. Javier Rodriguez-Laguna in this volume.

entropic involution) and maximally dissipative (with projection) ones give us a possibility to model hydrodynamic systems with various dissipation (viscosity) coefficients that are decoupled with time steps.

Large eddy simulation, filtering and subgrid modeling are very popular in fluid dynamics [13, 14, 15, 16, 17]. The idea is that small inhomogeneities should somehow equilibrate, and their statistics should follow the large scale details of the flow. Our goal is to restore a link between this approach and initial coarse-raining in statistical physics. Physically, this type of coarse-graining is transference the energy of small scale motion from macroscopic kinetic energy to microscopic internal energy. The natural framework for analysis of such transference provides physical kinetics, where initially exists no difference between kinetic and internal energy. This difference appears in the continuum mechanic limit. We proposed this idea several years ago, and an example for moment equations was published in [18]. Now the kinetic approach for filtering is presented. The general commutator expansion for all kind of linear or non-linear filters, with constant or with variable coefficients is constructed. The condition for stability of filtered equation is obtained.

The upper boundary for the filter width $\Delta$ that guaranties stability of the filtered equations is proportional to the square root of the Knudsen number. $\Delta/L \sim \sqrt{\mathrm{Kn}}$ (where $L$ is the characteristic macroscopic length). This scaling, $\Delta/L \sim \sqrt{\mathrm{Kn}}$, was discussed in [18] for moment kinetic equations because different reasons: if $\Delta/L \gg \sqrt{\mathrm{Kn}}$ then the Chapman–Enskog procedure for the way back from kinetics to continuum is not applicable, and, moreover, the continuum description is probably not valid, because the filtering term with large coefficient $\Delta/L$ violates the conditions of hydrodynamic limit. This important remark gives the frame for $\eta$ scaling. It is proven in this paper for the broad class of model kinetic equations. The entropic stability conditions presented below give the stability boundaries inside this scale.

Several other notions of coarse-graining were introduced and studied for dynamical systems during last hundred years. In this paper, we shall consider one of them, the coarse-graining by $\varepsilon$-motions ($\varepsilon$-orbits, or pseudo orbits) and briefly mention two other types: coarse-graining by rounding and by small random noise.

$\varepsilon$-motions describe dynamics of models with uncertainty. We never know our models exactly, we never deal with isolated systems, and the surrounding always uncontrollably affect dynamics of the system. This dynamics can be presented as a usual phase flow supplemented by a periodical *$\varepsilon$-fattening*: after time $\tau$, we add a $\varepsilon$-ball to each point, hence, points are transformed into sets. This periodical fattening expands all attractors: for the system with fattening they are larger than for original dynamics.

Interest to the dynamics of $\varepsilon$-motions was stimulated by the famous work of S. Smale [20]. This paper destroyed many naive dreams and expectations. For generic 2D system the phase portrait is the structure of attractors (sinks), repellers (sources), and saddles. For generic 2D systems all these attractors are either fixed point or closed orbits. Generic 2D systems are structurally

stable. It means that they do not change qualitatively after small perturbations. Our dream was to find a similar stable structure in generic systems for higher dimensions, but S. Smale showed it is impossible: Structurally stable systems are not dense! Unfortunately, in higher dimensions there are regions of dynamical systems that can change qualitatively under arbitrary small perturbations.

One of the reasons to study $\varepsilon$-motions (flow with fattening) and systems with sustained perturbations was the hope that even small errors coarsen the picture and can wipe some of the thin peculiarities off. And this hope was realistic, at least, partially [21, 22, 23]. The thin peculiarities that are responsible for appearance of regions of structurally unstable systems vanish after the coarse-graining via arbitrary small periodical fattening. All the models have some uncertainty, hence, the features of dynamics that are unstable under arbitrary small coarse-graining are unobservable.

Rounding is a sort of coarse-graining that appears automatically in computer simulations. It is very natural that in era of intensive computer simulation of complex dynamics the coarse-graining by rounding attracted special attention [24, 25, 26, 27, 28, 29, 30]. According to a very idealized popular dynamic model, rounding might be represented as restriction of shift in given time $\tau$ onto $\varepsilon$-net in phase space. Of courses, the restriction includes some perturbation of dynamics (Fig. 2). The formal definition of rounding action includes a tiling: around any point of the $\varepsilon$-net there is a cell, these cells form a tiling of the phase space, and rounding maps a cell into corresponding point of the $\varepsilon$-net. These cells have equal volumes if there are no special reasons to make their volumes different. If this volume is dynamically invariant then, for sufficiently large time of motion between rounding steps, all the mixing dynamical systems with rounding can be described by an universal object. This is a random dynamical system, the random map of a finite set: any point of the $\varepsilon$-net can be the image of a given point with probability $1/m$ (where $m$ is the number of points in the $\varepsilon$-net). The combinatorial theory of such *random graphs* is well–developed [31].

After rounding, some unexpected properties of dynamics appear. For example, even for transitive systems with strong mixing significant part of points of the $\varepsilon$-net becomes transient after rounding. Initially, attractor of such a continuous system is the whole phase space, but after rounding attractor of discrete dynamical system on the $\varepsilon$-net includes, roughly speaking, a half of its points (or, more precisely, the expectation of the number of transient points is $m(\mathrm{e}-1)/\mathrm{e}$, where $m$ is number of points, $\mathrm{e} = 2.7...$). In some circumstances, complicated dynamics has a tendency to collapse to trivial and degenerate behaviour as a result of discretizations [27]. For systems without conservation of volume, the number of periodic points after discretization is linked to the dimension of the attractor $d$. The simple estimates based on the random map analysis, and numerical experiments with chaotic attractors give $\sim \varepsilon^{-d}$ for the number of periodic points, and $\sim \varepsilon^{-d/2}$ for the scale of the expected period [26, 30]. The first of them is just the number of points in $\varepsilon$-net in

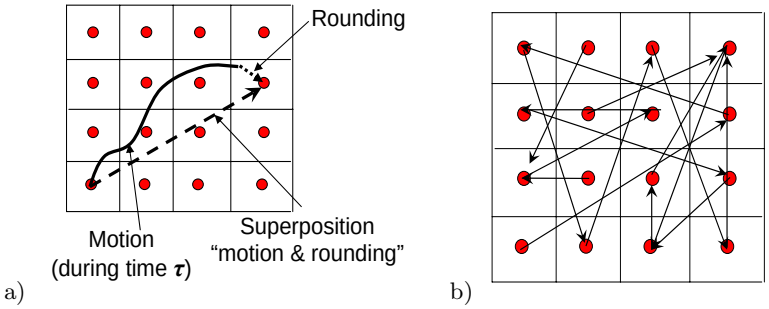



Fig. 2: Motion, rounding and "motion with rounding" for a dynamical system (a), and the universal result of motion with rounding: a random dynamical system (b).

$d$-dimensional compact, the second becomes clear after the following remark. Let us imagine a random walk in a finite set with $m$ elements (a $\varepsilon$-net). When the length of the trajectory is of order $\sqrt{m}$ then the next step returns the point to the trajectory with probability $\sim 1/\sqrt{m}$, and a loop appears with expected period $\sim \sqrt{m}$ (a half of the trajectory length). After $\sim \sqrt{m}$ steps the probability of a loop appearance is near 1, hence, for the whole system the expected period is $\sim \sqrt{m} \sim \varepsilon^{-d/2}$.

It is easy to demonstrate the difference between coarse-graining by fattening and coarse-graining by rounding. Let us consider a trivial dynamics on a connected phase space: let the shift in time be identical transformation. For coarse-graining by fattening the $\varepsilon$-motion of any point tends to cover the whole phase space for any positive $\varepsilon$ and time $t \to \infty$: periodical $\varepsilon$-fattening with trivial dynamics transforms, after time $n\tau$, a point into the sum of $n$ $\varepsilon$-balls. For coarse-graining by rounding this trivial dynamical system generates the same trivial dynamical system on $\varepsilon$-net: nothing moves.

Coarse-graining by small noise seems to be very natural. We add small random term to the right hand side of differential equations that describe dynamics. Instead of the Liouville equation for probability density the Fokker–Planck equation appears. There is no fundamental difference between various types of coarse-graining, and the coarse-graining by $\varepsilon$-fattening includes major results about the coarse-graining by small noise that are insensitive to most details of noise distribution. But the knowledge of noise distribution gives us additional tools. The *action functional* is such a tool for the description of fluctuations [32]. Let $X^\varepsilon(t)$ be a random process "dynamics with $\varepsilon$-small fluctuation" on the time interval $[0, T]$. It is possible to introduce such a functional $\mathbf{S}[\varphi]$ on functions $x = \varphi(t)$ $(t \in [0, T])$ that for sufficiently small $\varepsilon, \delta > 0$

$$\mathbf{P}\{\|X^\varepsilon - \varphi\| < \delta\} \approx \exp(-\mathbf{S}[\varphi]/\varepsilon^2).$$

Action functional is constructed for various types of random perturbations [32]. Introduction to the general theory of random dynamical systems with invariant measure is presented in [33].

In following sections, we consider two types of coarse-graining: the Ehrenfests' coarse-graining and its extension to a general principle of non-equilibrium thermodynamics, and the coarse-graining based on the uncertainty of dynamical models and $\varepsilon$-motions.

## 2 The Ehrenfests' Coarse-Graining

### 2.1 Kinetic Equation and Entropy

*Entropy conservation in systems with conservation of phase volume*

The Erenfest's coarse-graining was originally defined for conservative[3] systems. Usually, Hamiltonian systems are considered as conservative ones, but in all constructions only one property of Hamiltonian systems is used, namely, conservation of the phase volume $\mathrm{d}\Gamma$ (the Liouville theorem). Let $X$ be phase space, $v(x)$ be a vector field, $\mathrm{d}\Gamma = \mathrm{d}^n x$ be the differential of phase volume. The flow,

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v(x), \tag{1}$$

conserves the phase volume, if $\mathrm{div} v(x) = 0$. The continuity equation,

$$\frac{\partial f}{\partial t} = -\sum_i \frac{\partial (f v_i(x))}{\partial x_i}, \tag{2}$$

describes the induced dynamics of the probability density $f(x,t)$ on phase space. For incompressible flow (conservation of volume), the continuity equation can be rewritten in the form

$$\frac{\partial f}{\partial t} = -\sum_i v_i(x) \frac{\partial f}{\partial x_i}. \tag{3}$$

This means that the probability density is constant along the flow: $f(x, t + \mathrm{d}t) = f(x - v(x)\mathrm{d}t, t)$. Hence, for any continuous function $h(f)$ the integral

$$H(f) = \int_X h(f(x))\, \mathrm{d}\Gamma(x) \tag{4}$$

[3] In this paper, we use the term "conservative" as an opposite term to "dissipative:" conservative = with entropy conservation. Another use of the term "conservative system" is connected with energy conservation. For kinetic systems under consideration conservation of energy is a simple linear balance, and we shall use the first sense only.

does not change in time, provided the probability density satisfies the continuity equation (2) and the flow $v(x)$ conserves the phase volume. For $h(f) = -f \ln f$ integral (4) gives the classical Boltzmann–Gibbs–Shannon (BGS) entropy functional:

$$S(f) = -\int_X f(x) \ln(f(x))\, \mathrm{d}\Gamma(x). \tag{5}$$

For flows with conservation of volume, entropy is conserved: $\mathrm{d}S/\mathrm{d}t \equiv 0$ .

*Kullback entropy conservation in systems with regular invariant distribution*

Suppose the phase volume is not invariant with respect to flow (1), but a regular invariant density $f^*(x)$ (equilibrium) exists:

$$\sum_i \frac{\partial(f^*(x)v_i(x))}{\partial x_i} = 0. \tag{6}$$

In this case, instead of an invariant phase volume $\mathrm{d}\Gamma$, we have an invariant volume $f^*(x)\,\mathrm{d}\Gamma$. We can use (6) instead of the incompressibility condition and rewrite (2):

$$\frac{\partial(f(x,t)/f^*(x))}{\partial t} = -\sum_i v_i(x) \frac{\partial(f(x,t)/f^*(x))}{\partial x_i}. \tag{7}$$

The function $f(x,t)/f^*(x)$ is constant along the flow, the measure $f^*(x)\,\mathrm{d}\Gamma(x)$ is invariant, hence, for any continuous function $h(f)$ integral

$$H(f) = \int_X h(f(x,t)/f^*(x)) f^*(x)\, \mathrm{d}\Gamma(x) \tag{8}$$

does not change in time, if the probability density satisfies the continuity equation. For $h(f) = -f \ln f$ integral (8) gives the Kullback entropy functional [42]:

$$S_K(f) = -\int_X f(x) \ln\left(\frac{f(x)}{f^*(x)}\right) \mathrm{d}\Gamma(x). \tag{9}$$

This situation does not differ significantly from the entropy conservation in systems with conservation of volume. It is just a kind of change of variables.

*General entropy production formula*

Let us consider the general case without assumptions about phase volume invariance and existence of a regular invariant density (6). In this case, let a probability density $f(x,t)$ be a solution of the continuity equation (2). For the BGS entropy functional (5)

$$\frac{\mathrm{d}S(f)}{\mathrm{d}t} = \int_X f(x,t) \mathrm{div} v(x)\, \mathrm{d}\Gamma(x), \tag{10}$$

if the left hand side exists. This *entropy production formula* can be easily proven for small phase drops with constant density, and then for finite sums of such distributions with positive coefficients. After that, we obtain formula (10) by limit transition.

For a regular invariant density $f^*(x)$ (equilibrium) entropy $S(f^*)$ exists, and for this distribution $\mathrm{d}S(f)/\mathrm{d}t = 0$, hence,

$$\int_X f^*(x)\mathrm{div}v(x)\,\mathrm{d}\Gamma(x) = 0. \tag{11}$$

*Entropy production in systems without regular equilibrium*

If there is no regular equilibrium (6), then the entropy behaviour changes drastically. If volume of phase drops tends to zero, then the BGS entropy (5) and any Kullback entropy (9) goes to minus infinity. The simplest example clarifies the situation. Let all the solutions converge to unique exponentially stable fixed point $x = 0$. In linear approximation $\mathrm{d}x/\mathrm{d}t = Ax$ and $S(t) = S(0) + t\,\mathrm{tr}A$. Entropy decreases linearly in time with the rate $\mathrm{tr}A$ ($\mathrm{tr}A = \mathrm{div}v(x)$, $\mathrm{tr}A < 0$), time derivative of entropy is $\mathrm{tr}A$ and does not change in time, and the probability distribution goes to the $\delta$-function $\delta(x)$. Entropy of this distribution does not exist (it is "minus infinity"), and it has no limit when $f(x,t) \to \delta(x)$.

Nevertheless, time derivative of entropy is well defined and constant, it is $\mathrm{tr}A$. For more complicated singular limit distributions the essence remains the same: according to (10) time derivative of entropy tends to the average value of $\mathrm{div}v(x)$ in this limit distribution, and entropy goes linearly to minus infinity (if this average in not zero, of course). The order in the system increases. This behaviour could sometimes be interpreted as follows: the system is open and produces entropy in its surrounding even in a steady–state. Much more details are in review [41].[4]

*Starting point: a kinetic equation*

For the formalization of the Ehrenfests' idea of coarse-graining, we start from a formal kinetic equation

$$\frac{\mathrm{d}f}{\mathrm{d}t} = J(f) \tag{12}$$

with a concave entropy functional $S(f)$ that does not increase in time. This equation is defined in a convex subset $U$ of a vector space $E$.

[4] Applications of this formalism are mainly related to Hamiltonian systems in so-called force thermostat, or, in particular, isokinetic thermostat. These thermostats were invented in computational molecular dynamics for acceleration of computations, as a technical trick. From the physical point of view, this theory can be considered as a theory about a friction of particles on the space, the "ether friction." For isokinetic thermostats, for example, this "friction" decelerates some of particles, accelerates others, and keeps the kinetic energy constant.

Let us specify some notations: $E^T$ is the adjoint to the $E$ space. Adjoint spaces and operators will be indicated by $^T$, whereas the notation $^*$ is earmarked for equilibria and quasi-equilibria.

We recall that, for an operator $A : E_1 \to E_2$, the adjoint operator, $A^T : E_1^T \to E_2^T$ is defined by the following relation: for any $l \in E_2^T$ and $\varphi \in E_1$, $l(A\varphi) = (A^T l)(\varphi)$.

Next, $D_f S(f) \in E^T$ is the differential of the functional $S(f)$, $D_f^2 S(f)$ is the second differential of the functional $S(f)$. The quadratic functional $D_f^2 S(f)(\varphi, \varphi)$ on $E$ is defined by the Taylor formula,

$$S(f + \varphi) = S(f) + D_f S(f)(\varphi) + \frac{1}{2} D_f^2 S(f)(\varphi, \varphi) + o(\|\varphi\|^2). \tag{13}$$

We keep the same notation for the corresponding symmetric bilinear form, $D_f^2 S(f)(\varphi, \psi)$, and also for the linear operator, $D_f^2 S(f) : E \to E^T$, defined by the formula $(D_f^2 S(f)\varphi)(\psi) = D_f^2 S(f)(\varphi, \psi)$. In this formula, on the left hand side $D_f^2 S(f)$ is the operator, on the right hand side it is the bilinear form. Operator $D_f^2 S(f)$ is symmetric on $E$, $D_f^2 S(f)^T = D_f^2 S(f)$.

In finite dimensions the functional $D_f S(f)$ can be presented simply as a row vector of partial derivatives of $S$, and the operator $D_f^2 S(f)$ is a matrix of second partial derivatives. For infinite–dimensional spaces some complications exist because $S(f)$ is defined only for classical densities and not for all distributions. In this paper we do not pay attention to these details.

We assume strict concavity of $S$, $D_f^2 S(f)(\varphi, \varphi) < 0$ if $\varphi \neq 0$. This means that for any $f$ the positive definite quadratic form $-D_f^2 S(f)(\varphi, \varphi)$ defines a scalar product

$$\langle \varphi, \psi \rangle_f = -(D_f^2 S)(\varphi, \psi). \tag{14}$$

This *entropic scalar product* is an important part of thermodynamic formalism. For the BGS entropy (5) as well as for the Kullback entropy (9)

$$\langle \varphi, \psi \rangle_f = \int \frac{\varphi(x)\psi(x)}{f(x)} \, \mathrm{d}x. \tag{15}$$

The most important assumption about kinetic equation (12) is: entropy does not decrease in time:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = (D_f S(f))(J(f)) \geq 0. \tag{16}$$

A particular case of this assumption is: the system (12) is conservative and entropy is constant. The main example of such conservative equations is the Liouville equation with linear vector field $J(f) = -Lf = \{H, f\}$, where $\{H, f\}$ is the Poisson bracket with Hamiltonian $H$.

For the following consideration of the Ehrenfests' coarse-graining the underlying mechanical motion is not crucial, and it is possible to start from the formal kinetic equation (12) without any mechanical interpretation of vectors $f$. We develop below the coarse-graining procedure for general kinetic

equation (12) with non-decreasing entropy (16). After coarse-graining the entropy production increases: conservative systems become dissipative ones, and dissipative systems become "more dissipative."

### 2.2 Conditional Equilibrium Instead of Averaging in Cells

*Microdescription, macrodescription and quasi-equilibrium state*

Averaging in cells is a particular case of entropy maximization. Let the phase space be divided into cells. For the $i$th cell the population $M_i$ is

$$M_i = m_i(f) = \int_{\text{cell}_i} f(x)\,\mathrm{d}\Gamma(x).$$

The averaging in cells for a given vector of populations $M = (M_i)$ produces the solution of the optimization problem for the BGS entropy:

$$S(f) \to \max,\ m(f) = M, \tag{17}$$

where $m(f)$ is vector $(m_i(f))$. The maximizer is a function $f^*_M(x)$ defined by the vector of averages $M$.

This operation has a well-known generalization. In the more general statement, vector $f$ is a microscopic description of the system, vector $M$ gives a macroscopic description, and a linear operator $m$ transforms a microscopic description into a macroscopic one: $M = m(f)$. The standard example is the transformation of the microscopic density into the hydrodynamic fields (density–velocity–kinetic temperature) with local Maxwellian distributions as entropy maximizers (see, for example, [4]).

For any macroscopic description $M$, let us define the correspondent $f^*_M$ as a solution to optimization problem (17) with an appropriate entropy functional $S(f)$ (Fig. 3). This $f^*_M$ has many names in the literature: MaxEnt distribution, reference distribution (reference of the macroscopic description to the microscopic one), generalized canonical ensemble, conditional equilibrium, or *quasi-equilibrium*. We shall use here the last term.

The quasi-equilibrium distribution $f^*_M$ satisfies the obvious, but important identity of self-consistency:

$$m(f^*_M) = M, \tag{18}$$

or in differential form

$$m(D_M f^*_M) = 1,\ \text{i.e. } m((D_M f^*_M)a) \equiv a. \tag{19}$$

The last identity means that the infinitesimal change in $M$ calculated through differential of the quasi-equilibrium distribution $f^*_M$ is simply the infinitesimal change in $M$. For the second differential we obtain

$$m(D^2_M f^*_M) = 0,\ \text{i.e. } m((D^2_M f^*_M)(a,b)) \equiv 0. \tag{20}$$

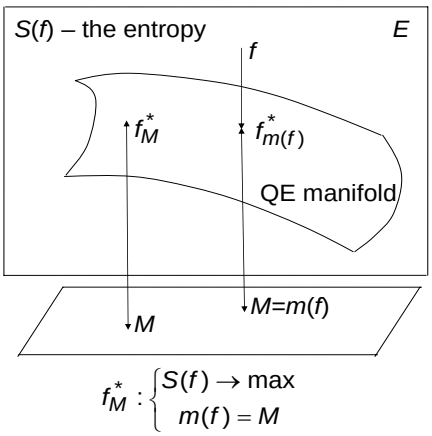



Fig. 3: Relations between a microscopic state $f$, a corresponding macroscopic state $M = m(f)$, and a quasi-equilibrium state $f^*_M$.

Following [4] let us mention that most of the works on nonequilibrium thermodynamics deal with quasi-equilibrium approximations and corrections to them, or with applications of these approximations (with or without corrections). This viewpoint is not the only possible but it proves very efficient for the construction of a variety of useful models, approximations and equations, as well as methods to solve them.

From time to time it is discussed in the literature, who was the first to introduce the quasi-equilibrium approximations, and how to interpret them. At least a part of the discussion is due to a different role the quasi-equilibrium plays in the entropy-conserving and in the dissipative dynamics. The very first use of the entropy maximization dates back to the classical work of G. W. Gibbs [47], but it was first claimed for a principle of informational statistical thermodynamics by E. T. Jaynes [48]. Probably, the first explicit and systematic use of quasiequilibria on the way from entropy-conserving dynamics to dissipative kinetics was undertaken by D. N. Zubarev. Recent detailed exposition of his approach is given in [49].

For dissipative systems, the use of the quasi-equilibrium to reduce description can be traced to the works of H. Grad on the Boltzmann equation [50]. A review of the informational statistical thermodynamics was presented in [51]. The connection between entropy maximization and (nonlinear) Onsager relations was also studied [52, 53]. Our viewpoint was influenced by the papers by L. I. Rozonoer and co-workers, in particular, [54, 55, 56]. A detailed exposition of the quasi-equilibrium approximation for Markov chains is given in the book [34] (Chap. 3, *Quasi-equilibrium and entropy maximum*, pp. 92-122), and for the BBGKY hierarchy in the paper [57].

The maximum entropy principle was applied to the description of the universal dependence of the three-particle distribution function $F_3$ on the two-particle distribution function $F_2$ in classical systems with binary interactions [58]. For a discussion of the quasi-equilibrium moment closure hierarchies for the Boltzmann equation [55] see the papers [59, 60, 61]. A very general discussion of the maximum entropy principle with applications to dissipative kinetics is given in the review [62]. Recently, the quasi-equilibrium approximation with some further correction was applied to the description of rheology of polymer solutions [64, 65] and of ferrofluids [66, 67]. Quasi-equilibrium approximations for quantum systems in the Wigner representation [70, 71] was discussed very recently [63].

We shall now introduce the quasi-equilibrium approximation in the most general setting. The coarse-graining procedure will be developed after that as a method for enhancement of the quasi-equilibrium approximation [5].

*Quasi-equilibrium manifold, projector and approximation*

A *quasi-equilibrium manifold* is a set of quasi-equilibrium states $f^*_M$ parameterized by macroscopic variables $M$. For microscopic states $f$ the correspondent quasi-equilibrium states are defined as $f^*_{m(f)}$. Relations between $f$, $M$, $f^*_M$, and $f^*_{m(f)}$ are presented in Fig. 3.

A *quasi-equilibrium approximation* for the kinetic equation (12) is an equation for $M(t)$:

$$\frac{\mathrm{d}M}{\mathrm{d}t} = m(J(f^*_M)). \tag{21}$$

To define $\dot{M}$ in the quasi-equilibrium approximation for given $M$, we find the correspondent quasi-equilibrium state $f^*_M$ and the time derivative of $f$ in this state $J(f^*_M)$, and then return to the macroscopic variables by the operator $m$. If $M(t)$ satisfies (21) then $f^*_{M(t)}$ satisfies the following equation

$$\frac{\mathrm{d}f^*_M}{\mathrm{d}t} = (D_M f^*_M)\left(\frac{\mathrm{d}M}{\mathrm{d}t}\right) = (D_M f^*_M)(m(J(f^*_M))). \tag{22}$$

The right hand side of (22) is the projection of vector field $J(f)$ onto the tangent space of the quasi-equilibrium manifold at the point $f = f^*_M$. After calculating the differential $D_M f^*_M$ from the definition of quasi-equilibrium (17), we obtain $\mathrm{d}f^*_M/\mathrm{d}t = \pi_{f^*_M} J(f^*_M)$, where $\pi_{f^*_M}$ is the *quasi-equilibrium projector*:

$$\pi_{f^*_M} = (D_M f^*_M)\, m = \left(D^2_f S\right)^{-1}_{f^*_M} m^T \left(m \left(D^2_f S\right)^{-1}_{f^*_M} m^T\right)^{-1} m. \tag{23}$$

It is straightforward to check the equality $\pi^2_{f^*_M} = \pi_{f^*_M}$, and the self-adjointness of $\pi_{f^*_M}$ with respect to entropic scalar product (14). In this scalar product, the quasi-equilibrium projector is the orthogonal projector onto the tangent space to the quasi-equilibrium manifold. The quasi-equilibrium projector for

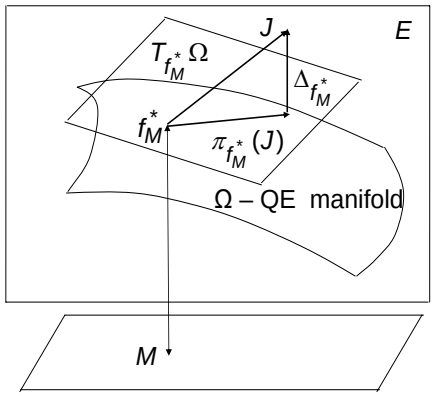



Fig. 4: Quasi-equilibrium manifold $\Omega$, tangent space $T_{f_M^*}\Omega$, quasi-equilibrium projector $\pi_{f_M^*}$, and defect of invariance, $\Delta_{f_M^*} = J - \pi_{f_M^*}(J)$.

a quasi-equilibrium approximation was first constructed by B. Robertson [68].

Thus, we have introduced the basic constructions: quasi-equilibrium manifold, entropic scalar product, and quasi-equilibrium projector (Fig. 4).

*Preservation of dissipation*

For the quasi-equilibrium approximation the entropy is $S(M) = S(f_M^*)$. For this entropy,

$$\frac{\mathrm{d}S(M)}{\mathrm{d}t} = \left(\frac{\mathrm{d}S(f)}{\mathrm{d}t}\right)_{f=f_M^*}, \tag{24}$$

Here, on the left hand side stands the macroscopic entropy production for the quasi-equilibrium approximation (21), and the right hand side is the microscopic entropy production calculated for the initial kinetic equation (12). This equality implies *preservation of the type of dynamics* [34, 35]:

- If for the initial kinetics (12) the dissipativity inequality (16) holds then the same inequality is true for the quasi-equilibrium approximation (21);
- If the initial kinetics (12) is conservative then the quasi-equilibrium approximation (21) is conservative also.

For example, let the initial kinetic equation be the Liouville equation for a system of many identical particles with binary interaction. If we choose as macroscopic variables the one-particle distribution function, then the quasi-equilibrium approximation is the Vlasov equation. If we choose as macroscopic variables the hydrodynamic fields, then the quasi-equilibrium approximation is the compressible Euler equation with self–interaction of liquid. Both of these equations are conservative and turn out to be even Hamiltonian systems [69].

*Measurement of accuracy*

Accuracy of the quasi-equilibrium approximation near a given $M$ can be measured by the *defect of invariance* (Fig. 4):

$$\Delta_{f_M^*} = J(f_M^*) - \pi_{f_M^*} J(f_M^*). \tag{25}$$

A dimensionless criterion of accuracy is the ratio $\|\Delta_{f_M^*}\|/\|J(f_M^*)\|$ (a "sine" of the angle between $J$ and tangent space). If $\Delta_{f_M^*} \equiv 0$ then the quasi-equilibrium manifold is an invariant manifold, and the quasi-equilibrium approximation is exact. In applications, the quasi-equilibrium approximation is usually not exact.

*The Gibbs entropy and the Boltzmann entropy*

For analysis of micro-macro relations some authors [77, 78] call entropy $S(f)$ the *Gibbs entropy*, and introduce a notion of the *Boltzmann entropy*. Boltzmann defined the entropy of a macroscopic system in a macrostate $M$ as the log of the volume of phase space (number of microstates) corresponding to $M$. In the proposed level of generality [34, 35], the Boltzmann entropy of the state $f$ can be defined as $S_{\mathrm{B}}(f) = S(f^*_{m(f)})$. It is entropy of the projection of $f$ onto quasi-equilibrium manifold (the "shadow" entropy). For conservative systems the Gibbs entropy is constant, but the Boltzmann entropy increases [35] (during some time, at least) for motions that start on the quasi-equilibrium manifold, but not belong to this manifold.

These notions of the Gibbs or Boltzmann entropy are related to micro-macro transition and may be applied to any convex entropy functional, not the BGS entropy (5) only. This may cause some terminological problems (we hope, not here), and it may be better just to call $S(f^*_{m(f)})$ the *macroscopic entropy*.

*Invariance equation and the Chapman–Enskog expansion*

The first method for improvement of the quasi-equilibrium approximation was the Chapman–Enskog method for the Boltzmann equation [79]. It uses the explicit structure of singularly perturbed systems. Many other methods were invented later, and not all of them use this explicit structure (see, for example review in [4]). Here we develop the Chapman–Enskog method for one important class of *model equations* that were invented to substitute the Boltzmann equation and other more complicated systems when we don't know the details of microscopic kinetics. It includes the well-known Bhatnagar–Gross–Krook (BGK) kinetic equation [38] , as well as wide class of generalized model equations [39].

As a starting point we take a formal kinetic equation with a small parameter $\epsilon$

$$\frac{\mathrm{d}f}{\mathrm{d}t} = J(f) = F(f) + \frac{1}{\epsilon}(f^*_{m(f)} - f). \tag{26}$$

The term $(f^*_{m(f)} - f)$ is non-linear because nonlinear dependency $f^*_{m(f)}$ on $m(f)$.

We would like to find a reduced description valid for macroscopic variables $M$. It means, at least, that we are looking for an invariant manifold parameterized by $M$, $f = f_M$, that satisfies the *invariance equation*:

$$(D_M f_M)(m(J(f_M))) = J(f_M). \tag{27}$$

The invariance equation means that the time derivative of $f$ calculated through the time derivative of $M$ ($\dot{M} = m(J(f_M))$) by the chain rule coincides with the true time derivative $J(f)$. This is the central equation for the model reduction theory and applications. First general results about existence and regularity of solutions to that equation were obtained by Lyapunov [83] (see review in [3, 4]). For kinetic equation (26) the invariance equation has a form

$$(D_M f_M)(m(F(f_M))) = F(f_M) + \frac{1}{\epsilon}(f^*_M - f_M), \tag{28}$$

because the self-consistency identity (18).

Due to presence of small parameter $\epsilon$ in $J(f)$, the zero approximation is obviously the quasi-equilibrium approximation: $f^{(0)}_M = f^*_M$. Let us look for $f_M$ in the form of power series: $f_M = f^{(0)}_M + \epsilon f^{(1)}_M + \ldots$; $m(f^{(k)}_M) = 0$ for $k \geq 1$. From (28) we immediately find:

$$f^{(1)}_M = F(f^{(0)}_M) - (D_M f^{(0)}_M)(m(F(f^{(0)}_M))) = \Delta_{f^*_M}. \tag{29}$$

It is very natural that the first term of the Chapman–Enskog expansion for model equations (26) is just the defect of invariance for the quasi-equilibrium approximation. Calculation of the following terms is also straightforward.

The correspondent first–order in $\epsilon$ approximation for the macroscopic equations is:

$$\frac{\mathrm{d}M}{\mathrm{d}t} = m(F(f^*_M)) + \epsilon m((D_f F(f))_{f^*_M} \Delta_{f^*_M}). \tag{30}$$

We should remind that $m(\Delta_{f^*_M}) = 0$. The last term in (28) vanishes in macroscopic projection for all orders.

The typical situation for the model equations (26) is: the vector field $F(f)$ is conservative, $(D_f S(f))F(f) = 0$. In that case, the first term $m(F(f^*_M))$ also conserves the correspondent Boltzmann (i.e. macroscopic, but not obligatory BGS) entropy $S(f^*_M)$. But the straightforward calculation of the Boltzmann entropy $S(f^*_M)$ production for the first-order Chapman–Enskog term in equation (30) gives us for conservative $F(f)$:

$$\frac{\mathrm{d}S(M)}{\mathrm{d}t} = \epsilon \langle \Delta_{f^*_M}, \Delta_{f^*_M} \rangle_{f^*_M} \geq 0. \tag{31}$$

where $\langle \bullet, \bullet \rangle_f$ is the entropic scalar product (14). The Boltzmann entropy production in the first Chapman–Enskog approximation is zero if and only if $\Delta_{f^*_M} = 0$, i.e. if at point $M$ the quasi-equilibrium manifold is locally invariant.

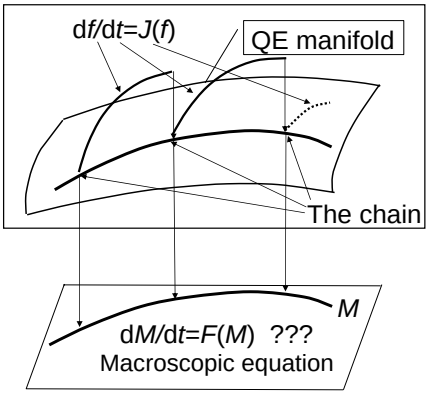



Fig. 5: The Ehrenfests' chain.

To prove (31) we differentiate the conservativity identity:

$$\begin{aligned}&(D_f S(f))F(f) \equiv 0 \\ &(D_f^2 S(F))(F(f), a) + (D_f S(f))((D_f F(f))a) \equiv 0 \\ &(D_f S(f))((D_f F(f))a) \equiv \langle F(f), a\rangle_f,\end{aligned} \tag{32}$$

use the last equality in the expression of the entropy production, and take into account that the quasi-equilibrium projector is orthogonal, hence

$$\langle F(f_M^*), \Delta_{f_M^*}\rangle_{f_M^*} = \langle \Delta_{f_M^*}, \Delta_{f_M^*}\rangle_{f_M^*}.$$

Below we apply the Chapman–Enskog method to the analysis of filtered BGK equation.

### 2.3 The Ehrenfests' Chain, Macroscopic Equations and Entropy production

*The Ehrenfests' Chain and entropy growth*

Let $\Theta_t$ be the time shift transformation for the initial kinetic equation (12):

$$\Theta_t(f(0)) = f(t).$$

The Ehrenfests' chain (Fig. 5) is defined for a given macroscopic variables $M = m(f)$ and a fixed time of coarse-graining $\tau$. It is a chain of quasi-equilibrium states $f_0, f_1, \ldots$:

$$f_{i+1} = f^*_{m(\Theta_\tau(f_i))}. \tag{33}$$

To get the next point of the chain, $f_{i+1}$, we take $f_i$, move it by the time shift $\Theta_\tau$, calculate the corresponding macroscopic state $M_{i+1} = m(\Theta_\tau(f_i))$, and find the quasi-equilibrium state $f^*_{M_{i+1}} = f_{i+1}$.

If the point $\Theta_\tau(f_i)$ is not a quasi-equilibrium state, then $S(\Theta_\tau(f_i)) < S(f^*_{m(\Theta_\tau(f_i))})$ because of quasi-equilibrium definition (17) and strict concavity of entropy. Hence, if the motion between $f_i$ and $\Theta_\tau(f_i)$ does not belong to the quasi-equilibrium manifold, then $S(f_{i+1}) > S(f_i)$, entropy in the Ehrenfests' chain grows. The entropy gain consists of two parts: the gain in the motion (from $f_i$ to $\Theta_\tau(f_i)$), and the gain in the projection (from $\Theta_\tau(f_i)$ to $f_{i+1} = f^*_{m(\Theta_\tau(f_i))}$). Both parts are non-negative. For conservative systems the first part is zero. The second part is strictly positive if the motion leaves the quasi-equilibrium manifold. Hence, we observe some sort of duality between entropy production in the Ehrenfests' chain and invariance of the quasi-equilibrium manifold. The motions that build the Ehrenfests' chain restart periodically from the quasi-equilibrium manifold and the entropy growth along this chain is similar to the Boltzmann entropy growth in the Chapman–Enskog approximation, and that similarity is very deep, as the exact formulas show below.

*The natural projector and macroscopic dynamics*

How to use the Ehrenfests' chains? First of all, we can try to define the *macroscopic kinetic equations* for $M(t)$ by the requirement that for any initial point of the chain $f_0$ the solution of these macroscopic equations with initial conditions $M(0) = m(f_0)$ goes through all the points $m(f_i)$: $M(n\tau) = m(f_n)$ $(n = 1, 2, \ldots)$ (Fig. 5) [5] (see also [4]). Another way is an "equation–free approach" [9] to the direct computation of the Ehrenfests' chain with a combination of microscopic simulation and macroscopic stepping.

For the definition of the macroscopic equations only the first link of the Ehrenfests' chain is necessary. In general form, for an ansatz manifold $\Omega$, projector $\pi : U \to \Omega$ of the vicinity of $\Omega$ onto $\Omega$, phase flow of the initial kinetic equation $\Theta_t$, and macroscopic phase flow $\tilde{\Theta}_t$ on $\Omega$ the matching condition is (Fig. 6):

$$\pi(\Theta_\tau(f)) = \tilde{\Theta}_\tau(f) \text{ for any } f \in \Omega. \tag{34}$$

We call this projector of the flow $\Theta$ onto an ansatz manifold $\Omega$ by fragments of trajectories of given duration $\tau$ the *natural projector* in order to distinguish it from the standard infinitesimal projector of vector fields on tangent spaces.

Let us look for the macroscopic equations of the form

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \Psi(M) \tag{35}$$

with the phase flow $\Phi_t$: $M(t) = \Phi_t M(0)$. For the quasi-equilibrium manifold and projector the matching condition (34) gives

$$m(\Theta_\tau(f^*_M)) = \Phi_\tau(M) \text{ for any macroscopic state } M. \tag{36}$$

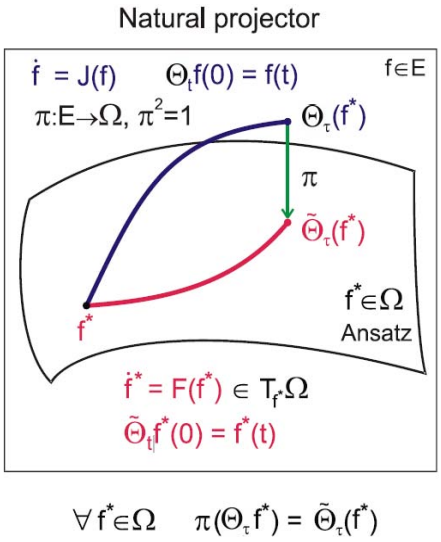



Fig. 6: Projection of segments of trajectories: The microscopic motion above the manifold $\Omega$ and the macroscopic motion on this manifold. If these motions begin in the same point on $\Omega$, then, after time $\tau$, projection of the microscopic state onto $\Omega$ should coincide with the result of the macroscopic motion on $\Omega$. For quasi-equilibrium $\Omega$, projector $\pi : E \to \Omega$ acts as $\pi(f) = f^*_{m(f)}$.

This condition is the equation for the macroscopic vector field $\Psi(M)$. The solution of this equation is a function of $\tau$: $\Psi = \Psi(M, \tau)$. For sufficiently smooth microscopic vector field $J(f)$ and entropy $S(f)$ it is easy to find the Taylor expansion of $\Psi(M, \tau)$ in powers of $\tau$. It is a straightforward exercise in differential calculus. Let us find the first two terms: $\Psi(M, \tau) = \Psi_0(M) + \tau\Psi_1(M) + o(\tau)$. Up to the second order in $\tau$ the matching condition (36) is

$$\begin{aligned} &m(J(f^*_M))\tau + m((D_f J(f))_{f=f^*_M}(J(f^*_M)))\frac{\tau^2}{2} \\ &= \Psi_0(M)\tau + \Psi_1(M)\tau^2 + (D_M\Psi_0(M))(\Psi_0(M))\frac{\tau^2}{2}. \end{aligned} \tag{37}$$

From this condition immediately follows:

$$\begin{aligned} \Psi_0(M) &= m(J(f^*_M)); \\ \Psi_1(M) &= \frac{1}{2}m[(D_f J(f))_{f=f^*_M}(J(f^*_M)) - (D_M J(f^*_M))(m(J(f^*_M)))] \\ &= m((D_f J(f))_{f=f^*_M}\Delta_{f^*_M}) \end{aligned} \tag{38}$$

where $\Delta_{f^*_M}$ is the defect of invariance (25). The macroscopic equation in the first approximation is:

$$\frac{\mathrm{d}M}{\mathrm{d}t} = m(J(f_M^*)) + \frac{\tau}{2} m((D_f J(f))_{f=f_M^*} \Delta_{f_M^*}). \tag{39}$$

It is exactly the first Chapman–Enskog approximation (30) for the model kinetics (26) with $\varepsilon = \tau/2$. The first term $m(J(f_M^*))$ gives the quasi-equilibrium approximation, the second term increases dissipation. The formula for entropy production follows from (39) [11]. If the initial microscopic kinetic (12) is conservative, then for macroscopic equation (39) we obtain as for the Chapman–Enskog approximation:

$$\frac{\mathrm{d}S(M)}{\mathrm{d}t} = \frac{\tau}{2} \langle \Delta_{f_M^*}, \Delta_{f_M^*} \rangle_{f_M^*}, \tag{40}$$

where $\langle \bullet, \bullet \rangle_f$ is the entropic scalar product (14). From this formula we see again a duality between the invariance of the quasi-equilibrium manifold and the dissipativity: entropy production is proportional to the square of the defect of invariance of the quasi-equilibrium manifold.

For linear microscopic equations ($J(f) = Lf$) the form of the macroscopic equations is

$$\frac{\mathrm{d}M}{\mathrm{d}t} = mL \left[ 1 + \frac{\tau}{2} (1 - \pi_{f_M^*}) L \right] f_M^*, \tag{41}$$

where $\pi_{f_M^*}$ is the quasi-equilibrium projector (23).

*The Navier–Stokes equation from the free flight dynamics*

The free flight equation describes dynamics of one-particle distribution function $f(\boldsymbol{x}, \boldsymbol{v})$ due to free flight:

$$\frac{\partial f(\boldsymbol{x}, \boldsymbol{v}, t)}{\partial t} = -\sum_i v_i \frac{\partial f(\boldsymbol{x}, \boldsymbol{v}, t)}{\partial x_i}. \tag{42}$$

The difference from the continuity equation (2) is that there is no velocity field $\boldsymbol{v}(\boldsymbol{x})$, but the velocity vector $\boldsymbol{v}$ is an independent variable. Equation (42) is conservative and has an explicit general solution

$$f(\boldsymbol{x}, \boldsymbol{v}, t) = f_0(\boldsymbol{x} - \boldsymbol{v}t, \boldsymbol{v}). \tag{43}$$

The coarse-graining procedure for (42) serves for modeling kinetics with an unknown dissipative term $I(f)$

$$\frac{\partial f(\boldsymbol{x}, \boldsymbol{v}, t)}{\partial t} = -\sum_i v_i \frac{\partial f(\boldsymbol{x}, \boldsymbol{v}, t)}{\partial x_i} + I(f). \tag{44}$$

The Ehrenfests' chain realizes a splitting method for (44): first, the free flight step during time $\tau$, than the complete relaxation to a quasi-equilibrium distribution due to dissipative term $I(f)$, then again the free flight, and so on. In this approximation the specific form of $I(f)$ is not in use, and the only

parameter is time $\tau$. It is important that this hypothetical $I(f)$ preserves all the standard conservation laws (number of particles, momentum, and energy) and has no additional conservation laws: everything else relaxes. Following this assumption, the macroscopic variables are: $M_0 = n(\boldsymbol{x}, t) = \int f \mathrm{d}\boldsymbol{v}$, $M_i = nu_i = \int v_i f \mathrm{d}\boldsymbol{v}$ $(i = 1, 2, 3)$, $M_4 = \frac{3nk_\mathrm{B}T}{m} + nu^2 = \int v^2 f \mathrm{d}\boldsymbol{v}$. The zero-order (quasi-equilibrium) approximation (21) gives the classical Euler equation for compressible non-isothermal gas. In the first approximation (39) we obtain the Navier–Stokes equations:

$$\begin{aligned}
\frac{\partial n}{\partial t} &= -\sum_i \frac{\partial (nu_i)}{\partial x_i}, \\
\frac{\partial (nu_k)}{\partial t} &= -\sum_i \frac{\partial (nu_k u_i)}{\partial x_i} - \frac{1}{m}\frac{\partial P}{\partial x_k} \\
&\quad + \underline{\frac{\tau}{2}\frac{1}{m}\sum_i \frac{\partial}{\partial x_i}\left[P\left(\frac{\partial u_k}{\partial x_i} + \frac{\partial u_i}{\partial x_k} - \frac{2}{3}\delta_{ki}\mathrm{div}u\right)\right]}, \\
\frac{\partial \mathcal{E}}{\partial t} &= -\sum_i \frac{\partial (\mathcal{E}u_i)}{\partial x_i} - \frac{1}{m}\sum_i \frac{\partial (Pu_i)}{\partial x_i} + \underline{\frac{\tau}{2}\frac{5k_\mathrm{B}}{2m^2}\sum_i \frac{\partial}{\partial x_i}\left(P\frac{\partial T}{\partial x_i}\right)},
\end{aligned} \tag{45}$$

where $P = nk_\mathrm{B}T$ is the ideal gas pressure, $\mathcal{E} = \frac{1}{2}\int v^2 f \,\mathrm{d}\boldsymbol{v} = \frac{3nk_\mathrm{B}T}{2m} + \frac{n}{2}u^2$ is the energy density per unite mass ($P = \frac{2m}{3}\mathcal{E} - \frac{mn}{3}u^2$, $T = \frac{2m}{3nk_\mathrm{B}}\mathcal{E} - \frac{m}{3k_\mathrm{B}}u^2$), and the underlined terms are results of the coarse-graining additional to the quasi-equilibrium approximation.

The dynamic viscosity in (45) is $\mu = \frac{\tau}{2}nk_\mathrm{B}T$. It is useful to compare this formula to the mean–free–path theory that gives $\mu = \tau_\mathrm{col}nk_\mathrm{B}T = \tau_\mathrm{col}P$, where $\tau_\mathrm{col}$ is the collision time (the time for the mean–free–path). According to these formulas, we get the following interpretation of the coarse-graining time $\tau$ for this example: $\tau = 2\tau_\mathrm{col}$.

The equations obtained (45) coincide with the first–order terms of the Chapman–Enskog expansion (30) applied to the BGK equations with $\tau_\mathrm{col} = \tau/2$ and meet the same problem: the Prandtl number (i.e., the dimensionless ratio of viscosity and thermal conductivity) is $\mathrm{Pr} = 1$ instead of the value $\mathrm{Pr} = \frac{2}{3}$ verified by experiments with perfect gases and by more detailed theory [80] (recent discussion of this problem for the BGK equation with some ways for its solution is presented in [81]).

In the next order in $\tau$ we obtain the stable post–Navier–Stokes equations instead of the unstable Burnett equations that appear in the Chapman–Enskog expansion [11, 76]. Here we can see the difference between two approaches.

*Persistence of invariance and mistake of differential pursuit*

L.M. Lewis called a generalization of the Ehrenfest's approach a "unifying principle in statistical mechanics," but he created other macroscopic equations: he produced the differential pursuit (Fig. 7a)

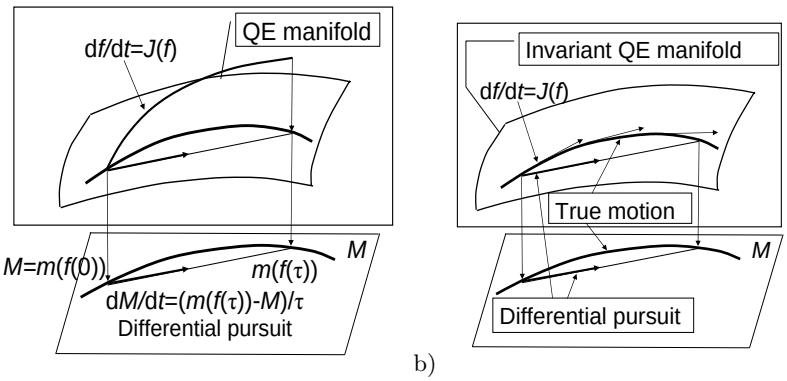



Fig. 7: Differential pursuit of the projection point (a). The mistake of differential pursuit (b): invariant manifold should preserve its invariance, but it does not!

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \frac{m(\Theta_\tau(f^*_M)) - M}{\tau} \tag{46}$$

from the full matching condition (34). This means that the macroscopic motion was taken in the first-order Tailor approximation, while for the microscopic motion the complete shift in time (without the Taylor expansion) was used. The basic idea of this approach is a non-differential time separation: the infinitesimal shift in macroscopic time is always such a significant shift for microscopic time that no Taylor approximation for microscopic motion may be in use. This sort of non-standard analysis deserves serious attention, but its realization in the form of the differential pursuit (46) does not work properly in many cases. If the quasi-equilibrium manifold is invariant, then the quasi-equilibrium approximation is exact and the Ehrenfests' chain (Fig. 5) just follows the quasi-equilibrium trajectory. But the differential pursuit does not follow the trajectory (Fig. 7b); this motion leaves the invariant quasi-equilibrium manifolds, and the differential pursuit does not approximate the Ehrenfests' chain, even qualitatively.

*Ehrenfests' coarse-graining as a method for model reduction*

The problem of model reduction in dissipative kinetics is recognized as a problem of time separation and construction of slow invariant manifolds. One obstacle on this way is that the slow invariant manifold is the thing that many people would like to find, but nobody knows exactly what it is. There is no conventional definition of *slow* invariant manifold without explicit small parameter that tends to zero. It seems now that the most reasonable way for such a definition is the analysis of induced dynamics of manifolds immersed into phase space. Fixed points of this dynamics are invariant manifolds, and asymptotically stable (stable and attracting) fixed points are slow invariant

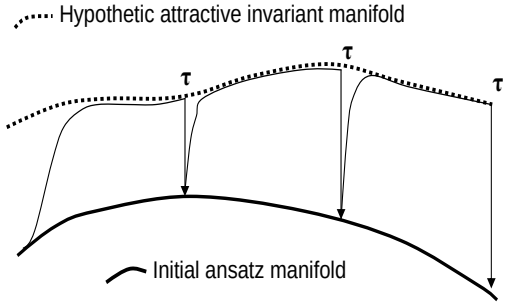



Fig. 8: Natural projector and attractive invariant manifolds. For large $\tau$, the natural projector gives the approximation of projection of the genuine motion from the attractive invariant manifold onto the initial ansatz manifold $\Omega$.

manifolds. This concept was explicitly developed very recently [84, 3, 4], but the basic idea was used in earlier applied works [35, 85].

The coarse-graining procedure was developed for *erasing* some details of the dynamics in order to provide entropy growth and uniform tendency to equilibrium. In this sense, the coarse-graining is opposite to the model reduction, because for the model reduction we try to find slow invariant manifolds as exactly, as we can. But unexpectedly the coarse-graining becomes a tool for model reduction without any "erasing."

Let us assume that for dissipative dynamics with entropy growth there exists an attractive invariant manifold. Let us apply the Ehrenfests' coarse-graining to this system for sufficiently large coarse-graining time $\tau$. For the most part of time $\tau$ the system will spend in a small vicinity of the attractive invariant manifold. Hence, the macroscopic projection will describe the projection of dynamics from the attractive invariant manifold onto ansatz manifold $\Omega$. As a result, we shall find a shadow of the proper slow dynamics without looking for the slow invariant manifold. Of course, the results obtained by the Taylor expansion (37–39) are not applicable for the case of large coarse-graining time $\tau$, at least, directly. Some attempts to utilize the idea of large $\tau$ asymptotic are presented in [4] (Ch. 12).

One can find a source of this idea in the first work of D. Hilbert about the Boltzmann equation solution [40] (a recent exposition and development of the Hilbert method is presented in [86] with many examples of applications). In the Hilbert method, we start from the local Maxwellian manifold (that is, quasi-equilibrium one) and iteratively look for "normal solutions." The normal solutions $f_{\mathrm{H}}(\boldsymbol{v}, n(\boldsymbol{x},t), \boldsymbol{u}(\boldsymbol{x},t), T(\boldsymbol{x},t))$ are solutions to the Boltzmann equation that depend on space and time only through five hydrodynamic fields. In the Hilbert method no final macroscopic equation arises. The next attempt to utilize this idea without macroscopic equations is the "equation free" approach [9, 87].

The Ehrenfests' coarse-graining as a tool for extraction of exact macroscopic dynamics was tested on exactly solvable problems [73]. It gives also a new approach to the fluctuation–dissipation theorems [72].

### 2.4 Kinetic Models, Entropic Involution, and the Second–Order "Euler Method"

*Time-step – dissipation decoupling problem*

Sometimes, the kinetic equation is much simpler than the coarse-grained dynamics. For example, the free flight kinetics (42) has the obvious exact analytical solution (43), but the Euler or the Navier–Stokes equations (45) seem to be very far from being exactly solvable. In this sense, the Ehrenfests' chain (33) (Fig. 5) gives a stepwise approximation to a solution of the coarse-grained (macroscopic) equations by the chain of solutions of the kinetic equations.

If we use the second-order approximation in the coarse-graining procedure (37), then the Ehrenfests' chain with step $\tau$ is the second–order (in time step $\tau$) approximation to the solution of macroscopic equation (39). It is very attractive for hydrodynamics: the second–order in time method with approximation just by broken line built from intervals of simple free–flight solutions. But if we use the Ehrenfests' chain for approximate solution, then the strong connection between the time step $\tau$ and the coefficient in equations (39) (see also the entropy production formula (40)) is strange. Rate of dissipation is proportional to $\tau$, and it seems to be too restrictive for computational applications: decoupling of time step and dissipation rate is necessary. This decoupling problem leads us to a question that is strange from the Ehrenfests' coarse-graining point of view: *how to construct an analogue to the Ehrenfests' coarse-graining chain, but without dissipation?* The *entropic involution* is a tool for this construction.

*Entropic involution*

The entropic involution was invented for improvement of the lattice–Boltzmann method [89]. We need to construct a chain with zero macroscopic entropy production and second order of accuracy in time step $\tau$. The chain consists of intervals of solution of kinetic equation (12) that is conservative. The time shift for this equation is $\Theta_t$. The macroscopic variables $M = m(f)$ are chosen, and the time shift for corresponding quasi-equilibrium equation is (in this section) $\widetilde{\Theta}_t$. The standard example is: the free flight kinetics (42,43) as a microscopic conservative kinetics, hydrodynamic fields (density–velocity–kinetic temperature) as macroscopic variables, and the Euler equations as a macroscopic quasi-equilibrium equations for conservative case (see (45), not underlined terms).

Let us start from construction of one link of a chain and take a point $f_{1/2}$ on the quasi-equilibrium manifold. (It is not an initial point of the link,

$f_0$, but a "middle" one.) The correspondent value of $M$ is $M_{1/2} = m(f_{1/2})$. Let us define $M_0 = m(\Theta_{-\tau/2}(f_{1/2}))$, $M_1 = m(\Theta_{\tau/2}(f_{1/2}))$. The dissipative term in macroscopic equations (39) is linear in $\tau$, hence, there is a symmetry between forward and backward motion from any quasiequilibrium initial condition with the second-order accuracy in the time of this motion (it became clear long ago [35]). Dissipative terms in the shift from $M_0$ to $M_{1/2}$ (that decrease macroscopic entropy $S(M)$) annihilate with dissipative terms in the shift from $M_{1/2}$ to $M_1$ (that increase macroscopic entropy $S(M)$). As the result of this symmetry, $M_1$ coincides with $\widetilde{\Theta}_\tau(M_0)$ with the second-order accuracy. (It is easy to check this statement by direct calculation too.)

It is necessary to stress that the second-order accuracy is achieved on the ends of the time interval only: $\widetilde{\Theta}_\tau(M_0)$ coincides with $M_1 = m(\Theta_\tau(f_0))$ in the second order in $\tau$

$$m(\Theta_\tau(f_0)) - \widetilde{\Theta}_\tau(M_0) = o(\tau^2).$$

On the way $\widetilde{\Theta}_t(M_0)$ from $M_0$ to $\widetilde{\Theta}_\tau(M_0)$ for $0 < t < \tau$ we can guarantee the first-order accuracy only (even for the middle point). It is essentially the same situation as we had for the Ehrenfests' chain: the second order accuracy of the matching condition (36) is postulated for the moment $\tau$, and for $0 < t < \tau$ the projection of the $m(\Theta_t(f_0))$ follows a solution of the macroscopic equation (39) with the first order accuracy only. In that sense, the method is quite different from the usual second–order methods with intermediate points, for example, from the Crank–Nicolson schemes. By the way, the middle quasi-equilibrium point, $f_{1/2}$ appears for the initiation step only. After that, we work with the end points of links.

The link is constructed. For the initiation step, we used the middle point $f_{1/2}$ on the quasi-equilibrium manifold. The end points of the link, $f_0 = \Theta_{-\tau/2}(f_{1/2})$ and $f_1 = \Theta_{\tau/2}(f_{1/2})$ don't belong to the quasi-equilibrium manifold, unless it is invariant. Where are they located? They belong a surface that we call a *film of non-equilibrium states* [74, 75, 4]. It is a trajectory of the quasi-equilibrium manifold due to initial microscopic kinetics. In [74, 75, 4] we studied mainly the positive semi-trajectory (for positive time). Here we need shifts in both directions.

A point $f$ on the film of non-equilibrium states is naturally parameterized by $M, \tau$: $f = q_{M,\tau}$, where $M = m(f)$ is the value of the macroscopic variables, and $\tau(f)$ is the time of shift from a quasi-equilibrium state: $\Theta_{-\tau}(f)$ is a quasi-equilibrium state. In the first order in $\tau$,

$$q_{M,\tau} = f^*_M + \tau \Delta_{f^*_M}, \tag{47}$$

and the first-order Chapman–Enskog approximation (29) for the model BGK equations is also here with $\tau = \epsilon$. (The two–times difference between kinetic coefficients for the Ehrenfests' chain and the first-order Chapman–Enskog approximation appears because for the Ehrenfests' chain the distribution walks linearly between $q_{M,0}$ and $q_{M,\tau}$, and for the first-order Chapman–Enskog approximation it is exactly $q_{M,\tau}$.)

For each $M$ and positive $s$ from some interval $]0, \varsigma[$ there exist two such $\tau_\pm(M,s)$ ($\tau_+(M,s) > 0$, $\tau_-(M,s) < 0$) that

$$S(q_{M,\tau_\pm(M,s)}) = S(M) - s. \tag{48}$$

Up to the second order in $\tau_\pm$

$$s = \frac{\tau_\pm^2}{2}\langle \Delta_{f_M^*}, \Delta_{f_M^*}\rangle_{f_M^*} + o(\tau_\pm^2) \tag{49}$$

(compare to (40)), and

$$\tau_+ = -\tau_- + o(\tau_-);\ |\tau_\pm| = \sqrt{\frac{s}{\langle \Delta_{f_M^*}, \Delta_{f_M^*}\rangle_{f_M^*}}}(1 + o(1)). \tag{50}$$

Equation (48) describes connection between entropy change $s$ and time coordinate $\tau$ on the film of non-equilibrium states, and (49) presents the first non-trivial term of the Taylor expansion of (48).

The *entropic involution* $I_S$ is the transformation of the film of non-equilibrium states:

$$I_S(q_{M,\tau_\pm}) = q_{M,\tau_\mp}. \tag{51}$$

This involution transforms $\tau_+$ into $\tau_-$, and back. For a given macroscopic state $M$, the entropic involution $I_S$ transforms the curve of non-equilibrium states $q_{M,\tau}$ into itself.

In the first order in $\tau$ it is just reflection $q_{M,\tau} \to q_{M,-\tau}$. A partial linearization is also in use. For this approximation, we define nonlinear involutions of straight lines parameterized by $\alpha$, not of curves:

$$I_S^0(f) = f^*_{m(f)} - \alpha(f - f^*_{m(f)}),\ \alpha > 0, \tag{52}$$

with condition of entropy conservation

$$S(I_S^0(f)) = S(f). \tag{53}$$

The last condition serves as equation for $\alpha$. The positive solution is unique and exists for $f$ from some vicinity of the quasi-equilibrium manifold. It follows from the strong concavity of entropy. The transformation $I_S^0$ (53) is defined not only on the film of non-equilibrium states, but on all distributions (microscopic) $f$ that are sufficiently closed to the quasi-equilibrium manifold.

In order to avoid the stepwise accumulation of errors in entropy production, we can choose a constant step in a conservative chain not in time, but in entropy. Let an initial point in macro-variables $M_0$ be given, and some $s > 0$ be fixed. We start from the point $f_0 = q_{M,\tau_-(M_0,s)}$. At this point, for $t = 0$, $S(m(\Theta_0(f_0)))) - s = S((\Theta_0(f_0)))$ ($\Theta_0 = \mathrm{id}$). Let the motion $\Theta_t(f_0)$ evolve until the equality $S(m(\Theta_t(f_0))) - s = S(\Theta_t(f_0))$ is satisfied next time. This time will be the time step $\tau$, and the next point of the chain is:

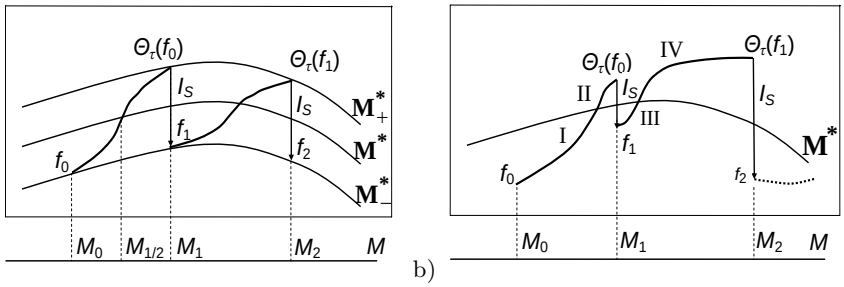



Fig. 9: The regular (a) and irregular (b) conservative chain. Dissipative terms for the regular chain give zero balance inside each step. For the irregular chain, dissipative term of part I (the first step) annihilates with dissipative term of part IV (the second step), as well, as annihilate dissipartive terms for parts II and III.

$$f_1 = I_S(\Theta_\tau(f_0)). \tag{54}$$

We can present this construction geometrically (Fig. 9a). The quasi-equilibrium manifold, $\mathbf{M}^* = \{q_{M,0}\}$, is accompanied by two other manifolds, $\mathbf{M}^*_\pm(\mathbf{s}) = \{q_{M,\tau_\pm(M,s)}\}$. These manifolds are connected by the entropic involution: $I_S\mathbf{M}^*_\pm(\mathbf{s}) = \mathbf{M}^*_\mp(\mathbf{s})$. For all points $f \in \mathbf{M}^*_\pm(\mathbf{s})$

$$S(f) = S(f^*_{m(f)}) - s.$$

The conservative chain starts at a point on $f_0 \in \mathbf{M}^*_-(\mathbf{s})$, than the solution of initial kinetic equations, $\Theta_t(f_0)$, goes to its intersection with $\mathbf{M}^*_+(\mathbf{s})$, the moment of intersection is $\tau$. After that, the entropic involution transfers $\Theta_\tau(f_0)$ into a second point of the chain, $f_1 = I_S(\Theta_\tau(f_0)) \in \mathbf{M}^*_-(\mathbf{s})$.

*Irregular conservative chain*

The regular geometric picture is nice, but for some generalizations we need less rigid structure. Let us combine two operations: the shift in time $\Theta_\tau$ and the entropic involution $I_S$. Suppose, the motions starts on a point $f_0$ on the film of non-equilibrium states, and

$$f_{n+1} = I_S(\Theta_\tau(f_n)). \tag{55}$$

This chain we call an *irregular conservative chain*, and the chain that moves from $\mathbf{M}^*_-(\mathbf{s})$ to $\mathbf{M}^*_+(\mathbf{s})$ and back, the regular one. For the regular chain the dissipative term is zero (in the main order in $\tau$) already for one link because this link is symmetric, and the macroscopic entropy ($S(M)$) loose for a motion from $\mathbf{M}^*_-(\mathbf{s})$ to $\mathbf{M}^*$ compensate the macroscopic entropy production on a way from $\mathbf{M}^*$ to $\mathbf{M}^*_+(\mathbf{s})$. For the irregular chain (55) with given $\tau$ (that may be constant) such a compensation occurs in two successive links (Fig. 9b) in main order in $\tau$.

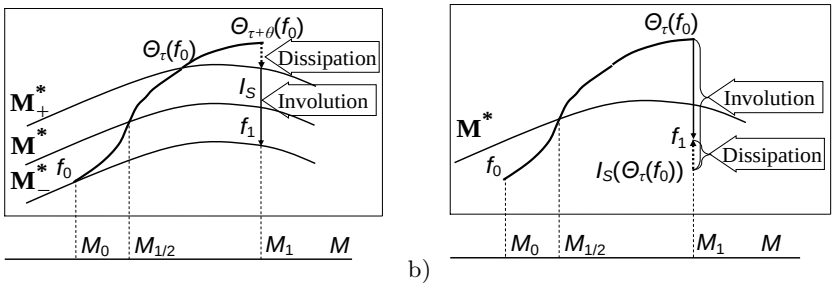



Fig. 10: Realization of dissipative chain by the extra time $\vartheta$ on the base of a regular conservative chain (a), and by the incomplete involution on the base of an irregular conservative chain (b).

Warning. *The doubled–step rule:* In the conservative irregular chains (that are more convenient in use), the second order accuracy emerges in steps that consist from *two successive links*.

*Kinetic modeling for non-zero dissipation. 1. Extension of regular chains*

The conservative chain of kinetic curves approximates the quasi-equilibrium dynamics. A typical example of quasi-equilibrium equations (21) is the Euler equation in fluid dynamics. Now, we combine conservative chains construction with the idea of the dissipative Ehrenfests' chain in order to create a method for kinetic modeling of dissipative hydrodynamics ("macrodynamics") (39) with arbitrary kinetic coefficient that is decoupled from the chain step $\tau$:

$$\frac{\mathrm{d}M}{\mathrm{d}t} = m(J(f^*_M)) + \kappa(M)m[(D_f J(f))_{f=f^*_M} \Delta_{f^*_M}]. \tag{56}$$

Here, a kinetic coefficient $\kappa(M) \geq 0$ is a non-negative function of $M$. The entropy production for (10) is:

$$\frac{\mathrm{d}S(M)}{\mathrm{d}t} = \kappa(M)\langle \Delta_{f^*_M}, \Delta_{f^*_M}\rangle_{f^*_M}. \tag{57}$$

Let us start from a regular conservative chain and deform it. A chain that approximates solutions of (56) can be constructed as follows (Fig. 10a). The motion starts from $f_0 \in \mathbf{M}^*_-(\mathbf{s})$, goes by a kinetic curve to intersection with $\mathbf{M}^*_+(\mathbf{s})$, as for a regular conservative chain, and, after that, follows the same kinetic curve an extra time $\vartheta$. This motion stops at the moment $\tau + \vartheta$ at the point $\Theta_{\tau+\vartheta}(f_0)$ (Fig. 10a). The second point of the chain, $f_1$ is the unique solution of equation

$$m(f_1) = m(\Theta_{\tau+\vartheta}(f_0)),\ f_1 \in \mathbf{M}^*_-(\mathbf{s}). \tag{58}$$

The time step is linked with the kinetic coefficient:

$$\kappa = \frac{\vartheta}{2} + o(\tau + \vartheta). \tag{59}$$

For entropy production we obtain the analogue of (40)

$$\frac{\mathrm{d}S(M)}{\mathrm{d}t} = \frac{\vartheta}{2}\langle \Delta_{f^*_M}, \Delta_{f^*_M}\rangle_{f^*_M} + o(\tau + \vartheta). \tag{60}$$

All these formulas follow from the first–order picture. In the first order of the time step,

$$\begin{aligned} q_{M,\tau} &= f^*_M + \tau\Delta_{f^*_M};\\ I_S(f^*_M + \tau\Delta_{f^*_M}) &= f^*_M - \tau\Delta_{f^*_M};\\ f_0 &= f^*_{M_0} - \frac{\tau}{2}\Delta_{f^*_{M_0}};\\ \Theta_t(f_0) &= f^*_{M(t)} + \left(t - \frac{\tau}{2}\right)\Delta_{f^*_{M_0}}, \end{aligned} \tag{61}$$

and up to the second order of accuracy (that is, again, the first non-trivial term)

$$S(q_{M,\tau}) = S(M) + \frac{\tau^2}{2}\langle \Delta_{f^*_M}, \Delta_{f^*_M}\rangle_{f^*_M}. \tag{62}$$

For a regular conservative chains, in the first order

$$f_1 = f^*_{M(\tau)} - \frac{\tau}{2}\Delta_{f^*_{M_0}}. \tag{63}$$

For chains (58), in the first order

$$f_1 = f^*_{M(\tau+\vartheta)} - \frac{\tau}{2}\Delta_{f^*_{M_0}}. \tag{64}$$

*Kinetic modeling for non-zero dissipation. 2. Deformed involution in irregular chains*

For irregular chains, we introduce dissipation without change of the time step $\tau$. Let us, after entropic involution, shift the point to the quasi-equilibrium state (Fig. 10) with some entropy increase $\sigma(M)$. Because of entropy production formula (57),

$$\sigma(M) = \tau\kappa(M)\langle \Delta_{f^*_M}, \Delta_{f^*_M}\rangle_{f^*_M}. \tag{65}$$

This formula works, if there is sufficient amount of non-equilibrium entropy, the difference $S(M_n) - S(f_n)$ should not be too small. In average, for several (two) successive steps it should not be less than $\sigma(M)$. The Ehrenfests' chain gives a limit for possible value of $\kappa(M)$ that we can realize using irregular chains with overrelaxation:

$$\kappa(M) < \frac{\tau}{2}. \tag{66}$$

Let us call the value $\kappa(M) = \frac{\tau}{2}$ the *Ehrenfests' limit*. Formally, it is possible to realize a chain of kinetic curves with time step $\tau$ for $\kappa(M) > \frac{\tau}{2}$ on the other side of the Ehrenfests' limit, without overrelaxation (Fig. 11).

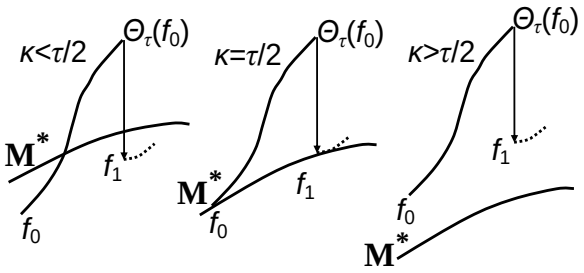



Fig. 11: The Ehrenfests' limit of dissipation: three possible links of a dissipative chain: overrelaxation, $\kappa(M) < \frac{\tau}{2}$ ($\langle\sigma\rangle = s_\tau - 2\sqrt{s_\tau\langle s_0\rangle}$), Ehrenfests' chain, $\kappa(M) = \frac{\tau}{2}$ ($\sigma = s_\tau$), and underrelaxation, $\kappa(M) > \frac{\tau}{2}$ ($\langle\sigma\rangle = s_\tau + 2\sqrt{s_\tau\langle s_0\rangle}$).

Let us choose the following notation for non-equilibrium entropy: $s_0 = S(M_0) - S(f_0)$, $s_1 = S(M_1) - S(f_1)$, $s_\tau(M) = \frac{\tau^2}{2}\langle\Delta_{f^*_M}, \Delta_{f^*_M}\rangle_{f^*_M}$. For the three versions of steps (Fig. 11) the entropy gain $\sigma = s(f_1) - S(I_S(\Theta_\tau(f_0)))$ in the main order in $\tau$ is:

- For overrelaxation ($\kappa(M) < \frac{\tau}{2}$) $\sigma = s_\tau + s_0 - s_1 - 2\sqrt{s_\tau s_0}$;
- For the Ehrenfests' chain (full relaxation, $\kappa(M) = \frac{\tau}{2}$) $s_0 = s_1 = 0$ and $\sigma = s_\tau$;
- For underrelaxation ($\kappa(M) > \frac{\tau}{2}$) $\sigma = s_\tau + s_0 - s_1 + 2\sqrt{s_\tau s_0}$.

After averaging in successive steps, the term $s_0 - s_1$ tends to zero, and we can write the estimate of the average entropy gain $\langle\sigma\rangle$: for overrelaxation $\langle\sigma\rangle = s_\tau - 2\sqrt{s_\tau\langle s_0\rangle}$ and for underelaxation $\langle\sigma\rangle = s_\tau + 2\sqrt{s_\tau\langle s_0\rangle}$.

In the really interesting physical problems the kinetic coefficient $\kappa(M)$ is non-constant in space. Macroscopic variables $M$ are functions of space, $\kappa(M)$ is also a function, and it is natural to take a space-dependent step of macroscopic entropy production $\sigma(M)$. It is possible to organize the involution (incomplete involution) step at different points with different density of entropy production step $\sigma$.

*Which entropy rules the kinetic model?*

For linear kinetic equations, for example, for the free flight equation (42) there exist many concave Lyapunov functionals (for dissipative systems) or integrals of motion (for conservative systems), see, for example, (4).

There are two reasonable conditions for entropy choice: additivity with respect to joining of independent systems, and trace form (sum or integral of some function $h(f, f^*)$). These conditions select a one-parametric family [43, 44], a linear combination of the classical Boltzmann–Gibbs–Shannon entropy with $h(f) = -f\ln f$ and the Burg Entropy with $h(f) = \ln f$, both in the Kullback form:

$$S_\alpha = -\alpha \int f \ln \frac{f}{f^*}\,\mathrm{d}\Gamma(x) + (1-\alpha)\int f^* \ln \frac{f}{f^*}\,\mathrm{d}\Gamma(x),$$

where $1 \geq \alpha \geq 0$, and $f^*\mathrm{d}\Gamma$ is invariant measure. Singularity of the Burg term for $f \to 0$ provides the positivity preservation in all entropic involutions.

If we weaken these conditions and require that there exists such a monotonic (nonlinear) transformation of entropy scale that in one scale entropy is additive, and in transformed one it has a trace form, then we get additionally a family of Renyi–Tsallis entropies with $h(f) = \frac{1-f^q}{1-q}$ [44] (these entropies and their applications are discussed in details in [45]).

Both the Renyi–Tsallis entropy and the Burge entropy are in use in the entropic lattice Boltzmann methods from the very beginning [89, 46]. The connection of this entropy choice with Galilei invariance is demonstrated in [46].

*Elementary examples*

In the most popular and simple example, the conservative formal kinetic equations (12) is the free flight equation (42). Macroscopic variables $M$ are the hydrodynamic fields: $n(\boldsymbol{x}) = \int f(\boldsymbol{x},\boldsymbol{v})\,\mathrm{d}\boldsymbol{v}$, $n(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x}) = \int \boldsymbol{v} f(\boldsymbol{x},\boldsymbol{v})\,\mathrm{d}\boldsymbol{v}$, $3n(\boldsymbol{x})k_\mathrm{B}T/2m = \frac{1}{2}\int \boldsymbol{v}^2 f(\boldsymbol{x},\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} - \frac{1}{2}n(\boldsymbol{x})\boldsymbol{u}^2(\boldsymbol{x})$, where $m$ is particle mass. In 3D at any space point we have five independent variables.

For a given value of five macroscopic variables $M = \{n, \boldsymbol{u}, T\}$ (3D), the quasi-equilibrium distribution is the classical local Maxwellian:

$$f^*_M(\boldsymbol{x},\boldsymbol{v}) = n\left(\frac{2\pi k_\mathrm{B}T}{m}\right)^{-3/2} \exp\left(-\frac{m(\boldsymbol{v}-\boldsymbol{u})^2}{2k_\mathrm{B}T}\right), \tag{67}$$

The standard choice of entropy for this example is the classical Boltzmann–Gibbs–Shannon entropy (5) with entropy density $s(\boldsymbol{x})$. All the involution operations are performed pointwise: at each point $\boldsymbol{x}$ we calculate hydrodynamic moments $M$, the correspondent local Maxwellian (67) $f^*_M$, and find the entropic inversion at this point with the standard entropy. For dissipative chains, it is useful to take the dissipation (the entropy density gain in one step) proportional to the $S(M) - S(f)$, and not with fixed value.

The special variation of the discussed example is the free flight with finite number of velocities: $f(\boldsymbol{x},\boldsymbol{v}) = \sum_i f_i(\boldsymbol{x})\delta(\boldsymbol{v}-\boldsymbol{v}_i)$. Free flight does not change the set of velocities $\{\boldsymbol{v}_1, \ldots\ldots \boldsymbol{v}_n\}$. If we define entropy, then we can define an equilibrium distribution for this set of velocity too. For the entropy definition let us substitute $\delta$-functions in expression for $f(\boldsymbol{x},\boldsymbol{v})$ by some "drops" with unite volume, small diameter, and fixed density that may depend on $i$. After that, the classical entropy formula unambiguously leads to expression:

$$s(\boldsymbol{x}) = -\sum_i f_i(\boldsymbol{x})\left(\ln \frac{f_i(\boldsymbol{x})}{f_i^0} - 1\right). \tag{68}$$

This formula is widely known in chemical kinetics (see elsewhere, for example [34, 35, 36]). After classical work of Zeldovich [37] (1938), this function is recognized as a useful instrument for analysis of chemical kinetic equations. Vector of values $f^0 = f_i^0$ gives us a "particular equilibrium:" for $M = m(f^0)$ the conditional equilibrium ($s \to \max$, $M = m(f^0)$) is $f^0$. With entropy (68) we can construct all types of conservative and dissipative chains for discrete set of velocities. If we need to approximate the continuous local equilibria and involutions by our discrete equilibria and involutions, then we should choose a particular equilibrium distribution $\sum_i f_i^0 \delta(\boldsymbol{v} - \boldsymbol{v}_i)$ in velocity space as an approximation to the Maxwellian $f^{*0}(\boldsymbol{v})$ with correspondent value of macroscopic variables $M^0$ calculated for the discrete distribution $f^0$: $n = \sum_i f_i^0$, ... This approximation of distributions should be taken in the weak sense. It means that $\boldsymbol{v}_i$ are nodes, and $f_i^0$ are weights for a cubature formula in 3D space with weight $f^{*0}(\boldsymbol{v})$:

$$\int p(\boldsymbol{v}) f^{*0}(\boldsymbol{v})\, \mathrm{d}\boldsymbol{v} \approx \sum_i p(\boldsymbol{v}_i) f_i^0. \tag{69}$$

There exist a huge population of cubature formulas in 3D with Gaussian weight that are optimal in various senses [95]. Each of them contains a hint for a choice of nodes $\boldsymbol{v}_i$ and weights $f_i^0$ for the best discrete approximation of continuous dynamics. Applications of this entropy (68) to the lattice Boltzmann models are developed in [93].

There is one more opportunity to use entropy (68) and related involutions for discrete velocity systems. If for some of components $f_i = 0$, then we can find the correspondent *positive* equilibrium, and perform the involution in the whole space. But there is another way: if for some of velocities $f_i = 0$, then we can reduce the space, and find an equilibrium for non-zero components only, for the shortened list of velocities. These *boundary equilibria* play important role in the chemical thermodynamic estimations [96].

This approach allows us to construct systems with variable in space set of velocities. There could be "soft particles" with given velocities, and the density distribution in these particles changes only when several particles collide. In 3D for the possibility of a non-trivial equilibrium that does not obligatory coincide with the current distribution we need more than 5 different velocity vectors, hence, a non-trivial collision ($\approx$ entropic inversion) is possible only for 6 one-velocity particles. If in a collision participate 5 one-velocity particles or less, then they are just transparent and don't interact at all. For more moments, if we add some additional fields (stress tensor, for example), the number of velocity vectors that is necessary for non-trivial involution increases.

*Lattice Boltzmann models: lattice is not a tool for discretization*

In this section, we presented the theoretical backgrounds of kinetic modeling. These problems were discussed previously for development of lattice Boltz-

mann methods in computational fluid dynamics. The "overrelaxation" appeared in [88]. In papers [90, 91] the overrelaxation based method for the Navier–Stokes equations was further developed, and the entropic involution was invented in [89]. Due to historical reasons, we propose to call it the *Karlin–Succi* involution. The problem of computational stability of entropic lattice Boltzmann methods was systematically analyzed in [93, 94]. $H$-theorem for lattice Boltzmann schemes was presented with details and applications in [92]. For further discussion and references we address to [19].

In order to understand links from the Ehrenfests' chains to the lattice Boltzmann models, let us take the model with finite number of velocity vectors and entropy (68). Let the velocities from the set $\{\boldsymbol{v}_1, \ldots\ldots \boldsymbol{v}_n\}$ be automorphisms of some lattice $\mathbf{L}$: $\mathbf{L} + \boldsymbol{v}_i = \mathbf{L}$. Then the restriction of free flight in time $\tau$ on the functions on the lattice $\tau\mathbf{L}$ is exact. It means that the free flight shift in time $\tau$, $f(x, v) \mapsto f(x - v\tau, v)$ is defined on functions on the lattice, because $\boldsymbol{v}_i\tau$ are automorphisms of $\tau\mathbf{L}$. The entropic involution (complete or incomplete one) acts pointwise, hence, the restriction of the chains on the lattice $\tau\mathbf{L}$ is exact too. In that sense, the role of lattice here is essentially different from the role of grid in numerical methods for PDE. All the discretization contains in the velocity set $\{\boldsymbol{v}_1, \ldots\ldots \boldsymbol{v}_n\}$, and the accuracy of discretization is the accuracy of cubature formulas (69).

The lattice $\tau\mathbf{L}$ is a tool for presentation of velocity set as a subset of $\mathbf{L}$ automorphism group. At the same time, it is a perfect screen for presentation of the chain dynamics, because restriction of that dynamics on this lattice is an autonomous dynamic of lattice distribution. (Here we meet a rather rare case of exact model reduction.)

The boundary conditions for the lattice Boltzmann models deserve special attention. There were many trials of non-physical conditions until the proper (and absolutely natural) discretization of well-known classical kinetic boundary conditions (see, for example, [80]) were proposed [97]. It is necessary and sufficient just to describe scattering of particles on the boundary with maximal possible respect to the basic physics (and given proportion between elastic collisions and thermalization).

## 3 Coarse-Graining by Filtering

The most popular area for filtering applications in mathematical physics is the Large Eddy Simulation (LES) in fluid dynamics [17]. Perhaps, the first attempt to turbulence modeling was done by Boussinesq in 1887. After that, Taylor (1921, 1935, 1938) and Kolmogorov (1941) have provided the bases of the statistical theory of turbulence. The Kolmogorov theory of turbulence self similarity inspired many attempts of so-called subgrid-scale modeling (SGS model): only the large scale motions of the flow are solved by filtering out the small and universal eddies. For the dynamic subgrid-scale models a filtering

step is required to compute the SGS stress tensor. The filtering a hydrodynamic field is defined as convoluting the field functions with a filtering kernel, as it is done in electrical engineering:

$$\overline{\{n, n\boldsymbol{u}, nT\}}(\boldsymbol{x}) = \int G(\boldsymbol{x}-\boldsymbol{y})\{n, n\boldsymbol{u}, nT\}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}. \tag{70}$$

Various filter kernels are in use. Most popular of them are:

1. The box filter $G(x) = H(\Delta/2 - |x|)/\Delta$;
2. The Gaussian filter $G(x) = \frac{1}{\Delta}\sqrt{6/\pi}\exp\left(-6x^2/\Delta^2\right)$,

where $\Delta$ is the filter width (for the Gaussian filter, $\Delta/2 = \sqrt{3}\sigma$, this convention corresponds to 91.6% of probability in the interval $[-\Delta/2, +\Delta/2]$ for the Gaussian distribution) , $H$ is the Heaviside function, $G(\boldsymbol{x}) = \prod_i G(x_i)$.

In practical applications, implicit filtering is sometimes done by the grid itself. This filtering by grids should be discussed in context of the Whittaker–Nyquist–Kotelnikov–Shannon sampling theory [98, 99]. Bandlimited functions (that is, functions which Fourier transform has compact support) can be exactly reconstructed from their values on a sufficiently fine grid by the Nyquist-Shannon interpolation formula and its multidimensional analogues. If, in 1D, the Fourier spectrum of $f(x)$ belongs to the interval $[-k_{\max}, k_{\max}]$, and the grid step $h$ is less than $\pi/k_{\max}$ (it is, twice less than the minimal wave length), then this formula gives the exact representation of $f(x)$ for all points $x$:

$$f(x) = \sum_{n=-\infty}^{+\infty} f(nh)\mathrm{sinc}\left(\pi\left[\frac{x}{h} - n\right]\right), \tag{71}$$

where $\mathrm{sinc}(x) = \frac{\sin x}{x}$. That interpolation formula implies an exact differentiation formula in the nodes:

$$\left.\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right|_{x=nh} = 2\pi\sum_{k=1}^{\infty}(-1)^{k+1}\frac{f((n+k)h) - f((n-k)h)}{2kh}. \tag{72}$$

Such "long tail" exact differentiation formulas are useful under assumption about bounded Fourier spectrum.

As a background for SGS modeling, the *Boussinesq hypothesis* is widely used. This hypothesis is that the turbulent terms can be modeled as directly analogues to the molecular viscosity terms using a "turbulent viscosity." Strictly speaking, no hypothesis are needed for equation filtering, and below a sketch of *exact filtering theory* for kinetic equations is presented. The idea of *reversible* regularization without apriory closure assumptions in fluid dynamics was proposed by Leray [13]. Now it becomes popular again [100, 102, 103].

### 3.1 Filtering as Auxiliary Kinetics

*Idea of filtering in kinetics*

The variety of possible filters is too large, and we need some fundamental conditions that allow to select physically reasonable approach.

Let us start again from the formal kinetic equation (12) $\mathrm{d}f/\mathrm{d}t = J(f)$ with concave entropy functional $S(f)$ that does not increase in time and is defined in a convex subset $U$ of a vector space $E$.

The filter transformation $\Phi_\Delta : U \to U$, where $\Delta$ is the filter width, should satisfy the following conditions:

1. Preservation of conservation laws: for any basic conservation law of the form $C[f] = \text{const}$ filtering does not change the value $C[f]$: $C[\Phi_\Delta(f)] = C[f]$. This condition should be satisfied for the whole probability or for number of particles (in most of classical situations), momentum, energy, and filtering should not change the center of mass, this is not so widely known condition, but physically obvious consequence of Galilei invariance.
2. The Second Law (entropy growth): $S(\Phi_\Delta(f)) \geq S(f)$.

It is easy to check the conservation laws for convoluting filters (70), and here we find the first benefit from the kinetic equation filtering: for usual kinetic equations and all mentioned conservation laws functionals $C[f]$ are linear, and the conservation preservation conditions are very simple linear restrictions on the kernel $G$ (at least, far from the boundary). For example, for the Boltzmann ideal gas distribution function $f(\boldsymbol{x}, \boldsymbol{v})$, the number of particles, momentum, and energy conserve in filtering $\overline{f(\boldsymbol{x}, \boldsymbol{v})} = \int G(\boldsymbol{x} - \boldsymbol{y}) f(\boldsymbol{y}, \boldsymbol{v})\, \mathrm{d}\boldsymbol{y}$, if $\int G(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1$; for the center of mass conservation we need also a symmetry condition $\int \boldsymbol{x} G(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 0$. It is necessary to mention that usual filters extend the support of distribution, hence, near the boundary the filters should be modified, and boundary can violate the Galilei invariance, as well, as momentum conservation. We return to these problems in this paper later.

For continuum mechanics equations, energy is not a linear functional, and operations with filters require some accuracy and additional efforts, for example, introduction of spatially variable filters [101]. Perhaps, the best way is to lift the continuum mechanics to kinetics, to filter the kinetic equation, and then to return back to filtered continuum mechanics. On kinetic level, it becomes obvious how filtering causes the redistribution of energy between internal energy and mechanical energy: energy of small eddies and of other small-scale inhomogeneities partially migrates into internal energy.

*Filtering semigroup*

If we apply the filtering twice, it should lead just just to increase of the filter width. This natural semigroup condition reduces the set of allowed filters significantly. The approach based on filters superposition was analyzed by Germano [15] and developed by many successors. Let us formalize it in a form

$$\Phi_{\Delta'}(\Phi_\Delta(f)) = \Phi_{\Delta''}(f), \tag{73}$$

where $\Delta''(\Delta', \Delta)$ is a monotonic function, $\Delta'' \geq \Delta'$ and $\Delta'' \geq \Delta$.

The semigroup condition (73) holds for the Gaussian filter with $\Delta''^2 = \Delta'^2 + \Delta^2$, and does not hold for the box filter. It is convenient to parameterize

the semigroup $\{\Phi_\Delta | \Delta \geq 0\}$ by an additive parameter $\eta \geq 0$ ("auxiliary time"): $\Delta = \Delta(\eta)$, $\Phi_\eta \circ \Phi_{\eta'} = \Phi_{\eta+\eta'}$, $\Phi_0 = \mathrm{id}$. Further we use this parameterization.

*Auxiliary kinetic equation*

The filtered distribution $f(\eta) = \Phi_\eta(f_0)$ satisfies differential equation

$$\frac{\mathrm{d}f(\eta)}{\mathrm{d}\eta} = \phi(f(\eta)), \text{ where } \phi(f) = \lim_{\eta\to 0} \frac{\Phi_\eta(f) - f}{\eta}. \tag{74}$$

For Gaussian filters this equation is the simplest diffusion equation $\mathrm{d}f(\eta)/\mathrm{d}\eta = \Delta f$ (here $\Delta$ is the Laplace operator).

Due to physical restrictions on possible filters, auxiliary equation (74) has main properties of kinetic equations: it respects conservation laws and the Second Law. It is also easy to check that in the whole space (without boundary effects) diffusion, for example, does not change the center of mass.

So, when we discuss filtering of kinetics, we deal with two kinetic equations in the same space, but in two times $t$ and $\eta$: initial kinetics (12) and filtering equation (74). Both have the same conservation laws and the same entropy.

### 3.2 Filtered Kinetics

*Filtered kinetic semigroup*

Let $\Theta_t$ be the semigroup of the initial kinetic phase flow. We are looking for kinetic equation that describes dynamic of filtered distribution $\Phi_\eta f$ for given $\eta$. Let us call this equation with correspondent dynamics the *filtered kinetics*. It is the third kinetic equation in our consideration, in addition to the initial kinetics (12) and the auxiliary filtering kinetics (74). The natural phase space for this filtered kinetics is the set of filtered distributions $\Phi_\eta(U)$. For the phase flow of the filtered kinetics we use notation $\Psi_{(\eta)\,t}$ This filtered kinetics should be the exact shadow of the true kinetics. It means that the motion $\Psi_{(\eta)\,t}(\Phi_\eta f_0)$ is the result of filtering of the true motion $\Theta_t(f_0)$: for any $f_0 \in U$ and $t > 0$

$$\Psi_{(\eta)\,t}(\Phi_\eta f_0) = \Phi_\eta(\Theta_t(f_0)). \tag{75}$$

This equality means that

$$\Psi_{(\eta)\,t} = \Phi_\eta \circ \Theta_t \circ \Phi_{-\eta} \tag{76}$$

The transformation $\Phi_{-\eta}$ is defined on the set of filtered distributions $\Phi_\eta(U)$, as well as $\Psi_{(\eta)\,t}$ is. Now it is necessary to find the vector field

$$\psi_{(\eta)}(f) = \left.\frac{\mathrm{d}\Psi_{(\eta)\,t}(f)}{\mathrm{d}t}\right|_{t=0}$$

on the base of conditions (75), (76). This vector field is the right-hand side of the filtered kinetic equations

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \psi_{(\eta)}(f). \tag{77}$$

From (76) immediately follows:

$$\frac{\mathrm{d}\psi_{(\eta)}(f)}{\mathrm{d}\eta} = (D_f\phi(f))\psi_{(\eta)}(f) - (D_f\psi_{(\eta)}(f))\phi(f) = [\psi_{(\eta)}, \phi](f), \tag{78}$$

where $[\psi, \phi]$ is the Lie bracket of vector fields.

In the first approximation in $\eta$

$$\psi_{(\eta)}(f) = J(f) + \eta((D_f\phi(f))J(f) - (D_fJ(f))\phi(f)) = J(f) + \eta[J, \phi](f), \tag{79}$$

the Taylor series expansion for $\psi_{(\eta)}(f)$ is

$$\psi_{(\eta)}(f) = J(f) + \eta[J, \phi](f) + \frac{\eta^2}{2}[[J, \phi], \phi](f) + \ldots + \frac{\eta^n}{n!}[\ldots[J, \underbrace{\phi], \ldots\phi}_{n}](f) + \ldots \tag{80}$$

We should stress again that filtered equations (77) with vector field $\psi_{(\eta)}(f)$ that satisfies (78) is exact and presents just a shadow of the original kinetics. Some problems may appear (or not) after truncating the Taylor series (80), or after any other approximation.

So, we have two times: physical time $t$ and auxiliary filtering time $\eta$, and four different equations of motion in these times:

- initial equation (12) (motion in time $t$),
- filtering equation (74) (motion in time $\eta$),
- filtered equation (77) (motion in time $t$),
- and equation for the right hand side of filtered equation (78) (motion in time $\eta$).

*Toy example: advection + diffusion*

Let us consider kinetics of system that is presented by one scalar density in space (concentration), with only one linear conservation law, the total number of particles.

In the following example the filtering equation (74) is

$$\frac{\partial f(\boldsymbol{x}, \eta)}{\partial \eta} = \Delta f(\boldsymbol{x}, \eta)\ (= \phi(f)). \tag{81}$$

The differential of $\phi(f)$ is simply the Laplace operator $\Delta$. The correspondent 3D heat kernel (the fundamental solution of (81)) is

$$K(\eta, \boldsymbol{x} - \boldsymbol{\xi}) = \frac{1}{(4\pi\eta)^{3/2}} \exp\left(-\frac{(\boldsymbol{x} - \boldsymbol{\xi})^2}{4\eta}\right). \tag{82}$$

After comparing this kernel with the Gaussian filter we find the filter width $\Delta = \sqrt{24\eta}$.

Here we consider the diffusion equation (81) in the whole space with zero conditions at infinity. For other domains and boundary conditions the filtering kernel is the correspondent fundamental solution.

The equation for the right hand side of filtered equation (78) is

$$\frac{\mathrm{d}\psi_{(\eta)}(f)}{\mathrm{d}\eta} = \Delta(\psi_{(\eta)}(f)) - (D_f\psi_{(\eta)}(f))(\Delta f)\ (= [\psi,\phi](f)). \tag{83}$$

For the toy example we select the advection + diffusion equation

$$\frac{\partial f(\boldsymbol{x},t)}{\partial t} = \kappa\Delta f(\boldsymbol{x},t) - \mathrm{div}(\boldsymbol{v}(\boldsymbol{x})f(\boldsymbol{x},t))\ (= J(f)). \tag{84}$$

where $\kappa > 0$ is a given diffusion coefficient, $\boldsymbol{v}(\boldsymbol{x})$ is a given velocity field. The differential $D_fJ(f)$ is simply the differential operator from the right hand side of (84), because this vector field is linear. After simple straightforward calculation we obtain the first approximation (79) to the filtered equation:

$$\begin{aligned}
[J,\phi](f) &= \Delta(J(f)) - (D_fJ(f))(\Delta f) = -\Delta[\mathrm{div}(\boldsymbol{v}f)] + \mathrm{div}(\boldsymbol{v}\Delta f) \\
&= \mathrm{div}[\boldsymbol{v}\Delta f - \Delta(\boldsymbol{v}f)] = -\mathrm{div}\left[f\Delta\boldsymbol{v} + 2\sum_r \frac{\partial\boldsymbol{v}}{\partial x_r}\frac{\partial f}{\partial x_r}\right] \\
&= -\mathrm{div}(f\Delta\boldsymbol{v}) - \sum_i \frac{\partial}{\partial x_i}\left[\sum_r \left(\frac{\partial v_i}{\partial x_r} + \frac{\partial v_r}{\partial x_i}\right)\frac{\partial f}{\partial x_r}\right. \\
&\quad \left. - \sum_r \frac{\partial f}{\partial x_r}\left(\frac{\partial v_i}{\partial x_r} - \frac{\partial v_r}{\partial x_i}\right)\right] \\
&= -\sum_i \frac{\partial}{\partial x_i}\left[\sum_r \left(\frac{\partial v_i}{\partial x_r} + \frac{\partial v_r}{\partial x_i}\right)\frac{\partial f}{\partial x_r}\right] - \sum_r \frac{\partial}{\partial x_r}\left(f\frac{\partial \mathrm{div}\boldsymbol{v}}{\partial x_r}\right).
\end{aligned} \tag{85}$$

The resulting equations in divergence form are

$$\begin{aligned}
\frac{\partial f(\boldsymbol{x},t)}{\partial t} &= J(f) + \eta[J,\phi](f) \\
&= -\mathrm{div}\left(-\kappa\nabla f + (\boldsymbol{v} + \eta\Delta\boldsymbol{v})f + 2\eta\sum_r \frac{\partial\boldsymbol{v}}{\partial x_r}\frac{\partial f}{\partial x_r}\right) \\
&= \mathrm{div}((\kappa - 2\eta\boldsymbol{S}(\boldsymbol{x}))\nabla f(\boldsymbol{x},t)) - \mathrm{div}((\boldsymbol{v}(\boldsymbol{x}) + \eta\nabla\mathrm{div}\boldsymbol{v}(\boldsymbol{x}))f(\boldsymbol{x},t)),
\end{aligned} \tag{86}$$

where $\boldsymbol{S}(\boldsymbol{x}) = (S_{ij}) = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right)$ is the strain tensor. In filtered equations (86) the additional diffusivity tensor $-2\eta\boldsymbol{S}(\boldsymbol{x})$ and the additional velocity $\eta\nabla\mathrm{div}\boldsymbol{v}(\boldsymbol{x})$ are present. The additional diffusivity tensor $-2\eta\boldsymbol{S}(\boldsymbol{x})$ may be not positive definite. The positive definiteness of the diffusivity tensor $\kappa - 2\eta\boldsymbol{S}(\boldsymbol{x})$ may be also violated. For arbitrary initial condition $f_0(\boldsymbol{x})$ it may cause some instability problems, but we should take into account that the

filtered equations (86) are defined on the space of filtered functions for given filtering time $\eta$. On this space the negative diffusion ($\partial_t f = -\Delta f$) is possible during time $\eta$. Nevertheless, the approximation of exponent (80) by the linear term (79) can violate the balance between smoothed initial conditions and possible negative diffusion and can cause some instabilities.

Some numerical experiments with this model (86) for incompressible flows ($\mathrm{div}\boldsymbol{v} = 0$) are presented in [103].

Let us discuss equation (83) in more details. We shall represent it as the dynamics of the filtered advection flux vector $\boldsymbol{\Pi}$. The filtered equation for any $\eta$ should have the form: $\partial f/\partial t = -\mathrm{div}(-\kappa\nabla f + \boldsymbol{\Pi}(f))$, where

$$\boldsymbol{\Pi}(f) = \left( \sum_{j_1,j_2,j_3 \geq 0} \boldsymbol{a}_{j_1 j_2 j_3}(\boldsymbol{x},\eta)\partial_x^{j_1 j_2 j_3} \right) f(\boldsymbol{x}), \tag{87}$$

where

$$\partial_x^{j_1 j_2 j_3} = \left(\frac{\partial}{\partial x_1}\right)^{j_1} \left(\frac{\partial}{\partial x_2}\right)^{j_2} \left(\frac{\partial}{\partial x_3}\right)^{j_3} \tag{88}$$

For coefficients $\boldsymbol{a}_{j_1 j_2 j_3}(\boldsymbol{x},\eta)$ equation (83) is

$$\frac{\partial \boldsymbol{a}_{j_1 j_2 j_3}(\boldsymbol{x},\eta)}{\partial \eta} = \Delta \boldsymbol{a}_{j_1 j_2 j_3}(\boldsymbol{x},\eta) \tag{89}$$
$$+ 2\frac{\partial \boldsymbol{a}_{j_1-1\, j_2 j_3}(\boldsymbol{x},\eta)}{\partial x_1} + 2\frac{\partial \boldsymbol{a}_{j_1 j_2-1\, j_3}(\boldsymbol{x},\eta)}{\partial x_2} + 2\frac{\partial \boldsymbol{a}_{j_1 j_2 j_3-1}(\boldsymbol{x},\eta)}{\partial x_3}.$$

The initial conditions are: $\boldsymbol{a}_{000}(\boldsymbol{x},0) = \boldsymbol{v}(\boldsymbol{x})$, $\boldsymbol{a}_{j_1 j_2 j_3}(\boldsymbol{x},0) = 0$ if at least one of $j_k > 0$. Let us define formally $\boldsymbol{a}_{j_1 j_2 j_3}(\boldsymbol{x},\eta) \equiv 0$ if at least one of $j_k$ is negative.

We shall consider (89) in the whole space with appropriate conditions at infinity. There are many representation of solution to this system. Let us use the Fourier transformation:

$$\frac{\partial \hat{\boldsymbol{a}}_{j_1 j_2 j_3}(\boldsymbol{k},\eta)}{\partial \eta} = -k^2 \hat{\boldsymbol{a}}_{j_1 j_2 j_3}(\boldsymbol{k},\eta) \tag{90}$$
$$+2i(k_1 \hat{\boldsymbol{a}}_{j_1-1\, j_2 j_3}(\boldsymbol{k},\eta) + k_2 \hat{\boldsymbol{a}}_{j_1 j_2-1\, j_3}(\boldsymbol{k},\eta) + k_3 \hat{\boldsymbol{a}}_{j_1 j_2 j_3-1}(\boldsymbol{k},\eta)).$$

Elementary straightforward calculations give us:

$$\hat{\boldsymbol{a}}_{j_1 j_2 j_3}(\boldsymbol{k},\eta) = (2i\eta)^{|j|} \mathrm{e}^{-k^2\eta} \frac{k_1^{j_1} k_2^{j_2} k_3^{j_3}}{j_1! j_2! j_3!} \hat{\boldsymbol{v}}(\boldsymbol{k}), \tag{91}$$

where $|j| = j_1 + j_2 + j_3$. To find this answer, we consider all monotonic paths on the integer lattice from the point $(0,0,0)$ to the point $(j_1, j_2, j_3)$. In concordance with (90), every such a path adds a term

$$\frac{(2i\eta)^{|j|}}{|j|!} \mathrm{e}^{-k^2\eta} k_1^{j_1} k_2^{j_2} k_3^{j_3} \hat{\boldsymbol{v}}(\boldsymbol{k})$$

to $\hat{\boldsymbol{a}}_{j_1j_2j_3}(\boldsymbol{k},\eta)$. The number of these paths is $|j|!/(j_1!j_2!j_3!)$.

The inverse Fourier transform gives

$$\boldsymbol{a}_{j_1j_2j_3}(\boldsymbol{x},\eta) = (2\eta)^{|j|-3/2}\frac{\partial_x^{j_1j_2j_3}}{j_1!j_2!j_3!}\int \exp -\frac{(\boldsymbol{x}-\boldsymbol{y})^2}{4\eta}\boldsymbol{v}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}. \qquad (92)$$

Finally, for $\boldsymbol{\Pi}$ we obtain

$$\boldsymbol{\Pi}(f) \qquad (93)$$
$$= \sum_{j_1,j_2,j_3\geq 0}\frac{(2\eta)^{|j|-3/2}}{j_1!j_2!j_3!}\left(\partial_x^{j_1j_2j_3}\int \exp -\frac{(\boldsymbol{x}-\boldsymbol{y})^2}{4\eta}\boldsymbol{v}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\right)\partial_x^{j_1j_2j_3}f(\boldsymbol{x}).$$

By the way, together with (93) we received the following formula for the Gaussian filtering of products [103]. If the semigroup $\Phi_\eta$ is generated by the diffusion equation (81), then for two functions $f(\boldsymbol{x}), g(\boldsymbol{x})$ in $R^n$ (if all parts of the formula exist):

$$\Phi_\eta(fg) = \sum_{j_1,j_2,\dots j_n\geq 0}\frac{(2\eta)^{|j|-n/2}}{j_1!j_2!\dots j_n!}(\partial_x^{j_1j_2\dots j_n}\Phi_\eta(f))(\partial_x^{j_1j_2\dots j_n}\Phi_\eta(g)). \qquad (94)$$

Generalization of this formula for a broader class of filtering kernels for convolution filters is described in [16]. This is simply the Taylor expansion of the Fourier transformation of the convolution equality $\Psi_t = \Phi\circ\Theta_t\circ\Phi^{-1}$, where $\Phi$ is the filtering transformation (see (76)).

For filtering semigroups all such formulas are particular cases of the commutator expansion (80), and calculation of all orders requires differentiation only. This case includes non-convolution filtering semigroups also (for example, solutions of the heat equations in a domain with given boundary conditions, it is important for filtering of systems with boundary conditions), as well as semigroups of non-linear kinetic equation.

*Nonlinear filtering toy example*

Let us continue with filtering of advection + diffusion equation (84) and accept the standard assumption about incompressibility of advection flow $\boldsymbol{v}$: $\mathrm{div}\boldsymbol{v} = 0$. The value of density $f$ does not change in motion with the advection flow, and for diffusion the maximum principle exists, hence, it makes sense to study bounded solutions of (84) with appropriate boundary conditions, or in the whole space. Let us take $\max f < A$. This time we use the filtering semigroup

$$\frac{\partial f(\boldsymbol{x},\eta)}{\partial\eta} = -\mathrm{div}(-(A-f)\nabla f) = (A-f)\Delta f(\boldsymbol{x},\eta) - (\nabla f)^2\ (=\phi(f)). \qquad (95)$$

This semigroup has slightly better properties of reverse filtering (at least, no infinity in values of $f$). The first-order filtered equation (79) for this filter is (compare to (85)):

$$\frac{\partial f(\boldsymbol{x},t)}{\partial t} = J(f) + \eta[J,\phi](f) \qquad (96)$$
$$= -\mathrm{div}[-\kappa\nabla(f+\eta(\nabla f)^2) + 2\eta(A-f)\boldsymbol{S}\nabla f + \boldsymbol{v}f].$$

Here, $\boldsymbol{S}$ is the strain tensor, the term $-2\eta(A-f)\boldsymbol{S}$ is the additional (nonlinear) tensor diffusivity, and the term $\eta\kappa\nabla(\nabla f)^2$ describes the flux from areas with high $f$ gradient. Because this flux vanishes near critical points of $f$, it contributes to creation of a patch structure.

In the same order in $\eta$, it is convenient to write:

$$\frac{\partial f(\boldsymbol{x},t)}{\partial t} = -\mathrm{div}[-(\kappa - 2\eta(A-f)\boldsymbol{S})\nabla(f+\eta(\nabla f)^2) + \boldsymbol{v}f].$$

The nonlinear filter changes not only the diffusion coefficient, but the correspondent thermodynamic force also: instead of $-\nabla f$ we obtain $-\nabla(f + \eta(\nabla f)^2)$. This thermodynamic force depends on $f$ gradient and can participate in the pattern formation.

*LES + POD filters*

In the title, LES stands for Large Eddy Simulation, as it is before, and POD stands for Proper Orthogonal Decomposition. POD [104] is an application of principal component analysis [105] for extraction of main components from the flow dynamics. The basic procedure is quite simple. The input for POD is a finite set of flow images (a sample) $\{f_1,\dots,f_n\}$. These images are functions in space, usually we have the values of these function on a grid. In the space of functions an inner product is given. The first choice gives the $L_2$ inner product $\int fg\,\mathrm{d}x$, or energetic one, or one of the Sobolev's space inner products. The mean point $\psi_0 = \sum_i f_i/n$ minimizes the sum of distance squares $\sum_i (f_i - \psi_0)^2$. The first principal component $\psi_1$ minimizes the sum of distance squares from points $f_i$ to a straight line $\{\psi_0 + \alpha\psi_1 \mid \alpha \in R\}$, the second principal component, $\psi_2$, is orthogonal to $\psi_1$ and minimizes the sum of distance squares from points $f_i$ to a plain $\{\psi_0 + \alpha_1\psi_1 + \alpha_2\psi_2 \mid \alpha_{1,2} \in R\}$, and so on. Vectors of principal components $\psi_i$ are the eigenvectors of the sample covariance matrix $\Sigma$, sorted by decreasing eigenvalue $\lambda_i$, where

$$\Sigma = \frac{1}{n}\sum_i (f_i - \psi_0)\otimes(f_i-\psi_0)^T = \frac{1}{n}\sum_i |f_i-\psi_0\rangle\langle f_i - \psi_0|. \qquad (97)$$

The projection of a field $f$ on the plane of the $k$ first principal components is $\psi_0 + P_k(f-\psi_0)$, where $P_k$ is the orthogonal projector on the space spanned by the first $k$ components:

$$P_k(\phi) = \sum_{1\le j\le k} \psi_j(\psi_j,\phi). \qquad (98)$$

The average square distance from the sample points $f_i$ to the plane of the $k$ first principal components is

$$\sigma_k^2 = \sum_{j>k} \lambda_j = \mathrm{tr}\Sigma - \sum_{1\le j\le k} \lambda_j \quad \left(\mathrm{tr}\Sigma = \frac{1}{n}\sum_i (f_i - \psi_0)^2\right). \tag{99}$$

This number, $\sigma_k$, measures the error of substitution of the typical (in this sample) field $f$ by its projection on the plane of the $k$ first principal components. The relative average squared error is $\sigma_k^2/\mathrm{tr}\Sigma$.

Among many applications of POD in fluid dynamics at least two have direct relations to the coarse-graining:

- Postprocessing, that is, analysis of an experimentally observed or numerically computed flow regime in projection on the finite-dimensional space of the first principal components;
- Creation of "optimal" Galerkin approximations (Galerkin POD, [106]). In this approach, after finding principal components from sampled images of flow, we project the equations on the first principal components, and receive a reduced model.

In addition to radical and irreversible step from initial equations to Galerkin POD, we can use POD filtering semigroup. It suppresses the component of field orthogonal to selected $k$ first principal components, but makes this reversibly. The filtering semigroup is generated by auxiliary equation

$$\frac{\mathrm{d}f(\eta)}{\mathrm{d}\eta} = \phi(f(\eta)) = -(1 - P_k)(f - \psi_0). \tag{100}$$

The filter transformation in explicit form is

$$U_\eta(f) = \psi_0 + (P_k + \mathrm{e}^{-\eta}(1 - P_k))(f - \psi_0). \tag{101}$$

with explicit reverse transformation $U_{-\eta}$.

For equations of the form (12) $\dot{f} = J(f)$, the POD filtered equations are

$$\frac{\mathrm{d}f}{\mathrm{d}t} = (D_f U_\eta(f))_{U_{-\eta}(f)}(J(U_{-\eta}(f))) = (P_k + \mathrm{e}^{-\eta}(1 - P_k))(J(U_{-\eta}f)). \tag{102}$$

These equations have nonconstant in space coefficients, because $P_k$ is combined from functions $\psi_i$. They are also non-local, because $P_k$ includes integration, but this non-locality appears in the form of several inner products (moments) only. Of course, this approach can be combined with usual filtering, projector operator technic from statistical physics [2], nonlinear Galerkin approximations [107], and non-linear principal manifold approaches [108].

*Main example: the BGK model kinetic equation*

The famous BGK model equation substitutes the Boltzmann equation in all cases when we don't care about exact collision integral (and it is rather often, because usually it is difficult to distinguish our knowledge about exact collision kernel from the full ignorance).

For the one-particle distribution function $f(\boldsymbol{x}, \boldsymbol{v}, t)$ the BGK equation reads:

$$\frac{\partial f(\boldsymbol{x}, \boldsymbol{v}, t)}{\partial t} + \sum_i v_i \frac{\partial f(\boldsymbol{x}, \boldsymbol{v}, t)}{\partial x_i} = \frac{1}{\tau_{\rm col}}(f^*_{m(f)}(\boldsymbol{x}, \boldsymbol{v}) - f(\boldsymbol{x}, \boldsymbol{v}, t)), \qquad (103)$$

where $m(f) = M(t)$ is the cortege of the hydrodynamic fields that corresponds to $f(\boldsymbol{x}, \boldsymbol{v}, t)$, and $f^*_{m(f)}$ is the correspondent local Maxwellian. Let us rescale variables $x, v, t$: we shall measure $x$ in some characteristic macroscopic units $L$, $v$ in units of thermal velocity $v_T$ for a characteristic temperature, $t$ in units $L/v_T$. Of course, there is no exact definition of the "characteristic time" or length, but usually it works if not take it too serious. After rescaling, the BGK equation remains the same, only the parameter becomes dimensionless:

$$\frac{\partial f(\boldsymbol{x}, \boldsymbol{v}, t)}{\partial t} + \sum_i v_i \frac{\partial f(\boldsymbol{x}, \boldsymbol{v}, t)}{\partial x_i} = \frac{1}{\rm Kn}(f^*_{m(f)}(\boldsymbol{x}, \boldsymbol{v}) - f(\boldsymbol{x}, \boldsymbol{v}, t)), \qquad (104)$$

where ${\rm Kn} = l/L$ is the dimensionless Knudsen number (and $l$ is the mean–free–path). It is the small parameter in the kinetics – fluid dynamics transition. If the ${\rm Kn} \gtrsim 1$ then the continuum assumption of fluid mechanics is no longer a good approximation and kinetic equations must be used.

It is worth to mention that the BGK equation is *non-linear*. The term $f^*_{m(f)}$ depends non-linearly on moments $m(f)$, and, hence, on the distribution density $f$ too. And $f^*_{m(f)}$ is the only term in (103) that don't commute with the Laplace operator from the filtering equation (81). All other terms do not change after filtering.

According to (79), in the first order in $\eta$ the filtered BGK equation is

$$\frac{\partial f}{\partial t} + \sum_i v_i \frac{\partial f}{\partial x_i} \qquad (105)$$
$$= \frac{1}{\rm Kn}(f^*_{m(f)} - f) + \frac{\eta}{\rm Kn}(D^2_M f^*_M)_{M=m(f)}(\nabla M, \nabla M)_{M=m(f)}.$$

The last notation may require some explanations: $(D^2_M f^*_M)$ is the second differential of $f^*_M$, for the BGK model equation it is a quadratic form in $R^5$ that parametrically depends on moment value $M = \{M_0, M_1, M_2, M_3, M_4\}$. In the matrix form, the last expression is

$$(D^2_M f^*_M)_{M=m(f)}(\nabla M, \nabla M)_{M=m(f)} \qquad (106)$$
$$= \sum_{r=1}^{3} \sum_{i,j=0}^{4} \left( \frac{\partial^2 f^*_M}{\partial M_i M_j} \right)_{M=m(f)} \frac{\partial M_i}{\partial x_r} \frac{\partial M_j}{\partial x_r}.$$

This expression depends on the macroscopic fields $M$ only. From identity (20) it follows that the filtering term gives no inputs in the quasi-equilibrium approximation, because $m(D^2_M f^*_M) = 0$.

This fact is a particular case of the general *commutation relations* for general quasi-equilibrium distributions. Let a linear operator $\mathbf{B}$ acts in the space of distributions $f$, and there exists such a linear operator $\mathbf{b}$ which acts in the space of macroscopic states $M$ that $m\mathbf{B} = \mathbf{b}m$. Then

$$m(\mathbf{B}f^*_{m(f)} - (D_f f^*_{m(f)})(\mathbf{B}f)) = 0. \tag{107}$$

This means that the macroscopic projection of the Lie bracket for the vector fields of equations $\partial_\eta f = \mathbf{B}f$ (a field $\phi$) and $\partial_t f = f^*_{m(f)} - f$ (a field $\theta$) is zero: $m([\theta, \phi]) = 0$.[5] These commutation relations follow immediately from the self-consistency identities (18), (19), if we use relations $m\mathbf{B} = \mathbf{b}m$ to carry $m$ through $\mathbf{B}$. In the case of BGK equation, relations (107) hold for any linear differential or pseudodifferential operator $\mathbf{B} = Q(\boldsymbol{x}, \partial/\partial\boldsymbol{x})$ that acts on functions of $\boldsymbol{x}$. In this case, $\mathbf{b} = \mathbf{B}$, if we use the same notation for differentiation of functions and of vector-functions.

Relations (107) imply a result that deserves special efforts for physical understanding: the filtered kinetic equations in zero order in the Knudsen number produce the classical Euler equations for filtered hydrodynamic fields without any trace of the filter terms. At the same time, direct filtering of the Euler equation adds new terms.

To obtain the next approximation we need the Chapman–Enskog method for equation (105). We developed a general method for all equations of this type (29), and now apply this method to the filtered BGK equation. Let us take in (26) $\epsilon = \mathrm{Kn}$, $F(f) = F_0(f) + F_{\mathrm{filt}}(f)$, where $F_0 = -\boldsymbol{v}\partial/\partial\boldsymbol{x}$ is the free flight operator and

$$F_{\mathrm{filt}}(f) = \frac{\eta}{\mathrm{Kn}}(D^2_M f^*_M)_{M=m(f)}(\nabla M, \nabla M)_{M=m(f)}. \tag{108}$$

In these notations, for the zero term in the Chapman–Enskog expansion we have $f^{(0)}_M = f^*_M$, and for the first term

$$\begin{aligned} f^{(1)}_M &= f^{\mathrm{NS}}_M + f^{\mathrm{filt}}_M = \Delta^{\mathrm{NS}}_{f^*_M} + \Delta^{\mathrm{filt}}_{f^*_M}, \\ f^{\mathrm{NS}}_M &= \Delta^{\mathrm{NS}}_{f^*_M} = F_0(f^*_M) - (D_M f^*_M)(m(F_0(f^*_M))) \\ f^{\mathrm{filt}}_M &= \Delta^{\mathrm{filt}}_{f^*_M} = F_{\mathrm{filt}}(f) \ \ (\text{because } m(F_{\mathrm{filt}}(f)) = 0), \end{aligned} \tag{109}$$

where NS stands for Navier–Stokes. The correspondent continuum equations (30) are

$$\frac{\mathrm{d}M}{\mathrm{d}t} = m(F_0(f^*_M)) + \mathrm{Kn}\, m(F_0(\Delta^{\mathrm{NS}}_{f^*_M} + \Delta^{\mathrm{filt}}_{f^*_M})). \tag{110}$$

Here, the first term includes non-dissipative terms (the Euler ones) of the Navier–Stokes equations, and the second term includes both the dissipative terms of the Navier–Stokes equations and the filtering terms. Let us collect all the classical hydrodynamic terms together:

[5] The term $-f$ gives zero input in these Lie brackets for any linear operator $\mathbf{B}$.

$$\frac{\partial M(\boldsymbol{x},t)}{\partial t} = \underbrace{\ldots\ldots\ldots\ldots}_{\text{NS terms}} + \mathrm{Kn}\, m\left(\boldsymbol{v}\frac{\partial}{\partial \boldsymbol{x}}F_{\text{filt}}(f)\right) \tag{111}$$

$$= \underbrace{\ldots\ldots\ldots\ldots}_{\text{NS terms}} + \eta m\left(\boldsymbol{v}\frac{\partial}{\partial \boldsymbol{x}}\sum_{r=1}^{3}\sum_{i,j=0}^{4}\frac{\partial^2 f_M^*}{\partial M_i M_j}\frac{\partial M_i}{\partial x_r}\frac{\partial M_j}{\partial x_r}\right),$$

The NS terms here are the right hand sides of the Navier–Stokes equations for the BGK kinetics (45) (with $\tau = 2\tau_{\text{col}} = 2\mathrm{Kn}$). Of course, (111) is one of the tensor viscosity – tensor diffusivity models. Its explicit form for the BGK equation and various similar model equations requires several quadratures:

$$\mathbf{C}_{ij} = m\left(\boldsymbol{v}\frac{\partial^2 f_M^*}{\partial M_i M_j}\right) \tag{112}$$

(for the Maxwell distributions $f_M^*$ that are just Gaussian integrals).

*Entropic stability condition for the filtered kinetic equations*

Instability of filtered equations is a well-known problem. It arises because the reverse filtering is an ill-posed operation, the balance between filter and reverse filter in (76) may be destroyed by any approximation, as well as a perturbation may move the hydrodynamic field out of space of pre-filtered fields. (And the general filtered equations are applicable for sure in that space only.)

Analysis of entropy production is the first tool for stability check. This is a main thermodynamic realization of the Lyapunov functions method (invented in physics before Lyapunov).

The filtration term $F_{\text{filt}}(f)$ (108) in the filtered BGK equation (105) does not produce the Boltzmann (i.e. macroscopic) entropy $S(f^*_{m(f)})$, but is not conservative. In more details:

1. $(D_M S(f_M^*))(m(F_{\text{filt}}(f^*_{m(f)}))) \equiv 0$, because $m(F_{\text{filt}}(f)) \equiv 0$;
2. $(D_f S(f))_{f^*_{m(f)}}(F_{\text{filt}}(f^*_{m(f)})) = (D_M S(f_M^*))(m(F_{\text{filt}}(f^*_{m(f)}))) \equiv 0$;
3. $(D_f S(f))_{f^*_{m(f)}}(F_{\text{filt}}(f)) = (D_f S(f))_{f^*_{m(f)}}(F_{\text{filt}}(f^*_{m(f)})) \equiv 0$, because $F_{\text{filt}}(f)$ depends on $f^*_{m(f)}$ only;
4. But for any field $F_{\text{filt}}(f)$ that depends on $f^*_{m(f)}$ only, if the conservativity identity (32) $(D_f S(f))_f(F_{\text{filt}}(f)) \equiv 0$ is true even in a small vicinity of quasi-equilibria, then $F_{\text{filt}}(f) \equiv 0$. Hence, the non-trivial filter term $F_{\text{filt}}(f)$ cannot be conservative, the whole field $F(f) = F_0(f) + F_{\text{filt}}(f)$ is not conservative, and we cannot use the entropy production formula (31).

Instead of (31) we obtain

$$\frac{\mathrm{d}S(M)}{\mathrm{d}t} = \mathrm{Kn}\langle \Delta^{\text{NS}}_{f_M^*}, \Delta^{\text{NS}}_{f_M^*}\rangle_{f_M^*} + \eta\langle \Delta^{\text{NS}}_{f_M^*}, \Delta^{\text{filt}}_{f_M^*}\rangle_{f_M^*}. \tag{113}$$

The *entropic stability condition* for the filtered kinetic equations is:

$$\mathrm{d}S(M)/\mathrm{d}t \geq 0, \text{ i.e. } \mathrm{Kn}\langle \Delta^{\mathrm{NS}}_{f^*_M}, \Delta^{\mathrm{NS}}_{f^*_M}\rangle_{f^*_M} + \eta\langle \Delta^{\mathrm{NS}}_{f^*_M}, \Delta^{\mathrm{filt}}_{f^*_M}\rangle_{f^*_M} \geq 0. \tag{114}$$

There exists a plenty of convenient sufficient conditions, for example,

$$\eta \leq \mathrm{Kn}\frac{|\langle \Delta^{\mathrm{NS}}_{f^*_M}, \Delta^{\mathrm{filt}}_{f^*_M}\rangle_{f^*_M}|}{\langle \Delta^{\mathrm{NS}}_{f^*_M}, \Delta^{\mathrm{NS}}_{f^*_M}\rangle_{f^*_M}}; \quad \text{or} \quad \eta \leq \mathrm{Kn}\sqrt{\frac{\langle \Delta^{\mathrm{filt}}_{f^*_M}, \Delta^{\mathrm{filt}}_{f^*_M}\rangle_{f^*_M}}{\langle \Delta^{\mathrm{NS}}_{f^*_M}, \Delta^{\mathrm{NS}}_{f^*_M}\rangle_{f^*_M}}}. \tag{115}$$

The upper boundary for $\eta$ that guaranties stability of the filtered equations is proportional to Kn. For the Gaussian filter width $\Delta$ this means $\Delta = L\sqrt{24\eta} \sim \sqrt{\mathrm{Kn}}$ (where $L$ is the characteristic macroscopic length). This scaling, $\Delta/L \sim \sqrt{\mathrm{Kn}}$, was discussed in [18] for moment kinetic equations because different reasons: if $\Delta/L \gg \sqrt{\mathrm{Kn}}$ then the Chapman–Enskog procedure is not applicable, and, moreover, the continuum description is probably not valid, because the filtering term with large coefficient $\eta$ violates the conditions of hydrodynamic limit. This important remark gives the frame for $\eta$ scaling, and (114), (115) give the stability boundaries inside this scale.

## 4 Errors of Models, $\varepsilon$-trajectories and Stable Properties of Structurally Unstable Systems

### 4.1 Phase Flow, Attractors and Repellers

*Phase flow*

In this section, we return from kinetic systems to general dynamical systems, and lose such specific tools as entropy and quasi-equilibrium. Topological dynamics gives us a natural language for general discussion of limit behavior and relaxation of general dynamical systems [109]. We discuss a general dynamical system as a semigroup of homeomorphisms (phase flow transformations): $\Theta(t, x)$ is the result of shifting point $x$ in time $t$.

Let the phase space $X$ be a compact metric space with the metrics $\rho$,

$$\Theta : [0, \infty[ \times X \to X \tag{116}$$

be a continuous mapping for any $t \geq 0$; let mapping $\Theta(t, \cdot) : X \to X$ be homeomorphism of $X$ into subset of $X$ and let these homeomorphisms form a one-parameter semigroup:

$$\Theta(0, \cdot) = \mathrm{id}, \ \Theta(t, \Theta(t', x)) = \Theta(t + t', x) \tag{117}$$

for any $t, t' \geq 0, x \in X$.

Below we call the semigroup of mappings $\Theta(t, \cdot)$ a *semiflow of homeomorphisms* (or, for short, semiflow), or simply system (116). We assume that the continuous map $\Theta(t, x)$ is continued to negative time $t$ as far as it is possible with preservation of the semigroup property (117). For phase flow we use also notations $\Theta_t$ and $\Theta_t(x)$. For any given $x \in X$, *x-motion* is a function of time $\Theta(t, x)$, $x$-motion is the *whole motion* if the function is defined on the whole axis $t \in ]-\infty, \infty[$. The image of $x$-motion is the *x-trajectory*.

*Attractors and repellers*

First of all, for the description of limit behaviour we need a notion of an $\omega$-limit set.

A point $p \in X$ is called $\omega$- ($\alpha$-)*-limit point* of the $x$-motion (correspondingly of the whole $x$-motion), if there is such a sequence $t_n \to \infty$ ($t_n \to -\infty$) that $\Theta(t_n, x) \to p$ as $n \to \infty$. The totality of all $\omega$- ($\alpha$-)-limit points of $x$-motion is called its $\omega$- ($\alpha$-)*-limit set* and is denoted by $\omega(x)$ ($\alpha(x)$).

A set $W \subset X$ is called *invariant* set, if, for any $x \in W$, the $x$-motion is whole and the whole $x$-trajectory belongs to $W$.

The sets $\omega(x)$, $\alpha(x)$ (the last in the case when $x$-motion is whole) are nonempty, closed, connected, and invariant.

The set of all $\omega$-limit points of the system $\omega_\Theta = \bigcup_{x \in X} \omega(x)$ is nonempty and invariant, but may be disconnected and not closed. The sets $\omega(x)$ might be considered as attractors, and the sets $\alpha(x)$ as repellers (attractors for $t \to -\infty$). The system of these sets represents all limit behaviours of the phase flow.

Perhaps, the most constructive idea of attractor definition combines pure topological (metric) and measure points of view. A *weak attractor* [113] is a closed (invariant) set $A$ such that the set $\mathcal{B}(A) = \{x \mid \omega(x) \subset A\}$ (a basin of attraction) has strictly positive measure. A *Milnor attractor* [112] is such a weak attractor that there is no strictly smaller closed $A' \subsetneqq A$ so that $\mathcal{B}(A)$ coincides with $\mathcal{B}(A')$ up to a set of measure zero. If $A$ is a Milnor attractor and for any closed invariant proper subset $A' \subsetneqq A$ the set $\mathcal{B}(A')$ has zero measure, then we say that $A$ is a *minimal Milnor attractor*.

Below in this section we follow a purely topological (metric) point of view, but keep in mind that its combination with measure–based ideas create a richer theory.

*The dream of applied dynamics*

Now we can formulate the "dream of applied dynamics." There is such a finite number of invariant sets $A_1, \ldots A_n$ that:

- Any attractor or repeller is one of the $A_i$;
- The following relation between sets $A_1, \ldots A_n$ is acyclic: $A_i \succeq A_j$ if there exists such $x$ that $\alpha(x) = A_i$ and $\omega(x) = A_j$;
- The system $A_1, \ldots A_n$ with the preorder $A_i \succeq A_j$ does not change qualitatively under sufficiently small perturbations of the dynamical system: all the picture can be restored by a map that is close to id.

For generic two-dimensional systems this dream is the reality: there is a finite number of fixed points and closed orbits such that any motion goes to one of them at $t \to \infty$, and to another one at $t \to -\infty$ for a whole motion.

The multidimensional analogues of generic two-dimensional systems are the Morse–Smale systems. For them all attractors and repellers are fixed points or closed orbits. The relation $A_i \succeq A_j$ for them is the *Smale order*.

But the class of the Morse–Smale systems is too narrow: there are many systems with more complicated attractors, and some of these systems are structurally stable and do not change qualitatively after sufficiently small perturbations.[6] It is necessary to take into account that typically some of motions have smaller attractors (for example, in $A_i$ exists a dense set of closed orbits), and $\omega(x) = A_j$ not for all, but for almost all $x$. Finally, the "dream of applied dynamics" was destroyed by S. Smale [20]. He demonstrated that "structurally stable systems are not dense." It means that even the last item of this dream contradicts the multidimensional reality.

### 4.2 Metric Coarse-Graining by $\varepsilon$-motions

*$\varepsilon$-motions*

The observable picture must be structurally stable. Any real system exists under the permanent perturbing influence of the external world. It is hardly possible to construct a model taking into account all such perturbations. Besides that, the model describes the internal properties of the system only approximately. The discrepancy between the real system and the model arising from these two circumstances is different for different models. So, for the systems of celestial mechanics it can be done very small. Quite the contrary, for chemical engineering this discrepancy can be if not too large but not such small to be neglected. Structurally unstable features or phase portrait should be destroyed by such an unpredictable divergence of the model and reality. The perturbations "conceal" some fine details of dynamics, therefore these details become irrelevant to analysis of real systems.

There are two traditional approaches to the consideration of perturbed motions. One of them is to investigate the motion in the presence of small sustained perturbations [119, 120, 122], the other is the study of fluctuations under the influence of small stochastic perturbations [32, 33]. In this section, we join mainly the first direction.

A small unpredictable discrepancy between the real system and the dynamical model can be simulated by periodical "fattening." For a set $A \subset X$ its $\varepsilon$-fattening is the set

$$A_\varepsilon = \{x \mid \rho(x, y) < \varepsilon \text{ for all } y \in A\}. \tag{118}$$

Instead of one $x$-motion we consider motion of a set, $A(t) = \Theta_t A$, and combine this motion with periodical $\varepsilon$-fattening for a given period $\tau$. For superposition of $\Theta_\tau$ with $\varepsilon$-fattening we use the notation $\mathbf{\Theta}_\tau^\varepsilon$:

$$\mathbf{\Theta}_\tau^\varepsilon A = (\Theta_\tau A)_\varepsilon \tag{119}$$

For $t \in [n\tau, (n+1)\tau[$ We need to generalize this definition for $t \in [n\tau, (n+1)\tau[$:

[6] Review of modern dynamics is presented in [110, 111]

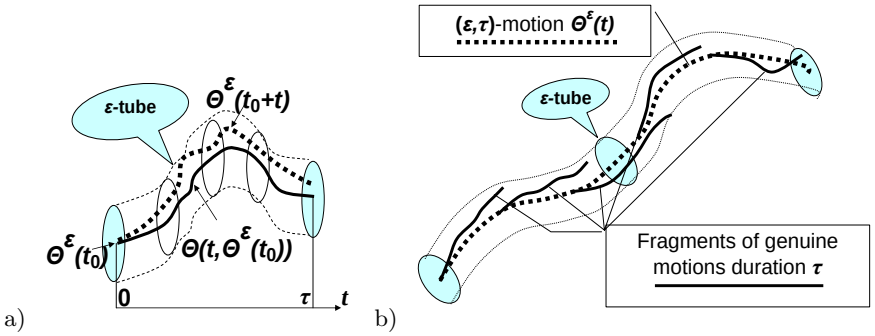



Fig. 12: An $(\varepsilon,\tau)$-motion $\Theta^\varepsilon(t_0+t)$ $(t\in[0,\tau])$ in the $\varepsilon$-tube near a genuine motion $\Theta(t,\Theta^\varepsilon(t_0))$ $(t\in[0,\tau])$ duration $\tau$ (a), and an $(\varepsilon,\tau)$-motion $\Theta^\varepsilon(t)$ with fragments of genuine motions duration $\tau$ in the $\varepsilon$-tube near $\Theta^\varepsilon(t)$ (b).

$$\mathbf{\Theta}^\varepsilon_t A = \Theta_{t-n\tau}((\mathbf{\Theta}^\varepsilon_\tau)^n A). \tag{120}$$

Analysis of these motions of sets gives us the information about dynamics with $\varepsilon$-uncertainty in model. Single-point sets are natural initial conditions for such motions.

One can call this coarse-graining the *metric coarse-graining*, and the Erenfest's coarse-graining for dynamics of distribution function might be called the *measure coarse-graining*. The concept of *metric–measure spaces* (*mm*-spaces [123]) gives the natural framework for analysis of various sorts of coarse-graining.

It is convenient to introduce individual $\varepsilon$-motions. A function of time $\Theta^\varepsilon(t)$ with values in $X$, defined at $t\geq 0$, is called $(\varepsilon,x)$-motion $(\varepsilon>0)$, if $\Theta^\varepsilon(0)=x$ and for any $t_0\geq 0$, $t\in[0,\tau]$ the inequality $\rho(\Theta^\varepsilon(t_0+t),\Theta(t,\Theta^\varepsilon(t_0)))<\varepsilon$ holds. In other words, if for an arbitrary point $\Theta^\varepsilon(t_0)$ one considers its motion due to phase flow of dynamical system, this motion will diverge $\Theta^\varepsilon(t_0+t)$ from no more than at $\varepsilon$ for $t\in[0,\tau]$. Here $[0,\tau]$ is a certain interval of time, its length $\tau$ is not very important (it is important that it is fixed), because later we shall consider the case $\varepsilon\to 0$. For a given $\tau$ we shall call the $(\varepsilon,x)$-motion $(\varepsilon,x,\tau)$-motion when reference to $\tau$ is necessary. On any interval $[t_0,t_0+\tau]$ an $(\varepsilon,x,\tau)$-motion deviates from a genuine motion not further than on distance $\varepsilon$ if these motions coincide at time moment $t_0$ (Fig. 12a). If a genuine motion starts from a point of an $(\varepsilon,x,\tau)$-trajectory, it remains in the $\varepsilon$-tube near that $(\varepsilon,\tau)$-motion during time $\tau$ (Fig. 12b).

*Limit sets of $\varepsilon$-motions*

Let us study the limit behaviour of the coarse-grained trajectories $\mathbf{\Theta}^\varepsilon_t A$, and than take the limit $\varepsilon\to 0$. For systems with complicated dynamics, this limit may differ significantly from the limit behaviour of the original system for $\varepsilon=0$. This effect of the perturbation influence in the zero limit is a "smile of

a Cheshire cat:" the cat tends to disappear, leaving only its smile hanging in the air.

For any $\Theta^\varepsilon(t)$ the $\omega$-limit set $\omega(\Theta^\varepsilon)$ is the set of all limit points of $\Theta^\varepsilon(t)$ at $t \to \infty$. For any $x \in X$ a set $\omega^\varepsilon(x)$ is a totality of all $\omega$-limit points of all $(\varepsilon, x)$-motions:

$$\omega^\varepsilon(x) = \bigcup_{\Theta^\varepsilon(0)=x} \omega(\Theta^\varepsilon).$$

For $\varepsilon \to 0$ we obtain the set

$$\omega^0(x) = \bigcap_{\varepsilon>0} \omega^\varepsilon(x).$$

Firstly, it is necessary to notice that $\omega^\varepsilon(x)$ does not always tend to $\omega(x)$ as $\varepsilon \to 0$: the set $\omega^0(x)$ may not coincide with $\omega(x, k)$.

The sets $\omega^0(x)$ are closed and invariant. Let $x \in \omega^0(x)$. Then for any $\varepsilon > 0$ there exists periodical $(\varepsilon, x)$-motion (This is a version of Anosov's $C^0$-closing lemma [114, 111]).

The function $\omega^0(x)$ is *upper semicontinuous*. It means that for any sequence $x_i \to x$ all limit points of all sequences $y_i \in \omega^0(x_i)$ belong to $\omega^0(x)$.

In order to study the limit behaviour for all initial conditions, let us join all $\omega^0(x)$:

$$\omega^0 = \bigcup_{x\in X} \omega^0(x) = \bigcup_{x\in X} \bigcap_{\varepsilon>0} \omega^\varepsilon(x) = \bigcap_{\varepsilon>0} \bigcup_{x\in X} \omega^\varepsilon(x). \tag{121}$$

The set $\omega^0$ is closed and invariant. If $y \in \omega^0$ then $y \in \omega^0(y)$. If $Q \subset \omega^0$ and $Q$ is connected, then $Q \subset \omega^0(y)$ for any $y \in Q$.[7]

The $\varepsilon$-motions were studied earlier in differential dynamics, in connection with the theory of Anosov about $\varepsilon$-trajectories and its applications [114, 115, 116, 117, 118]. For systems with hyperbolic attractors an important *$\varepsilon$-motion shadowing property* was discovered: for a given $\eta > 0$ and sufficiently small $\varepsilon > 0$ for any $\varepsilon$-motion $\Theta^\varepsilon(t)$ there exists a motion of the non-perturbed system $\Theta(t, x)$ that belongs to $\eta$ - neighborhood of $\Theta^\varepsilon(t)$:

$$\rho(\Theta^\varepsilon(\phi(t)), \Theta(t, x)) < \eta,$$

for $t > 0$ and some monotonous transformation of time $\phi(t)$ $(t-\phi(t) = O(\varepsilon t))$. The sufficiently small coarse-graining changes nothing in dynamics of systems with this shadowing property, because any $\varepsilon$-motion could be approximated uniformly by genuine motions on the whole semiaxis $t \in [0, \infty[$.

*Preorder and equivalence generated by dynamics*

Let $x_1, x_2 \in X$. Let us say $x_1 \gtrsim_\Theta x_2$ if for any $\varepsilon > 0$ there exists such a $(\varepsilon, x_1)$-motion $\Theta^\varepsilon(t)$ $(\Theta^\varepsilon(0) = x_1)$ that $\Theta^\varepsilon(t_0) = x_2$ for some $t_0 \geq 0$.

[7] For all proofs here and below in this section we address to [22, 23].

Let $x_1, x_2 \in X$. Say that points $x_1$ and $x_2$ are $\Theta$-*equivalent* (denotation $x_1 \sim_\Theta x_2$), if $x_1 \succsim_\Theta x_2$ and $x_2 \succsim_\Theta x_1$.

The relation $\succsim_\Theta$ is a closed $\Theta$-invariant preorder relation on $X$:

- It is reflexive: $x \succsim_\Theta x$ for all $x \in X$;
- It is transitive: $x_1 \succsim_\Theta x_2$ and $x_2 \succsim_\Theta x_3$ implies $x_1 \succsim_\Theta x_3$;
- The set of pairs $(x_1, x_2)$, for which $x_1 \sim_\Theta x_2$ is closed in $X$;
- If $x_1 \succsim_\Theta x_2$ then $\Theta(t, x_1) \succsim_\Theta \Theta(t, x_2)$ for any $t > 0$.

The necessary and sufficient conditions for the preorder $\succsim_\Theta$ relation are as follows: $x_1 \succsim_\Theta x_2$ if and only if either $x_2 \in \omega^0(x_1)$ or $x_2 = \Theta(t, x_1)$ for some $t \geq 0$. Therefore,

$$\omega^0(x) = \{y \in \omega^0 \mid x \succsim_\Theta y\} \tag{122}$$

The relation $\sim_\Theta$ is a closed $\Theta$-invariant equivalence relation:

- The set of pairs $(x_1, x_2)$, for which $x_1 \sim_\Theta x_2$ is closed in $X$;
- If $x_1 \sim x_2$ and $x_1 \neq x_2$, then $x_1$- and $x_2$-motions are whole and $\sim_\Theta \Theta(t, x_2)$ for any $t \in ]-\infty, \infty[\ \Theta(t, x_1)$.

If $x_1 \neq x_2$, then $x_1 \sim_\Theta x_2$ if and only if $\omega^0(x_1) = \omega^0(x_2)$ , $x_1 \in \omega^0(x_1)$, and $x_2 \in \omega^0(x_2)$.

Compare with [32], where analogous theorems are proved for relations defined by action functional for randomly perturbed dynamics.

*The coarsened phase portrait*

We present the results about the coarsened phase portrait as a series of theorems.

Let us remind, that topological space is called *totally disconnected* if there exist a base of topology, consisting of sets which are simultaneously open and closed. Simple examples of such spaces are discrete space and Cantor's discontinuum. In a totally disconnected space all subsets with more than one element are disconnected. Due to the following theorem, in the coarsened phase portrait we have a totally disconnected space instead of finite set of attractors mentioned in the naive dream of applied dynamics.

**Theorem 1.** *The quotient space $\omega^0/\sim_\Theta$ is compact and totally disconnected.*

The space $\omega^0/\sim_\Theta$ with the factor-relation $\succsim_\Theta$ on it is the *generalized Smale diagram* with the *generalized Smale order* on it [22, 23].

Attractors and basins of attraction are the most important parts of a phase portrait. Because of (122), all attractors are *saturated downwards*. The set $Y \subset \omega^0$ is saturated downwards, if for any $y \in Y$,

$$\{x \in \omega^0 \mid y \succsim_\Theta x\} \subset Y.$$

Every saturated downwards subset in $\omega^0$ is saturated also for the equivalence relation $\sim_\Theta$ and includes with any $x$ all equivalent points. The following theorem states that coarsened attractors $Y$ (open in $\omega^0$ saturated downwards subsets of $\omega^0$) have open coarsened basins of attraction $\mathcal{B}^0(Y)$.

**Theorem 2.** *Let $Y \subset \omega^0$ be open (in $\omega^0$) saturated downwards set. Then the set $\mathcal{B}^0(Y) = \{x \in X \mid \omega^0(x) \subset Y\}$ is open in $X$.*

There is a natural expectation that $\omega$-limit sets can change by jumps on boundaries of basins of attraction only. For the coarsened phase portrait it is true.

**Theorem 3.** *The set $B$ of all points of discontinuity of the function $\omega^0(x)$ is the subset of first category in $X$.*[8] *If $x \in B$ then $\Theta(t, x) \in B$ for all $t$ when $\Theta(t, x)$ is defined.*

**Theorem 4.** *Let $x \in X$ be a point of discontinuity of the function $\omega^0(x)$. Then there is such open in $\omega^0$ saturated downwards set $W$ that $x \in \partial\mathcal{B}^0(W)$.*

The function $\omega^0(x)$ is upper semicontinuous, hence, in any point $x^*$ of its discontinuity the *lower semicontinuity* is broken: there exist a point $y^* \in \omega^0(x^*)$, a number $\eta > 0$, and a sequence $x_i \to x^*$ such that

$$\rho(y^*, y) > \eta \text{ for any } y \in \omega^0(x_i) \text{ and all } i.$$

The classical Smale order for hyperbolic systems was defined on a finite totality $A_1, \dots A_n$ of basic sets that are closed, invariant, and transitive (i.e. containing a dense orbit). $A_i \succ A_j$ if there exists such $x \in X$ that $x$-trajectory is whole, $\alpha(x) \subset A_i$, $\omega(x) \subset A_j$. Such special trajectories exist in the general case of coarsened dynamical system also.

**Theorem 5.** *Let $X$ be connected, $\omega^0$ be disconnected. Then there is such $x \in X$ that $x$-motion is whole and $x \notin \omega^0$. There is also such partition of $\omega^0$ that*

$$\omega^0 = W_1 \cup W_2,\ W \cap W_2 = \varnothing,\ \alpha_f(x) \subset W_1,\ \omega^0(x) \subset W_2,$$

*and $W_{1,2}$ are open and, at the same time, closed subsets of $\omega^0$ (it means that $W_{1,2}$ are preimages of open–closed subsets of the quotient space $\omega^0/\sim_\Theta$).*

This theorem can be applied, by descent, to connected closures of coarsened basins of attraction $\mathcal{B}^0(Y)$ (see Theorem 2).

Theorems 1–5 give us the picture of coarsened phase portrait of a general dynamical system, and this portrait is qualitatively close to phase portraits of structurally stable systems: rough 2D systems, the Morse–Smale systems and the hyperbolic Smale systems. For proofs and some applications we address to [22, 23].

[8] A set of first category, or a meagre set is a countable union of nowhere dense sets. In a complete metric space a complement of a meagre set is dense (the Baire theorem).

*Stability of the coarsened phase portrait under smooth perturbations of vector fields*

In order to analyze stability of this picture under the perturbation of the vector field (or the diffeomorphism, for discrete time dynamics) it is necessary to introduce $C^k$ $\varepsilon$-fattening in the space of smooth vector fields instead of periodic $\varepsilon$-fattening of phase points. We shall discuss a $C^k$-smooth dynamical system $\Theta$ on a compact $C^m$-manifold $M (0 \le k \le m)$. Let $\Theta_t$ be the semigroup of phase flow transformations (shifts in time $t \ge 0$) and $U_\varepsilon(\Theta)$ be the set of phase flows that corresponds to a closed $\varepsilon$-neighborhood of system $\Theta_t$ in the $C^k$-norm topology of vector fields. The positive semi-trajectory of phase point $x$ is a set $\Theta(x) = \{\Theta_t(x) \colon t \ge 0\}$. The $C^k$ $\varepsilon$-fattened semi-trajectory is $\Theta^\varepsilon(x) = \bigcup_{\Phi \in U_\varepsilon(\Theta)} \Phi(x)$. Let us take this set with all limits for $t \to \infty$. It is the closure $\overline{\Theta^\varepsilon(x)}$. After that, let us take the limit $\varepsilon \to 0$: $P_x = \bigcap_{\varepsilon > 0} \overline{\Theta^\varepsilon(x)}$ (it is an analogue of $\Theta(x) \cup \omega_0(x)$ from our previous consideration for general dynamical systems). Following [21] let us call this set $P_x$ a *prolongation* of the semi-trajectory $\Theta(x)$.

A trajectory of a dynamical system is said to be *stable under $C^k$ constantly-acting perturbations* if its prolongation is equal to its closure: $P_x = \overline{\Theta(x)}$

For a given dynamical system let $L(\Theta)$ denote the union of all trajectories that are stable in the above sense and let $\mathbb{L}_1$ be the set of all dynamical systems $\Theta$ for which $L(\Theta)$ is dense in phase space: $\overline{L(\Theta)} = M$. All structurally stable systems belong to $\mathbb{L}_1$. The main result of [21] is as follows:

**Theorem 6.** *The set $\mathbb{L}_1$ is a dense $\mathbb{G}_\delta$ in the space of $C^k$ dynamical systems with the $C^k$ norm.*[9]

So, for almost all smooth dynamical systems almost all trajectories are stable under smooth constantly-acting perturbations: this type of stability is typical.

## 5 Conclusion

Two basic ideas of coarse-graining are presented. In the Ehrenfests' inspired approach the dynamics of distributions with averaging is studied. In the metric approach the starting point of analysis is dynamics of sets with periodical $\varepsilon$-fattening.

The main question of the Ehrenfests' coarse-graining is: where should we take the coarse-graining time $\tau$? There are two limit cases: $\tau \to 0$ and $\tau \to \infty$ (physically, $\infty$ here means the time that exceeds all microscopic time scales). The first limit, $\tau \to 0$, returns us to the quasi-equilibrium approximation. The

[9] In a topological space a $\mathbb{G}_\delta$ set is a countable intersection of open sets. A complement of a dense $\mathbb{G}_\delta$ set is a countable union of nowhere dense sets. It is a set of first category, or a meagre set.

second limit is, in some sense, exact (if it exists). Some preliminary steps in the study of this limit are made in [74, 75, 4]. On this way, the question about proper values of the Prandtl number, as well, as many other similar questions about kinetic coefficients, has to be solved.

The constructed family of chains between conservative (with the Karlin–Succi involution) and maximally dissipative (with Ehrenfests' projection) ones give us a possibility to model hydrodynamic systems with various dissipation (viscosity) coefficients that are decoupled with time steps. The *collision integral* is successfully substituted by combinations of the involution and projection.

The direct descendant of the Ehrenfests' coarse-graining, the kinetic approach to filtering of continuum equations, seems to be promising and physically reasonable: if we need to include the small eddies energy into internal energy, let us lift the continuum mechanics to kinetics where all the energies live together, make there the necessary filtering, and then come back. Two main questions: when the obtained filtered continuum mechanics is stable, and when there is way back from filtered kinetics to continuum mechanics, have unexpectedly the similar answer: the filter width $\Delta$ should be proportional to the square root of the Knudsen number. The coefficient of this proportionality is calculated from the entropic stability conditions.

The metric coarse-graining by $\varepsilon$-motions in the limit $\varepsilon \to 0$ gives the stable picture with the totally disconnected system of basic sets that form sources and sinks structure in the phase space. Everything looks nice, but now we need algorithms for effective computation and representation of the coarsened phase portrait even in modest dimensions 3-5 (for discrete time systems in dimensions 2-4).

It is necessary to build a bridge between theoretical topological picture and applied computations. In some sense, it is the main problem of modern theory of dynamical systems to develop language and tools for constructive analysis of arbitrary dynamics. Of course, the pure topological point of view is unsufficient, and we need an interplay between measure and topology of dynamical systems, perhaps, with inclusion of some physical and probabilistic ideas.

*Acknowledgement.* A couple of years ago, Wm. Hoover asked me to explain clearly the difference between various types of coarse-graining. This paper is the first attempt to answer. I am grateful to H.C. Öttinger, and L. Tatarinova for scientific discussion and collaboration. Long joint work with I.V. Karlin was very important for my understanding of the Ehrenfests' coarse-graining and lattice Boltzmann models.

## References


1. P. Ehrenfest, T. Ehrenfest-Afanasyeva: The Conceptual Foundations of the Statistical Approach in Mechanics, In: *Mechanics Enziklopädie der Mathematischen Wissenschaften*, vol. 4. (Leipzig 1911). Reprinted: P. Ehrenfest, T. Ehrenfest-Afanasyeva, The Conceptual Foundations of the Statistical Approach in Mechanics (Dover, Phoneix 2002)
2. H. Grabert: *Projection operator techniques in nonequilibrium statistical mechanics* (Springer, Berlin Heidelberg New York 1982)
3. A.N. Gorban, I.V. Karlin, A.Yu. Zinovyev: Constructive methods of invariant manifolds for kinetic problems. *Phys. Reports* **396**, 197–403 (2004) Preprint online: `http://arxiv.org/abs/cond-mat/0311017`.
4. A.N. Gorban, I.V. Karlin: *Invariant manifolds for physical and chemical kinetics*, Lect. Notes Phys., vol. 660 (Springer, Berlin, Heidelberg, New York 2005)
5. A.N. Gorban, I.V. Karlin, P. Ilg, H.C. Öttinger: Corrections and enhancements of quasi–equilibrium states. *J. Non–Newtonian Fluid Mech.* **96**, 203–219 (2001)
6. K.G. Wilson, J. Kogut: The renormalization group and the $\epsilon$-expansion. *Phys. Reports* **12C**, 75–200 (1974)
7. O. Pashko, Y. Oono: The Boltzmann equation is a renormalization group equation. *Int. J. Mod. Phys. B* **14**, 555–561 (2000)
8. Y. Hatta, T. Kunihiro: Renormalization group method applied to kinetic equations: roles of initial values and time. *Annals Phys.* **298**, 24–57 (2002)
9. I.G. Kevrekidis, C.W. Gear, J.M. Hyman, P.G. Kevrekidis, O. Runborg, C. Theodoropoulos: Equation-free, coarse-grained multiscale computation: enabling microscopic simulators to perform system-level analysis. *Comm. Math. Sci.* **1**, 715–762 (2003)
10. A.J. Chorin, O.H. Hald, R. Kupferman: Optimal prediction with memory. *Physica D* **166**, 239–257 (2002)
11. A.N. Gorban, I.V. Karlin, H.C. Öttinger, L.L. Tatarinova: Ehrenfests' argument extended to a formalism of nonequilibrium thermodynamics. *Phys.Rev.E* **63**, 066124 (2001)
12. A.N. Gorban, I.V. Karlin: Uniqueness of thermodynamic projector and kinetic basis of molecular individualism. *Physica A* **336**, 391–432 (2004)
13. J. Leray: Sur les movements dun fluide visqueux remplaissant lespace. *Acta Mathematica* **63**, 193-248 (1934)
14. J. Smagorinsky: General Circulation Experiments with the Primitive Equations: I. The Basic Equations. *Mon. Weather Rev.* **91**, 99–164 (1963)
15. M. Germano: Turbulence: the filtering approach. *J. Fluid Mech.* **238**, 325–336 (1992)
16. D. Carati, G.S. Winckelmans, H. Jeanmart: On the modelling of the subgrid-scale and filtered-scale stress tensors in large-eddy simulation. *J. Fluid Mech.* **441**, 119–138 (2001)
17. M. Lesieur, O. Métais, P. Comte: *Large-Eddy Simulations of Turbulence* (Cambridge University Press 2005)
18. S. Ansumali, I.V. Karlin, S. Succi: Kinetic Theory of Turbulence Modeling: Smallness Parameter, Scaling and Microscopic Derivation of Smagorinsky Model. *Physica A*, **338**, 379–394 (2004)
19. S. Succi: *The lattice Boltzmann equation for fluid dynamics and beyond* (Clarendon Press, Oxford 2001)

20. S. Smale: Structurally stable systems are not dense, *Amer. J. Math.* **88**, 491–496 (1966)
21. V.A. Dobrynskiĭ, A.N. Sharkovskiĭ: Genericity of the dynamical systems almost all orbits of which are stable under sustained perturbations. *Soviet Math. Dokl.* **14**, 997–1000 (1973)
22. A.N. Gorban: Slow relaxations and bifurcations of omega-limit sets of dynamical systems, PhD Thesis in Physics & Math. (Differential Equations & Math.Phys), Kuibyshev, Russia (1980)
23. A.N. Gorban: Singularities of Transition Processes in Dynamical Systems: Qualitative Theory of Critical Delays, *Electronic Journal of Differential Equations*, Monograph **05**, (2004) `http://ejde.math.txstate.edu/Monographs/05/abstr.html` (Includes English translation of [22].)
24. B.A. Huberman, W.F. Wolff: Finite precision and transient behavior. *Phys. Rev. A* **32**, 3768–3770 (1985)
25. C. Beck, G. Roepsorff: Effects of phase space discretization on the long–time behavior of dynamical systems. *Physica D* **25**, 173–180 (1987)
26. C. Grebogi, E. Ott, J.A. Yorke: Roundoff–induced periodicity and the correlation dimension of chaotic attractors. *Phys. Rev. A* **38**, 3688–3692 (1988)
27. P. Diamond, P. Kloeden, A. Pokrovskii, A. Vladimirov: Collapsing effect in numerical simulation of a class of chaotic dynamical systems and random mappings with a single attracting centre. *Physica D* **86**, 559–571 (1995)
28. L. Longa, E.M.F. Curado, A. Oliveira: Rounoff–induced coalescence of chaotic trajectories. *Phys. Rev. E* **54**, R2201–R2204 (1996)
29. P.-M. Binder, J.C. Idrobo: Invertibility of dynamical systems in granular phase space. *Phys. Rev. E* **58**, 7987–7989 (1998)
30. C. Dellago, Wm.G. Hoover: Finite–precision stationary states at and away from equilibrium. *Phys. Rev. E* **62**, 6275–6281 (2000)
31. B. Bollobas: *Random Graphs*, Cambridge Studies in Advanced Mathematics (Cambridge University Press 2001)
32. M.I. Freidlin, A.D. Wentzell: *Random Perturbations of Dynamical Systems*, Grundlehren der mathematischen Wissenschaften, vol. 260 (Springer, Berlin, Heidelberg, New York 1998)
33. L. Arnold: *Random Dynamical Systems*, Springer Monographs in Mathematics, vol. 16, (Springer, Berlin, Heidelberg, New York 2002)
34. A.N. Gorban: *Equilibrium encircling. Equations of chemical kinetics and their thermodynamic analysis* (Nauka, Novosibirsk 1984)
35. A.N. Gorban, V.I. Bykov, G.S. Yablonskii: *Essays on chemical relaxation* (Nauka, Novosibirsk 1986)
36. G.S.Yablonskii, V.I.Bykov, A.N. Gorban, V.I.Elokhin: *Kinetic Models of Catalytic Reactions*, Series Comprehensive Chemical Kinetics, vol. 32, ed. by R.G. Compton, (Elsevier, Amsterdam 1991)
37. Y.B. Zeldovich: Proof of the Uniqueness of the Solution of the Equations of the Law of Mass Action, In: *Selected Works of Yakov Borisovich Zeldovich,* ed. by J.P. Ostriker, vol. 1, 144–148 (Princeton University Press, Princeton 1996)
38. P.L. Bhatnagar, E.P. Gross, M. Krook: A model for collision processes in gases. I. Small amplitude processes in charged and neutral one–component systems. *Phys. Rev.*, **94**, 511–525 (1954)
39. A.N. Gorban, I.V. Karlin: General approach to constructing models of the Boltzmann equation. *Physica A*, **206**, 401–420 (1994)

40. D. Hilbert: Begründung der kinetischen Gastheorie. *Math. Annalen* **72**, 562–577 (1912)
41. D. Ruelle: Smooth Dynamics and New Theoretical Ideas in Nonequilibrium Statistical Mechanics. *J. Stat. Phys.* **95**, 393-468 (1999)
42. S. Kullback: *Information theory and statistics* (Wiley, New York 1959)
43. A.N. Gorban, I.V. Karlin: Family of additive entropy functions out of thermodynamic limit. *Phys. Rev. E* **67**, 016104 (2003)
44. P. Gorban: Monotonically equivalent entropies and solution of additivity equation. *Physica A* **328**, 380-390 (2003)
45. S. Abe, Y. Okamoto (Eds.), *Nonextensive statistical mechanics and its applications* (Springer, Berlin Heidelberg New York 2001)
46. B.M. Boghosian, P.J. Love, P.V. Coveney, I.V. Karlin, S. Succi, J. Yepez: Galilean-invariant lattice-Boltzmann models with $H$-theorem. *Phys. Rev. E* 68, 025103(R) (2003)
47. G.W. Gibbs: Elementary Principles of Statistical Mechanics (Dover, Phoenix 1960)
48. E.T. Jaynes: Information theory and statistical mechanics, in: *Statistical Physics. Brandeis Lectures*, vol. 3, ed. by K. W. Ford, 160–185 (Benjamin, New York 1963)
49. D. Zubarev, V. Morozov, G. Röpke: *Statistical mechanics of nonequilibrium processes*, vol. 1 (Akademie Verlag, Berlin 1996), vol. 2 (Akademie Verlag, Berlin 1997)
50. H. Grad: On the kinetic theory of rarefied gases. *Comm. Pure and Appl. Math.* **2**, 331–407 (1949)
51. J.T. Alvarez-Romero, L.S. García-Colín: The foundations of informational statistical thermodynamics revisited. *Physica A* **232**, 207–228 (1996)
52. R.E. Nettleton, E.S. Freidkin: Nonlinear reciprocity and the maximum entropy formalism. *Physica A* **158**, 672–690 (1989)
53. N.N. Orlov, L.I. Rozonoer: The macrodynamics of open systems and the variational principle of the local potential. *J. Franklin Inst.* **318**, 283–314 and 315–347 (1984)
54. A.M. Kogan, L.I. Rozonoer: On the macroscopic description of kinetic processes. *Dokl. AN SSSR* **158**, 566–569 (1964)
55. A.M. Kogan: Derivation of Grad–type equations and study of their properties by the method of entropy maximization. *Prikl. Matem. Mech.* **29**, 122–133 (1965)
56. L.I. Rozonoer: Thermodynamics of nonequilibrium processes far from equilibrium. In: *Thermodynamics and Kinetics of Biological Processes*, 169–186 (Nauka, Moscow 1980)
57. J. Karkheck, G. Stell: Maximization of entropy, kinetic equations, and irreversible thermodynamics. *Phys. Rev. A* **25**, 3302–3327 (1984)
58. N.N. Bugaenko, A.N. Gorban, I.V. Karlin: Universal Expansion of the Triplet Distribution Function. *Teoret. i Matem. Fisika* **88**, 430–441 (1991) (Transl.: Theoret. Math. Phys. 977–985 (1992))
59. A.N. Gorban, I.V. Karlin: Quasi–equilibrium approximation and non–standard expansions in the theory of the Boltzmann kinetic equation. In: *Mathematical Modelling in Biology and Chemistry. New Approaches*, ed. by R. G. Khlebopros, 69–117 (Nauka, Novosibirsk, 1991)

60. A.N. Gorban, I.V. Karlin: Quasi–equilibrium closure hierarchies for the Boltzmann equation. *Physica A* **360**, 325–364 (2006) (Includes translation of the first part of [59])
61. C.D. Levermore: Moment Closure Hierarchies for Kinetic Theories *J. Stat. Phys.* **83**, 1021–1065 (1996)
62. R. Balian, Y. Alhassid, H. Reinhardt: Dissipation in many–body systems: A geometric approach based on information theory. *Phys. Reports* **131**, 1–146 (1986)
63. P. Degond, C. Ringhofer: Quantum moment hydrodynamics and the entropy principle. *J. Stat. Phys.* **112**, 587–627 (2003)
64. P. Ilg, I.V. Karlin, H.C. Öttinger: Canonical distribution functions in polymer dynamics: I. Dilute solutions of flexible polymers. *Physica A* **315**, 367–385 (2002)
65. P. Ilg, I.V. Karlin, M. Kröger, H.C. Öttinger: Canonical distribution functions in polymer dynamics: II Liquid–crystalline polymers. *Physica A* **319**, 134–150 (2003)
66. P. Ilg, M. Kröger: Magnetization dynamics, rheology, and an effective description of ferromagnetic units in dilute suspension, *Phys. Rev. E* **66**, 021501 (2002); Erratum, *Phys. Rev. E* **67**, 049901(E) (2003)
67. P. Ilg, I.V. Karlin: Combined micro–macro integration scheme from an invariance principle: application to ferrofluid dynamics. *J. Non–Newtonian Fluid Mech.* 120, 33–40 (2004)
68. B. Robertson: Equations of motion in nonequilibrium statistical mechanics. *Phys. Rev.* **144**, 151–161 (1966)
69. P.J. Morrison: Hamiltonian description of the ideal fluid. *Rev. Mod. Phys.* **70**, 467-521 (1998)
70. E. Wigner: On the quantum correction for thermodynamic equilibrium. *Phys. Rev.* **40**, 749–759 (1932)
71. A.O. Caldeira, A.J. Leggett: Influence of damping on quantum interference: An exactly soluble model. *Phys. Rev. A* **31**, 1059–1066 (1985)
72. A.N. Gorban, I.V. Karlin: Reconstruction lemma and fluctuation–dissipation theorem. *Revista Mexicana de Fisica* **48**, 238–242 (2002)
73. A.N. Gorban, I.V. Karlin: Macroscopic dynamics through coarse–graining: A solvable example. *Phys. Rev. E* **56** 026116 (2002)
74. A.N. Gorban, I.V. Karlin: Geometry of irreversibility. in: *Recent Developments in Mathematical and Experimental Physics*, vol. C, ed. by F. Uribe, 19–43 (Kluwer, Dordrecht 2002)
75. A.N. Gorban, I.V. Karlin: *Geometry of irreversibility: The film of nonequilibrium states*, Preprint IHES/P/03/57, Institut des Hautes Études Scientifiques in Bures-sur-Yvette (France) (2003) Preprint on-line: `http://arXiv.org/abs/cond-mat/0308331`
76. I.V. Karlin, L.L. Tatarinova, A.N. Gorban, H.C. Öttinger: Irreversibility in the short memory approximation. *Physica A* **327**, 399–424 (2003)
77. J.L. Lebowitz: Statistical Mechanics: A Selective Review of Two Central Issues. *Rev. Mod. Phys.* **71**, S346 (1999)
78. S. Goldstein, J.L. Lebowitz: On the (Boltzmann) Entropy of Nonequilibrium Systems. *Physica D* **193**, 53–66 (2004)
79. S. Chapman, T. Cowling: *Mathematical theory of non-uniform gases*, Third edition (Cambridge University Press 1970)

80. C. Cercignani: *The Boltzmann equation and its applications*, (Springer, Berlin Heidelberg New York 1988)
81. L. Mieussens, H. Struchtrup: Numerical Comparison of Bhatnagar–Gross–Krook models with proper Prandtl number. *Phys. Fluids* **16**, 2797–2813 (2004)
82. R.M. Lewis: A unifying principle in statistical mechanics. *J. Math. Phys.* **8** , 1448–1460 (1967)
83. A.M. Lyapunov: *The general problem of the stability of motion* (Taylor & Francis, London 1992)
84. L.B. Ryashko, E.E. Shnol: On exponentially attracting invariant manifolds of ODEs *Nonlinearity* **16**, 147–160 (2003)
85. C. Foias, M.S. Jolly, I.G. Kevrekidis, G.R. Sell, E.S. Titi: On the computation of inertial manifolds. *Phys. Lett. A* **131**, 433–436 (1988)
86. Y. Sone: *Kinetic theory and fluid dynamics* (Birkhäuser, Boston 2002)
87. C.W. Gear, T.J. Kaper, I.G. Kevrekidis, A. Zagaris: Projecting to a slow manifold: singularly perturbed systems and legacy codes. *SIAM J. Appl. Dynamical Systems* **4**, 711–732 (2005)
88. F. Higuera, S. Succi, R. Benzi: Lattice gas-dynamics with enhanced collisions. *Europhys. Lett.* **9**, 345–349 (1989)
89. I.V. Karlin, A.N. Gorban, S. Succi, V. Boffi: Maximum entropy principle for lattice kinetic equations. *Phys. Rev. Lett.* **81**, 6–9 (1998)
90. H. Chen, S. Chen, W. Matthaeus: Recovery of the Navier–Stokes equation using a lattice–gas Boltzmann Method. *Phys. Rev. A* **45**, R5339–R5342 (1992)
91. Y.H. Qian, D. d'Humieres, P. Lallemand: Lattice BGK models for Navier–Stokes equation. *Europhys. Lett.* **17**, 479–484 (1992)
92. S. Succi, I.V. Karlin, H. Chen: Role of the $H$ theorem in lattice Boltzmann hydrodynamic simulations. *Rev. Mod. Phys.* **74**, 1203–1220 (2002)
93. I.V. Karlin, A. Ferrante, H.C. Öttinger: Perfect entropy functions of the Lattice Boltzmann method. *Europhys. Lett.* **47**, 182–188 (1999)
94. S. Ansumali, I.V. Karlin: Entropy function approach to the lattice Boltzmann method. *J. Stat. Phys.* **107** 291–308 (2002)
95. R. Cools: An encyclopaedia of cubature formulas. *J. of Complexity* **19**, 445–453 (2003)
96. A. Gorban, B. Kaganovich, S. Filippov, A. Keiko, V. Shamansky, I. Shirkalin: *Thermodynamic Equilibria and Extrema: Analysis of Attainability Regions and Partial Equilibrium* (Springer, Berlin Heidelberg New York 2006)
97. S. Ansumali, I.V. Karlin: Kinetic Boundary condition for the lattice Boltzmann method. *Phys. Rev. E* **66**, 026311 (2002)
98. J.R. Higgins: *Sampling Theory in Fourier and Signal Analysis: Foundations* (Clarendon, Oxford 1996)
99. J.R. Higgins, R.L. Stens: *Sampling Theory in Fourier and Signal Analysis: Advanced Topics* (Clarendon, Oxford 1999)
100. J.G.M. Kuerten, B.J. Geurts, A.W. Vreman, M. Germano: Dynamic inverse modeling and its testing in large-eddy simulations of the mixing layer. *Phys. Fluids* **11** 3778–3785 (1999)
101. A.W. Vreman: The adjoint filter operator in large-eddy simulation of turbulent flow. *Phys. Fluids* **16**, 2012–2022 (2004)
102. B.J. Geurts, D.D. Holm: Nonlinear regularization for large-eddy simulation. *Phys. Fluids* **15**, L13–L16 (2003)
103. P. Moeleker, A. Leonard: Lagrangian methods for the tensor-diffusivity subgrid model. *J. Comp. Phys.* **167**, 1–21 (2001)

104. G. Berkooz, P. Holmes, J.L. Lumley: The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Rev. Fluid Mech.* **25**, 539–575 (1993)
105. I.T. Jolliffe: *Principal component analysis* (Springer, Berlin Heidelberg New York 1986)
106. K. Kunisch, S. Volkwein: Galerkin Proper Orthogonal Decomposition Methods for a General Equation in Fluid Dynamics. *SIAM J Numer. Anal.* **40**, 492-515 (2002)
107. M. Marion, R. Temam: Nonlinear Galerkin methods. *SIAM J. Numer. Anal.* **26**, 1139–1157 (1989)
108. A. Gorban, A. Zinovyev: Elastic Principal Graphs and Manifolds and their Practical Applications. *Computing* **75**, 359–379 (2005)
109. G.D. Birkhoff: *Dynamical systems* (AMS Colloquium Publications, Providence 1927) Online: `http://www.ams.org/online_bks/coll9/`
110. B. Hasselblatt, A. Katok, (Eds.): *Handbook of Dynamical Systems* (Elsevier 2002)
111. A. Katok, B. Hasselblat: *Introduction to the Modern Theory of Dynamical Systems*, Encyclopedia of Math. and its Applications, vol. 54 (Cambridge University Press 1995)
112. J. Milnor: On the concept of attractor. *Comm. Math. Phys.* **99**, 177–195 (1985)
113. P. Ashwin and J.R. Terry: On riddling and weak attractors. *Physica D* **142**, 87–100 (2000)
114. D.V. Anosov: About one class of invariant sets of smooth dynamical systems. In: *Proceedings of International conference on non-linear oscillation,* vol. 2, 39–45 (Kiev 1970)
115. P. Walters: On the pseudoorbit tracing property and its relationship to stability. In: *Lect. Notes Math.* vol. 668, 231–244 (Springer, Berlin Heidelberg New York 1978)
116. J.E. Franke, J.F. Selgrade: Hyperbolicity and chain recurrence. *J. Different. Equat.* **26**, 27–36 (1977)
117. H. Easton: Chain transitivity and the domain of influence of an invariant set. In: *Lect. Notes Math.*, vol. 668, 95–102 (Springer, Berlin Heidelberg New York 1978)
118. Y. Sinai: Gibbs measures in ergodic theory. *Russ. Math. Surveys* **166**, 21–69 (1972)
119. I.G. Malkin: On the stability under uniformly influencing perturbations. *Prikl. Matem. Mech.* **8**, 241–245 (1944)
120. V.E. Germaidze, N.N. Krasovskii: On the stability under sustained perturbations. *Prikl. Matem. Mech.* **21**, 769–775 (1957)
121. I.G. Malkin: *The motion stability theory* (Nauka, Moscow 1966)
122. A. Strauss, A.J. Yorke: Identifying perturbations which preserved asymptotic stability. *Proc. Amer. Math. Soc.* **22**, 513–518 (1969)
123. M. Gromov: *Metric structures for Riemannian and non-Riemannian spaces*, Progress in Mathematics, 152 (Birkhauser Boston, Inc., Boston 1999)

# Renormalization Group Methods for Coarse-Graining of Evolution Equations

A. Degenhard[1] and J. Rodríguez-Laguna[2]

[1] Universität Bielefeld, Fakultät für Physik, Universitätstraße 25, 33615 Bielefeld, Germany, andreas@physik.uni-bielefeld.de
[2] Condensed Matter Sector, Scuola Internazionale Superiore di Studi Avanzati (SISSA), 34014 Trieste, Italia, jrlaguna@sissa.it

**Summary.** We review some applications of the renormalization group (RG) to the coarse-graining of evolution equations. These techniques allow simulations with a lower computational cost. The idea behind real space RG is discussed in the first part, along with the background for its relation to coarse-graining. The rest of the article deals with their application to the selection of relevant degrees of freedom in discretized partial evolution equations, both linear and non-linear, and to the evolution of many-body systems, focusing on stochastic lattice models for reaction-diffusion.

## 1 Introduction and Basic Formalism

In this section we provide a brief introduction to the historical development and the basic concepts of the renormalization group (RG). A reader familiar with the foundations of RG theory is invited to skip this section.

### 1.1 A Short History

In 1966 the basic formulation of the ideas underlying the renormalization group (RG) concept were published in a paper by Leo P. Kadanoff [1]. Within this pioneering work a method was proposed allowing to extract the critical behaviour without computing the partition function explicitly. The introduced method was then referred to as the *block spin picture* and applications were carried out in various fields in the physical sciences [2]. A particularly nice illustration is provided in the original context of critical phenomena [3, 4]. The main difficulty in the quantitative understanding of critical behaviour are the infinitely many degrees of freedom relevant in this cooperative phenomenon. Self-similarity at the *critical point*, i.e.: fluctuations over an infinite hierarchy of length scales need to be taken into account. For a description of the characteristic critical long range behaviour it then suffices to consider the

critical system to consist of "large blocks" interacting among themselves in a way that resembles that of the former small scale variables. Spin variables are often considered as a simple example, yielding the term "block spin". By forming block spins one "simplifies" the system by removing a large number of degrees of freedom expected to be irrelevant for macroscopic physics. The block spins approach is, therefore, a coarse graining of the original system.

In 1971 Kenneth G. Wilson converted this idea into an efficient computational method. In [5] he reformulated Kadanoff's block spin approach in infinitesimal form, yielding a set of differential equations termed as *renormalization group equations*. This reformulation allowed for a careful mathematical analysis and to present the universal character of the method. In [6] he made usage of the momentum space description of the block spin picture to analyze Ginzburg-Landau's model. Taking the momentum space concept further, Wilson solved the Kondo problem, which dealt with the effect of a magnetic impurity on the conduction band electrons of a metal [7]. He remarked that it was the first example where the renormalization group program had been carried out in full. His solution was based on the division into shells of the whole lattice around the impurity (center), where the shells were integrated outwards iteratively. This method is now termed Wilson's RG approach.

Following Wilson's first application of RG ideas to critical phenomena and the Kondo problem, various variants of the Wilson renormalization group were introduced in both momentum and position space, the latter known as real space RG (RSRG). These techniques include the *Migdal-Kadanoff bond moving techniques* [8, 9], the Monte Carlo renormalization group (MCRG) initially suggested by Ma [10] and further developed as the *large-cell renormalization transformation* by Friedman and Felsteiner [11] and also Lewis [12], the real space dynamic renormalization group (RSDRG) in which the coarse grained description preserves the information about the long range and slowly varying degrees of freedom [13, 14] and finally the dynamic renormalization group (DRG) formulated in momentum space [15, 16].

Applications of Wilson's RG and its real space variants range from problems in percolations [17], polymer research [18], material sciences [19] and quantum field theory [20, 21] to strongly correlated electron systems and spin chains [22]. However, in particular the application to the latter exhibited an unexpected failure of the Wilson renormalization group. Investigating this failure it was found by S.R. White and R.M Noack that taking into account the low energy states of the block spins is not sufficient [23]. Contrarily they demonstrated, that it is essential to capture the most probable states in constituting global ground state, thereby defining an example for a *target state* [24]. It was shown by these authors that the optimal states to represent the target state are computed as the eigenstates of the density matrix instead of the Hamiltonian itself. This approach has given birth to the density matrix renormalization group (DMRG) algorithm. Here we summarize the basic thoughts in the next section, whereas in the last chapters of this review we discuss recent advances and modifications of the DMRG.

### 1.2 The Failure of RSRG for Quantum Lattice Problems

A simple example to illustrate this failure was suggested by Wilson and later explored by S.R. White and R.M. Noack [23]. The model considered is a spinless particle in a one dimensional lattice, with the following hamiltonian with fixed boundary conditions

$$H_{1D} = \begin{cases} 2 & \text{if } i = j\,, \\ -1 & \text{if } |i-j| = 1 \end{cases} \qquad \text{and } 0 \text{ otherwise}\,. \tag{1}$$

The problem, which is fairly simple, is to find the ground state of this system. The standard block spin RG approach considers a coarse grained lattice of block spins. This is equivalent to integrating out the fine scale details, so each block will only contain low energy information. The basic block spin procedure divides the lattice into two parts and finds the ground states for each of them. Afterwards the global ground state is found within the linear subspace spanned by the ground states for each block. The higher energy states for the block are neglected, considering that they do not contribute much to the global low energy properties.

However, this procedure yields quite poor results. The problem, in this case, stems in the *fixed boundary conditions*. Each block state is zero in the boundary. But the right boundary of the left block (respectively, the left boundary of the right block) is the *center* of the global system, which is forced to contain a high energy kink in that position.

In more general terms, the problem is to have *isolated* blocks. Somehow, the blocks must have some knowledge about its environment in order to choose the best states, which need not be the lowest energy states for the block. From a technical point of view this can be achieved in a number of ways, as shown in figure 1. Four different types of blocking schemes are illustrated.

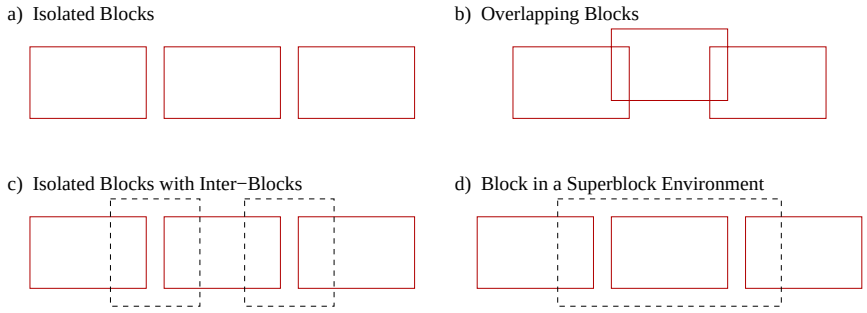



Fig. 1: Four different types of standard blocking schemes.

In case a) the blocks are isolated, leading to failure. Case b) shows an early solution: to let the blocks overlap. In c) a more advanced scheme is pictured, in

which neighbouring blocks are connected by means of additional inter-blocks. Finally in d) the superblock concept is illustrated, in which each single block is incorporated within a larger superblock. This latter is the principle underlying the DMRG algorithm which is investigated in the last section of this review.

### 1.3 Renormalization Group Transformations

The previous section has provided insight into the reasons for the failure of RG schemes. This section is more constructive, and it explains how to build proper renormalization group transformations (RGTs). The definition of an RGT depends on the problem at hand, although various *recipes* exist. In particular we discuss how *inter-block correlations* introduced in the preceding section are to be taken into account.

RG is not universal scheme in which one always works out a RGT by means of a general algorithm, which will then be applied to the problem at hand. The effective usage of the RG requires a correct choice of the RGT depending of the physics of the system. A simple example of a RGT is to average over the sites in the block to construct an effective site. This particular RGT has the practical advantage of linearity. But it is by no means the only option. It is this ambiguity that makes it obvious that a RGT is not just a mathematical procedure, because the particular transformation will depend on the intuition of the physicist. This may be a reason why it is often difficult to formulate the general framework of RG in a rigorous mathematical context.

The optimal and most natural definition of a RGT is to make it as physically and as mathematically well defined as possible, which in general turns out to be an unattainable goal. In most of the interesting cases, where the interaction part of the *Lagrangian* $\mathfrak{L}$ turns out to be very complicated, one is forced to use rather crude transformations. The aim of this work is to define RGTs which are motivated by physical considerations and are also mathematically well defined procedures. This has clearly not been achieved by a straightforward improvement of already existing transformations and for this purpose a new mathematical formalism is introduced in this section. It is thus necessary to give a definition of an RGT as general as possible. Such a universal formulation allows to search for the optimum among a larger number of possible RGTs.

Figure 2 gives an example of this basic idea concerning the definition of RGT, which is sometimes referred to as *coarse graining*. Of course figure 2 only provides a graphical representation of this change of scales and we have to define the precise mapping prescription. This is precisely the part where by physical assumptions we get a mathematical description, which in most cases we can work out approximately or numerically.

To capture as many approaches as possible we define a renormalization group transformation (RGT) in terms of a general map:

$$R: \ (\{\sigma_l\}, \mathbf{k}) \longrightarrow (\{\mu_m\}, \mathbf{k}') \ , \tag{2}$$

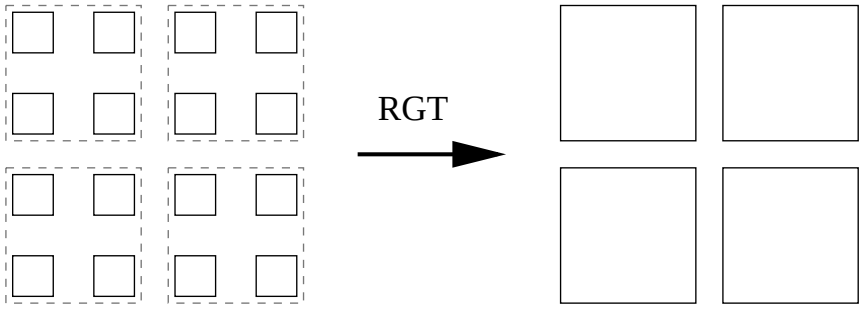



Fig. 2: Visualization of a RG transformation (RGT).

on a set of physical variables $\{\sigma_l\}$ and a further set of parameters $\mathbf{k} = (k_1, k_2, \dots)$, where $\{l\}$ and $\{m\}$ are not necessarily equal indexing sets. The $\{\mu_m\}$ denote the new variables and $\mathbf{k}'$ contains the *effective* parameters.

The fundamental requirement for the $R$ transformation is to keep the physics of the problem invariant. This in turn includes several constraints as for example the conservation of the symmetries, the maintenance of the structure of the Hamiltonian and the values of physical observables, such as the free energy. Providing such a transformation in rigorous mathematical terms is not only a difficult task: sometimes it is impossible to obey all the constraints at the same time. This is the reason for the enormous variety of transformations developed in the last decades with corresponding results ranging from totally unphysical to exact physical numbers.

It should also be mentioned that the methods for implementation of constraints can be very different. The aim of this work is to propose new transformations which should include nearly all constraints and we will point out the missing ones and discuss the effect of the approximations. Here we adopt a notation in which we will express the foregoing statements. The RGT is always introduced by applying it on a functional dependence $\mathcal{O}(\{\sigma_l\}, \mathbf{k})$ given by physics. By applying the transformation $R$, this function $\mathcal{O}$ should not change and therefore leaves the physics invariant. Thus we obtain the first class of constraints as

$$R\left[\mathcal{O}\left(\{\sigma_l\}, \mathbf{k}\right)\right] \;=\; \mathcal{O}\left(\{\mu_m\}, \mathbf{k}'\right) \;. \tag{3}$$

In some known RGT this functional dependence is given by the Lagrangian or the Hamiltonian themselves, and the sets of physical variables will then correspond to the original and the block variables.

In an infinite system the number of original variables is infinite, and so is the number of block variables. Therefore, in the thermodynamical limit there is no reduction in the number of degrees of freedom. This leads us to study the *RG-flow* of the system [25], i.e.: to extract information about just how does the description of the physical system changes with the blocking operation. This

change is restricted to a few parameters –called *running coupling constants*– in the most favourable cases. In those cases, the thermodynamic limit is analyzed via the *fixed points* of the RGT. The examples we shall analyze are not of this kind, and the number of *running coupling constants* shall be too large for a fixed point analysis to be practical. Therefore, we shall attach to the finite-size picture and provide examples of effective reduction of degrees of freedom and coarse-graining.

## 2 RSRG for the Selection of Relevant Degrees of Freedom

Soon after the considerable success of the RG for equilibrium systems, generalizations of the method were introduced to handle *non-equilibrium systems*, based on Monte Carlo approaches [10] or perturbative series expansions, such as the the dynamic RG (DRG) [26], inherently defined as a Fourier space technique, which was used to study the dynamics of the Burgers equation [27, 28] and the related KPZ equation [29, 30]. However, the impact of dimensionality on applying the DRG method to the KPZ or the Burgers equation is crucial [31, 32].

### 2.1 Projection Operator Concept

The attempt to adapt RSRG techniques for systems far from equilibrium started with a work by N. Goldenfeld, A. McKane and Q. Hou in which the usage of RSRG methods to solve partial differential equations (PDEs) was investigated numerically [33, 34]. Utilizing the operator concept of the RSRG and replacing the equilibrium Hamiltonian operator by the time evolution operator of a partial differential equation, a coarse-graining procedure was devised which did not depend on the details of the dynamical system. However, Goldenfeld et al. pointed out that the approach is not as well-defined and systematic as it should. In particular, the geometric construction of the coarse-to-fine operator assumes the relevant degrees of freedom to be distributed within the long wavelength fluctuations. This is not necessarily the case for the evolution of many systems with non-linear dynamics.

Contrarily to DRG in momentum space, we will show non-perturbative approaches for effective coarse-graining, allowing to compute the relevant degrees of freedom which are difficult to examine when the modes mixing is non-trivial. In this section we consider systems that are described by partial differential equations (PDEs) of the form

$$\partial_t \phi \;=\; \tilde{H}\phi, \tag{4}$$

with $\phi = \phi(x,t)$ a function of 1D space $x$ and time $t$ and the operator $\tilde{H}$ acting as the generator of the evolution. Spatial and temporal dependence of

the discretized function $f$ will in turn be denoted by indexing sets $i, i+1, \ldots$ and $t, t+1, \ldots$ respectively. Sites which are not necessarily neighbours are denoted as $i, j$ and by using capital letters $I, J$ we refer to the sites of the effective lattice. Quantities defined in this effective vector space are equipped with a prime.

To exemplify our analysis for linear and quadratic evolution operators, $H$ and $Q$ respectively, we employ the following discretization of equation (4)

$$\phi_{i,t+\Delta t} \;=\; \sum_{j=1}^{N}(\mathbb{1}_{i,j} + \Delta t \,\cdot\, H_{i,j})\phi_{j,t} \;+\; \Delta t \sum_{j,k=1}^{N} Q_{i,j,k}\phi_{j,t}\phi_{k,t} \;:=\; f_i\,[\boldsymbol{\phi}_t] \tag{5}$$

with $i \in \{1, \ldots, N\}$ and $N$ denoting the number of sites in the lattice. The spatial lattice spacing is denoted as $\Delta x$ and the discrete temporal integration interval as $\Delta t$. The field at time $t$ is represented by $\boldsymbol{\phi}_t \in V^N$ and $H : V^N \to V^N$ is the discretized evolution operator operating on the N-dimensional vector space $V^N$. Here we aim to approximate the time-evolution in a lower-dimensional effective vector space $V^M$ $(M < N)$. Let $G : V^N \to V^M$ denote the *truncation operator* which takes an element $\boldsymbol{\phi}_t$ of the full vector space $V^N$ and returns an element $\boldsymbol{\phi'}_t$ of the truncated space $V^M$ [35] . Here capital indexing letters $I \in \{1, \ldots, M\}$ refer to lattice sites in the effective vector space $V^M$. The application of the *embedding operator* $G^p : V^M \to V^N$ to an effective field configuration $\boldsymbol{\phi'}$ yields a field configuration $G^p\boldsymbol{\phi'}$ on the original lattice. Ideally we would choose $G^p = G^{-1}$. Then applying this operator to $\boldsymbol{\phi'}$ we would recover $\boldsymbol{\phi}$. However, due to our assumption $M < N$, $G$ has a non-trivial kernel and the true inverse does not exist. One therefore employs the *pseudo-inverse* by defining the embedding operator $G^p$, satisfying the Moore-Penrose conditions [36]

$$GG^pG = G\,, \qquad G^pGG^p = G^p\,, \qquad (G^pG)^\dagger = G^pG\,, \qquad (GG^p)^\dagger = GG^p\,.$$

For a given truncation operator $G$ these equations are solved if the embedding operator $G^p$ is chosen as the *singular value decomposition* (SVD) pseudo-inverse of $G$. Then $GG^p$ is the identity operator on $V^M$ while $\mathcal{R} := G^pG$ is the projection operator on the subspace $V^{\mathrm{rel}} \subset V^N$ spanned by the relevant degrees of freedom. We like to point out that this operator is fully determined by its kernel and we may formally write

$$\mathcal{R} \;=\; G^pG \;=\; \sum_{i=1}^{M} |\mathbf{v}_i\rangle\,\langle\mathbf{v}_i|\;, \tag{6}$$

where $\{\mathbf{v}_i\}_{i=1}^M$ is an orthonormal basis of $V^{\mathrm{rel}}$. We will refer to such a set as a set of *target states*. Using the Dirac notation for vectors, $|\mathbf{v}_i\rangle\,\langle\mathbf{v}_i|$ denotes the projection operator on the subspace spanned by the single vector $\mathbf{v}_i$. The projector $\mathcal{R}$ is termed the *reduction operator* and the ratio $\lambda = N/M$ defines

the *reduction factor* $\lambda$.
For $M < N$ the functional $f'\left[\boldsymbol{\phi}'_t\right] : V^M \hookrightarrow V^M$ is defined as in (5) on an effective lattice with a reduced number of lattice sites $M$ and

$$\phi'_{I,t+\Delta t} \;=\; \sum_{J=1}^{M} \left(\mathbb{1}_{I,J} + \Delta t \cdot H'_{I,J}\right) \phi'_{J,t} \;+\; \Delta t \sum_{J,K=1}^{M} Q'_{I,J,K} \phi'_{J,t} \phi'_{K,t} \;. \quad (7)$$

In equation (7) $H'$ and $Q'$ denote the effective linear and quadratic evolution operators defined as [37]

$$H'_{I,J} \;:=\; G_{I,i}\, H_{i,j}\, G^p_{j,J} \qquad \text{and} \qquad Q'_{I,J,K} \;:=\; G_{I,i}\, Q_{i,j,k}\, G^p_{j,J}\, G^p_{k,K} \;. \quad (8)$$

Inserting (7) and (8) into equation (5) an approximate field evolution equation on a coarse grained lattice is obtained by

$$\phi_{i,t+\Delta t} \;\approx\; \Big[\, G^p \left(\mathbb{1} \;+\; \Delta t \cdot H' \;+\; \Delta t \cdot Q'\right) G\, \boldsymbol{\phi}_t \,\Big]_i \quad , \quad (9)$$

describing the evolution of the field $\boldsymbol{\phi}$ under a reduced number of degrees of freedom. Equation (9) defines a RG transformation (RGT) within this operator formalism. Carrying out one RGT is called a RG step (RGS) and according to the concept developed in this section is equivalent to an approximate field evolution using less degrees of freedom. Equation (9) defines the real-space analogue of a RGT established within the DRG method. The field itself provides the set of parameters used to establish a RGT [38]. Furthermore, equation (9) fuses the coarse-graining and the time evolution procedure.

### 2.2 Geometric Reduction of the Degrees of Freedom

To employ the RGT (9) for practical use, information regarding the operators $G^p$ and $G$ is requested. In [37] a general geometric approach is introduced using overlapping blocks or cells as discussed in section 1.2. For regular arranged partitions of cells one considers an interval together with a regular partition into $n$ equal cells in each dimension, denoted by $C_i^n \equiv [\frac{i-1}{n}, \frac{i}{n}]$. The truncation operator $G$ is defined by [37]

$$G_{Ii} \;=\; \frac{\text{Overlap between cells } C_i \text{ and } C_I.}{\text{Measure of cell } C_I.} \quad , \quad (10)$$

where $C_i$ denotes a block or cell of the previous geometric partition and $C_I$ is part of the coarser one. The idea is to employ overlapping blocks to perform single degree truncation operators, i.e. to truncate the number of degrees of freedom from $N$ to $N-1$. Such single step sudden transformation are given analytically by

$$G_{Ii} \;=\; \delta_{I,i} \frac{N-I}{N} + \delta_{I,i-1} \frac{I}{N} \;. \quad (11)$$

Concatenating these operators allows to perform coarse graining operations of any desired size [37]. Figure 3 (a) illustrates the overlapping of two successive regular partitions. The dashed overlaying and coarse grained partition contains one degree of freedom less in every dimension. The degrees of freedom which are retained by the truncation matrix $G$ are plotted in figure 3 (b). They are the $E^N$ vectors given by the columns of $G$. Each of the discrete

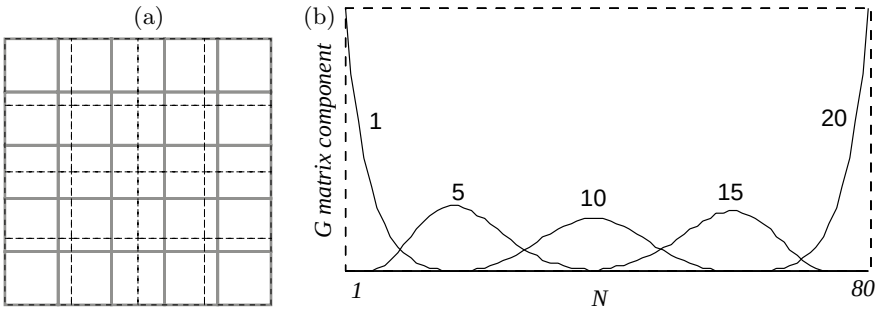



Fig. 3: Two overlapping partitions are depicted in (a). The thicker grey lines refer to the previous partition whereas the overlaying dashed lines are referring to the coarse grained one. Some of the degrees of freedom which are retained by the quasi-static truncation operator proceeding from 80 → 20 sites (b). Cells 1, 5, 10, 15 and 20 are depicted. Notice that the "cells" are now overlapping and have slightly Gaussian nature.

functions depicted in figure 3 may be considered to represent a relevant degree of freedom when truncating with the matrix $G$ from 80 to 20 lattice sites, i.e. a reduction factor $\lambda = 4$. Although the functions representing the degrees of freedom are now overlapping, they conserve a true real-space nature. It should be noticed that the width of the leftmost and rightmost cells is smaller than the one at the middle of the interval. A consequence is the quite exact representation of the boundary conditions.

### 2.3 Non-Geometric Reduction of the Degrees of Freedom

In this section we provide a general concept for reducing the degrees of freedom in evolutionary systems without geometrically coarse-graining the lattice equations. Including the temporal characteristics of the particular partial differential equation (PDE) into the construction of the embedding and truncation operators we distinguish between a short-time regime and a long-time regime.

**Short-time evolution.** Within the short-time regime we describe the evolution of the field as a perturbation of the initial field configuration at time $t = 0$. Evolving the initial field $\boldsymbol{\phi}_0$ for $M$ time steps $\Delta t$ with respect to

$\tilde{H} = (H + Q)$ the dynamics within this short-time interval is conserved by the set of vectors

$$\mathcal{S} := \{\boldsymbol{\phi}_0, \tilde{H}\boldsymbol{\phi}_0, \tilde{H}^2\boldsymbol{\phi}_0, \ldots, \tilde{H}^{M-1}\boldsymbol{\phi}_0\} . \tag{12}$$

Using the set of vectors in (12) as the columns and rows of the linear operators $G$ and $G^p$ (previously orthonormalized), these can be considered as projection maps from and into the space $V^M$ respectively.
If $M \leq N$ the relation

$$\boldsymbol{\phi}_t = G^p \left(\mathbb{1} + \Delta t \tilde{H}'\right)^M G\,\boldsymbol{\phi}_0 \qquad \text{with} \quad \tilde{H}' = H' + Q' \tag{13}$$

governs an exact evolution of the field $\boldsymbol{\phi}_t$ on the effective coarse-grained lattice for $t < M\Delta t$. In this case the states in (12) span a subspace $V^M$ of the full vector-space $V^N$ conserving the relevant degrees of freedom for the short-time evolutionary regime. In the considered short-time regime, we may rewrite equation (13) as [35]

$$\boldsymbol{\phi}_t = \left[\, G^p \left(\mathbb{1} + \Delta t \tilde{H}'\right) G \,\right]^M \boldsymbol{\phi}_0 \tag{14}$$

which determines a RG flow in the short time regime and a RGS is defined by relation (9).
However, if $t \geq M\Delta t$ equation (14) is an approximation to the evolved field $\boldsymbol{\phi}_t$. In this case the approach is only applicable if no relevant scale interference in the evolutionary process occurs. This is unlikely the case for longer times in nonlinear dynamical systems and relation (14) becomes a crude approximation giving rise to numerical instabilities.

**Long-time evolution.** Nonlinear evolution processes exhibit most of their characteristics in the long-time regime. The asymptotic form of the field or surface configuration in growth phenomena [29] or the formation of turbulent states out of spiral waves [39] are only two examples. This demands for the construction of embedding and truncation operators $G^p$ and $G$ by minimizing the error in the approximate field evolution equation (9) for all times of the evolution process. To accomplish this task an error operator $\mathcal{E}$ is introduced as

$$\mathcal{E}\boldsymbol{\phi}_t := (\mathbb{1} + \Delta t \tilde{H})\boldsymbol{\phi}_t - (G^p)(\mathbb{1} + \Delta t \tilde{H}')G\,\boldsymbol{\phi}_t \ . \tag{15}$$

which has to be minimized within a proper operator norm, a concept taken from equilibrium operator RSRG techniques [24, 40]. Using the established notation within our established operator RSRG, $(\mathbb{1} + \Delta t H)\boldsymbol{\phi}_t$ is called the target state [24] and $(G^p)(\mathbb{1} + \Delta t \tilde{H}')G\,\boldsymbol{\phi}_t$ an optimal representation of the target state [24] in the subspace $V^M \subset V^N$.
We are interested in the limit $\mathcal{E}\boldsymbol{\phi}_t \to 0$ for $t \gg 0$ subject to any field configuration $\boldsymbol{\phi}_0$. To exclude the explicit field dependence from the minimization procedure we rewrite equation (15) in operator form as

$$\mathcal{E} \;=\; (\mathbb{1} + \Delta t H) \;-\; G^p (1 + \Delta t \tilde{H}') G \;\; . \tag{16}$$

and minimize $\mathcal{E}$ in the matrix notation according to the Frobenius norm [†]. Inserting the definitions (8) we rewrite equation (16) using the Frobenius norm $||_F$ as

$$|\mathcal{E}|_F \;:=\; \Big\{ \mathbb{1} - (G^p G) + \Delta t \Big[ H - (G^p G)\, \tilde{H}\, (G^p G) \Big] \Big\}^2 \;\; . \tag{17}$$

According to equation (17) the minimization of the error operator $\mathcal{E}$ only depends on the composed operator $G^p G$ since it governs the field evolution under a reduced number of degrees of freedom. With respect to (6) and rewriting $G$ and $G^p$ in terms of a Singular Value Decomposition

$$G \;=\; \sum_{i=1}^{M} \lambda_i \, |\mathbf{u}_i\rangle \, \langle \mathbf{v}_i| \qquad \text{and} \qquad G^p \;=\; \sum_{i=1}^{M} \lambda_i^{-1} \, |\mathbf{v}_i\rangle \, \langle \mathbf{u}_i| \;\; , \tag{18}$$

where $\{|\mathbf{v}_i\rangle\}_{i=1}^{M}$ and $\{|\mathbf{u}_i\rangle\}_{i=1}^{M}$ are respectively sets of $N$-dimensional and $M$-dimensional vectors we have

$$\mathcal{R} \;=\; G^p G \;=\; \sum_{i=1}^{M} |\mathbf{v}_i\rangle \, \langle \mathbf{v}_i| \;=\; \mathbb{1} - \sum_{i=M+1}^{N} |\mathbf{v}_i\rangle \, \langle \mathbf{v}_i| \;\; . \tag{19}$$

The reduction operator $\mathcal{R}$ is composed of $M$ states in the vector space $V^N$ thereby projecting out those states representing the less relevant degrees of freedom in the evolution process. Successive dimensional reduction of the target space then allows for an iterative construction for $\mathcal{R}$ as [35]

$$\mathcal{R}^M \;=\; \mathcal{R}^{M+1} - |\mathbf{v}_{N+1-M}\rangle \, \langle \mathbf{v}_{N+1-M}| \;=\; \mathcal{R}^{M+1} - \mathcal{P}_{|\mathbf{v}_{N+1-M}\rangle} \;\; , \tag{20}$$

with $\mathcal{R}^0 = \mathbb{1}$. In equation (20) we have denoted by $\mathcal{P}_{|\mathbf{v}_{N+1-M}\rangle}$ the projection operator onto the state $|\mathbf{v}_{N+1-M}\rangle$ which is the target to calculate in the $M$th minimization procedure. Applying the iterative scheme (20) we write the minimization procedure for the state $|\mathbf{v}_{N+1-M}\rangle$ as

$$|\mathcal{E}^M|_F \;=\; \mathbb{1} - \mathcal{R}^M + \Delta t \Big( \tilde{H} \;-\; \mathcal{R}^M \, \tilde{H} \, \mathcal{R}^M \Big) \;. \tag{21}$$

The error operator is therefore minimized by adjusting the components of the target states $|\mathbf{v}_i\rangle$ for $\mathcal{P}_{|\mathbf{v}_i\rangle}$. Figure 4 compares the conventional coarse graining operator with a reduction operator adapted to the Burgers' nonlinearity [27]. Whereas the coarse graining operator has only diagonal blocks different from zero, the computed reduction operator shows off-diagonal elements different

[†] In principle every matrix norm can be used, although the Frobenius norm can be easily related to concepts from linear algebra like singular value decomposition. It is defined to be the sum of the squares of all the entries in a matrix.

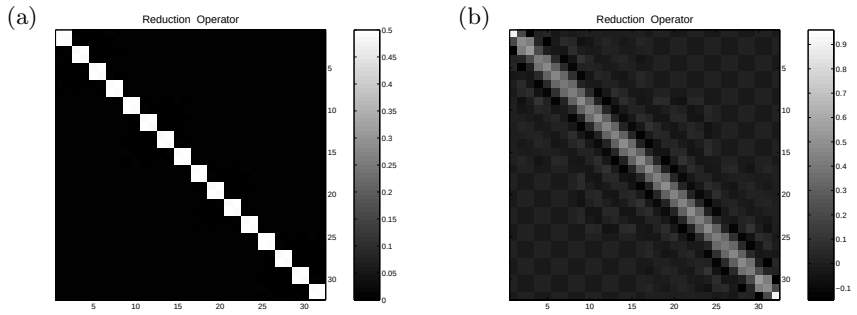



Fig. 4: Comparison between conventional coarse graining (a) and the minimization procedure (b) for Burgers' nonlinearity coupling constant 0.1 and a reduction factor $\lambda = 2$.

from zero, indicating long range correlations taking small scale details into account [41].

**Long-time evolution with subgrid correction.** Recently the method was compared to the theory of numerical homogenization [42], a technique essentially based on multiresolution analysis [43]. Numerical homogenization has repeatedly been used to explore possible reductions of various PDEs [44]. In multiresolution analysis a function $f \in L^2(\mathbb{R})$, i.e. an element in the space of square-integrable functions, is decomposed into *large scale contributions* (or *averages*) and *details* (or *fluctuations*) [43]. Starting from the large scale contribution one recovers the original function $f$ by increasing the resolution, i.e. by successively adding the details corresponding to the next finer scale as follows [43]: A multiresolution analysis (MRA) of $L^2(\mathbb{R})$ is a nested sequence of closed subspaces $(V_j)_{j\in\mathbb{Z}}$ of $L^2(\mathbb{R})$, i.e. $\cdots \subset V_1 \subset V_0 \subset V_{-1} \subset \cdots$. A further requirement is that all spaces $V_j$ are scaled versions of the central space $V_0$. Thus, $V_j$ can be considered as the subspace of functions that contain information down to the scale $2^j$. The scale is denoted by the subscript $j$ and becomes coarser with an increasing $j$. Because of the nesting sequence it is possible to define the orthogonal compliment $W_j$ of $V_j$ in $V_{j-1}$ according to the decomposition $V_{j-1} = V_j \oplus W_j$. Further we denote the projection operators onto $V_j$ and $W_j$ by $P_j$ and $Q_j$ so that for all $f \in V_{j-1}$ we write $P_j f = s_f \in V_j$ and $Q_j f = d_f \in W_j$. By defining

$$U_{j-1} := \begin{pmatrix} Q_j \\ P_j \end{pmatrix} \; : \; V_{j-1} \; \rightarrow \; V_j \; \oplus \; W_j \tag{22}$$

$f$ is decomposed into its details and averages as

$$U_{j-1} f = \begin{pmatrix} \boldsymbol{d}_f \\ \boldsymbol{s}_f \end{pmatrix} \in \; V_j \; \oplus \; W_j \tag{23}$$

Consider a linear algebraic system $H_{j-1}\boldsymbol{f} = \boldsymbol{g}$ with $\boldsymbol{f}, \boldsymbol{g} \in V_{j-1}$. Application of $U_{j-1}$ and $U_{j-1}^{\dagger}$ yields

$$U_{j-1}H_{j-1}U_{j-1}^{\dagger} = \begin{pmatrix} A_j & B_j \\ C_j & T_j \end{pmatrix} : \; V_j \; \oplus \; W_j \rightarrow V_j \; \oplus \; W_j \; , \tag{24}$$

where $T_j = P_j H_{j-1} P_j$ projects $H_{j-1}$ onto a coarser scale, thereby resembling a RG step. In case the A-block possesses an inverse[†] $A_j^{-1}$ one derives using equations (23) and (24)

$$R_j \boldsymbol{s}_f = \boldsymbol{s}_g - C_j A_j^{-1} \boldsymbol{d}_g \; . \tag{25}$$

Here the *reduced operator* $R_j$ is defined as the Schur complement of the block $T_j$ of the matrix $H_{j-1}$ according to

$$R_j = T_j - C_j A_j^{-1} B_j \; . \tag{26}$$

Equation (25) describes the large scale component of the solution regarding the original algebraic system. In this sense this equation is referred to as the *homogenized equation* and the reduction procedure is termed a *homogenization step*. The second term on the right hand side of (26) is a correction to the projection $T_j$ of the original operator $H_{j-1}$ onto a coarser scale. Applying numerical homogenization recursively produces reduced equations on increasingly coarser scales. In each homogenization step the role of the above operator $H_{j-1}$ is played by the reduced operator $R_j$ of the previous step, corresponding to an effective evolution operator.

To compute a homogenized version of a PDE we aim to approximately replace the evolution operator $H$ in a PDE by its reduced or effective counterpart (26). For the equations of the type (4) with $\tilde{H}$ is a linear evolution operator, decomposition in averages and details yields

$$\frac{\partial}{\partial t} \begin{pmatrix} \boldsymbol{d}_\phi(t) \\ \boldsymbol{s}_\phi(t) \end{pmatrix} = \begin{pmatrix} A & B \\ C & T \end{pmatrix} \begin{pmatrix} \boldsymbol{d}_\phi(t) \\ \boldsymbol{s}_\phi(t) \end{pmatrix} \; . \tag{27}$$

Employing the concept of a reduced operator as the Shur complement of the coarse grained projector we approximate solutions of (27) by solutions of [42]

$$\frac{\partial}{\partial t} \boldsymbol{s}_\phi(t) = R \boldsymbol{s}_\phi(t) \; . \tag{28}$$

The concept of the Schur complement can be utilized to define reduction operators accounting for the effect of the neglected degrees of freedom [42]. An example illustrating fine scale corrections is given by the *advection equation*

$$\frac{\partial}{\partial t} \phi(x,t) \; = \; -a(x) \cdot \frac{\partial}{\partial x} \phi(x,t) \qquad \text{with} \quad 0 \leq x \leq 1, \, a(x) > 0$$

describing wave propagation in the positive $x$-direction and $a(x)$ a non-constant function in x. Here $a(x)$ can be regarded as the velocity of a wave packet at position $x$. As an example for $a(x)$ consider the slit condition, where $a(x) = 1/6$ only for $0.45 < x < 0.55$, otherwise $a(x) = 1$. The target state $|\mathbf{v}_{27}\rangle$ that minimizes the error function (21) is shown in figure 5 for a constant (a) and a non-constant velocity profile (b), the latter adapted to the slit.

[†] For example, if $H_{j-1}$ is positive definite the same holds for $A_j$ and $A_j^{-1}$ exists [44]).

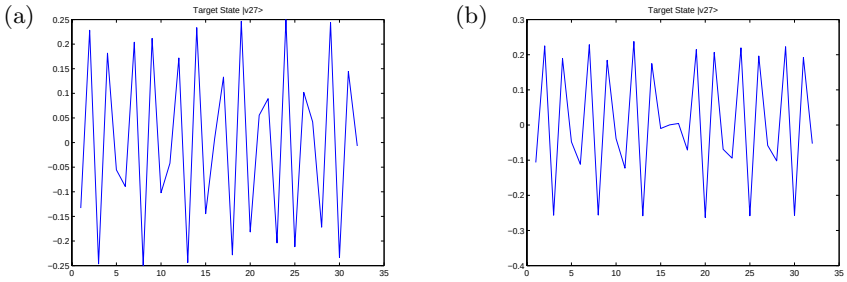



Fig. 5: Selected target states for the advection equation with (a) constant and (b) non-constant velocity $a(x)$, the latter obviously adapted to the spatial dependence.

## 3 DMRG and the Time-Evolution of Strongly Correlated Many-Body Systems

Many physical systems may not be modelled with partial differential equations or related mathematical tools, since they may not be described as a small set of interacting classical fields. Most of them belong to the class of strongly correlated many-body systems. The high correlations prevent us from using local field averages as suitable coarse-grained descriptors. Examples range from non-equilibrium reaction-diffusion equations in low dimension [45], magnetic and superconducting systems [46, 40] to quantum chromodynamics [47].

This chapter presents the idea behind one of the main tools for the extraction of the relevant degrees of freedom in those systems: the Density Matrix Renormalization Group (DMRG), along with its application to the computation of the evolution of strongly correlated many-body systems. DMRG techniques are introduced in a rather non-standard way, making extensive use of the Matrix Product Ansatz. All the essential features needed for understanding the success of the technique are exposed, but because of lack of space we do not delve into the technical details necessary to implement an efficient DMRG algorithm. The interested reader will find them in [48], [49], [50] or [51].

### 3.1 Second Quantization Many-Body Formalism

This section introduces the second quantization formalism, which is one of the most widely used tools in modern physics, through a simple example. Let us consider a 1D lattice (see [52]) with $N$ sites, any of which may be either occupied (1) or empty (0). Occupation is related to the presence of some type of chemical species $A$ which (a) is injected at one of the boundaries with a given rate, (b) diffuses through the system, and (c) reacts with itself, annihilating: $A + A \rightarrow \emptyset$.

Those three events are *probabilistic*, i.e.: a particle at a given site may jump to any of its neighbour sites, if they are empty, with a certain probability $K_d$ per unit time, which we call the *diffusion rate*. The probability (per unit time) of spontaneous appearance of a particle at the chosen injection site shall be denoted by $K_s$ and the probability (per unit time) of annihilation of two neighbouring particles by $K_a$. The three processes are exemplified in figure 6.

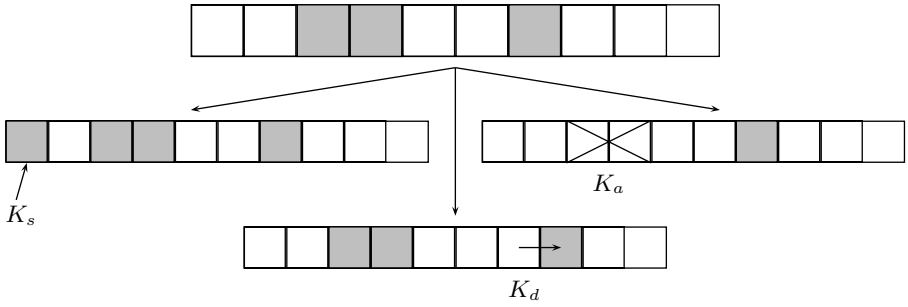



Fig. 6: The three elementary probabilistic processes for our system. The shaded boxes denote occupied sites of a one dimensional lattice.

Let us consider the long-time behaviour of the system. The first natural approach is *mean field theory*, i.e.: to neglect fluctuations. In this case we arrive at the *mass-action law* [45], which assumes that diffusion achieves a perfect mixing. In low dimensional systems diffusion is unable to provide such a perfect homogenization, and fluctuations remain essential.

There are two possible states for each site, thus $2^N$ possible states for a lattice of $N$ sites. Each of these states ($|00\cdots0\rangle$, $|00\cdots1\rangle$... $|11\cdots1\rangle$) may be considered to be a basis vector of a Hilbert space $\Omega = \mathbb{C}^{2^N}$. Any vector from that space may be regarded as a probability distribution for our problem provided that all its components (in the canonical basis) are real, positive and add up to 1. E.g.: $|P\rangle = 1/3\,|001\rangle + 2/3\,|110\rangle$. Any such probability distribution evolves under the given probabilistic rules. This evolution is governed by a linear operator in $\Omega$ which we may call the *master operator* $\mathcal{H}$. Thus, $\mathcal{H}$ is a $2^N \times 2^N$ matrix and the evolution equation is called the *master equation* [53]:

$$\partial_t\,|P(t)\rangle = -\mathcal{H}\,|P(t)\rangle \tag{29}$$

This master equation must conserve probability, i.e.: the components of $|P(t)\rangle$ must add up to 1 for all time. Therefore, the components of $\mathcal{H}\,|P(t)\rangle$ must add up to 0 for all time or, equivalently, all columns of the master operator should add up to zero.

Furthermore, $\mathcal{H}$ may be non-symmetric. If, given a couple of states $i$ and $j$, $\mathcal{H}_{ij} \neq \mathcal{H}_{ji}$, it simply means physically that the probability of transitions $i \to j$ and $j \to i$ are different. In our example, e.g., the state $|11\rangle$ may decay

to $|00\rangle$ through an annihilation process, although the inverse transition is not allowed.

In order to write down the master operator $\mathcal{H}$ for the problem illustrated in figure 6, it is convenient to use algebraic language. Let us consider, for each site $i$, the *creator* $a_i^+$ and *destructor* $a_i$ operators. The first one returns an occupied site $|1\rangle$ when acting on an empty site $|0\rangle$ and returns 0 when acting on a site which is already full. The second one, $a_i$, is its adjoint. We shall also employ the *occupancy* (or number) $n_i$ and *vacancy* $v_i$ operators, which project respectively on the occupied and the empty states for the $i$-th site. Within the basis $|0\rangle$ and $|1\rangle$ we define the one–site operators

$$a = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \quad a^+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad n = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad v = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{30}$$

The operators listed in (30) are not independent and relations among them are given by: $a^+ = (a)^\dagger$, $n = a^+ a$ and $v = I_2 - n$ with $I_2$ the identity operator for a single site (comprising two possible states). Consequently, the matrix representation of (e.g.) $n_i$ for the $i$-th particle of a $N$–sites system may be written as a tensor product:

$$n_i = I_2 \otimes \cdots I_2 \otimes n \otimes I_2 \otimes \cdots I_2 \tag{31}$$

with $n$ in the $i$-th position.

Let us consider the annihilation process for a system with two sites. The only state amenable to annihilation is $|11\rangle$. Thus, all basis states should be taken to zero but $|11\rangle$. When the master operator acts on this state, it should return the rate of change induced into the probability vector. The weight of state $|11\rangle$ should diminish and the weight of state $|00\rangle$ should increase in the same quantity, so as to conserve probability. We, thus, introduce an *annihilation operator* whose only nontrivial action is

$$\mathcal{A}\,|11\rangle = |00\rangle - |11\rangle \tag{32}$$

This operator fulfills all the properties we considered necessary to construct a master operator. In terms of the elementary operators we introduced previously, the annihilation operator acting on two sites $i$ and $j$ can be written as

$$\mathcal{A}_{ij} = a_i a_j - n_i n_j \tag{33}$$

This notation assumes that $a_i$ and all other operators act on the whole $N$–sites lattice, although their action is nontrivial only at their corresponding site. In matrix representations this is achived by a suitable tensor product with identity matrices.

Let us define a diffusion operator between two sites. This operator only acts nontrivially if one of the sites is empty and the other is occupied. In this case, it swaps the positions of *particle* and *hole*. Thus, for two particles, it

acts on $|01\rangle$ yielding $|10\rangle - |01\rangle$, and symmetrically on $|10\rangle$. Using eq. (30) we conclude that the expression in terms of the elementary operators is given by

$$\mathcal{D}_{ij} = a_i^+ a_j - v_i n_j + a_i a_j^+ - n_i v_j \tag{34}$$

Finally, the one-site operator describing the *source* is given by:

$$\mathcal{S}_i = a_i^+ - v_i \tag{35}$$

Using the representations (33), (34) and (35), the master operator for the process is constructed as

$$\mathcal{H} = K_s \mathcal{S}_1 + \sum_i \left( K_d \mathcal{D}_{i,i+1} + K_a \mathcal{A}_{i,i+1} \right) \equiv K_s \mathcal{S}_1 + \sum_i h_{i,i+1} \tag{36}$$

where we have defined a "link" operator between two neighbouring sites as $h_{i,i+1} \equiv K_d \mathcal{D}_{i,i+1} + K_a \mathcal{A}_{i,i+1}$, providing a useful shorthand notation.

The master equation may be solved given an appropriate initial condition. In general terms, this is a hard problem, since the number of degrees of freedom grows exponentially with the lattice size.

**Measurements on a state.** Let us consider a probability vector $|P(t)\rangle$ and any observable we wish to measure (e.g.: average particle number, density correlations between two points, etc.). These observables can be implemented by suitable operators. For example, the particle number is measured by $\mathcal{N} = \sum_i n_i$. When this operator acts on $|P(t)\rangle$, each component (in the canonical basis) is multiplied by its contribution to the total average. In order to obtain the final number, all such weights should be added. This final addition may be formally obtained employing a *summation state* $\langle s|$, with all components equal to 1 in the canonical basis, such that

$$\langle \mathcal{N} \rangle = \frac{\langle s | \mathcal{N} | \mathcal{P} \rangle}{\langle s | P \rangle} \tag{37}$$

Of course, $\langle s|P\rangle = 1$, and there is no need to introduce it, but it is convenient because this way the formula holds even if the states $|P\rangle$ or $\langle s|$ are not appropriately normalized. The identity operator provides the simplest example. In this case, equation (37) reduces to $\langle s|P(t)\rangle = 1$ for all time, i.e.: probability conservation.

**Eigenstates of the master operator.** Being in general a non-symmetric matrix, each eigenvalue is attached to a set of right and left eigenvectors, which need not coincide. Probability conservation directly yields a left eigenstate for the eigenvalue zero: $\langle s|$. Since all columns of $\mathcal{H}$ add up to zero, $\langle s| \mathcal{H} = 0$.

The right eigenvectors associated to this eigenvalue are *equilibrium states*, since $\partial_t |P\rangle = -\mathcal{H} |P\rangle = 0$. If this right null eigenvector is not unique, it means that the evolution process is not *ergodic*, i.e.: different initial states may lead to different stationary states at infinite time [53]. This phenomenon is also related to *spontaneous symmetry breaking* [45].

Let us consider a right eigenvector of the master operator: $\mathcal{H}\,|\psi_i\rangle = \lambda_i\,|\psi_i\rangle$, where $\lambda_i$ may be complex. If we denote $|P(0)\rangle = |\psi_i\rangle$ and allow it to evolve we find

$$|P(t)\rangle = \exp(-\lambda_i t)\,|\psi_i\rangle \tag{38}$$

Thus, the real part of the eigenvalue is the inverse of the relaxation time of the given state. If any eigenvalue had a negative real part, its corresponding eigenvector would grow up exponentially, making it unable to contribute to a probability vector. General mathematical arguments prove that there can not be such *unphysical* eigenvalues. Moreover, a purely imaginary eigenvalue may not exist either for a finite number of states (see page 54 of [45] or 110 ff. of [54]).

**Numerical paths into our problem.** The physics of the proposed non-equilibrium problem at large times can be explored using different techniques. The most straightforward one is, probably, direct Monte-Carlo simulation. Its problems related to convergence, such as *critical slow-down* are explained, e.g. in [55]. The integration of the master equation involves solving a system of $2^N$ coupled differential equations, looking for the long-term behaviour of random initial states. Another possible approach is to obtain the lowest eigenvalues and corresponding eigenstates of the master operator, which is a non-symmetric $2^N$ matrix. Although this last approach seems to be out of our computational reach for large $N$, with the aid of real space renormalization group techniques it may yield the most accurate results with the minimum computational effort.

We have disregarded the analytical approaches to the solution. A few of these systems are exactly integrable [45], and they provide valuable hints for the development of approximate methods.

### 3.2 The Matrix Product Ansatz

The Hilbert space of a many-body problem is usually too large for exact diagonalization techniques to be useful. Therefore, alternative approaches have been developed throughout this century. The first one is *perturbation theory*, i.e.: the assumption that the theory is, in a certain sense, close to a *free* theory, which may be a good starting point for our calculations. Perturbation theory has provided impressively good results for quantum electrodynamics and other theories in which correlations are not *strong*.

The second main technique is the use of *variational methods*. While perturbation theory is highly dependent on the chosen free theory, variational methods are highly dependent on the chosen *Ansatz*, which is usually suggested by physical insight. In this section an Ansatz is described which has proved rather successful for 1D and quasi-1D calculations, paradoxically requiring no previous knowledge about the physics of the system.

Let us consider a many-body 1D system with $N$ sites, each one with (e.g.) two possible states. Any wavefunction may be written as

$$\left|\Psi^N\right\rangle = \sum_{s_1,\cdots,s_N} C_{s_1\cdots s_N} \left|s_1\right\rangle \otimes \cdots \otimes \left|s_N\right\rangle \tag{39}$$

The number of $C_{s_1\cdots s_N}$ coefficients grows as $2^N$. As a variational Ansatz, the previous expression would be perfect but useless: too many parameters are involved. Thus, an alternative is required.

Consider a small set of $m_{N-1}$ states representing the Hilbert space of sites 1 to $N-1$: $\left|\Psi^{N-1}_{j_{N-1}}\right\rangle$, where $m_{N-1} \ll 2^{N-1}$ is much smaller than the dimension of the full Hilbert space. Let us assume that, for some reason, we consider these states to be the most relevant in order to represent the target state with $N$ sites. Now the following Ansatz may be adopted for the full $N$ sites system

$$\left|\Psi^N\right\rangle = \sum_{s_N,j_{N-1}} B_{(s_N,j_{N-1})} \left|s_N\right\rangle \otimes \left|\Psi^{N-1}_{j_{N-1}}\right\rangle \tag{40}$$

Pairs of indices in a parenthesis should be considered as a single index, ranging over possibile combinations. Thus, $B_{(s_N,j_{N-1})}$ must be thought as a vector of $2m_{N-1}$ components, containing the *weights* of each tensor product. Adopting the *Russian dolls* view, typical in RG, it seems natural to expand the $\left|\Psi^{(N-1)}_{j_{N-1}}\right\rangle$ in the same way:

$$\left|\Psi^{N-1}_{j_{N-1}}\right\rangle = \sum_{s_{N-1},j_{N-2}} B_{j_{N-1},(s_{N-1},j_{N-2})} \left|s_{N-1}\right\rangle \otimes \left|\Psi^{N-2}_{j_{N-2}}\right\rangle \tag{41}$$

Now $B_{j_{N-1},(s_{N-1},j_{N-2})}$ is not a vector but a *matrix* of dimensions $m_{N-1} \times 2m_{N-2}$. This relation may be now recoursed for all sites in order to obtain the full fledged *Matrix Product Ansatz* (MPA) ([59], [56], [60]):

$$\left|\Psi^N\right\rangle = \sum B^{(N)}_{(s_N,j_{N-1})} B^{(N-1)}_{j_{N-1},(s_{N-1},j_{N-2})} \cdots B^{(3)}_{j_3,(s_3,j_2)} B^{(2)}_{j_2,(s_2,s_1)} \left|s_N\right\rangle \otimes \cdots \otimes \left|s_1\right\rangle \tag{42}$$

The sum extends over all repeated indices, i.e.: $\{s_1 \ldots s_N\}$ and $\{j_2 \ldots j_{N-1}\}$. The $B^k$ matrices shall be called *truncation* matrices. The columns of $B^k$ represent the relevant degrees of freedom for the system of $k$ sites in the tensor basis of the states of site $k$ with the relevant degrees of freedom of the system with $k-1$ sites.

We may represent equation (42) formally as

$$\left|\Psi^N\right\rangle = \sum B^{(N)} B^{(N-1)} \cdots B^{(2)} \left|s_N \cdots s_1\right\rangle \tag{43}$$

With this form, our many-body problem is reduced to finding the values of the matrix entries for all the $B^{(k)}$ such that certain magnitude, which is always quadratic in them, is minimum. I.e.: a variational approach using the entries of the matrices as variational parameters.

There are two main difficulties: (a) The constraints for the $B^{(k)}$ entries: their columns must make up an orthonormal set; (b) The high number of variational parameters. Let us consider all the $m_k$ to be equal to a certain average $m$. The dimensions of each matrix would be $m \times 2m$. Therefore, the total number of parameters is $\approx 2Nm^2$. Both issues are solved in the most successful practical implementation of the MPA, the DMRG, which is explained in the following section.

How accurate is this Ansatz? In practice, if $m$ is high, our subspace should be large enough to contain a good approximation to our target state. This last assertion is not gratuitous: some theoretical background may be found in [56, 57, 58]. If the correlation function decays exponentially, then the MPA gives a good approximation with a low $m$. If it decays as a power law, results have worse quality [59].

So, summarizing, the Matrix Product Ansatz is the assumption that our target state may be written as a product of consecutive truncation matrices, and the idea of using their elements as variational parameters. Many real space RG algorithms for many-body problems are based on this idea, from Wilson's approach to the Kondo problem [7, 60] up to the DMRG algorithm and its near-future extensions, described in 3.4.

### 3.3 The DMRG Implementation

The full technical details of a working DMRG program are out of the scope of these notes. The ideas exposed in this section are sufficient to write down a DMRG algorithm, but not a state-of-the-art one. The reader interested in the technical details is referred to [48]. The application to non-equilibrium phenomena may be traced back to the works of Hieida [61] and Carlon et al. [62].

The MPA may, in principle, be implemented by any minimization algorithm which respects the orthonormality of the columns of the $B$ matrices. In practice, the best known approach is to minimize for each $B$ matrix, keeping the rest of them fixed, and to iterate this procedure until convergence.

The DMRG algorithm employs this iteration on *two* series of MPA matrices. One of them *advances* leftwards and the other one rightwards. They will be denoted by $B_{p\to p+1}$ and $B_{p+1\to p}$. The principal idea is to store not only the $B$ matrices, but also the matrix representations of other operators, which are needed so as to perform only *local* computations, i.e.: to improve matrix $B_{p\to p+1}$ using only information which is local to site $p$. Otherwise, we might have to use all other $B$'s to improve matrix $B_{p\to p+1}$, thus rendering the number of operations for a single iteration at least $O(N)$.

Let us consider our 1D system divided into a *left block*, a *right block* and two sites between them, e.g. sites $p+1$ and $p+2$, as it is shown in figure 7.

Each block has an *active site*, i.e.: the rightmost site of the left block and the leftmost site of the right one. This is the site with a "dangling link", which may be used to link the block to another site or block. The active site for the

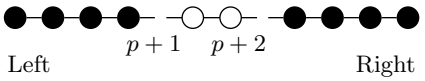



Fig. 7: Splitting of a system into left and right parts, with two sites between them.

left block, in our case, is site $p$. Each block is represented by a set of $m$ states, and a set of operators: its *master operator* and the *creation operator* at the active site. For the left block in our example, e.g., they are

$$H_p^L = K_s \mathcal{S}_1 + \sum_{i<p} h_{i-1,i} \qquad \text{and} \qquad a_{L,p}^+ \tag{44}$$

The states for the blocks are never stored in practice, only the matrix elements of the restriction of the operators to the subspace spanned by them, i.e.:

$$(H^L)_{k,l} = \left\langle \psi_k^L \right| H_p^L \left| \psi_l^L \right\rangle \qquad \text{and} \qquad (a_{L,p}^+)_{k,l} = \left\langle \psi_k^L \right| a_L^+ \left| \psi_l^L \right\rangle \ . \tag{45}$$

The terminology for *left* and *right blocks* refers to lattice geometry, not to left and right eigenstates of the master operator. We would also like to remark that the destruction, occupancy and vacancy operators may be obtained trivially from $a^+$.

Let us assume that we are given this set of matrices somehow. Now, a DMRG step consists of the following processes [51, 23, 24]:

**Superblock construction.** The *superblock* is the name for the whole system when it is rebuilt from its constituent blocks. In this stage we will explain how to write down the master operator for the superblock.

Let us remark that any operator acting on either one of the blocks or one of the sites may be promoted to act on the whole superblock by multiplying it tensorially with identity matrices of the appropriate size. From now on we shall assume all operators to be promoted to act on the appropriate Hilbert space.

We have the matrix representations of $H_p^L$, $H_p^R$, $a_{p,L}^+$ and $a_{p+2,R}^+$, all of them $m \times m$ matrices. The superblock master operator has dimension $4m^2 \times 4m^2$, and may be written as:

$$H_{SB} = H_p^L \otimes I_{4m} + I_{4m} \otimes H_p^R + h_{L,p+1} + h_{p+1,p+2} + h_{p+2,R} \tag{46}$$

$I_{4m}$ is the identity matrix of dimension $4m \times 4m$. The link $h_{p,p+1}$ is easy to write down:

$$h_{p+1,p+2} = I_m \otimes (K_d \mathcal{D}_{1,2} + K_a \mathcal{A}_{1,2}) \otimes I_m \tag{47}$$

where $\mathcal{D}_{1,2}$ and $\mathcal{A}_{1,2}$ are the 2 sites operators, given by eq. 34 and eq. 33. The procedure for $h_{L,p+1}$, which is the link of the left block to the is the following: (a) to obtain, from $a_{p,L}^+$, the full set of operators for site $p$: $a_{p,L}$,

$v_{p,L}$ and $n_{p,L}$; (b) to multiply them tensorially with $I_{4n}$ to promote them to superblock operators; (c) promote accordingly the single site operators to become superblock operators for site $p+1$. Now, follow equations 34 and 33. Of course, the same procedure applies to $h_{p+2,R}$.

**Target state obtention.** The left $\langle\Psi_L|$ and right $|\Psi_R\rangle$ ground states of the superblock master operator which we have built are an approximation to the *exact* ones, which are the target of our calculation. More accurately, the eigenvalues of the superblock master operator are *variational* estimates to the exact ones within the subspace spanned by the $2m^2$ chosen states. Since probability conservation is a consequence of the construction of the master operator, a zero eigenvalue must exist. If the operator is nondegenerate, as it is in our case, its left associated eigenvector will be the *summation state* $\langle s|$. Its right eigenstate will be, of course, the equilibrium state of the system. The first *excited states*, i.e. the right eigenvectors with smallest eigenvalues, correspond to the natural long time contaminations of the equilibrium state, since they are the slowest ones to disappear.

If only the left and right ground states are targetted, then there is no need to diagonalize the superblock master operator, which is a $2m^2$ non-symmetric matrix. In this case it suffices to solve the homogeneous linear equation $H_{SB}\,|\Psi_R\rangle = 0$ and $\langle\Psi_L|\,H_{SB} = 0$. In general, non-symmetric matrix diagonalization is a hard problem [36, 63] which one must face only when the excited states are required.

**Density matrix.** The procedure might end here, with an estimation of the ground state, but in order to find the optimal set of $B$'s it must be iterated. The procedure to perform this gives its name to the technique.

We have found an estimate for the left and right target states: $\langle\Psi_L|$ and $|\Psi_R\rangle$. Our next step will be to split the system so as the central sites are $p+2$ and $p+3$, i.e.: the left block *swallows* the $p$-th site. We may now try to find the states in that block which perform best at reproducing both $\langle\Psi_L|$ and $|\Psi_R\rangle$, i.e.: the most relevant states of the new block. The way to do it is through a least squares procedure, and the procedure is easily understood in terms of *density matrices*.

Density matrices appear naturally in quantum and statistical mechanics [64]. If we denote the (normalized) states for the *new* left block by $|l\rangle$ and those for the rest of the system by $|r\rangle$, then any state $|\xi\rangle$ for the full system may be written as

$$|\xi\rangle = \sum_{l,r} C_{l,r}\,|l\rangle \otimes |r\rangle \tag{48}$$

Its associated density matrix is defined by:

$$(\rho)_{(l_1,r_1),(l_2,r_2)} = C_{l_1,r_1} \cdot C_{l_2,r_2} \tag{49}$$

This matrix is self-adjoint, has unit trace and is positive-definite [64]. It is a density matrix, but it is also a *projector*. Therefore, its only non-zero eigenvalue is a unique 1. We now *project* on the left side through the operation:

$$(\rho^l)_{l_1,l_2} = \sum_r \rho_{l_1,l_2,r,r} \tag{50}$$

This second density matrix shares most properties with the previous one, but has an important statistical interpretation. Its eigenvalues are always in the range $[0,1]$, and they all add up to 1. The eigenvalue of a given eigenvector represents the "weight" it has in the construction of the full state $|\xi\rangle$. Thus, if we wish to retain the $m$ states which provide a best fit to $|\xi\rangle$, all we have to do is to fetch the $m$ highest eigenvalue states. The sum of the neglected eigenvalues yields a measure of the error in the truncation.

Now a criterion has been provided in order to choose the $m$ most relevant states for the left block (out of the $2m$ that we had), once we know which global state to *target*. In our case we would like to focus on *two* states instead of one, because both the left and right ground states of the effective master operator should be *well fitted*. The solution is to write the density matrix for each of them, $\rho_L^l$ and $\rho_R^l$ and then find the highest eigenvectors of the linear combination $\rho_C^l = 1/2(\rho_L^l + \rho_R^l)$.

Now the new $B_{p\to p+1}$ is built using the most relevant states of the associated density matrix as columns.

**Truncation.** Once the matrix $B_{p\to p+1}$ has been updated, we should move to matrix $B_{p+1\to p+2}$. Once the movement procedure has been exposed, the RG-cycle will be closed.

In order to update matrix $B_{p+1\to p+2}$ we need the appropriate operators so as to be able to build up the superblock. This amounts to have $a^+_{p+2,L}$, the creation operator for the $(p+2)$-th site and and $H_{p+1,L}$, the master operator for the block containing the first $p+1$ sites. The matrix representations of these operators may be found by applying the basis changing matrix $B_{p\to p+1}$ on operators employed in the last step:

$$\begin{aligned} a^+_{p+1,L} &= B^+_{p\to p+1}(I_m \otimes a^+)B_{p\to p+1} \\ H^p_L &= B^+_{p\to p+1}\left(H^{p-1}_L \otimes I_2 + h_{L,p+1}\right)B_{p\to p+1} \end{aligned} \tag{51}$$

**System Sweeping.** Now the procedure may advance and update matrix $B_{p\to p+1}$. The *sweeping* procedure continues until it reaches the end of the system. Then, the process is *reversed*, updating the $B_{p+1\to p}$ matrices in turn.

Figure 8 illustrates the sweeping procedure. In our example, we have made the left block grow at the expense of the right one. The left blocks are improved this way, while the right ones stay invariant. The process finishes when the right block takes its minimum possible size. Then the sense is reversed and it

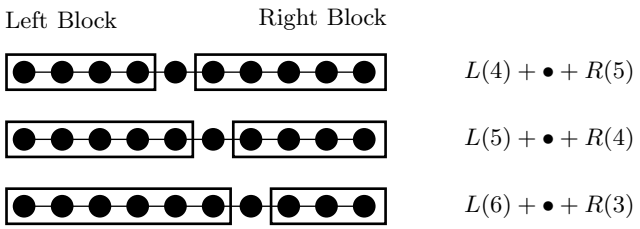



Fig. 8: Three RG-steps which start a sweeping process. The total system is composed of $N = 10$ lattice sites.

is the turn for the right block to grow and to improve its states. A complete *sweeping run* is illustrated in figure 9.

Fig. 9: A sweeping run including improvement of the left and right blocks.

The iteration of such *sweeps* is repeated until convergence, and is terminated when a desired accuracy has been achieved. The number of sweeps required for convergence does not scale with the lattice size, and in practical applications is always smaller than 10.

**Measurements.** When convergence has been attained, we wish to be able to perform measurements of arbitrary observables. Combinations of the creation operators for the active sites, stored at each block, make up a broad set of possible local observables. A caution is needed: all measurements should be performed using equation 37, i.e.: inserting the desired operator between the summation state (the left ground state) and a probability vector, such as the right ground state for equilibrium measurements.

**Warmup.** An initial set of operators is needed in order to start the sweeping stage. This set is obtained during the *warmup*. We start with the smallest possible blocks, consisting of one site. At this step the construction of the operators is trivial: the creator operator is given by the $2 \times 2$ matrix $a^+$ and the block master operator is null for the right block and $K_s\mathcal{S}_1$ for the left one.

The warmup proceeds by inflating the left and the right blocks, thus duplicating the number of stored states. If this number exceeds $m$, i.e.: the desired number of states per block, then truncation is necessary. This truncation is done in a very straightforward way: retaining the lowest eigenstates of the block master operator.

**More Technical Issues.** In order to be able to write down a state-of-the-art DMRG algorithm a few hints will be needed. The reader is referred to [48] for the details.

(a) The superblock master operator should never be diagonalized exactly. A Lanczos or Arnoldi technique should be used instead. These are, basically, sophisticated versions of the power method which use a seed (either random or, if possible, informed) which is improved in successive steps.

(b) The target states may be *recycled* from one RG-step to the next one. This procedure, which is called *wavefunction transformations* in the jargon, provides a good seed for the Lanczos or Arnoldi method.

(c) The superblock master operator should never be written down in full. Only its action on the states is important, and this may be found without computing all its matrix elements.

**Quantum Many-Body Problems.** DMRG was born [24] as a tool to analyze the low energy properties of quantum many-body systems, and it is in this field where most of its applications have been implemented. The problem of finding the equilibrium state of a probabilistic lattice model is mathematically equivalent to the obtention of the ground state of a quantum hamiltonian. This second problem is usually simpler in practice, since hamiltonians are always hermitian and, therefore, easier to diagonalize.

Usual quantum hamiltonians studied with DMRG are spin systems (e.g. Ising model in a transverse field, XY, Heisenberg), fermionic systems (e.g. t-J or Hubbard models), impurity problems (e.g. Kondo lattices) and even bosonic problems (phonons). It has been successfully extended to quantum chemistry problems and to momentum space. An excellent review of this type of applications may be found in [49].

### 3.4 Recent Developments of DMRG

This section shows some of the recent developments in the field of DMRG.

**Real Time Evolution.** In the previous sections we have considered the long time evolution of a random non-equilibrium system, which we have argued is equivalent to the ground state of a quantum system. In this section we discuss briefly how to find the real time evolution of a given state (i.e.: probability distribution) expressed in the MPA form with DMRG.

Mainly, two different types of problems may be analyzed: (a) The system is prepared in a certain state and, after that, allowed to evolve freely, or (b) the system is in its ground state, but a perturbation term is added to the probability rules (master operator). In any case, let us assume that the original state has been built in DMRG form, as a set of $B_{p\to p+1}$ and $B_{p+1\to p}$ matrices.

As it is exposed in [65], there is an efficient algorithm to apply the master operator to a DMRG state. Let us consider that we have an estimate of the target state $|\Psi\rangle$ in a certain RG-step, centered on sites $p+1$ and $p+2$. Now we may apply on it the part of the evolution operator, $\exp(-H\Delta t)$ which

regards only these sites, making use of the Suzuki-Trotter formula. The new target state is now transformed, using the *wavefunction transformations*, to the basis corresponding to sites $p+2$ and $p+3$, and the operation is iterated. If the time steps are taken with due care, at the end of each RG-sweep the target state will be evolved in $\Delta t$.

**Towards multidimensional DMRG.** The MPA is one-dimensional in nature, and so is the DMRG. It has been extended to work on tree-like structures for a long time [66]. In order to work on a multidimensional system, DMRG proceeds by preparing a pseudo-1D system, with many active sites per block. The performance of DMRG decreases with the number of links between the left and right blocks. In the 2D case, for example, of size $L \times L$, the number of links is $O(L)$. The presence of loops in the system connectivity also makes the efficience decrease as, e.g. in the case of periodic boundary conditions [24, 57].

Only recently the necessary ideas to write down a truly multidimensional DMRG have been put forward. A glimpse of the method was given in [67], in which the system was divided not into a left and a right parts, but into a *patch* and its surroundings. Thus, the number of broken links did not scale with the system size, but it was applied only to a single particle problem in quantum mechanics.

Cirac and coworkers explained in [57] the failure of DMRG on systems with periodic boundary conditions in terms of quantum information, and devised a modification of DMRG which provided much better results. In the usual MPA formula, equation (42), the two edge sites are treated in a different footing, although there is no physical difference between them and the bulk. In their extension, the MPA is modified so as its rhs becomes:

$$\sum B^{(N)}_{j_N,(s_N,j_{N-1})} B^{(N-1)}_{j_{N-1},(s_{N-1},j_{N-2})} \cdots B^{(2)}_{j_2,(s_2,j_1)} B^{(1)}_{j_1,(s_1,j_N)} \, |s_N\rangle \otimes \cdots \otimes |s_1\rangle \tag{52}$$

In the 2D case, see [68], the necessary extension of the MPA is more dramatic. In this case, the matrix for a site with $k$ neighbours is substituted by a tensor with $k$ indices. Each of them must be contracted with one index of the tensor sitting in the corresponding nearby site. Now the algorithm proceeds by minimizing the energy with respect to the tensor entries, with certain constraints. In [68] an algorithm to carry out this procedure in practice is explained, although much work is still needed in this project before it reaches the same efficiency as DMRG for 1D problems.

This approach was suggested by considerations stemming from quantum information theory, although it may be described, as we have done, without reference to it. The relation of DMRG, MPA and their extensions to AKLT states and entanglement is explored in [57, 58, 68].

## 4 Conclusions

In this review article we have given a cursory look on some RG techniques related to the coarse-graining of evolution equations, both for systems described by partial differential equations and strongly correlated many-body systems. We have overlooked all the applications stemming from momentum-space RG, such as the dynamical renormalization group.

The applications to the obtention of the most relevant degrees of freedom in the case of partial differential equations were described in section 2, where certain possible approaches were described: a geometric one, which is closest to standard coarse-graining techniques, and some physically inspired ones, e.g. the short-time evolution, related to the Lanczos diagonalization scheme, or the minimization of a certain evolution error, which provides insights into the long-term dynamics. In section 3, the application of the matrix product Ansatz (MPA) and its main technical implementation, the density matrix renormalization group (DMRG) were exposed in the context of strongly-correlated many body systems. In order to make the exposition widely available, we explained the basic ideas of the method without recourse to quantum problems, but in a much easier to understand non-equilibrium lattice model, which is mathematically equivalent in many senses.

As it has been exposed, RG methods are a natural approach to the understanding of complex spatially distributed systems, since they provide a toolbox for the selection of local degrees of freedom which are most relevant in order to describe the global dynamics. Renormalization group transformations (RGT) play the role of derivatives when dealing with problems involving different scales. And, as calculus is not an all-embracing recipe, neither is RG theory. Expertise in calculus costed humankind about two centuries of research. Therefore, it should not be a surprise to realize that RG methods, despite their successes, are still in a very early stage of development, and that much work from physicists of different branches and mathematicians is still needed in order to have a global picture of their applicability.

*Acknowledgement.* The authors would like to thank the organizers of the workshop *Model Reduction and Coarse-Graining Approaches for Multiscale Phenomena* where this work originated within an inspiring atmosphere. For clarifying and constructive scientific discussions, special thanks are given to G. Sierra, M.A. Martín-Delgado, S. Santalla, R. Cuerno, S.R. White, N.D. Goldenfeld and A.N. Gorban.

## References

1. L.P. Kadanoff: Scaling Laws for Ising Models near $T_c$. *Physics* **2**, 263–272 (1966)
2. T.W. Burkhardt, J.M.J. van Leeuwen: *Real-Space Renormalization* (Springer, Berlin Heidelberg New York 1982)

3. J.M. Yeomans: *Statistical Mechanics of Phase Transitions* (Oxford Science Publications, Oxford 2000)
4. Th. Niemeijer, J.M.J. van Leeuwen: In: *Phase Transitions and Critical Phenomena*, vol 6, ed by C. Domb, M.S. Green, 425–505 (Academic, New York 1976)
5. K.G. Wilson: Renormalization group and critical phenomena I. Renormalization group and the Kadanoff scaling picture. *Phys. Rev. B* **4**, 3174–3183 (1971)
6. K.G. Wilson: Renormalization group and critical phenomena II. Phase-cell analysis of critical behaviour. *Phys. Rev. B* **4**, 3184–3205 (1971)
7. K.G. Wilson: The renormalization group: critical phenomena and the Kondo problem. *Rev. Mod. Phys.* **47**, 773–840 (1975)
8. A.A. Migdal: Phase transitions in gauge and spin-lattice systems. *Sov. Phys. JETP* **42**, 743-746 (1976)
9. L.P. Kadanoff: Notes on Migdal's recursion formulas. *Ann. Phys.* **100**, 359–394 (1976)
10. S. Ma: Renormalization Group by Monte Carlo Methods. *Phys. Rev. Lett.* **37**, 461–464 (1976)
11. Z. Friedman, J. Felsteiner: Kadanoff block transformation by the Monte Carlo technique. *Phys. Rev. B* **15**, 5317–5319 (1977)
12. A.L. Lewis: Lattice renormalization group and the thermodynamic limit. *Phys. Rev. B* **16**, 1249–1252 (1977)
13. G.F. Mazenko, M.J. Nolan, O.T. Valls: Application of the Real-Space Renormalization Group to Dynamic Critical Phenomena *Phys. Rev. Lett.* **41**, 500–503 (1978)
14. G.F. Mazenko, J.E. Hirsch, M.J. Nolan, O.T. Valls: Dynamical Correlation Functions in the Two-Dimensional Kinetic Ising Model: A Real-Space Renormalization-Group Approach. *Phys. Rev. Lett.* **44**, 1083 (1980)
15. A.-L. Barabási, H.E. Stanley: *Fractal Concepts in Surface Growth* (Cambridge University Press 1995)
16. E. Frey, U. C. Täuber, T. Hwa: Mode-coupling and renormalization group results for the noisy Burgers equation. *Phys. Rev. E* **53**, 4424 (1996)
17. D. Stauffer, A. Aharony: *Introduction to percolation theory* (Taylor & Francis, London 1998)
18. T.W. Burkhardt, J.M.J. van Leeuwen: *Polymers Near Surfaces* (World Scientific, Singapore 1993)
19. N. Goldenfeld, B.P. Athreya, J.A. Dantzig: Renormalization group approach to multiscale modelling in materials science. *J. Stat. Phys.* (2005) in review
20. K. Gawędzki, A. Kupiainen: A rigorous block spin approach to massless lattice theories. *Commun. Math. Phys.* **77**, 31–64 (1980)
21. K. Gawędzki, A. Kupiainen: Block Spin Renormalization Group for Dipole Gas and $(\nabla\Phi)^4$. *Ann. Phys.* **147**, 198–243 (1983)
22. J.W. Bray, S.T. Chui: Computer renormalization-group calculations of $2k_F$ and $4k_F$ correlation functions of the one-dimensional Hubbard model *Phys. Rev. B* **19**, 4876–4882 (1979)
23. S.R. White: Density matrix formulation for quantum renormalization groups. *Phys. Rev. Lett.* **69**, 2863–2866 (1992)
24. S.R. White: Density-matrix algorithms for quantum renormalization groups. *Phys. Rev. B* **48**, 10345–10356 (1993)
25. N.D. Goldenfeld: *Lectures on phase transitions and the renormalization group* (Perseus Books, Reading, Massachusetts 1992)

26. P.C. Hohenberg, B.I. Halperin: Theory of dynamic critical phenomena. *Rev. Mod. Phys.* **49**, 435–479 (1977)
27. J.M. Burgers: *The Nonlinear Diffusion Equation* (Riedel, Boston 1974)
28. D. Forster, D. Nelson, M. Stephen: Large-distance and long-time properties of a randomly stirred fluid. *Phys. Rev. A* **16**, 732–749 (1977)
29. M. Kardar, G. Parisi, Y.-C. Zhang: Dynamic Scaling of Growing Interfaces *Phys. Rev. Lett.* **56**, 889–892 (1986)
30. E. Frey, U.C. Taeuber: Two-loop renormalization group analysis of the Burgers-Kardar-Parisi-Zhang equation. *Phys. Rev. E* **50**, 1024–1045 (1994)
31. T. Nattermann, L.-H. Tang: Kinetic surface roughening. I. The Kardar-Parisi-Zhang equation in the weak-coupling regime. *Phys. Rev. A* **45**, 7156–7161 (1992)
32. J.P. Bouchaud, M.E. Cates: Self-consistent approach to the Kardar-Parisi-Zhang equation. *Phys. Rev. E* **47**, R1455–R1458 (1993)
33. N.D. Goldenfeld, A. McKane, Q. Hou: Block Spins for Partial Differential Equations. *J. Stat. Phys.* **93**, 699–714 (1998)
34. Q. Hou, N.D. Goldenfeld, A. McKane: Renormalization group and perfect operators for stochastic differential equations. *Phys. Rev. E* **63**, 36125 (2001)
35. A. Degenhard, J. Rodríguez-Laguna: Towards the Evaluation of the relevant degrees of freedom in nonlinear partial differential equations. *J. Stat. Phys.* **106**, 1093–1120 (2001)
36. G.H. Golub, C.F. Van Loan: *Matrix computations* (Johns Hopkins Univ. Press, Baltimore 1996)
37. A. Degenhard, J. Rodríguez-Laguna: Real-space renormalization-group approach to field evolution equations. *Phys. Rev. E* **65**, 036703 (2001)
38. A. Degenhard: A non-perturbative real-space renormalization group scheme. *J. Phys. A: Math. Gen.* **33**, 6173–6185 (2000)
39. Y. Kuramoto: *Chemical Oscillations, Waves and Turbulence* (Springer, Berlin Heidelberg New York 1984)
40. J. Gonzalez, M.A. Martín-Delgado, G. Sierra, A.H. Vozmediano: New and old real-space renormalization group methods. In: *Quantum Electron Liquids and High-$T_c$ Superconductivity*, Lecture Notes in Physics 38, (Springer, Berlin Heidelberg New York 1995)
41. A. Degenhard, J. Rodríguez-Laguna: Projection Operators for Nonlinear Evolutionary Dynamics. *SIAM Multiscale Modeling and Simulation* **4**, 641-663 (2005)
42. A. Degenhard, S. Getfert, J. Rodríguez-Laguna: Reduction Schemes for Multiscale Evolutionary Dynamics. *Model Reduction and Coarse-Graining for Multiscale Phenomena*, abstract (Leicester University 2005)
43. I. Daubechies: *Ten Lectures on Wavelets* (SIAM, Philadelphia 1992)
44. G. Beylkin, N. Coult: A multiresolution strategy for the reduction of elliptic PDEs and eigenvalue problems. *Appl. Comp. Harmon. Anal.* **5**, 129–155 (1998)
45. G.M. Schütz: Exactly solvable models for many-body systems far from equilibrium. In: *Phase transitions and critical phenomena*, vol 19, C. Domb and J. Lebowitz (eds.), 1–251 (Academic press 2000)
46. A. Auerbach: Interacting Electrons and Quantum Magnetism (Springer, Berlin Heidelberg New York 1994)
47. M. Creutz: *Quarks, gluons and lattices* (Cambridge University Press 1983)
48. I. Peschel, X. Wang, M. Kaulke, K. Hallberg (eds.): *Density Matrix Renormalization* (Springer, Berlin Heidelberg New York 1999)

49. K. Hallberg: Density matrix renormalization: a review of the method and its applications. Preprint available at `cond-mat/0303557` (2003)
50. U. Schollwöck: The density-matrix renormalization group. *Rev. Mod. Phys.* **77**, 259–315 (2005)
51. J. Rodríguez-Laguna: Real Space Renormalization Group Techniques and Applications. Ph.D. thesis, Universidad Complutense de Madrid, Madrid, Spain (2002) Preprint available at `condmat/0207340`
52. A. Degenhard, J. Rodríguez-Laguna, S. Santalla: Density Matrix Renormalization Group Approach to Nonequilibrium Phenomena. *SIAM Multiscale Modeling and Simulation*, **3**, 89–105 (2004)
53. R. Zwanzig: *Nonequilibrium Statistical Mechanics* (Oxford University Press 2001)
54. N.G. van Kampen: *Stochastic processes in phyiscs and chemistry* (North-Holland Publ. Co., Amsterdam 1981)
55. K. Binder, D.W. Heermann: *Monte Carlo simulation in statistical physics* (Springer, Berlin Heidelberg New York 1992)
56. M. Fannes, B. Nachtergaele, R.F. Werner: Finitely correlated states in spin chains. *Comm. Math. Phys.* **144**, 443–490 (1992)
57. F. Verstraete, D. Porras, J. I. Cirac: Density Matrix Renormalization Group and Periodic Boundary Conditions: A Quantum Information Perspective *Phys. Rev. Lett.* **93**, 227205 (2004)
58. F. Verstraete, J.I. Cirac: Matrix product states represent ground states faithfully. Preprint available at `cond-mat/0505140` (2005)
59. S. Rommer, S. Östlund: Class of ansatz wave functions for one-dimensional spin systems and their relation to the density matrix renormalization group. *Phys. Rev. B* **55**, 2164–2181 (1997)
60. F. Verstraete, A. Weichselbaum, U. Schollwck, J. I. Cirac, J. von Delft: Variational matrix product state approach to quantum impurity models. Preprint available at `cond-mat/0504305` (2005)
61. Y. Hieida: Application of the Density Matrix Renormalization Group Method to a Non-Equilibrium Problem. *J. Phys. Soc. Jpn.* **67**, 369-372 (1998)
62. E. Carlon, M. Henkel, U. Schollwöck: Density matrix renormalization group and reaction-diffusion processes *Eur. Phys. J. B* **12**, 99–114 (1999)
63. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery: *Numerical Recipes in C* (Cambridge University Press 1997) Also at `http://www.nr.com`
64. R.P. Feynman: *Statistical mechanics: a set of lectures* (Benjamin, Reading, MA 1972)
65. S.R. White, A.E. Feiguin: Real time evolution using the density matrix renormalization group. *Phys. Rev. Lett.* **93**, 076401 (2004)
66. M.A. Martín-Delgado, J. Rodriguez-Laguna, G. Sierra: A density matrix renormalization group study of excitons in dendrimers. *Phys. Rev. B* **65**, 155116 (2002)
67. M.A. Martín-Delgado, J. Rodríguez-Laguna, G. Sierra: Single block renormalization group: quantum mechanical problems. *Nucl. Phys. B* **601**, 569–590 (2001)
68. F. Verstraete, J.I. Cirac: Renormalization group for quantum many-body systems in two and higher dimension. Preprint available at `cond-mat/0407066` (2004)

# A Stochastic Process Behind Boltzmann's Kinetic Equation and Issues of Coarse Graining

H. C. Öttinger

ETH Zürich, Department of Materials, Polymer Physics, HCI H 543, CH-8093 Zürich, Switzerland, hco@mat.ethz.ch, http://www.polyphys.mat.ethz.ch/

**Summary.** We consider a stochastic process behind Boltzmann's kinetic equation that is obtained by identifying the infinitesimal generator of a Markov process. Within an intrinsically probabilistic interpretation, the nonlinear nature of Boltzmann's equation, which can be regarded as an expression of self-consistency, can be achieved via weakly interacting Markov processes. Whereas the Markov process associated with the Boltzmann equation is known as a powerful tool both to study fundamental mathematical issues of existence and to solve practical engineering problems, we here consider fundamental physical issues of coarse graining and, in particular, the role of diffusion in hydrodynamic equations. As a provocative conclusion, it may be less important *to solve* the Boltzmann equation than *to coarse grain* it.

## 1 Motivation and Problem

The passage from the reversible equations of motion for a many-particle system to the irreversible Boltzmann equation has been a highly controversial topic for more than a century. Irreversibility arises as a result of coarse graining from the many-particle system to the single-particle distribution function. The resulting Boltzmann equation is a milestone in understanding gas dynamics and irreversibility. From the perspective of physical understanding, coarse graining to obtain the Boltzmann equation is a much bigger achievement than solving the equations of motion for a many-particle system.

Once we have realized that it may be more important to coarse grain equations than to solve them, the question arises how we shall proceed with Boltzmann's equation. Shall we finally solve it or shall we further coarse grain it? Human instincts seem to favor solutions. Solution techniques, such as Grad's thirteen-moment method or the first-order Chapman-Enskog method, are commonly employed to produce hydrodynamic equations [1, 2].

By passing from Boltzmann's kinetic equation to hydrodynamic equations, one coarse grains from the scale of the particle size ($10^{-10}$ m) to the scale

of cars or airplanes (10 m). Statistical nonequilibrium thermodynamics suggests that new irreversible processes should arise in such an enormous coarse graining step [3]. Rapid reversible motions are turned into irreversible fluctuations accompanied by friction. Irreversibility is not produced once and for all times—additional irreversibility arises in any step of coarse graining. The controversial step behind the derivation of the Boltzmann equation has to be repeated again and again, and the usual solution techniques are inappropriate for that purpose. It is hence important to understand systematic coarse graining techniques [3, 4, 5, 6].

In view of the fact that there exist important intermediate length and time scales in a gas, namely the mean free path ($10^{-7}$ m) and the collision time, it is natural to expect that additional irreversible processes arise in passing from the Boltzmann equation to hydrodynamics. Indeed, one realizes that diffusion effects are introduced into the hydrodynamic equations when multicollision events are taken into account. By *solving* Boltzmann's equation we obtain the Navier-Stokes-Fourier equations of hydrodynamics which contain the viscosity and the thermal conductivity, but not the diffusion coefficient. More general equations of hydrodynamics [7, 8, 9] containing all transport coefficients on an equal footing are obtained only after *coarse graining* Boltzmann's equation. These issues provide our motivation to look at the Boltzmann equation from the illuminating perspective of stochastic processes in this paper.

Boltzmann's kinetic equation describes the time-evolution of the single-particle distribution function $p_t(\boldsymbol{r}, \boldsymbol{p})$, that is, the probability density for finding a particle in a rarefied gas at the time $t$ with momentum $\boldsymbol{p}$ at the position $\boldsymbol{r}$ within a volume $V$. The question that we first need to address in this paper is how one can construct a stochastic process $(\boldsymbol{R}_t, \boldsymbol{P}_t)_{t\geq 0}$ in $V^3 \times \mathbb{R}^3$ (equipped with the Borel $\sigma$-algebra) for which, at any time $t \geq 0$, $p_t(\boldsymbol{r}, \boldsymbol{p})$ is the probability density of the random variable $(\boldsymbol{R}_t, \boldsymbol{P}_t)$. We refer to such a stochastic process as a *Boltzmann process*, and we look for the most natural construction.

Any stochastic process implies information not only about random variables at given times, but also about correlations between random variables at different times. More specifically, a Boltzmann process implies all finite-dimensional marginal distributions $P_{t_1 \ldots t_n}$, which are probability measures on $(V^3 \times \mathbb{R}^3)^n$, for all $n$ and arbitrary sequences of times $0 \leq t_1 < t_2 < \ldots < t_n$.

Conversely, the collection of all finite-dimensional marginal distributions is sufficient to characterize the distribution of a stochastic process (according to the Kolmogorov extension theorem; see, for example, Sect. 2.3.1 of [10] or Sect. 12.1 of [11]). Of course, these finite-dimensional marginal distributions need to be compatible to characterize a process, that is, if we "integrate out" the particle position and momentum at one of the times then we obtain the corresponding lower-dimensional marginal distribution. Only the distribution of the stochastic process is unique; the random variable at any time can be altered on any subset with zero probability.

In general, the construction of a family of compatible finite-dimensional marginal distributions is a major task. Two special classes of stochastic processes, however, can be constructed quite elegantly: Markov and Gaussian processes. Only Boltzmann processes of these categories are considered in this paper.

As a benefit of constructing a stochastic process behind the nonlinear Boltzmann equation, we expect to obtain some new perspectives on old problems (which is much more than old wine in new bottles). In spite of the impressive progress on the initial value problem [12], the development of a rigorous mathematical theory of the Boltzmann equation is far from complete. A summary of some rigorous achievements based on the stochastic perspective and a number of unsolved problems can be found in a recent paper [13]. In the last 20 years, an intrinsically probabilistic approach has considerably enhanced our understanding of the Boltzmann equation. This approach can be traced back to the famous work of Kac [14] who formulated a linear master equation for an ensemble of interacting gas particles. In the limit of large ensembles, initially independent particles remain independent for all times ("propagation of chaos"), and the probability density for each particle satisfies the nonlinear Boltzmann equation; this nonlinearity arises from a mean field effect. Deeper insight into the "propagation of chaos" was provided in an influential contribution by McKean [15], in particular, through an analogy to the central limit theorem. Important steps toward a rigorous justification of the "propagation of chaos" for spatially homogeneous systems were made by Grünbaum [16], Tanaka [17], and Sznitman [18]. The situation for spatially inhomogeneous systems has been summarized in a broad educational introduction to probabilistic tools for the Boltzmann equation [19], and even the rate of convergence to equilibrium has been estimated for the inhomogeneous case [13].

In polymer kinetic theory, the analogous stochastic perspective [10] turned out to be beneficial to clarify a number of fundamental problems and to develop efficient computer simulation techniques. Moreover, to understand how the irreversible nature of the Boltzmann equation arises, different concepts of convergence for random variables are known to be of crucial importance [20] so that a fully stochastic analysis might be useful to understand coarse graining. We here develop some thoughts on the procedures to introduce the nonlinearity into the Boltzmann equation, on fundamental coarse graining issues, on thermodynamic admissibility, and on approximate kinetic equations.

In the following, we first describe some key elements of the theory of Markov processes, most importantly, the infinitesimal generator. For guidance and comparison, we summarize the situation for the role model of nonlinear Fokker-Planck equations, for which a well-elaborated theory, a whole range of applications, and appropriate simulation tools exist in the applied literature (both in the natural sciences and in engineering). After presenting Boltzmann's equation in the usual form, we consider the infinitesimal generator of the naturally associated Markov process. Even at equilibrium, the construction of a Gaussian Boltzmann process requires an approximation

which corresponds to the use of Fokker-Planck equations in kinetic theory. The calculation of the self-diffusion coefficient for the Boltzmann process illustrates how one can benefit from the formulation of a full stochastic process in kinetic theory and how an additional irreversible process can arise in the passage from Boltzmann's equation to hydrodynamics.

## 2 Markov Processes

Loosely speaking, a stochastic process or a stochastic dynamical system satisfies the Markov property if the future state of the system at any time $t' > t$ is independent of the past behavior of the system at times $t'' < t$, given the present state at time $t$. In other words, the further evolution of the process depends on its history only through the current state of the process. This property, formulated by the Russian mathematician Andrei Andreievich Markov (1856–1922) in 1906, plays an important role in the theory of stochastic differential equations for which, like for ordinary differential equations, the future time evolution can be expressed in terms of the current state, independent of previous states of the system.

For Markov processes, the idea of two-time transition probabilities is a key concept. The finite-dimensional marginal distributions $P_{t_1 \ldots t_n}$ can be constructed from the initial distribution and the transition probabilities between any two successive times $t_{j-1}$ and $t_j$. The compatibility of the family of finite-dimensional marginal distributions is guaranteed by the Chapman-Kolmogorov equation which states that, for $t_1 < t_2 < t_3$, the transition probability from $t_1$ to $t_3$ is obtained from the transition probabilities from $t_1$ to $t_2$ and from $t_2$ to $t_3$ after integrating over all possible states at the intermediate time $t_2$.

The two-parameter family of transition probabilities still contains a large amount of redundant information, because each time step can be decomposed into a sequence of smaller time steps. By going to the limit of infinitesimal time steps one can obtain a one-parameter description of Markov processes and, at the same time, eliminate the remaining Chapman-Kolmogorov compatibility condition by avoiding redundant information. A useful concept in studying infinitesimally small time steps is the infinitesimal generator which, for general Markov processes $(X_t)_{t \geq 0}$, can be defined by

$$\mathcal{L}_t g(x) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \left[ E(g(X_{t+\Delta t}) | X_t = x) - g(x) \right], \tag{1}$$

where $E(\cdot|\cdot)$ is the conditional expectation. The infinitesimal generator $\mathcal{L}_t$ is a linear operator whose domain is the set of measurable functions $g$ for which the limit on the right side of Eq. (1) exists (see Sect. 2.4 of [21]). Of course, the domain must be sufficiently rich, for example, dense in the class of bounded functions (cf. Sect. XIII.9 of [22]). The infinitesimal generator can be determined from the transition probabilities and, conversely, the transition proba-

bilities can be reconstructed from the infinitesimal generator. When operating on a function $g$, the infinitesimal generator describes the rate of change of the average of $g$ under the condition that the process starts at a given position $x$ at time $t$. The one-parameter family of linear operators $\mathcal{L}_t$, $t \geq 0$, together with the initial distribution $P_0$, yields the most compact characterization of a general Markov process.

If $\mathcal{L}_t^\dagger$ denotes the adjoint of the operator $\mathcal{L}_t$, then the probability density $p_t$ is governed by the equation

$$\frac{\partial}{\partial t} p_t = \mathcal{L}_t^\dagger p_t. \tag{2}$$

This equation provides a recipe to construct a Markov process from the time-evolution equation for the probability density $p_t$: One first reads off the linear operator $\mathcal{L}_t^\dagger$ from some given time-evolution equation. Note that the operator $\mathcal{L}_t^\dagger$ must preserve the normalization and positivity of $p_t$. The adjoint operator $\mathcal{L}_t$ can then be used as an infinitesimal generator to define a Markov process.

Actually, Eq. (2) can also be used to reconstruct the transition probabilities from the infinitesimal generator if it is solved for a sufficiently rich set of initial conditions. When considered as an equation for transition probabilities, Eq. (2) is known as Kolmogorov's forward equation. In the special case of a second-order differential operator $\mathcal{L}_t^\dagger$, we obtain the Fokker-Planck equation.

The infinitesimal generator expresses the trend of a Markov process, and the remainder is a martingale [10]. For a rigorous treatment of nonlinear Markov processes it is useful to realize that all Markov processes can hence be specified by a martingale problem [19]. For example, a weak solution of the Boltzmann equation follows by averaging the associated martingale problem.

## 3 Nonlinear Fokker-Planck Equations

Fokker-Planck equations are usually given as time-evolution equations for probability densities $p_t(\boldsymbol{x})$ on $\mathbb{R}^d$,

$$\frac{\partial}{\partial t} p_t(\boldsymbol{x}) = -\frac{\partial}{\partial \boldsymbol{x}} \cdot [\boldsymbol{A}_t(\boldsymbol{x}) p_t(\boldsymbol{x})] + \frac{1}{2} \frac{\partial}{\partial \boldsymbol{x}} \frac{\partial}{\partial \boldsymbol{x}} : [\boldsymbol{D}_t(\boldsymbol{x}) p_t(\boldsymbol{x})], \tag{3}$$

where $\boldsymbol{A}_t(\boldsymbol{x})$ and $\boldsymbol{D}_t(\boldsymbol{x})$ are suitable column vector and matrix functions, respectively. The $d \times d$-matrix $\boldsymbol{D}_t(\boldsymbol{x})$ must be symmetric and positive-semidefinite, that is, it can be decomposed in the form

$$\boldsymbol{D}_t(\boldsymbol{x}) = \boldsymbol{B}_t(\boldsymbol{x}) \cdot \boldsymbol{B}_t(\boldsymbol{x})^T \tag{4}$$

with a $d \times d'$-matrix $\boldsymbol{B}_t(\boldsymbol{x})$ ($d'$ is conveniently but not necessarily chosen as the rank of $\boldsymbol{D}_t(\boldsymbol{x})$). By comparing the Fokker-Planck equation (3) to Kolmogorov's forward equation, we can identify the second-order differential operator $\mathcal{L}_t^\dagger$ and find the infinitesimal generator of a Markov process,

$$\mathcal{L}_t = \boldsymbol{A}_t(\boldsymbol{x}) \cdot \frac{\partial}{\partial \boldsymbol{x}} + \frac{1}{2}\boldsymbol{D}_t(\boldsymbol{x}) : \frac{\partial}{\partial \boldsymbol{x}}\frac{\partial}{\partial \boldsymbol{x}}. \tag{5}$$

A famous theorem states that the corresponding Markov process coincides with the solution of the Itô stochastic differential equation [10, 21, 23, 24]

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{A}_t(\boldsymbol{X}_t)\mathrm{d}t + \boldsymbol{B}_t(\boldsymbol{X}_t) \cdot \mathrm{d}\boldsymbol{W}_t, \tag{6}$$

where $\boldsymbol{W}_t$ is a $d'$-dimensional Wiener process. The equivalence between Fokker-Planck equations and stochastic differential equations is the basis for Brownian dynamics simulations in kinetic theory.

The Fokker-Planck equation corresponding to a stochastic differential equation is linear in the probability density. In a number of applications [10, 25], one is faced with the situation that the diffusion equation is nonlinear in the probability density $p_t$ because the drift term and/or the diffusion tensor in the Fokker-Planck equation depend on averages evaluated with $p_t$.

We here consider the situation in which the coefficients $\boldsymbol{A}_t(\boldsymbol{x})$ and $\boldsymbol{B}_t(\boldsymbol{x})$ in the stochastic differential equation (6) are replaced by $\boldsymbol{A}(\boldsymbol{x}, \langle \boldsymbol{g}_A(\boldsymbol{x}, \boldsymbol{X}_t)\rangle)$ and $\boldsymbol{B}(\boldsymbol{x}, \langle \boldsymbol{g}_B(\boldsymbol{x}, \boldsymbol{X}_t)\rangle)$, that is, they depend on expectations of one or several real-valued functions. Provided that the functions $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{g}_A$, and $\boldsymbol{g}_B$ satisfy Lipschitz conditions and suitable growth conditions [26], the existence of a unique solution of the modified stochastic differential equation can be established.

We next describe a law of large numbers that is of great importance for the numerical simulation of such processes with mean field interactions. Consider the system of $n$ stochastic differential equations,

$$\begin{aligned}\mathrm{d}\boldsymbol{X}_t^{(j)} = {} & \boldsymbol{A}\left(\boldsymbol{X}_t^{(j)}, \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{g}_A(\boldsymbol{X}_t^{(j)}, \boldsymbol{X}_t^{(k)})\right)\mathrm{d}t \\ & + \boldsymbol{B}\left(\boldsymbol{X}_t^{(j)}, \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{g}_B(\boldsymbol{X}_t^{(j)}, \boldsymbol{X}_t^{(k)})\right) \cdot \mathrm{d}\boldsymbol{W}_t^{(j)},\end{aligned} \tag{7}$$

for $j = 1, 2, \ldots, n$, where the processes $\boldsymbol{W}^{(j)}$ are independent $d'$-dimensional Wiener processes. In other words, the expectations in the drift and diffusion terms are replaced by ensemble averages. We furthermore assume that the initial conditions $\boldsymbol{X}_0^{(j)}$ are independent, identically distributed random variables.

If $\boldsymbol{X}_t$ is the solution of the stochastic differential equation with mean field interactions, where the distribution of the initial condition $\boldsymbol{X}_0$ coincides with the distribution of the random variables $\boldsymbol{X}_0^{(j)}$ and the coefficient functions satisfy Lipschitz and suitable growth conditions, then we have the following result for the mean-square limit:

$$\text{ms-}\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} g(\boldsymbol{X}_t^{(j)}) = \langle g(\boldsymbol{X}_t)\rangle, \tag{8}$$

for all bounded continuous functions $g : \mathbb{R}^d \to \mathbb{R}$ [26]. Convergence theorems under weaker conditions for the coefficient functions have been described in [27]. Typical fluctuations for finite $n$ are expected to be of the order of $n^{-1/2}$; a detailed discussion of these fluctuations can be found in [28]. In a pioneering paper on nonlinear diffusions, McKean [29] has shown that, under suitable smoothness conditions, one has the stronger mode of almost sure convergence in (8), so that one has a strong law of large numbers. These results for nonlinear Fokker-Planck equations have analogous counterparts in the theory of Boltzmann processes.

The processes $\boldsymbol{X}^{(j)}$ are often referred to as "weakly interacting stochastic processes." The weakness of the interaction between any two of these processes is obvious from the factor $1/n$ in front of the interaction terms in (7). For large $n$, any finite number of processes becomes essentially independent; this asymptotic independence property expresses the previously mentioned "propagation of chaos" [14, 27, 29, 30].

## 4 Boltzmann's Kinetic Equation

Boltzmann's kinetic equation for the single-particle distribution function can be written in the form

$$\frac{\partial p_t(\boldsymbol{r},\boldsymbol{p})}{\partial t} = -\left(\frac{\boldsymbol{p}}{m}\cdot\frac{\partial}{\partial \boldsymbol{r}} - \frac{\partial \phi^{(\mathrm{e})}(\boldsymbol{r})}{\partial \boldsymbol{r}}\cdot\frac{\partial}{\partial \boldsymbol{p}}\right) p_t(\boldsymbol{r},\boldsymbol{p}) + \left[\frac{\partial p_t(\boldsymbol{r},\boldsymbol{p})}{\partial t}\right]_{\mathrm{coll}}. \quad (9)$$

The first term on the right-hand side of this equation corresponds to the free flight of a single particle of mass $m$ in an external field with potential $\phi^{(\mathrm{e})}$. The second term describes the change of the single-particle distribution function due to the occurrence of collisions with the other particles in the system.

The collision term in (9) consists of gain and loss contributions. If a particle with the given momentum $\boldsymbol{p}$ collides with another particle with any momentum $\boldsymbol{p}'$, this leads to a decrease of $p_t(\boldsymbol{r},\boldsymbol{p})$. If two particles with momenta $\boldsymbol{q}$ and $\boldsymbol{q}'$ collide such that one of the particles acquires the momentum $\boldsymbol{p}$, this leads to an increase of $p_t(\boldsymbol{r},\boldsymbol{p})$. For a precise formulation, we introduce the transition rate $w(\boldsymbol{q},\boldsymbol{q}'|\boldsymbol{p},\boldsymbol{p}')$ for a pair of particles with initial momenta $\boldsymbol{p}$, $\boldsymbol{p}'$ to the final momenta $\boldsymbol{q}$, $\boldsymbol{q}'$. For an elastic collision, $w(\boldsymbol{q},\boldsymbol{q}'|\boldsymbol{p},\boldsymbol{p}')$ vanishes unless momentum and energy are conserved in a collision,

$$\boldsymbol{p}+\boldsymbol{p}' = \boldsymbol{q}+\boldsymbol{q}', \quad (10)$$

$$\boldsymbol{p}^2+\boldsymbol{p}'^2 = \boldsymbol{q}^2+\boldsymbol{q}'^2. \quad (11)$$

The two free parameters in the choice of $\boldsymbol{q}$ and $\boldsymbol{q}'$ [six components minus four conditions (10), (11)] are determined by the solid angle into which a colliding particle is scattered. Furthermore, the transition rate $w(\boldsymbol{q},\boldsymbol{q}'|\boldsymbol{p},\boldsymbol{p}')$ has the following symmetry properties corresponding to time reversal and particle exchange in the two-particle collisions:

$$w(\boldsymbol{q},\boldsymbol{q}'|\boldsymbol{p},\boldsymbol{p}') = w(\boldsymbol{p},\boldsymbol{p}'|\boldsymbol{q},\boldsymbol{q}'), \tag{12}$$

and

$$w(\boldsymbol{q},\boldsymbol{q}'|\boldsymbol{p},\boldsymbol{p}') = w(\boldsymbol{q}',\boldsymbol{q}|\boldsymbol{p}',\boldsymbol{p}). \tag{13}$$

We can now write

$$\left[\frac{\partial p_t(\boldsymbol{r},\boldsymbol{p})}{\partial t}\right]_{\text{coll}} = N\int w(\boldsymbol{q},\boldsymbol{q}'|\boldsymbol{p},\boldsymbol{p}')[p_t(\boldsymbol{r},\boldsymbol{q})\tilde{p}_t(\boldsymbol{r},\boldsymbol{q}') - p_t(\boldsymbol{r},\boldsymbol{p})\tilde{p}_t(\boldsymbol{r},\boldsymbol{p}')]\,\mathrm{d}^3p'\,\mathrm{d}^3q\,\mathrm{d}^3q', \tag{14}$$

where $N$ is the total number of particles in the system. Equation (9) with the collision integral (14) is Boltzmann's famous kinetic equation. In Eq. (14) we have distinguished between the probability density $p_t$ for a test particle and the probability density $\tilde{p}_t$ for the ensemble of background scatterers. To arrive at the Boltzmann equation one needs to assume the self-consistency condition $\tilde{p}_t = p_t$.

The assumption that the probability for a two-particle collision is proportional to the product $p_t(\boldsymbol{r},\boldsymbol{p})\tilde{p}_t(\boldsymbol{r},\boldsymbol{p}')$ is Boltzmann's famous *Stoßzahlansatz*. It has been assumed that the collisions are strictly local so that the same argument $\boldsymbol{r}$ appears in all single-particle distribution functions occurring in Boltzmann's equation. The transition rates $w$ for two-particle collisions are independent of the number of particles in the system; they only depend on the interaction potential and on the impact conditions.

## 5 Boltzmann Process

For a stochastic interpretation, the collision integral (14) in Boltzmann's kinetic equation (9) can be rewritten as

$$\left[\frac{\partial p_t(\boldsymbol{r},\boldsymbol{p})}{\partial t}\right]_{\text{coll}} = \int P(\boldsymbol{p}|\boldsymbol{q})\nu(\boldsymbol{q})p_t(\boldsymbol{r},\boldsymbol{q})\,\mathrm{d}^3q - \nu(\boldsymbol{p})p_t(\boldsymbol{r},\boldsymbol{p}), \tag{15}$$

with the collision rate

$$\nu(\boldsymbol{p}) = N\int w(\boldsymbol{q},\boldsymbol{q}'|\boldsymbol{p},\boldsymbol{p}')\tilde{p}_t(\boldsymbol{r},\boldsymbol{p}')\,\mathrm{d}^3p'\,\mathrm{d}^3q\,\mathrm{d}^3q' \tag{16}$$

and the single-collision transition probabilities

$$P(\boldsymbol{p}|\boldsymbol{q})\nu(\boldsymbol{q}) = N\int w(\boldsymbol{p},\boldsymbol{p}'|\boldsymbol{q},\boldsymbol{q}')\tilde{p}_t(\boldsymbol{r},\boldsymbol{q}')\,\mathrm{d}^3p'\,\mathrm{d}^3q'. \tag{17}$$

The dependence of $\nu(\boldsymbol{p})$ and $P(\boldsymbol{p}|\boldsymbol{q})$ on $t$ and $\boldsymbol{r}$ (through the single-particle distribution function $\tilde{p}_t$ for the background particles) is suppressed in our notation.

If the transition rates $w(\boldsymbol{q}, \boldsymbol{q}'|\boldsymbol{p}, \boldsymbol{p}')$ are expressed in terms of the differential cross section for particle collisions (see, for example, Exercise 171 of [3]), we obtain the collision rate

$$\nu(\boldsymbol{p}) = N \int \frac{|\boldsymbol{p}' - \boldsymbol{p}|}{m} \sigma_{\mathrm{tot}}(|\boldsymbol{p}' - \boldsymbol{p}|) \tilde{p}_t(\boldsymbol{r}, \boldsymbol{p}') \, \mathrm{d}^3 p' \tag{18}$$

in terms of the total cross section for scattering at a given relative velocity. For an interaction potential with infinite range, $\nu(\boldsymbol{p})$ would hence be infinite. To avoid an infinite total cross section $\sigma_{\mathrm{tot}}$ we need to truncate the potential at some large but finite distance. The effect of truncation can be removed from the final results, for example, for the diffusion coefficient, by letting the cutoff go to infinity. Also the difference of the gain and loss terms in the collision integral (15) for increasing cutoff remains finite because the collisions at large impact parameters have a minor influence on the momenta [for large impact parameters, the expression in square brackets in the formulation (14) of Boltzmann's collision integral hence goes to zero]. Nevertheless, to keep an underlying Markov process at all stages of the calculation, we eliminate the cutoff only from the final results. For rigorous mathematical procedures to eliminate the cutoff, see [31] and references therein.

For given $\tilde{p}_t$, Boltzmann's kinetic equation (9), together with the collision integral (14), implies a linear operator $\mathcal{L}_t^\dagger$ that preserves the normalization and positivity of $p_t$. We thus know that there exists a corresponding Markov process $(\boldsymbol{R}_t, \boldsymbol{P}_t)_{t \geq 0}$ with infinitesimal generator $\mathcal{L}_t$. The first-order differential operator in Eq. (9) describes the smooth deterministic evolution under the influence of an external potential,

$$\mathrm{d}\boldsymbol{R}_t = \frac{\boldsymbol{P}_t}{m} \mathrm{d}t, \tag{19}$$

and

$$\mathrm{d}\boldsymbol{P}_t = -\frac{\partial \phi^{(\mathrm{e})}(\boldsymbol{R}_t)}{\partial \boldsymbol{R}_t} \mathrm{d}t \tag{20}$$

between collisions, which happen with the rate $\nu(\boldsymbol{P}_t)$ defined in Eq. (18). Whenever a collision occurs, the value of $\boldsymbol{P}_t$ changes discontinuously in time according to the transition probabilities in Eq. (17). If one removes the cutoff, an infinite number of collisions occurs in finite time, but only few of them lead to significant changes of the particle momentum.

Once we have constructed a Markov process for given $\tilde{p}_t$, this process implies the one-dimensional marginal distributions $p_t$. We can hence try to establish the self-consistency condition $p_t = \tilde{p}_t$ to obtain the process associated with the standard Boltzmann equation. As in the theory of nonlinear Fokker-Planck equations, we consider a large ensemble of weakly interacting stochastic processes to realize the self-consistency in a natural way. Nonlinearity arises through the ensemble averages of

$$g_{\boldsymbol{rp}}(\boldsymbol{R}_t, \boldsymbol{P}_t) = \delta_\alpha(\boldsymbol{r} - \boldsymbol{R}_t) \delta_\beta(\boldsymbol{p} - \boldsymbol{P}_t), \tag{21}$$

which, on average, equals $p_t$. Because Dirac $\delta$ functions would be too singular in Eq. (21), we use smoothed versions of these generalized functions with width $\alpha$ in position space and width $\beta$ in momentum space. The smoothed equations are also known as the "mollified problem" (see, for example, [19]). To be specific, we may use isotropic Gaussians $\delta_\alpha$, $\delta_\beta$ with widths $\alpha$, $\beta$.

In the construction of a trajectory of a test particle, there occur collisions with the other members of the ensemble which are interpreted as background scatterers. There is an ambiguity about what should be done with the background particles in a collision. For theoretical analyses, it is convenient to leave them unaffected [32] (in the spirit of "true background particles") whereas, for practical calculations, a symmetric treatment of test and background particles is more efficient [33].

The usual formulation of Boltzmann's equation corresponds to the naive limits $\alpha \to 0$, $\beta \to 0$. Mathematically speaking, the choice of $\alpha$ and $\beta$ is linked to the number $n$ of weakly interacting processes; one first needs to perform the limit $n \to \infty$ before one can consider the limits $\alpha \to 0$, $\beta \to 0$. Only for sufficiently large ensembles, the ensemble averages can reproduce expectations. Physically speaking, however, the limits $\alpha \to 0$, $\beta \to 0$ cannot really be justified. The physical counterpart of the number of weakly interacting processes is the number of particles, which is certainly finite. We therefore need to keep also $\alpha$ and $\beta$ finite. For example, the length scale $\alpha$ should be at least of the order of the mean free path so that a test particle actually has a chance to interact with the ensemble of background scatterers. This discussion emphasizes some of the subtle aspects associated with the nonlinearity in Boltzmann's equation, and it suggests a careful discussion of the limits in the cutoff of the potential, the number of weakly interacting processes, and the smearing of the weak interactions in the position and velocity space. Smoothing should not be regarded as a mathematical evil but rather as an essential part of the physical picture. All these issues are crucial if one wants to establish and investigate a full stochastic process behind Boltzmann's nonlinear kinetic equation.

## 6 Gaussian Boltzmann Process

We now consider the Boltzmann equation in the absence of an external field. At equilibrium, the solution is a Maxwellian in momentum space, which is a Gaussian. Also the uniform distribution in position space may be considered as a degenerate special case of a Gaussian. If the one-dimensional marginal distributions are Gaussian then it is natural to ask whether the underlying stochastic process is Gaussian. More generally, one could consider anisotropic Gaussians which correspond to the entropic version of Grad's ten-moment method [34, 35, 36] (see also Sect. 7.4.3 of [3]). This implies that viscous phenomena can be described by Gaussian single-particle distribution functions, whereas heat flow must be absent. A further motivation to look at Gaussian

processes is provided by the idea to construct a coarse grained kinetic theory. According to the central limit theorem, the description of multicollision events naturally suggests the use of Gaussian processes.

The most general form of Gaussian Markov processes has been established in Exercise 2.85 of [10]. It turns out that the infinitesimal generator must be a second-order differential operator, so that we arrive at a Fokker-Planck equation with an equivalent stochastic differential equation. This stochastic differential equation must possess a drift term linear in the configurational variables and a diffusion term independent of the configurational variables. Drift and diffusion can, however, depend on time. If we insist on an inertial formulation, the most general Gaussian Boltzmann process is characterized by the following set of linear stochastic differential equations:

$$\mathrm{d}\boldsymbol{R}_t = \frac{\boldsymbol{P}_t}{m}\mathrm{d}t, \tag{22}$$

and

$$\mathrm{d}\boldsymbol{P}_t = -\zeta\left(\frac{\boldsymbol{P}_t}{m} - \boldsymbol{\kappa}_t\cdot\boldsymbol{R}_t\right)\mathrm{d}t + B_t\mathrm{d}\boldsymbol{W}_t. \tag{23}$$

The parameters in the equation of motion can be interpreted as a friction coefficient ($\zeta$), a velocity gradient tensor ($\boldsymbol{\kappa}_t$), and the amplitude of the noise ($B_t$), where the subscript $t$ indicates time dependence. Noise and friction must be related by the fluctuation-dissipation theorem [3],

$$B_t = \sqrt{2k_\mathrm{B}T_t\zeta}, \tag{24}$$

where $k_\mathrm{B}$ is Boltzmann's constant and $T_t$ is the time-dependent absolute temperature.

A system of linear stochastic differential equations is clearly inconsistent with Boltzmann's kinetic equation, even at equilibrium. For example, the Gaussian solution of Eq. (23) has continuous trajectories, in contrast to the characteristic jumps in momentum occurring in collisions according to the Boltzmann equation. But we can ask whether we can choose the parameters in Eq. (23) such that we mimic certain features of the Boltzmann equation as closely as possible.

To identify the parameters, we consider the Fokker-Planck equation for $p_t(\boldsymbol{r},\boldsymbol{p})$ associated with the stochastic differential equations (22) and (23). Kinetic equations of the Fokker-Planck type have a long tradition in kinetic gas theory [37, 38]. By comparing the equations for the first moment vector and the second moment tensor from this Fokker-Planck equation to those from the Boltzmann equation, we realize that $\boldsymbol{\kappa}_t$ characterizes a homogeneous velocity field,

$$V\int\frac{\boldsymbol{p}}{m}p_t(\boldsymbol{r},\boldsymbol{p})\,\mathrm{d}^3p = \boldsymbol{\kappa}_t\cdot\boldsymbol{r}, \tag{25}$$

and that the friction coefficient $\zeta$ is related to the fundamental relaxation time scale $\hat{\tau}$ of Grad's moment method,

$$\zeta = \frac{m}{2\hat{\tau}}. \tag{26}$$

If, in the context of coarse grained kinetic theories (from which hydrodynamic equations can be produced by solution procedures), we are interested also in noninertial formulations of Gaussian processes, then we can replace Eqs. (22) and (23) by

$$\mathrm{d}\boldsymbol{R}_t = \boldsymbol{\kappa}_t \cdot \boldsymbol{R}_t \mathrm{d}t + B'_t \mathrm{d}\boldsymbol{W}'_t, \tag{27}$$

and

$$\mathrm{d}\boldsymbol{P}_t = -\zeta \left( \frac{\boldsymbol{P}_t}{m} - \boldsymbol{\kappa}_t \cdot \boldsymbol{R}_t \right) \mathrm{d}t + \boldsymbol{\Theta}_t \cdot \boldsymbol{R}_t \mathrm{d}t + B_t \mathrm{d}\boldsymbol{W}_t, \tag{28}$$

where the Wiener process $\boldsymbol{W}'_t$ is independent of $\boldsymbol{W}_t$ and

$$B'_t = \sqrt{2k_\mathrm{B}T_t/\zeta}. \tag{29}$$

The deterministic contribution in Eq. (27) describes convection with the homogeneous average flow field, and the stochastic contribution describes the diffusion resulting from the fluctuations in the momenta. The new force term $\boldsymbol{\Theta}_t \cdot \boldsymbol{R}_t$ in Eq. (28) is introduced to account for inertial effects such as pressure and vorticity, where we have assumed that the Gaussian is centered at the origin (that is, $\langle \boldsymbol{R}_t \rangle = 0$, $\langle \boldsymbol{P}_t \rangle = 0$). The time evolution of the velocity field $\boldsymbol{v}_t = \boldsymbol{\kappa}_t \cdot \boldsymbol{r}$ is obtained from the conditional expectation of $\boldsymbol{P}_t/m$ for given $\boldsymbol{r}$ which suggests the identification (see Exercise 2.60 on p. 58 of [10])

$$\boldsymbol{\kappa}_t = \frac{1}{m} \langle \boldsymbol{P}_t \boldsymbol{R}_t \rangle \cdot \langle \boldsymbol{R}_t \boldsymbol{R}_t \rangle^{-1}. \tag{30}$$

Similarly, the time evolution of the homogeneous temperature is found from the conditional variance of $\boldsymbol{P}_t$ for given $\boldsymbol{r}$.

Equations (27) and (28) describe a particle on time scales large compared to the collision time scale, such that the fluctuations in position and momentum in Eqs. (27) and (28) become decorrelated. This formulation is similar in spirit to the modified kinetic theory of Klimontovich [39], who also introduced independent noise terms. However, Klimontovich keeps a fully inertial description which we believe to be inappropriate.

Equations (27) and (28) represent a truly coarse grained description of gas dynamics. When this coarse grained kinetic theory is used to derive hydrodynamic equations, a diffusive velocity contribution in addition to $\boldsymbol{v}_t$ arises naturally in the mass balance equation. The resulting equations can be brought into the form postulated by Brenner [7, 8, 9] (see also Section 2.2.5 of [3]), which contain the diffusion coefficient. It is hence worthwhile to calculate the diffusion coefficient from the Boltzmann process.

## 7 Application: Diffusion Coefficient

The self-diffusion coefficient $D$ can be calculated from the Boltzmann process by following a single particle through a sequence of collisions until the initial momentum and hence the correlation between displacements has decayed [40]. We here consider diffusion at equilibrium, that is, we use

$$\tilde{p}_t(\boldsymbol{r}, \boldsymbol{p}) = \frac{1}{V}\tilde{p}_{\mathrm{eq}}(\boldsymbol{p}) \tag{31}$$

with

$$\tilde{p}_{\mathrm{eq}}(\boldsymbol{p}) = (2\pi m k_{\mathrm{B}} T)^{-3/2} \exp\left\{-\frac{\boldsymbol{p}^2}{2mk_{\mathrm{B}}T}\right\}. \tag{32}$$

The required calculation of the mean-square displacement of a particle in a sequence of collisions is analogous to the one of the mean-square end-to-end distance of polymer molecules with persistence (see, for example, [40] or Section 2.3.1 of [41]).

After a large number of collisions, $N_{\mathrm{coll}}$, the mean square displacement of a particle is given by

$$\begin{aligned}\langle \Delta \boldsymbol{r}^2 \rangle &= \frac{1}{m^2} \sum_{j,k=0}^{N_{\mathrm{coll}}} \langle \boldsymbol{P}_j \cdot \boldsymbol{P}_k \tau_j \tau_k \rangle \\ &= \frac{N_{\mathrm{coll}}}{m^2} \left( \langle \boldsymbol{P}_0^2 \tau_0^2 \rangle + 2 \sum_{j=1}^{\infty} \langle \boldsymbol{P}_0 \cdot \boldsymbol{P}_j \tau_0 \tau_j \rangle \right),\end{aligned} \tag{33}$$

where $\boldsymbol{P}_j$ is the random particle momentum and $\tau_j$ the random time period between collisions $j$ and $j+1$. By expressing the conditional expectations of $\tau_j$ and $\tau_j^2$ for given $\boldsymbol{P}_j$ in terms of the collision rate $\nu(\boldsymbol{P}_j)$, we obtain

$$\begin{aligned}\langle \Delta \boldsymbol{r}^2 \rangle &= \frac{2N_{\mathrm{coll}}}{m^2} \sum_{j=0}^{\infty} \left\langle \frac{\boldsymbol{P}_0 \cdot \boldsymbol{P}_j}{\nu(\boldsymbol{P}_0)\nu(\boldsymbol{P}_j)} \right\rangle \\ &= \frac{2N_{\mathrm{coll}}}{m^2} \sum_{j=0}^{\infty} \int \frac{\boldsymbol{p} \cdot \boldsymbol{q}}{\nu(\boldsymbol{p})\nu(\boldsymbol{q})} P^j(\boldsymbol{p}|\boldsymbol{q}) \tilde{p}_{\mathrm{eq}}(\boldsymbol{q}) \, \mathrm{d}^3 p \, \mathrm{d}^3 q,\end{aligned} \tag{34}$$

where $P^j(\boldsymbol{p}|\boldsymbol{q})$ is the $j$-step transition matrix associated with the single-step matrix $P(\boldsymbol{p}|\boldsymbol{q})$ defined in Eq. (17). The summation of the geometric series of $P^j(\boldsymbol{p}|\boldsymbol{q})$ requires the calculation of an inverse matrix which, in view of continuous momentum labels, implies the solution of an integral equation. It is useful to introduce a time scale $\tau(\boldsymbol{p})$ for a particle with momentum $\boldsymbol{p}$ as an average inverse rate,

$$\tau(\boldsymbol{p})\boldsymbol{p} = \sum_{j=0}^{\infty} \int \frac{\boldsymbol{q}}{\nu(\boldsymbol{q})} P^j(\boldsymbol{q}|\boldsymbol{p}) \, \mathrm{d}^3 q, \tag{35}$$

which, by construction, satisfies the integral equation

$$\int \left[\tau(\boldsymbol{p}) - \tau(\boldsymbol{q})\frac{\boldsymbol{p}\cdot\boldsymbol{q}}{\boldsymbol{p}^2}\right] \nu(\boldsymbol{p}) P(\boldsymbol{q}|\boldsymbol{p})\, \mathrm{d}^3 q = 1. \tag{36}$$

After solving this integral equation for a characteristic time scale $\tau(\boldsymbol{p})$, we obtain from Eqs. (34) and (35)

$$\langle \Delta \boldsymbol{r}^2 \rangle = \frac{2N_{\mathrm{coll}}}{m^2} \int \boldsymbol{p}^2 \tau(\boldsymbol{p}) \frac{\tilde{p}_{\mathrm{eq}}(\boldsymbol{p})}{\nu(\boldsymbol{p})}\, \mathrm{d}^3 p. \tag{37}$$

Because the average time required for $N_{\mathrm{coll}}$ collisions is

$$t = N_{\mathrm{coll}} \int \frac{\tilde{p}_{\mathrm{eq}}(\boldsymbol{p})}{\nu(\boldsymbol{p})}\, \mathrm{d}^3 p, \tag{38}$$

our final result for the diffusion coefficient is

$$D = \frac{1}{3m^2} \int \boldsymbol{p}^2 \tau(\boldsymbol{p}) \frac{\tilde{p}_{\mathrm{eq}}(\boldsymbol{p})}{\nu(\boldsymbol{p})}\, \mathrm{d}^3 p \left[\int \frac{\tilde{p}_{\mathrm{eq}}(\boldsymbol{p})}{\nu(\boldsymbol{p})}\, \mathrm{d}^3 p\right]^{-1}. \tag{39}$$

For Maxwell molecules, for which $\nu(\boldsymbol{p})$ is independent of $\boldsymbol{p}$, the integral equation (36) becomes (see Exercise 174 of [3])

$$\frac{N}{V}\sqrt{\frac{2\hat{\phi}}{m}} \int \left[\tau(\boldsymbol{p}) - \tau(\boldsymbol{q})\frac{\boldsymbol{p}\cdot\boldsymbol{q}}{\boldsymbol{p}^2}\right] \tilde{p}_{\mathrm{eq}}(\boldsymbol{p}') \frac{\hat{b}(\theta')}{\sin\theta'} \left|\frac{\mathrm{d}\hat{b}(\theta')}{\mathrm{d}d\theta'}\right| \mathrm{d}\Omega'\, \mathrm{d}^3 p' = 1, \tag{40}$$

where the inverse of the function $\hat{b}(\theta')$ can be expressed in terms of a complete elliptic integral of the first kind and the final momentum $\boldsymbol{q}$ can be calculated from the initial momenta $\boldsymbol{p}$ and $\boldsymbol{p}'$ for the scattering into a given solid angle $\Omega'$ in the centre-of-mass system.

With the methods described in detail in Exercise 175 of [3], one realizes that the time scale $\tau$ is independent of $\boldsymbol{p}$ and given by

$$\tau 2\pi \frac{N}{V}\sqrt{\frac{2\hat{\phi}}{m}}\tilde{c}^{\mathrm{Maxw}} = 1 \tag{41}$$

with

$$\tilde{c}^{\mathrm{Maxw}} = \int_0^\infty \left[\sin\frac{\theta(\hat{b})}{2}\right]^2 \hat{b}\, \mathrm{d}\hat{b} \approx 0.42194. \tag{42}$$

By comparing the diffusion coefficient $D = k_{\mathrm{B}}T\tau/m$ following from Eq. (39) for Maxwell molecules with constant $\tau$ and $\nu$ to the viscosity (7.120) of [3], we obtain the Schmidt number

$$\frac{\eta}{\rho D} = \frac{4}{3}\frac{\tilde{c}^{\mathrm{Maxw}}}{c^{\mathrm{Maxw}}} \approx 0.645, \tag{43}$$

which we have evaluated with

$$c^{\mathrm{Maxw}} = \int_0^\infty [\sin\theta(\hat{b})]^2 \hat{b}\, \mathrm{d}\hat{b} \approx 0.87239. \tag{44}$$

The result (43) coincides with Eq. (23) of [42]. Note that the ratio in Eq. (43) is equal to the ratio of previously introduced time scales, $\hat{\tau}/\tau$.

## 8 Perspectives

We have constructed and discussed the Markov process naturally associated with Boltzmann's kinetic equation. Even for the motion of a test particle through a given ensemble of scatterers, the existence of a Markov process with finite collision rate requires a cutoff of the interaction potential at a finite distance. When nonlinearity is introduced into the Boltzmann equation, further regularization parameters are required. It is known that the nonlinearity of the Boltzmann equation can be reproduced by weakly interacting Markov processes.

In terms of Boltzmann processes, we found a natural possibility to coarse grain to the level of multicollision events, which correspond to length scales larger than the mean free path. In this procedure, diffusion arises and the knowledge about correlations on top of the one-dimensional marginal distributions of the process is crucial. Only the Boltzmann process allows us to perform proper coarse graining, and it leads to the modified hydrodynamic equations as postulated by Brenner [7, 8, 9]. By applying solution techniques to the Boltzmann equation itself we do not obtain the most general form of hydrodynamic equations. There is a general lesson about *solving versus coarse graining* to be learnt from this example.

The Boltzmann equation for the single-particle distribution function is known to be thermodynamically consistent [3, 43]. One could hence raise the question whether also the Boltzmann process is thermodynamically consistent. First of all, however, stochastic processes with mean-field interactions tend to violate the fluctuation dissipation theorem (see Section 4.2.4 of [10] and [44, 45]). More importantly, the theory of thermodynamically admissible stochastic processes is not even fully developed yet. Whereas the meaning of thermodynamic consistency for stochastic differential equations and the role of the fluctuation-dissipation theorem are well-understood, this is not the case for more general Markov processes. The idea of time scale separation, when formalized through projection-operator techniques, leads to the GENERIC structure with a Green-Kubo formula for the friction matrix and the fluctuation-dissipation theorem. We conjecture that additional statistical tools are required to handle escape problems and other rare events leading to frictional properties that can be expressed in terms of Kramers-type formulas for escape rates [46].

*Acknowledgement.* I'm grateful to Alain-Sol Sznitman for introducing me to the relevant mathematical literature and to Alexander Gorban for providing continuous motivation and stimulation.

## References

1. C. Cercignani: *The Boltzmann Equation and Its Applications*. Applied Mathematical Sciences, vol. 67 (Springer, Berlin Heidelberg New York 1988)
2. H. Struchtrup: *Macroscopic Transport Equations for Rarefied Gas Flows* (Springer, Berlin Heidelberg New York 2005)
3. H.C. Öttinger: *Beyond Equilibrium Thermodynamics* (Wiley, Hoboken 2005)
4. H.C. Öttinger: Derivation of two-generator framework of nonequilibrium thermodynamics for quantum systems. *Phys. Rev. E* **62**, 4720–4724 (2000)
5. A.N. Gorban, I.V. Karlin, H.C. Öttinger, L.L. Tatarinova: Ehrenfest's argument extended to a formalism of nonequilibrium thermodynamics. *Phys. Rev. E* **63**, 066124, 1–6 (2001)
6. A.N. Gorban: Basic types of coarse-graining. Preprint `cond-mat/0602024` (2006)
7. H. Brenner: Is the tracer velocity of a fluid continuum equal to its mass velocity? *Phys. Rev. E* **70**, 061201, 1–10 (2004)
8. H. Brenner: Kinematics of volume transport. *Physica A* **349**, 11–59 (2005)
9. H. Brenner: Navier-Stokes revisited. *Physica A* **349**, 60–132 (2005)
10. H.C. Öttinger: *Stochastic Processes in Polymeric Fluids: Tools and Examples for Developing Simulation Algorithms* (Springer, Berlin Heidelberg New York 1996)
11. H. Bauer: *Probability Theory and Elements of Measure Theory*, second edition (Academic Press, London 1981)
12. R.J. DiPerna, P.L. Lions: On the Cauchy problem for Boltzmann equations: Global existence and weak stability. *Ann. Math.* **130**, 321–366 (1989)
13. L. Desvillettes, C. Villani: On the trend to global equilibrium for spatially inhomogeneous kinetic systems: The Boltzmann equation. *Invent. Math.* **159**, 245–316 (2005)
14. M. Kac: Foundations of kinetic theory. In J. Neyman, editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. III, 171–197 (Univ. of California Press. Berkeley 1956)
15. H.P. McKean: Speed of approach to equilibrium for Kac's caricature of a Maxwellian gas. *Arch. Rational Mech. Anal.* **21**, 343–367 (1966)
16. F.A. Grünbaum: Propagation of chaos for the Boltzmann equation. *Arch. Rational Mech. Anal.* **42**, 323–345 (1971)
17. H. Tanaka: Probabilistic treatment of the Boltzmann equation of Maxwellian molecules. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **46**, 67–105 (1978)
18. A.S. Sznitman: Équations de type de Boltzmann, spatialment homogènes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **66**, 559–592 (1984)
19. C. Graham, S. Méléard: Probabilistic tools and Monte-Carlo approximations for some Boltzmann equations. In: *Présentation du CEMRACS 1999*, ed. by F. Coquel and S. Cordier, ESAIM Proceedings, vol. 10, 77–126 (EDP Sciences, Les Ulis, 2001) `www.edpsciences.org/articlesproc/Vol.10/contents.htm`.

20. O.E. Lanford III: Time evolution of large classical systems. In: *Dynamical Systems, Theory and Applications*, Batelle Seattle 1974 Rencontres, ed. by J. Moser, Lecture Notes in Physics, vol. 38, 1–111 (Springer, Berlin 1975)
21. L. Arnold: *Stochastic Differential Equations: Theory and Applications* (Wiley, New York 1974)
22. W. Feller: *An Introduction to Probability Theory and Its Applications*, vol. 2, second edition (Wiley, New York 1971)
23. C.W. Gardiner: *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, Springer Series in Synergetics, vol. 13, second edition (Springer, Berlin Heidelberg New York 1990)
24. G. Kallianpur: *Stochastic Filtering Theory*. Applications of Mathematics, vol. 13 (Springer, Berlin Heidelberg New York 1980)
25. T.D. Frank: *Nonlinear Fokker-Planck Equations*. Springer Series in Synergetics. (Springer, Berlin Heidelberg New York 2005)
26. K. Oelschläger: A martingale approach to the law of large numbers for weakly interacting stochastic processes. *Ann. Probab.* **12**, 458–479 (1984)
27. J. Gärtner: On the McKean-Vlassov limit for interacting diffusions. *Math. Nachr.* **137**, 197–248 (1988)
28. A.S. Sznitman: A fluctuation result for nonlinear diffusions. In: *Infinite Dimensional Analysis and Stochastic Processes*, ed. by S. Albeverio, Research Notes in Mathematics, vol. 124, 145–160 (Pitman, Boston 1985)
29. H.P. McKean: Propagation of chaos for a class of non-linear parabolic equations. In: *Stochastic Differential Equations*, ed. by A. K. Aziz, Lecture Series in Differential Equations, vol. 2, 177–194 (Van Nostrand, Amsterdam 1969)
30. A.S. Sznitman: Topics in propagation of chaos. In: *École d'Été de Probabilités de Saint-Flour*, ed by P. L. Hennequin, Lecture Notes in Mathematics, vol. 1464, 165–251 (Springer, Berlin Heidelberg New York 1991)
31. N. Fournier, S. Méléard: A Markov process associated with a Boltzmann equation without cutoff and for non-Maxwell molecules. *J. Stat. Phys.* **104**, 359–385 (2001)
32. K. Nanbu: Interrelations between various direct simulation methods for solving the Boltzmann equation. *J. Phys. Soc. Jpn.* **52**, 3382–3388 (1983)
33. G.A. Bird: *Molecular Gas Dynamics*, Oxford Engineering Science Series (Clarendon, Oxford 1976)
34. A.M. Kogan: Derivation of Grad's type equations and study of their relaxation properties by the method of maximization of entropy. *J. Appl. Math. Mech. 29*, 130–142 (1965)
35. R.M. Lewis: A unifying principle in statistical mechanics. *J. Math. Phys.* **8**, 1448–1459 (1967)
36. A.N. Gorban, I.V. Karlin: General approach to constructing models of the Boltzmann equation. *Physica A* **206**, 401–420 (1994)
37. J.G. Kirkwood: The statistical mechanical theory of transport processes. I. General theory. *J. Chem. Phys.* **14**, 180–201 (1946)
38. J.L. Lebowitz, H.L. Frisch, E. Helfand: Nonequilibrium distribution functions in a fluid. *Phys. Fluids* **3**, 325–338 (1960)
39. Yu.L. Klimontovich: On the need for and the possibility of a unified description of kinetic and hydrodynamic processes. *Theor. Math. Phys.* **92**, 909–921 (1992)
40. S. Hess: Boltzmann equation approach to polymer statistics. *Physica A* **112**, 287–302 (1982)

41. M. Rubinstein, R.H. Colby: *Polymer Physics*. (Oxford University Press 2003)
42. H. Matsumoto: Variable sphere molecular model for inverse power law and Lennard-Jones potentials in Monte Carlo simulations. *Phys. Fluids* **14**, 4256–4265 (2002)
43. H.C. Öttinger: GENERIC formulation of Boltzmann's kinetic equation. *J. Non-Equilib. Thermodyn.* **22**, 386–391 (1997)
44. M. Hütter, H.C. Öttinger: Modification of linear response theory for mean field approximations. *Phys. Rev. E* **54**, 2526–2530 (1996)
45. M. Hütter, I.V. Karlin, H.C. Öttinger: Dynamic mean-field models from a nonequilibrium thermodynamics perspective. *Phys. Rev. E* **68**, 016115, 1–9 (2003)
46. H.A. Kramers: Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* **7**, 284–304 (1940)

# Finite Difference Patch Dynamics for Advection Homogenization Problems

G. Samaey[1], D. Roose[1], and I. G. Kevrekidis[2]

[1] Department of Computer Science, K.U. Leuven, Celestijnenlaan 200A, 3000 Leuven, Belgium, {giovanni.samaey,dirk.roose}@cs.kuleuven.ac.be
[2] Department of Chemical Engineering and PACM, Princeton University, Princeton, USA, yannis.kevrekidis@princeton.edu

**Summary.** We consider problems in which there is a separation between the (microscopic) scale at which the available model is defined, and the (macroscopic) scale of interest. For time-dependent multi-scale problems of this type, an "equation-free" framework has been proposed, of which patch dynamics is an essential component. Patch dynamics is designed to perform numerical simulations of an unavailable macroscopic equation on macroscopic time and length scales; it only uses appropriately initialized simulations of the available microscopic model in a number of small boxes (patches), which cover a fraction of the space-time domain. We review some recent convergence results and demonstrate that the method allows to simulate advection-dominated problems accurately.

## 1 Introduction

In many problems of current interest, one is interested in the behaviour of a (physical, chemical) system on macroscopic length and time scales, while the only valid model is available at a more microscopic scale. For example, in polymer flow, it is often impossible to find a closed formula for the stress tensor in terms of the velocity field. Therefore, the macroscopic model (a partial differential equation, PDE) needs to be supplemented with a Monte Carlo simulation to estimate the stress tensor [19, 20, 27]. Similar problems arise in flow through porous media, where it is often hard to obtain an effective permeability coefficient analytically [2], or bacterial chemotaxis, where a PDE for the density can only be derived from an individual-based model under simplifying assumptions which cannot always be fully justified [10].

In this work, we consider situations where only a microscopic model is known,

$$\partial_t u(\boldsymbol{x}, t) = f(u(\boldsymbol{x}, t)), \tag{1}$$

in which $u(\boldsymbol{x}, t)$ represents the microscopic state variables, $\boldsymbol{x} \in D_m$ are the remaining microscopic independent variables, and $\partial_t$ denotes the time deriva-

tive. We assume that an equivalent macroscopic model exists, but cannot be obtained in closed form. We denote this model by

$$\partial_t U(\boldsymbol{X}, t) = F(U(\boldsymbol{X}, t)), \tag{2}$$

in which $U(\boldsymbol{X}, t)$ represents the macroscopic state variables, and $\boldsymbol{X} \in D_M$ and $t$ are the macroscopic independent variables. If one is only interested in the macroscopic solution $U(\boldsymbol{X}, t)$, one can construct a so-called *coarse-grained time-stepper* as proposed by Kevrekidis *et al* [21, 35]. We introduce a time-stepper for the microscopic evolution law (1),

$$u(\boldsymbol{x}, t + \mathrm{d}t) = s(u(\boldsymbol{x}, t); \mathrm{d}t), \tag{3}$$

where $\mathrm{d}t$ is the size of the (microscopic) time-step, and the aim is to obtain a coarse-grained time-stepper for the variables $U(\boldsymbol{X}, t)$ as

$$U(\boldsymbol{X}, t + \delta t) = \bar{S}(U(\boldsymbol{X}, t); \delta t), \tag{4}$$

where $\delta t$ denotes the size of the (coarse-grained) time-step, and the bar has been introduced to emphasize the fact that the time-stepper for the macroscopic variables is only an *approximation* of a time-stepper for (2), since this equation is not explicitly known.

To define a coarse-grained time-stepper (4), we need to introduce two operators that make the transition between microscopic and macroscopic variables. We define a *lifting operator*,

$$\mu : U(\boldsymbol{X}, t) \mapsto u(\boldsymbol{x}, t) = \mu(U(\boldsymbol{X}, t)), \tag{5}$$

which maps macroscopic to microscopic variables, and its complement, the *restriction operator*

$$\mathcal{M} : u(\boldsymbol{x}, t) \mapsto U(\boldsymbol{X}, t) = \mathcal{M}(u(\boldsymbol{x}, t)). \tag{6}$$

The restriction operator can often be determined as soon as the macroscopic variables are known. For instance, when the microscopic model consists of an evolving ensemble of many particles, the restriction typically consists of the computation of the low-order moments of the distribution (density, momentum, energy), which are considered as the macroscopic variables $U(\boldsymbol{X}, t)$. The assumption that a macroscopic equation exists for these low-order moments, implies that the higher-order moments become functionals of the low-order moments on time-scales which are fast compared to the overall system evolution (*slaving*).

The construction of the lifting operator is usually more involved. Again taking the example of a particle model, we need to define a mapping from a few low-order moments to an initial condition for each of the particles. We know that the higher-order moments of the resulting particle distribution should be functionals of the low-order moments, but unfortunately, these functionals are unknown (since the macroscopic evolution law is also unknown).

Several approaches have been suggested to address this problem. One could for instance initialize the higher-order moments randomly. This introduces a *lifting error*, and one then relies on the separation of time-scales to ensure that these higher-order moments relax quickly to a functional of the low-order moments (*healing*) [13, 24, 34]. We note that, in some cases, this approach can be shown to produce inaccurate results [22]. In fact, to initialize the higher-order moments correctly, one should perform a simulation of the microscopic system, with the additional constraint that the low-order moments should be kept fixed. How this can be done using only a time-stepper for the original microscopic system, is explained and analyzed in [11, 12, 23].

Given an initial condition for the macroscopic variables $U(\boldsymbol{X}, t^*)$ at some time $t^*$, we can then construct the time-stepper (4) in the following way:

1. **Lifting.** Using the lifting operator (5), create appropriate initial conditions $u(\boldsymbol{x}, t^*)$ for the microscopic time-stepper (3), consistent with the macroscopic variables.
2. **Simulation.** Use the time-stepper (3) to compute the microscopic state $u(\boldsymbol{x}, t)$ for $t \in [t^*, t^* + \delta t]$.
3. **Restriction.** Obtain the macroscopic state $U(\boldsymbol{X}, t^* + \delta t)$ from the microscopic state $u(\boldsymbol{x}, t^* + \delta t)$ using the restriction operator (6).

Assuming $\delta t = k\mathrm{d}t$, this can be written as

$$U(\boldsymbol{X}, t + \delta t) = \bar{S}(U(\boldsymbol{X}, t), \delta t) = \mathcal{M}(s^k(\mu(U(\boldsymbol{X}, t)), \mathrm{d}t)), \tag{7}$$

where we have represented the $k$ microscopic time-steps by a superscript on $s$. If the microscopic model is stochastic, one may need to perform multiple replica simulations to get an accurate result.

Here, we consider situations where the macroscopic model (2) is assumed to be a partial differential equation in one space dimension, so $\boldsymbol{X} = x$. For this type of problems, the *patch dynamics scheme* was proposed [21, 31, 32], which only performs appropriately initialized microscopic simulations in a small fraction of the space-time domain to reduce the computational cost. The general idea is the following. First, we construct a coarse time-stepper which only performs simulations of the microscopic model in a number of small boxes, which can be thought of as macroscopic mesh points. We initialize a microscopic simulation at time $t^*$ in each of the boxes (*lifting*); run the time-stepper (3) until $t = t^* + \delta t$ and compute the macroscopic variables in each of the boxes at time $t^* + \delta t$ (*restriction*). The resulting coarse-grained time-stepper is called the *gap-tooth scheme* [21, 32]. Because the microscopic time-stepper (3) takes very small time-steps of size $\mathrm{d}t$, the coarse-grained time-step $\delta t$ may still be very small compared to the slow time-scales of the macroscopic model (2). Therefore, we use the gap-tooth time-stepper to estimate the macroscopic time derivative and use this estimate to take a time-step of size $\Delta t \gg \delta t$.

The performance and accuracy of the patch dynamics scheme are currently under active investigation. Recently, we have studied the convergence properties patch dynamics scheme for a model diffusion homogenization problem.

We showed that the patch dynamics scheme approximates a finite difference scheme for the effective (homogenized) equation, using only the microscopic (homogenization) equation in a set of small boxes [30, 31, 32]. A major issue is the imposition of appropriate box boundary conditions. For example, when the macroscopic behaviour is governed by diffusion, we can impose the average gradient as a boundary condition [32], or we can take *arbitrary* boundary conditions, provided we surround the computational boxes by buffer boxes to reduce the artefacts [31]. This latter technique is especially suited when a (e.g. particle) code is given, with built-in boundary conditions which are impossible, or very difficult, to change. Roberts *et al.* are investigating boundary conditions that lead to higher order accurate schemes [28].

In this paper, we confine ourselves to homogenization problems for the purpose of convergence analysis. In this case, the microscopic model is a partial differential equation with coefficients that vary on a small spatial scale, while the macroscopic model is a partial differential equation for the effective behaviour on large spatial scales. However, we emphasize that the method can also be applied with, and is in fact designed for, the effective behaviour of truly microscopic models, such as kinetic Monte Carlo methods, or molecular dynamics.

We note that many numerical schemes have been devised for the homogenization problem. The earliest work dates back to Babuska [3] for elliptic problems and Engquist [8] for dynamic problems. Without the aim of being complete, we mention some recent multi-scale approaches to the homogenization problem. The multi-scale finite element method of Hou and Wu uses special basis functions to capture the correct microscopic behaviour [16, 17]. Schwab, Matache and Babuska have devised a generalized FEM method based on a two-scale finite element space [25, 33]. Other approaches include the use of wavelet projections [6, 9] and multi-grid cycles [26]. Runborg et al. [29] proposed a time-stepper based method that obtains the effective behaviour through short bursts of detailed simulations appropriately averaged over many shifted initial conditions. The simulations were performed over the whole domain, but the notion of effective behaviour is identical. In their recent work, E and Engquist and collaborators address the same problem of simulating only the macroscopic behaviour of a multiscale model, see e.g. [1, 7]. In their method, which is very similar in spirit, an unavailable macroscopic *flux* is estimated from appropriately initialized and constrained microscopic simulations, and used inside a macroscopic finite volume scheme.

The paper is organized as follows. In section 2 we discuss some model homogenization problems. Section 3 explains the patch dynamics scheme. We briefly review some theoretical convergence results in section 4. In section 5, we show that we can also approximate the macroscopic behaviour of hyperbolic homogenization problems. This is possible because we can approximate *any* desired finite difference scheme by an appropriate choice of the lifting step (the initialization of the small boxes). We note that the theoretical convergence analysis has not explicitly been done for this case. We conclude in section 6.

## 2 Model Problems

### 2.1 Parabolic Homogenization Problem

As a *microscopic problem*, we consider a parabolic partial differential equation,

$$\begin{aligned} \partial_t u_\epsilon(x,t) &= \partial_x \left( a\left(x/\epsilon\right) \partial_x u_\epsilon(x,t) \right), \\ u_\epsilon(x,0) &= u^0(x) \in L^2([0,1]), \\ u_\epsilon(0,t) &= u_\epsilon(1,t) = 0, \end{aligned} \tag{8}$$

where $a(y) = a\left(x/\epsilon\right)$ is uniformly elliptic and periodic in $y$ and $\epsilon$ is a small parameter. We choose homogeneous Dirichlet boundary conditions for simplicity.

On the macroscopic scale, we are interested in an *effective, homogenized* partial differential equation, in which the small-scale parameter $\epsilon$ has been eliminated. According to classical homogenization theory [4], the solution of (8) can be written as an asymptotic expansion in $\epsilon$,

$$u_\epsilon(x,t) = U(x,t) + \sum_{i=1}^{\infty} \epsilon^i \left( u_i(x, x/\epsilon, t) \right), \tag{9}$$

where the functions $u_i(x,y,t) \equiv u_i(x,x/\epsilon,t)$, $i = 1,2,\ldots$ are periodic in $y$. Here, $U(x,t)$ is the solution of the *homogenized equation*

$$\begin{aligned} \partial_t U(x,t) &= \partial_x \left( a^* \partial_x U(x,t) \right) \\ U(x,0) &= u^0(x) \in L^2([0,1]), \\ U(0,t) &= U(1,t) = 0. \end{aligned} \tag{10}$$

Here, $a^*$ is the constant effective coefficient, given by

$$a^* = \int_0^1 a(y) \left( 1 - \frac{\mathrm{d}}{\mathrm{d}y} \chi(y) \right) \mathrm{d}y, \tag{11}$$

and $\chi(y)$ is the periodic solution of

$$\frac{\mathrm{d}}{\mathrm{d}y} \left( a(y) \frac{\mathrm{d}}{\mathrm{d}y} \chi(y) \right) = \frac{\mathrm{d}}{\mathrm{d}y} a(y), \tag{12}$$

the so-called *cell problem*. The solution of (12) is only defined up to an additive constant, so we impose the extra condition

$$\int_0^1 \chi(y) \, \mathrm{d}y = 0.$$

We note that in one space dimension, an explicit formula is known for $a^*$ [4],

$$a^* = \left[ \int_0^1 \frac{1}{a(y)} \, \mathrm{d}y \right]^{-1}. \tag{13}$$

These asymptotic expansions have been rigorously justified in the classical book [4], see also [5]. Under the smoothness assumptions made on $a(x/\epsilon)$, one obtains *strong* convergence of $u_\epsilon(x,t)$ to $U(x,t)$ as $\epsilon \to 0$ in $L^2([0,1]) \times C([0,T))$. Indeed, we can write

$$\|u_\epsilon(x,t) - U(x,t)\|_{L_2([0,1])} \leq C_0 \epsilon, \tag{14}$$

uniformly in $t$.

### 2.2 Hyperbolic Homogenization Problem

We consider the following hyperbolic partial differential equation in one space dimension,

$$\begin{aligned} &\partial_t u_\epsilon(x,t) + \partial_x \left[ c\left(x/\epsilon\right) u_\epsilon(x,t) \right] = 0, \\ &\quad u_\epsilon(x,0) = u^0(x) \in L^2([0,1]), \qquad \partial_x u_\epsilon(0,t) = 0, \end{aligned} \tag{15}$$

where $c(y) = c\left(x/\epsilon\right) > 0$ is periodic in $y$ and $\epsilon$ is a small parameter. We choose a homogeneous Neumann boundary condition for simplicity.

As in the previous section, we are interested in an effective, homogenized partial differential equation on a macroscopic scale, where the dependence on the small scale parameter has been eliminated. According to classical homogenization theory [4, 5], the solution of (15) converges *weakly* in the limit of $\epsilon \to 0$ to the solution of

$$\begin{aligned} &\partial_t U(x,t) + \partial_x \left[ c^* U(x,t) \right] = 0, \\ &\quad U(x,0) = u^0(x) \in L^2([0,1]), \qquad \partial_x U(0,t) = 0, \end{aligned} \tag{16}$$

which describes the evolution of the averaged, effective behaviour. As in the parabolic case, the effective coefficient $c^*$ is given by the harmonic average,

$$c^* = \left[ \int_0^1 \frac{1}{c(y)} \, \mathrm{d}y \right]^{-1}. \tag{17}$$

## 3 Patch Dynamics

We devise a scheme for the evolution of the effective behaviour $U(x,t)$ of a general homogenization problem,

$$\partial_t u_\epsilon = f(u_\epsilon, \partial_x u_\epsilon, \ldots, \partial_x^d u_\epsilon, t; \epsilon), \tag{18}$$

where $\partial_t$ denotes again the time derivative, and $\partial_x^k$ denotes the $k$-th spatial derivative ($k = 1, \ldots, d$, where $k = 1$ is usually omitted). We assume that

a time integration code for this equation has already been written and is available with a number of *standard* boundary conditions, such as no-flux or Dirichlet. Further, we assume that the macroscopic equation is of the form

$$\partial_t U = F(U, \partial_x U, \ldots, \partial_x^d U, t). \tag{19}$$

Suppose we want to obtain the solution of (19) on the interval $[0, 1]$, using an equidistant, macroscopic mesh $\Pi(\Delta x) := \{0 = x_0 < x_1 = x_0 + \Delta x < \ldots < x_N = 1\}$. Given equation (19), we can define a method-of-lines space discretization,

$$\partial_t U_i(t) = F(U_i(t), D^1(U_i(t)), \ldots, D^d(U_i(t)), t), \qquad i = 0, \ldots, N. \tag{20}$$

where $U_i(t) \approx U(x_i, t)$ and $D^k(U_i(t))$ denotes a suitable finite difference approximation for the $k$-th spatial derivative. We subsequently discretize equation (20) in time using a time integration method of choice, e.g. forward Euler. We denote the resulting time-stepper as

$$U^{n+1} = S(U^n, t_n; \Delta t) = U^n + \Delta t\, F(U^n, t_n), \tag{21}$$

where $U^n = (U_0(t_n), \ldots, U_N(t_n))^T$ and $\Delta t$ denotes the macroscopic time-step. We have suppressed the dependence of $F(U^n, t_n)$ on the spatial derivatives for notational convenience. Note that, although we have used the forward Euler scheme here for concreteness, in principle any time discretization method can be used to solve equation (20).

Since equation (19) is assumed not to be known explicitly, we will use (21) for analysis purposes only. We construct a (patch dynamics) scheme to approximate (21). To this end, we consider a small interval (box, *tooth*) of size $h \ll \Delta x$ around each mesh point, and define the discrete solution $\bar{U}(t) = (\bar{U}_0(t), \ldots, \bar{U}_N(t))^T \in \mathbb{R}^{N+1}$ as being the average of the microscopic solution in the small boxes,

$$\bar{U}_i(t) = \mathcal{S}_h(u_\epsilon)(x_i, t) = (1/h) \int_{x_i - h/2}^{x_i + h/2} u_\epsilon(\xi, t) \mathrm{d}\xi, \qquad i = 0, \ldots, N. \tag{22}$$

We denote an approximation of $\bar{U}(t)$ at $t = t_n$ as $\bar{U}^n$.

The patch dynamics scheme is now constructed as follows. We introduce a larger *buffer* box of size $H > h$ around each mesh point (see figure 1.) In each box of size $H$, we perform a time integration over a time interval of size $\delta t$ using the microscopic model (18), and restrict to macroscopic variables. The results are used to estimate the macroscopic time derivative. We provide each microscopic simulation with the following initial and boundary conditions.

**Initial condition.** We define the initial condition by constructing a local Taylor expansion, based on the (given) box averages $\bar{U}_i^n$, $i = 0, \ldots, N$, at mesh point $x_i$ and time $t_n$,

$$\bar{u}_\epsilon^i(x, t_n) = \sum_{k=0}^{d} D_i^k(\bar{U}^n) \frac{(x - x_i)^k}{k!}, \qquad x \in [x_i - \frac{H}{2}, x_i + \frac{H}{2}], \tag{23}$$

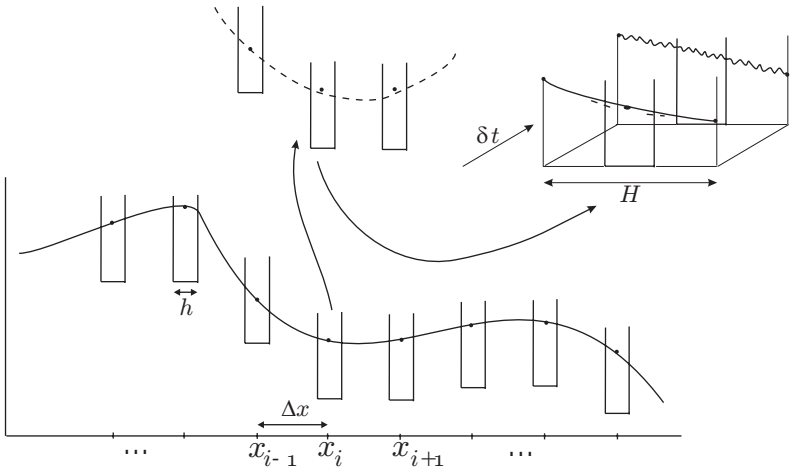



Fig. 1: A schematic representation of the gap-tooth scheme with buffer boxes. We choose a number of boxes of size $h$ around each macroscopic mesh point $x_i$ and define a local Taylor approximation as initial condition in each box. Simulation is performed inside the larger (buffer) boxes of size $H$, where some boundary conditions are imposed.

where $d$ is the order of the macroscopic equation (19). The coefficients $D_i^k(\bar{U}^n)$, $k > 0$ are the same finite difference approximations for the $k$-th spatial derivative that would be used in the comparison scheme (20), whereas $D_i^0(\bar{U}^n)$ is chosen such that

$$\frac{1}{h}\int_{x_i-h/2}^{x_i+h/2} \bar{u}_\epsilon^i(\xi, t_n)\mathrm{d}\xi = \bar{U}_i^n. \tag{24}$$

**Boundary conditions.** The time integration of the microscopic model in each box should provide information on the evolution of the *global* problem at that location in space. It is therefore crucial that the boundary conditions are chosen such that the solution inside each box evolves *as if it were embedded in the larger domain*. To this end, we introduce a larger box of size $H > h$ around each macroscopic mesh point. The simulation can subsequently be performed using any of the *built-in* boundary conditions of the microscopic code. Lifting and (short-term) evolution (using *arbitrary* available boundary conditions) are performed in the larger box; yet the restriction is done by processing the solution (here taking its average) over the inner, small box only. The goal of the additional computational domains, the *buffers*, is to buffer the solution inside the small box from the artificial disturbance caused by the (repeatedly updated) boundary conditions. This can be accomplished over *short enough* time intervals, provided the buffers are *large enough*; analyzing the method is tantamount to making these statements quantitative.

**The algorithm.** The complete algorithm to obtain an estimate of the macroscopic time derivative at time $t_n$ is given below:

1. **Lifting.** At time $t_n$, construct the initial condition $\bar{u}^i_\epsilon(x, t_n)$, $i = 0, \ldots, N$ using the box averages $\bar{U}^n_i$, as defined in (23).
2. **Simulation.** Compute the box solution $\bar{u}^i_\epsilon(x, t)$, $t > t_n$, by solving equation (18) in the interval $[x_i - H/2, x_i + H/2]$ with *some* boundary conditions up to time $t_{n+\delta} = t_n + \delta t$. The boundary conditions can be anything that the microscopic code allows.
3. **Restriction.** Compute the average $\bar{U}^{n+\delta}_i = 1/h \int_{x_i - h/2}^{x_i + h/2} \bar{u}^i_\epsilon(\xi, t_{n+\delta}) \mathrm{d}\xi$ over the *inner, small box only.*
4. **Estimation.** We estimate the time derivative at time $t_n$ as

$$\bar{F}^d(\bar{U}^n, t_n; h, \delta t, H) = \frac{\bar{U}^{n+\delta} - \bar{U}^n}{\delta t}, \tag{25}$$

   where we have added a superscript $d$ to denote the highest spatial derivative that has been initialized in the lifting step. We also made explicit the dependence of the estimate on $H$ and $\delta t$.

Since the first three steps constitute a gap-tooth time-step, we call the estimator (25) a *gap-tooth time derivative estimator*. It can be used in any ODE time integration code. For example, a forward Euler patch dynamics scheme would be

$$\bar{U}^{n+1} = \bar{U}^n + \Delta t \; \bar{F}^d(\bar{U}^n, t_n; h, \delta t, H). \tag{26}$$

For more details, including a discussion of the additional issues that need to be addressed for truly microscopic models, we refer to [31]. We emphasize that an initialization according to equation (23) has the important advantage that one can choose a suitable finite difference approximation for each derivative independently, as opposed to the method described in [21, 32], which automatically leads to central finite differences. This property is crucial, and will allow us to approximate advection-dominated equations more effectively.

## 4 Convergence Results

In this section, we briefly review some theoretical convergence results that were obtained for the parabolic homogenization problem (8), see [31] for details. In this case, we know that the order of the macroscopic equation $d = 2$.

### 4.1 Consistency Analysis

For the effective equation (10), one can write a finite-difference/forward Euler time-stepper as follows,

$$\begin{aligned} U^{n+\delta} &= S(U^n, t_n; \delta t) \\ &= U^n + \delta t\, F(U^n, D^1(U^n), D^2(U^n), t_n) \\ &= U^n + \delta t \left[a^*\, D^2(U^n)\right]. \end{aligned} \tag{27}$$

.

We compare the gap-tooth time-derivative estimator with the effective time derivative. For concreteness, we impose Dirichlet boundary conditions at the boundaries of the boxes, which will clearly introduce artefacts on the estimated time derivative. The subsequent theorem shows that these artefacts can be made arbitrarily small by increasing the buffer size $H$ [31].

**Theorem 1 (Consistency)** *Let $\bar{F}^2(\bar{U}^n, t_n; h, \delta_t, H)$ be a gap-tooth time-stepper for the homogenization problem (8). Then, assuming $U^n = \bar{U}^n$, we have,*

$$\left\|\bar{F}^2(\bar{U}^n, t_n; \delta t, H) - a^* D^2(U^n)\right\| \leq C_4 \underbrace{\frac{\epsilon}{\sqrt{h}\delta t}}_{\text{micro-scales}} + C_5 \underbrace{\left(1 + \frac{h^2}{\delta t}\right)}_{\text{averaging}} \underbrace{\left(1 - \exp(-a^*\pi^2 \frac{\delta t}{H^2})\right)}_{\text{boundary conditions}} \tag{28}$$

Formula (28) shows the main consistency properties of the gap-tooth estimator. The error decays exponentially as a function of buffer size, but the optimal accuracy of the estimator is limited by the presence of the microscopic scales. Therefore, we need to make a trade-off to determine an optimal choice for $H$ and $\delta t$. The smaller $\delta t$, the smaller $H$ can be used to reach optimal accuracy (and thus the smaller the computational cost), but smaller $\delta t$ implies a larger optimal error.

It is shown numerically in [31] that the convergence result does not depend crucially on the type of boundary conditions. E.g. for no-flux boundary conditions, we obtain qualitatively the same result. However, if we know how the macroscopic solution behaves at the boundaries of the boxes, we can use this knowledge to eliminate the buffers. For the diffusion problem, we have shown that we do not need buffer regions if we constrain the macroscopic gradient at the boundaries [32]. However, in general it is very difficult to find and implement such constraints for a given microscopic simulator.

### 4.2 Stability

Theorem 1 establishes the consistency of the gap-tooth scheme. To obtain convergence, we also need stability. In [7], E and Engquist state that the heterogeneous multiscale method is stable if the corresponding comparison scheme is stable, see [7, Theorem 5.5]. This theorem would also apply to our case. However, due to the a priori assumption that the numerical approximation remains bounded, it may be of little practical value. Here, we circumvent

some of these difficulties by studying the stability properties of the scheme *numerically*. This can be done by computing the eigenvalues of the time derivative estimator as a function of $H$.

Consider the homogenization diffusion equation (8) with the diffusion coefficient $a(x/\epsilon) = 1.1 + \sin(2\pi x/\epsilon)$. The homogenized equation is given by (10) with $a^* = 0.45825686$.

We define the concrete patch dynamics scheme to be a forward Euler scheme,

$$\bar{U}^{n+1} = U^n + \Delta t\ \bar{F}^2(\bar{U}^n, t_n; \delta t, H), \tag{29}$$

with the box initialization defined by (23) with second order central finite differences. In this case, the comparison finite difference scheme for the macroscopic equation is given by

$$U^{n+1} = U^n + \Delta t\ F(U^n, t_n) = U^n + a^* \Delta t\ \frac{U^n_{i+1} - 2U^n_i + U^n_{i-1}}{\Delta x^2} \tag{30}$$

The time derivative operator $F(U^n, t_n)$ in the comparison scheme (30) has eigenvalues

$$\lambda_k = -\frac{4a^*}{\Delta x^2} \sin^2(\pi k \Delta x), \tag{31}$$

which, using the forward Euler scheme as time-stepper, results in the stability condition

$$\max_k |1 + \lambda_k \Delta t| \leq 1 \qquad \text{or} \qquad \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2} a^*$$

It can easily be checked that the operator $\bar{F}^2(U^n, t_n; \delta t, H)$ is linear, so we can interpret the evaluation of $\bar{F}^2(U^n, t_n; \delta t, H)$ as a matrix-vector product. We can therefore use any matrix-free linear algebra technique to compute the eigenvalues of $\bar{F}^2(U^n, t_n; \delta t, H)$, e.g. Arnoldi [14]. We choose to compute $\bar{F}^2(U^n, t_n; \delta t, H)$ and $F(U^n, t_n)$ on the domain $[0, 1]$ with Dirichlet boundary conditions, on a mesh of width $\Delta x = 0.05$ and with an inner box width of $h = 2 \cdot 10^{-3}$. We choose $\delta t = 5 \cdot 10^{-6}$ and compute the eigenvalues of $\bar{F}^2(U^n, t_n; \delta t, H)$ as a function of $H$. The results are shown in figure 2. When the buffer size is too small, the eigenvalues of the gap-tooth estimator are closer to 0 than the corresponding eigenvalues of the finite difference scheme. This is because the microscopic simulation approaches a steady state quickly (due to the Dirichlet boundary conditions), instead of following the true system evolution in a larger domain. With increasing buffer size $H$, the eigenvalues of $\bar{F}^2(U^n, t_n; \delta t, H)$ approximate those of $F(U^n, t_n)$, which is an indication of consistency for larger $H$. Since all eigenvalues are negative and the most negative eigenvalue for $\bar{F}^2(U^n, t_n; \delta t, H)$ is always smaller in absolute value than the corresponding eigenvalue of $F(U^n, t_n)$, the patch dynamics scheme is always stable if the comparison scheme is stable.

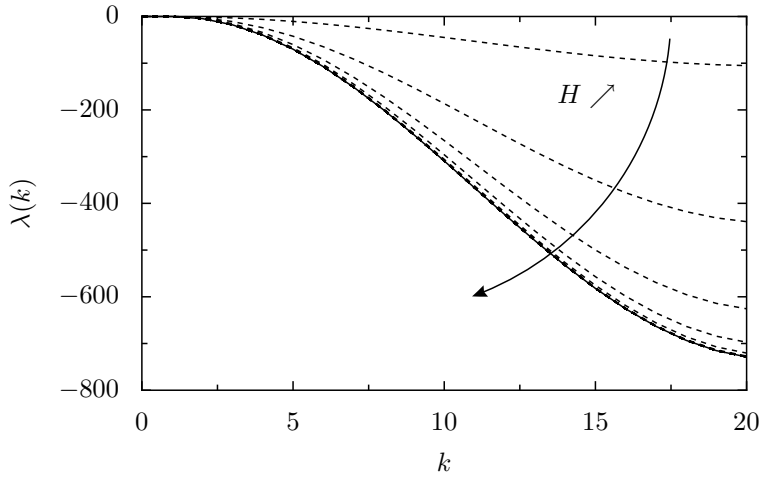



Fig. 2: Spectrum of the estimator $\bar{F}^2(U^n, t_n; \delta t, H)$ (dashed) for the model equation (8) for $H = 2 \cdot 10^{-3}, 4 \cdot 10^{-3}, \ldots, 2 \cdot 10^{-2}$ and $\delta t = 5 \cdot 10^{-6}$, and the eigenvalues (31) of $F(U^n, t_n)$ (solid).

### 4.3 Numerical Illustration

We illustrate the theory with a diffusion homogenization problem. Consider the model problem (8) with

$$a(x/\epsilon) = 1.1 + \sin(2\pi x/\epsilon), \qquad \epsilon = 1 \cdot 10^{-5} \tag{32}$$

as a microscopic problem on the domain $[0, 1]$ with homogeneous Dirichlet boundary conditions and initial condition $u(x, 0) = 1 - 4(x - 1/2)^2$. The corresponding macroscopic equation is given by equation (10), with $a^* = 0.45825686$. This problem has also been used as a model example in [1, 32]. To solve this microscopic problem, we use a second order finite difference discretization with mesh width $\delta x = 1 \cdot 10^{-7}$ and `lsode` [15] as time-stepper. The concrete gap-tooth scheme for this example is again defined by taking second order central finite differences.

We first perform a numerical experiment to show the convergence behaviour in terms of buffer width. Once a suitable buffer width has been determined, we perform a long term simulation.

**Buffer width.** We first compare a gap-tooth step with $h = 2 \cdot 10^{-3}$ and $\Delta x = 1 \cdot 10^{-1}$ with the reference estimator $a^* D^2(\hat{U}^n)$, in which the effective diffusion coefficient is known to be $a^* = 0.45825686$. Figure 3 shows the error with respect to the finite difference time derivative as a function of $H$ (left) and $\delta t$ (right). It is clear that the convergence is in agreement with Theorem 1. We see that smaller values of $\delta t$ result in larger values for the optimal error, but the convergence towards this optimal error is faster.

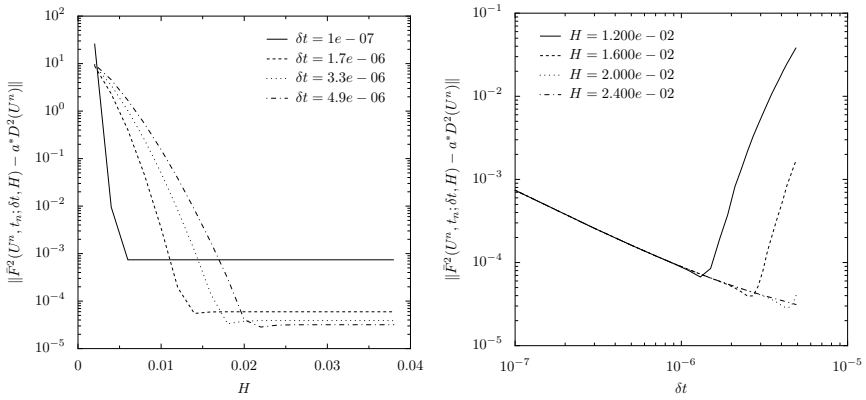



Fig. 3: Error of the gap-tooth estimator $\bar{F}^2(U^n, t_n; \delta t, H)$ with respect to the finite difference time derivative $a^* D^2(U^n)$ on the same mesh. Left: Error with respect to $H$ for fixed $\delta t$. Right: Error with respect to $\delta t$ with fixed $H$.

**Long term simulation.** We now perform a long term simulation and compare the results with a long term simulation using the comparison scheme. The properties for the macroscopic scheme are chosen to be $\Delta x = 1 \cdot 10^{-1}$ and $\Delta t = 1 \cdot 10^{-3}$. As gap-tooth parameters, we choose $H = 8 \cdot 10^{-3}$, $\delta t = 1 \cdot 10^{-6}$ and $h = 1 \cdot 10^{-4}$. Thus, simulations are performed in only 8 % of the spatial domain, and 0.1% of the time domain. The results are shown in figure 4. We also compare the results of the patch dynamics scheme to a reference solution of the effective equation, which is obtained using the comparison scheme on a much finer grid ($\Delta x = 5 \cdot 10^{-3}$ and $\Delta t = 1 \cdot 10^{-6}$). We see that the solution is well approximated, and that the error of the patch dynamics scheme with respect to the finite difference comparison scheme is an order of magnitude smaller than the total error with respect to the reference solution.

## 5 Numerical Results for Advection Problems

Consider equation (15) with

$$c(x/\epsilon) = 1/(3 + \sin(2\pi x/\epsilon)), \qquad \epsilon = 1 \cdot 10^{-5}. \tag{33}$$

The effective equation is then given by (16) with $c^* = 1/3$. The available microscopic simulation code is an upwind/forward Euler time-stepper on a grid with size $\delta x = 5 \cdot 10^{-10}$ and a time-step $\mathrm{d}t = 5 \cdot 10^{-11}$. We take the size of the small boxes to be $h = 5 \cdot 10^{-4}$.

We first investigate how the accuracy of the scheme is influenced by the buffer size $H$ and the gap-tooth time-step $\delta t$. Once a good set of method

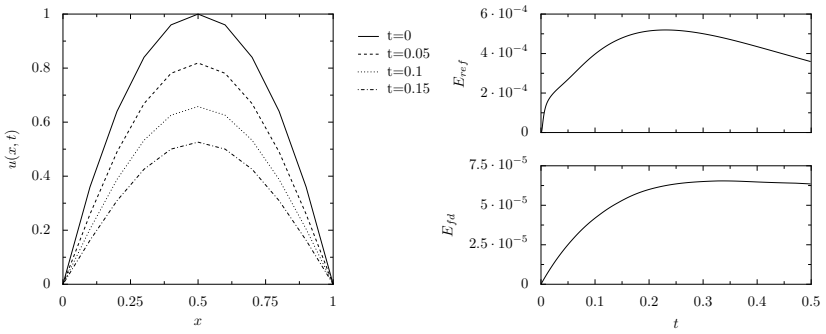



Fig. 4: Left: Snapshots of the solution of the homogenization diffusion equation using the patch dynamics scheme at certain moments in time. Right: error with respect to the "exact" solution of the effective equation (top) and a finite difference comparison scheme (bottom). The total error is dominated by the error of the finite difference scheme.

parameters is found, we perform a long-term simulation. We construct patch dynamics schemes to mimic the upwind, third-order upwind-biased and central fourth-order spatial discretizations.

### 5.1 Consistency

To determine the buffer size $H$ and the gap-tooth time-step $\delta t$, we perform a numerical simulation for this model on the domain $[-H/2, +H/2]$, with $H = h + 5i \cdot 10^{-9}$ for $i = 1, \ldots, 20$ on the time interval $[0, \delta t]$ with $\delta t = j \cdot 10^{-9}$, $j = 1, \ldots, 100$ and the linear initial condition

$$u_\epsilon(x, 0) = D^1 x + D^0 = 3.633x + 0.9511.$$

The results are shown in figure 5(left). We notice two differences with respect to the parabolic case. First, it is clear that we do not need very large buffer regions. Indeed, the advective nature of equation (15) ensures that information travels with finite speed. The consequence is that, as soon as the time-step is too short for the boundary information to reach the interior of the domain, the buffer size $H$ will not have any influence on the accuracy of the result.

The second difference is that the error decreases monotonically with decreasing $\delta t$, whereas the theoretical result for diffusion indicates that we would have an error term of the form $O(\epsilon/\delta t)$. This discrepancy is due to additional numerical inaccuracies during the restriction step, which are caused by the weak convergence towards the homogenized equation in the hyperbolic case. Figure 5 shows how $u_\epsilon(x, t)$ varies as a function of time. We see that the microscopic solution develops oscillations which grow in amplitude with time. Recall that the macroscopic quantity at time $t = \delta t$ is computed as the spatial

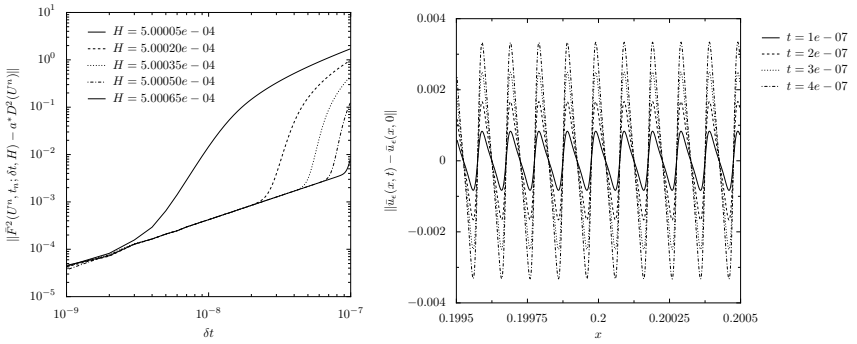



Fig. 5: Left: Error of the gap-tooth estimator with respect to the macroscopic time derivative $c^* D^1$. Right: Difference between the solution inside the box at time $t$ and the initial condition at time $t = 0$.

average of the solution $u_\epsilon(x, \delta t)$ over a box of size $h$. We need to approximate this spatial average using a quadrature formula, in which we can only use the solution on the numerical grid points as quadrature points. Thus, we may expect a decrease of accuracy in the computation of the box average for increasing values of $\delta t$. We numerically verified this intuitive reasoning by increasing $\epsilon$. The box solution then becomes less oscillatory, and we observed that the accuracy of the restriction was increased.

Based on these results, we choose $H = h + 1 \cdot 10^{-7}$ and $\delta t = 5 \cdot 10^{-9}$. Since our macroscopic schemes will use $\Delta x = O(10^{-2})$ and $\Delta t = O(10^{-2})$, the method results in gains of the order of 100 in space and $10^6$ in time. However, we need to mention that, for realistic microscopic problems, part of this spectacular gain will be lost because we need to initialize the microscopic system consistently (the lifting step) using only a few low-order moments, which may require additional microscopic simulations [11, 12, 23].

### 5.2 First Order Upwind Scheme

We perform a numerical simulation for this model on the domain $[0, 1]$ with periodic boundary conditions. As an initial condition, we choose

$$u^0(x) = (\sin(\pi x))^{100}, \tag{34}$$

which is a typical initial condition to study spatial discretizations for the advection equation [18]. We use a macroscopic mesh of size $\Delta x = 1 \cdot 10^{-2}$ and a time-step $\Delta t = 1 \cdot 10^{-2}$, and we define our macroscopic comparison scheme as an upwind/forward Euler scheme

$$U_i^{n+1} = U_i^n - \Delta t \; c^* \frac{U_i^n - U_{i-1}^n}{\Delta x}. \tag{35}$$

The corresponding patch dynamics scheme is defined by the algorithm in section 3, where the initial condition (23) is defined by taking $d = 1$ and

$$D_i^1(\bar{U}^n) = \frac{\bar{U}_i^n - \bar{U}_{i-1}^n}{\Delta x}, \qquad D_i^0(\bar{U}^n) = \bar{U}_i^n. \tag{36}$$

The resulting time derivative estimator is used with a forward Euler time-stepper. The results are shown in figure 6. The patch dynamics scheme clearly

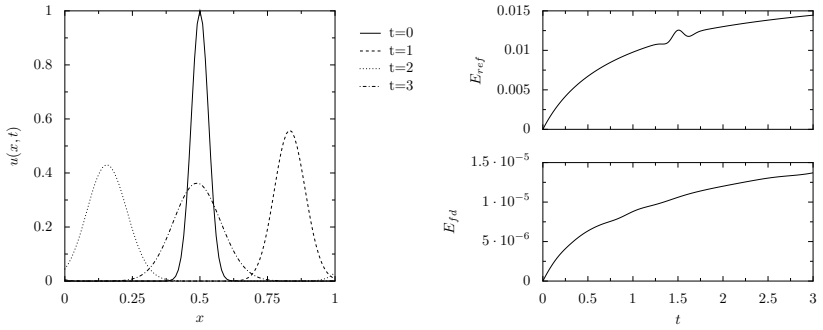



Fig. 6: Left: Snapshots of the solution of the homogenization advection equation (15) with coefficient (33) using the upwind patch dynamics scheme at certain moments in time. Right: $L_2$-error with respect to the "exact" solution of the effective equation (16) (top) and the finite difference comparison scheme (35) (bottom). The total error is dominated by the error of the finite difference scheme.

has the same properties as the comparison finite difference scheme. It is very diffusive, but maintains positivity. The left figure shows the $L_2$-error of patch dynamics with respect to the finite difference scheme, and with respect to an "exact" solution of the effective equation, which was obtained using the upwind scheme on a very fine mesh with $\Delta x = 1 \cdot 10^{-4}$ and $\Delta t = 1 \cdot 10^{-4}$. We see that the error of the patch dynamics scheme is completely dominated by the finite difference error.

### 5.3 Third-Order Upwind-Biased Scheme

Next, we design a patch dynamics algorithm to mimic the third-order upwind-biased scheme as a spatial discretization, which we combine with the classical fourth-order Runge–Kutta time integration method. In this case, the macroscopic time derivative is given by

$$F(U_i^n, t_n) = \frac{c^*}{\Delta x}\left(-\frac{1}{6}U_{i-2}^n + U_{i-1}^n - \frac{1}{2}U_i^n - \frac{1}{3}U_{i+1}^n\right). \tag{37}$$

The Runge–Kutta method requires some auxiliary evaluations of the time derivative operator,

$$
\begin{aligned}
k_1 &= F(U_i^n, t_n) \\
k_2 &= F(U_i^n + \frac{\Delta t}{2} k_1, t_n + \frac{\Delta t}{2}) \\
k_3 &= F(U_i^n + \frac{\Delta t}{2} k_2, t_n + \frac{\Delta t}{2}) \\
k_4 &= F(U_i^n + \Delta t\ k_3, t_n + \Delta t)
\end{aligned}
\tag{38}
$$

and the time-stepper $U^{n+1} = S(U^n, t_n; \Delta_t)$ is then defined as

$$
U^{n+1} = U^n + \Delta t \ \left( \frac{1}{6} k_1 + \frac{1}{3} k_2 + \frac{1}{3} k_3 + \frac{1}{6} k_4 \right) \tag{39}
$$

The corresponding patch dynamics scheme is defined by the algorithm in section 3, where the initial condition (23) is defined by taking $d = 1$ and

$$
\begin{aligned}
D_i^1(\bar{U}^n) &= \frac{1}{\Delta x} \left( \frac{1}{6} \bar{U}_{i-2}^n - \bar{U}_{i-1}^n + \frac{1}{2} \bar{U}_i^n + \frac{1}{3} \bar{U}_{i+1}^n \right), \\
D_i^0(\bar{U}^n) &= \bar{U}_i^n.
\end{aligned}
\tag{40}
$$

The resulting time derivative estimator is subsequently used inside the fourth-order Runge–Kutta method.

We perform a numerical simulation on a macroscopic mesh with size $\Delta x = 2 \cdot 10^{-2}$ and $\Delta t = 2 \cdot 10^{-2}$. The results are shown in figure 7. The patch

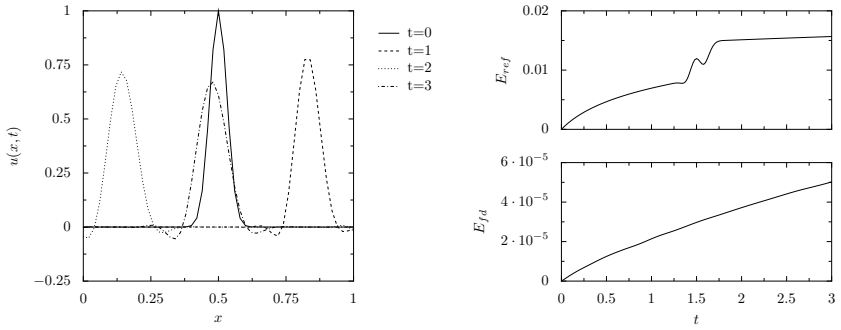



Fig. 7: Left: Snapshots of the solution of the homogenization advection equation (15) with coefficient (33) using the upwind-biased patch dynamics scheme at certain moments in time. Right: $L_2$-error with respect to the "exact" solution of the effective equation (16) (top) and the finite difference comparison scheme (39)-(37) (bottom). The total error is dominated by the error of the finite difference scheme.

dynamics scheme clearly has the same properties as the comparison finite difference scheme. It is less diffusive than the upwind scheme, but some artificial oscillations are introduced. The left figure shows the $L_2$-error of patch

dynamics with respect to the finite difference scheme, and with respect to an "exact" solution of the effective equation, which was obtained using the upwind scheme on a very fine mesh with $\Delta x = 1 \cdot 10^{-4}$ and $\Delta t = 1 \cdot 10^{-4}$. Again, we see that the error in approximating the exact solution is completely dominated by the error of the macroscopic scheme, while the errors due to estimation are negligible.

### 5.4 Fourth-Order Central Scheme

Finally, we design a patch dynamics algorithm to mimic a fourth-order central scheme as a spatial discretization, which we combine again with the classical fourth-order Runge–Kutta time integration method. In this case, the macroscopic time derivative is given by

$$F(U_i^n, t_n) = \frac{c^*}{\Delta x}\left(-\frac{1}{12}U_{i-2}^n + \frac{2}{3}U_{i-1}^n - \frac{2}{3}U_{i+1}^n + \frac{1}{12}U_{i+2}^n\right), \qquad (41)$$

and the time-integration method is again given by (38)-(39). The corresponding patch dynamics scheme is defined by the algorithm in section 3, where the initial condition (23) is defined by taking $d = 1$ and

$$\begin{aligned} D_i^1(\bar{U}^n) &= \frac{1}{\Delta x}\left(\frac{1}{12}\bar{U}_{i-2}^n - \frac{2}{3}\bar{U}_{i-1}^n + \frac{2}{3}\bar{U}_{i+1}^n - \frac{1}{12}\bar{U}_{i+2}^n\right), \\ D_i^0(\bar{U}^n) &= \bar{U}_i^n. \end{aligned} \qquad (42)$$

The resulting time derivative estimator is subsequently used inside the fourth-order Runge–Kutta method.

We perform a numerical simulation on a macroscopic mesh with size $\Delta x = 2 \cdot 10^{-2}$ and $\Delta t = 2 \cdot 10^{-2}$. The results are shown in figure 8. The patch dynamics scheme clearly has the same properties as the comparison finite difference scheme. It is much less diffusive than the upwind scheme, but many artificial oscillations are introduced, which is typical behaviour for central schemes. The left figure shows the $L_2$-error of patch dynamics with respect to the finite difference scheme, and with respect to an "exact" solution of the effective equation, which was obtained using the upwind scheme on a very fine mesh with $\Delta x = 1 \cdot 10^{-4}$ and $\Delta t = 1 \cdot 10^{-4}$. Again, we see that the error in approximating the exact solution is completely dominated by the error of the macroscopic scheme, while the errors due to estimation are negligible.

### 5.5 Advection Coefficients with Macro-Scale Variations

As a second example, we consider equation (15) with

$$c(x/\epsilon) = 1/(3 + \sin(2\pi x/\epsilon) + \sin(2\pi x)), \qquad \epsilon = 1 \cdot 10^{-5}. \qquad (43)$$

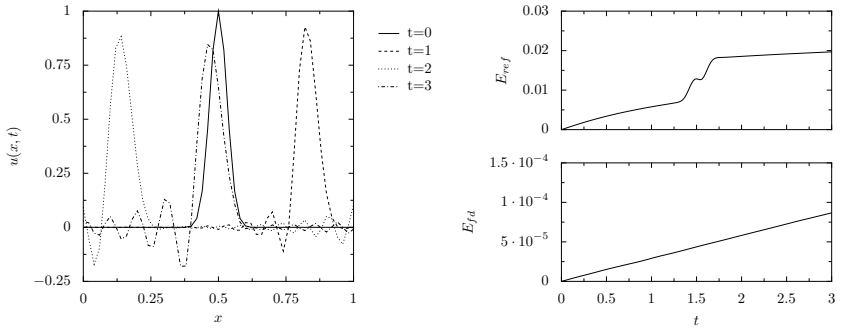



Fig. 8: Left: Snapshots of the solution of the homogenization advection equation (15) with coefficient (33) using the central fourth-order patch dynamics scheme at certain moments in time. Right: $L_2$-error with respect to the "exact" solution of the effective equation (16) (top) and the finite difference comparison scheme (39)-(41) (bottom). The total error is dominated by the error of the finite difference scheme.

The effective equation is then given by (16) with $c^* = 1/(3 + \sin(2\pi x))$. The available microscopic simulation code is an upwind/forward Euler time-stepper on a grid with size $\delta x = 5 \cdot 10^{-10}$ and a time-step $\mathrm{d}t = 5 \cdot 10^{-11}$. We take the size of the small boxes to be $h = 5 \cdot 10^{-4}$.

We choose $H = h + 2 \cdot 10^{-7}$ and $\delta t = 5 \cdot 10^{-9}$ as method parameters, and we perform a patch dynamics simulation using a macroscopic mesh size $\Delta x = 2 \cdot 10^{-2}$ and $\Delta t = 5 \cdot 10^{-3}$ using the upwind initialization (36), combined with forward Euler in time.

The simulations show that the patch dynamics scheme is a good approximation to a finite difference approximation of equation (16) *in non-conservative form*. In particular, the correct comparison scheme would be

$$U_i^{n+1} = U_i^n - \Delta t \left( c^*(x_i) \frac{U_i^n - U_{i-1}^n}{\Delta x} + U_i^n \partial_x c^*(x_i) \right), \tag{44}$$

which is not entirely the same as the classical finite volume upwind scheme

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x} \left( c^*(x_{i+1/2}) U_i^n - c^*(x_{i-1/2}) U_{i-1}^n \right). \tag{45}$$

In particular, the scheme (44) is not conservative.

The results are shown in figure 9. Again, we note that the first-order upwind scheme is very diffusive, and that the error of the patch dynamics scheme with respect to the finite difference approximation (44) is 3 orders of magnitude smaller than the error with respect to the exact solution. Moreover, the error with respect to the finite difference scheme is an order of magnitude smaller than the error with respect to the finite volume scheme (45), which is consistent with the statements made above.

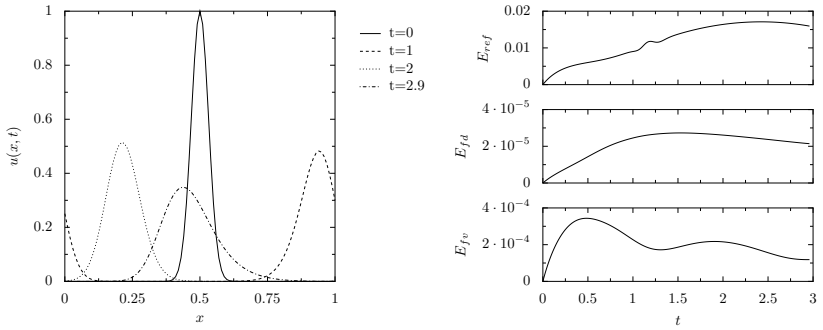



Fig. 9: Left: Snapshots of the solution of the homogenization advection equation (15) with coefficient (43) using the first-order upwind patch dynamics scheme at certain moments in time. Right: error with respect to the "exact" solution of the effective equation (16) (top), the finite difference comparison scheme (44) (middle) and the finite volume scheme (45) (bottom). The total error is dominated by the error of the finite difference scheme, and the error with respect to the finite volume scheme is significantly larger that the error with respect to (44).

## 6 Conclusions

In this paper, we reviewed the patch dynamics scheme and showed its basic convergence properties on model hyperbolic and parabolic homogenization problems. We illustrated that the scheme is capable of reproducing the correct macroscopic behaviour, even when the macroscopic equation is not of diffusion-type, and demonstrated that the required buffer size depends severely on the properties of the effective equation. Specifically, in the case of a macroscopic transport equation, the buffers can be very small compared to the diffusion case.

We wish to stress the fact that patch dynamics is an approximation to a finite difference scheme of the macroscopic equation *in non-conservative form*, which is most apparent in the case of coefficients that vary on a macroscopic scale. However, we note that there is no guarantee that the patch dynamics scheme will be conservative, even if the corresponding finite difference scheme is, since the extra errors that are induced might (and will) destroy conservation in the numerical solutions. When numerical conservation is important (e.g. if the macroscopic solution would develop sharp fronts), we will therefore need to resort to a finite volume formulation of the patch dynamics scheme. This variant is currently under active investigation.

## References

1. A. Abdulle, W.E: Finite difference heterogeneous multi-scale method for homogenization problems. *Journal of Computational Physics* **191** (1), 18–39 (2003)

2. S. Attinger: Generalized coarse-graining procedures for flow in porous media. *Computational Geosciences* **7**, 253–273 (2003)
3. I. Babuska: Homogenization and its applications. In: *SYNSPADE*, ed. by B. Hubbard, 89–116 (1975)
4. A. Bensoussan, J.L. Lions, G. Papanicolaou: *Asymptotic analysis of periodic structures*, vol. 5 of *Studies in Mathematics and its Applications*. (North-Holland, Amsterdam 1978)
5. D. Cioranescu, P. Donato: *An introduction to homogenization* (Oxford University Press 1999)
6. M. Dorobantu, B. Engquist: Wavelet-based numerical homogenization. *SIAM J. Numer. Anal.* **35** (2), 540–559 (1998)
7. W. E, B. Engquist: The heterogeneous multi-scale methods. *Comm. Math. Sci.* **1** (1), 87–132 (2003)
8. B. Engquist: Computation of oscillatory solutions to hyperbolic differential equations, In: *Springer Lecture Notes Math.*, vol. 1270, 10–22 (1987)
9. B. Engquist, O. Runborg: Wavelet-based numerical homogenization with applications. In *Multiscale and Multiresolution Methods*, vol. 20 of *Lecture Notes in Computational Science and Engineering*, 97–148 (Springer, Berlin Heidelberg New York 2002)
10. R. Erban, H.G. Othmer: From individual to collective behavior in bacterial chemotaxis. *SIAM J. on Applied Mathematics* **65** (2), 361–391 (2004)
11. C.W. Gear, T.J. Kaper, I.G. Kevrekidis, A. Zagaris: Projecting to a slow manifold: Singularly perturbed systems and legacy codes. *SIAM J. on Applied Dynamical Systems* **4** (3) 711–732 (2005)
12. C.W. Gear, I.G. Kevrekidis: Constraint-defined manifolds: a legacy code approach to low-dimensional computation. *J. Sci. Comp.* (2004) In press
13. C.W. Gear, I.G. Kevrekidis, C. Theodoropoulos: "Coarse" integration/bifurcation analysis via microscopic simulators: micro-Galerkin methods. *Computers and Chemical Engineering* **26**, 941–963 (2002)
14. G.H. Golub and C.F. Van Loan: *Matrix computations (3rd ed.)* (Johns Hopkins University Press, Baltimore, MD, USA 1996)
15. A.C. Hindmarsh: ODEPACK, a systematized collection of ODE solvers. In: *Scientific Computing*, ed. by R.S. Stepleman et al, 55–64 (North-Holland, Amsterdam 1983)
16. T.Y. Hou, X.H. Wu: A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.* **134**, 169–189 (1997)
17. T.Y. Hou, X.H. Wu: Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients. *Mathematics of Computation*, **68** (227), 913–943 (1999)
18. W. Hundsdorfer, J.G. Verwer: *Numerical solution of time-dependent advection-diffusion-reaction equations*, vol. 33 of *Springer Series in Computational Mathematics* (Springer, Berlin Heidelberg New York 2003)
19. B. Jourdain, T. Lelièvre, C. Le Bris: Numerical analysis of micro-macro simulations of polymeric fluid flows: a simple case. *Mathematical Models and Methods in Applied Sciences* **12** (9), 1205–1243 (2002)
20. R. Keunings: Micro-macro methods for the multiscale simulation of viscoelastic flows using molecular methods of kinetic theory. In: *Rheology Reviews*, ed. by D.M. Binding, K. Walters,67–98 (British Society of Rheology, 2004)

21. I.G. Kevrekidis, C.W. Gear, J.M. Hyman, P.G. Kevrekidis, O. Runborg, C. Theodoropoulos: Equation-free, coarse-grained multiscale computation: enabling microscopic simulators to perform system-level tasks. *Comm. Math. Sciences*: **1** (4), 715–762 (2003)
22. P. Van Leemput, K. Lust, I.G. Kevrekidis: Coarse-grained numerical bifurcation analysis of lattice boltzmann models. *Physica D*, **210** (1–2), 58–76 (2005)
23. P. Van Leemput, W. Vanroose, D. Roose: Initialization of a lattice-Boltzmann model with constrained runs. *J Comput. Phys.* (2005) Submitted.
24. A.G. Makeev, D. Maroudas, A.Z. Panagiotopoulos, I.G. Kevrekidis: Coarse bifurcation analysis of kinetic Monte Carlo simulations: a lattice-gas model with lateral interactions. *J. Chem. Phys.* **117** (18), 8229–8240 (2002)
25. A.M. Matache, I. Babuska, C. Schwab. Generalized p-FEM in homogenization. *Numerische Mathematik* **86** (2), 319–375 (2000)
26. N. Neuss, W. Jäger, G. Wittum: Homogenization and multigrid. *Computing* **66** (1), 1–26 (2001)
27. H.C. Öttinger: *Stochastic processes in polymeric fluids* (Springer, Berlin Heidelberg New York 1996)
28. A.J. Roberts, I.G. Kevrekidis: Higher order accuracy in the gap-tooth scheme for large-scale solutions using microscopic simulators. E-print: math.DS/0410310 arxiv.org. (2004)
29. O. Runborg, C. Theodoropoulos, I.G. Kevrekidis: Effective bifurcation analysis: a time-stepper based approach. *Nonlinearity* **15**, 491–511 (2002)
30. G. Samaey, I.G. Kevrekidis, D. Roose: Damping factors for the gap-tooth scheme. In: *Multiscale Modeling and Simulation*, ed. by S. Attinger and P. Koumoutsakos, vol. 36 of *Lecture Notes in Computational Science and Engineering*, 93–102 (Springer, Berlin Heidelberg New York 2004)
31. G. Samaey, I.G. Kevrekidis, D. Roose: Patch dynamics with buffers for homogenization problems. *Journal of Computational Physics*, 2005 In press.
32. G. Samaey, D. Roose, I.G. Kevrekidis. The gap-tooth scheme for homogenization problems. *SIAM Multiscale Modeling and Simulation* **4** (1), 278–306 (2005)
33. C. Schwab, A.M. Matache: Generalized FEM for homogenization problems. *Multiscale and Multiresolution methods*, vol. 20 of *Lecture Notes in Computational Science and Engineering*, 197–238 (Springer, Berlin Heidelberg New York 2002)
34. C.I. Siettos, M.D. Graham, I.G. Kevrekidis: Coarse Brownian dynamics for nematic liquid crystals: bifurcation, projective integration and control via stochastic simulation. *J. Chem. Phys.* **118** (22), 10149–10157 (2003) can be obtained as cond-mat/0211455 at arxiv.org.
35. C. Theodoropoulos, Y.H. Qian, I.G. Kevrekidis: Coarse stability and bifurcation analysis using time-steppers: a reaction-diffusion example. *Proc. Natl. Acad. Sci.* **97**, 9840–9845 (2000)

# Coarse-Graining the Cyclic Lotka-Volterra Model: SSA and Local Maximum Likelihood Estimation

C. P. Calderon[1], G. A. Tsekouras[2,3], A. Provata[2], and I. G. Kevrekidis[1,4]

[1] Department of Chemical Engineering, Princeton University, Princeton, New Jersey 08544-5263, USA
[2] Institute of Physical Chemistry, National Research Center "Demokritos", 15310 Athens, Greece
[3] Physics Department, University of Athens, Panepistimioupolis, 10679 Athens, Greece
[4] Corresponding Author: yannis@arnold.Princeton.edu

**Summary.** When the output of an atomistic simulation (such as the Gillespie stochastic simulation algorithm, SSA) can be approximated as a diffusion process, we may be interested in the dynamic features of the deterministic (drift) component of this diffusion. We perform traditional scientific computing tasks (integration, steady state and closed orbit computation, and stability analysis) on such a drift component using a SSA simulation of the Cyclic Lotka-Volterra system as our illustrative example. The results of short bursts of appropriately initialized SSA simulations are used to fit local diffusion models using Aït-Sahalia's transition density expansions [1, 2, 3] in a maximum likelihood framework. These estimates are then coupled with standard numerical algorithms (such as Newton-Raphson or numerical integration routines) to help design subsequent SSA experiments. A brief discussion of the validity of the local diffusion approximation of the SSA simulation (a jump process) is included.

## 1 Introduction

Reactive particle dynamic models arise in scientific fields ranging from physical and chemical processes to systems biology [33, 34, 37, 41, 38, 19, 17]. Incorporating successive levels of detail in the modeling quickly leads to models that are analytically intractable, necessitating computational exploration. Gillespie's Stochastic Simulation Algorithm (SSA) and its variants [41, 20, 19] have gained popularity in recent years for modeling so-called mixed reacting systems; the approach provides a middle ground between detailed molecular dynamics and lumped, Ordinary Differential Equation (ODE) descriptions of chemical kinetics, incorporating fluctuations. Knowing the kinetic scheme underlying such a simulation allows one to write, at the infinite particle limit,

the corresponding kinetic ODE. At intermediate particle numbers ($N_{mol}$), the SSA has been approximated with the continuous "chemical Langevin equation" [19].

In what follows we will assume that the results of an SSA simulation can be successfully approximated through a continuous diffusion process. Explicit knowledge of the drift and noise components of such a process allows one to easily analyze certain features of the overall behavior; one might, for example, be interested in the bifurcation behavior of the "underlying" drift component of the model, including the number and stability of its steady states and their parametric dependence. In our work we assume that the only available simulation tool is a "black box" SSA simulator, in which the mechanistic rules have been correctly incorporated, but which we, as users, do not know: we can only observe the SSA simulator *output*. We want to perform a quantitative computational study of the underlying drift component. Since we cannot derive it in closed form (not knowing the evolution rules), we want to perform this study using the least possible simulation with the SSA code. The approach we use follows the so-called "equation-free" framework [28, 27]: in this framework traditional numerical algorithms become protocols for designing short bursts of numerical experiments with the SSA code. The quantities necessary for numerical computation with the unavailable model (time derivatives, the action of Jacobians) are estimated locally by processing the "fine scale" SSA simulations. In this work we extract such numerical information via parametric local diffusion models using the transition density expansions proposed by Aït-Sahalia [2, 3]. The numerical procedures we illustrate can also be used, in principle, for different types of "fine scale" models if their output happens to be well approximated by diffusion processes.

The article is organized as follows: In Section 2 we describe our illustrative model system. We then quickly outline the basic ideas underlying equation-free numerics (Section 3), and discuss our estimation procedure (Section 4). Our computational results are presented in Section 5, and we conclude with a discussion including goodness-of-fit issues.

## 2 The Lattice Lotka-Volterra Model

Our Cyclic Lotka-Volterra [36, 14] illustrative example consists of a three-species ($X$, $Y$ and $S$) nonlinear kinetic scheme of the following form [36]:

$$\begin{aligned} X + Y &\xrightarrow{k_1} 2Y \\ Y + S &\xrightarrow{k_2} 2S \\ S + X &\xrightarrow{k_3} 2X. \end{aligned}$$

In the remainder of the paper we will refer to it simply as LV. In the deterministic limit, this kinetic scheme gives rise to a set of three coupled nonlinear ODEs for the evolution of the concentrations $X, Y$ and $S$.

$$\begin{aligned}\frac{dX}{dt} &= -k_1XY + k_3XS \\ \frac{dY}{dt} &= -k_2YS + k_1YX \\ \frac{dS}{dt} &= -k_3SX + k_2SY\end{aligned} \tag{1}$$

The total concentration $(X+Y+S)$ is constant over time; setting (without loss of generality) this constant to unity and eliminating $S$

$$X + Y + S = 1, \rangle\rangle\rangle \Longrightarrow \rangle\rangle\rangle S = 1 - X - Y$$

reduces the system to

$$\begin{aligned}\frac{dX}{dt} &= X\left[k_3 - k_3X - (k_1 + k_3)Y\right] \\ \frac{dY}{dt} &= Y\left[-k_2 + (k_1 + k_2)X + k_2Y\right].\end{aligned} \tag{2}$$

For every (positive) value of $k_1, k_2$ and $k_3$ four fixed points exist: three trivial and one non-trivial steady state:

$$\begin{aligned}&X_s = 0,\, Y_s = 0,\, S = 1 \quad (\text{system invaded by S}) \\ &X_s = 1,\, Y_s = 0,\, S = 0 \quad (\text{system invaded by X}) \\ &X_s = 0,\, Y_s = 1,\, S = 0 \quad (\text{system invaded by Y}) \\ &X_s = \frac{k_2}{K},\, Y_s = \frac{k_3}{K}, S_s = \frac{k_1}{K} \quad (\text{nontrivial fixed point})\end{aligned}$$

where

$$K = k_1 + k_2 + k_3. \tag{3}$$

An interesting feature of the phase space of the deterministic model is the existence of a one-parameter family of closed orbits surrounding a "center" (see Figure 3). The neutral stability of these orbits affects, as we will see below, the fixed point algorithms used to converge on them. The system is simulated through both the ODEs (2) and through an SSA implementation of the kinetic scheme (1) using $k_1, k_2 = 0.5$ and $k_3 = 0.7$ throughout.

## 3 Equation Free Computation

The basic premise underlying equation-free modeling and computation is that we have available a "black box" fine-scale dynamic simulator, and we believe

that an *effective* evolution equation exists (closes) for some set of (coarse-grained) *outputs* or *observables* of the fine scale simulation. As discussed in more detail in [27, 23], one can numerically solve this (unavailable explicitly) equation through linking traditional numerical methods with the fine scale code; in particular, the classical continuum algorithms become protocols for the design of short, appropriately initialized numerical experiments with the fine scale code. The process starts by identifying the appropriate coarse-grained observables (sometimes also called order parameters); typically these variables are low-order moments of microscopically/stochastically evolving particle distributions (e.g. concentrations for chemically reacting systems, like our example). In general, good coarse-grained observables are not known, and data analysis techniques to identify them from computational or experimental observations are the subject of intense current research [39, 8, 12]. If the unavailable "effective" equations are deterministic and reasonably smooth, short runs of the fine-scale simulator are used to estimate *time derivatives* of the coarse-grained observables; initializing fine scale simulations consistent with nearby values of the coarse-grained observables gives estimates of directional derivatives (again assuming appropriate smoothness), and can be linked with matrix-free iterative linear algebra techniques (e.g. [26]). When an explicit evolution equation is available, these quantities, necessary in numerical computation, are obtained through function or Jacobian evaluations of the model formulas; here, they are estimated *on demand* from short computational experiments with the fine scale solver. If the underlying effective equation is stochastic, e.g. a diffusion, then the results of the short simulation bursts must be used to estimate both the drift and the noise components of the effective model - this is the case we study here. We will illustrate, using the SSA LV example, how certain types of computations can be accelerated by appealing to classical numerical methods.

## 4 Estimation Procedure

In what follows, we will assume that species concentrations are good observables, and that the true process (the LV SSA simulation) can be adequately approximated by a diffusion process, that is, a stochastic differential equation (SDE) of the form:

$$d\mathbf{X_t} = \boldsymbol{\mu}(\mathbf{X_t};\theta)dt + \boldsymbol{\Sigma}(\mathbf{X_t};\theta)d\mathbf{W_t}. \tag{4}$$

Here $\mathbf{X_t}$ is a stochastic process which is meant to model the evolution of the observable(s), $\mathbf{W_t}$ represents a vector of standard Brownian motions, and the functions $\boldsymbol{\mu}(\mathbf{X_t};\theta)$ and $\boldsymbol{\Sigma}(\mathbf{X_t};\theta)$ are the drift and diffusion coefficients of the process. In the classical parametric setup, one assumes that the parameterized function families to which the drift and diffusion coefficient functions belong are known, and that the parameter vector $\theta$ is finite dimensional. In practice

one rarely knows a class of functions which can be used to describe the *global* dynamics of the observables; in the equation free computations below, however, we simulate the true process for only relatively short bursts of time. It therefore makes sense to (locally) consider the following SDE:

$$d\mathbf{X_t} = \Big(\mathbf{A} + \mathbf{B}(\mathbf{X_t} - \mathbf{X_o})\Big)dt + \Big(\mathbf{C} + \mathbf{D}(\mathbf{X_t} - \mathbf{X_o})\Big)d\mathbf{W_t}. \tag{5}$$

where $\mathbf{W_t}$,$\mathbf{X_t}$,$\mathbf{X_o}$, $\mathbf{A}$ and $\mathbf{C} \in \mathbb{R}^d$ and $\mathbf{B}$ and $\mathbf{D} \in \mathbb{R}^{d\times d}$ (the $d$ Brownian motions are assumed independent; the vector multiplying them, by slight abuse of notation, contains the nonzero elements of the diagonal matrix $\mathbf{\Sigma}$; extending to the correlated case is straightforward).

This simple model is based on the fact that we expect smooth evolution of moments of the observables, while at the same time taking into account the state dependence of the noise (neglecting this dependence can cause bias in the estimation of the drift). The parameters of this local linear model are estimated through techniques associated with maximum likelihood estimation (MLE). The motivation for using MLE techniques stems from the fact that under certain regularity conditions [42] such estimators are (asymptotically) efficient as regards the variance of the estimated parameter distribution. In addition, the asymptotic parameter distributions associated with MLE can sometimes be worked out analytically, or approximated through Monte Carlo simulations; this knowledge can guide the selection of the sample size necessary for a given desired accuracy in coarse-grained computations [11].

### 4.1 Maximum Likelihood Estimation for Discretely Observed Diffusions

We now recall a few basic facts about MLE estimation; standard references include, e.g. [21, 24, 42]. It is assumed throughout that the *exact* distribution associated with the parametric model admits a continuous density whose logarithm is well defined almost everywhere and is three times continuously differentiable with respect to the parameters [30].

MLE is based on maximizing the log-likelihood ($\mathcal{L}_\theta$) with respect to the parameter vector (for our model $\theta \equiv [\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}]$):

$$\mathcal{L}_\theta \equiv \log\Big(f(\mathbf{x};\theta)\Big). \tag{6}$$

In the above equation, $\mathbf{x}$ corresponds to a matrix of observations $\in \mathbb{R}^{d\times M}$ where $d$ is the dimension of the state and $M$ is the length of the time series; $f(\mathbf{x};\theta)$ corresponds to the probability of making observation $\mathbf{x}$. For a single sample path of a discretely observed diffusion known to be initialized at $\mathbf{x}_0$, $f(\mathbf{x};\theta)$ can be evaluated as [21]:

$$f(\mathbf{x};\theta) = \delta_{x_0} \prod_{m=1}^{M-1} f(\mathbf{x}_m|\mathbf{x}_{m-1};\theta). \tag{7}$$

In this equation $f(\mathbf{x}_m|\mathbf{x}_{m-1};\theta)$ represents the conditional probability (transition density) of observing $\mathbf{x}_m$ given the observation $\mathbf{x}_{m-1}$ for a given $\theta$ and $\delta_{x_0}$ is the Dirac distribution. In our applications, we search for the parameter vector that is best over *all* observations (we have an ensemble of $N$ paths of length $M$). In this case our expression for the log-likelihood (given the data and transition density) takes the form:

$$\mathcal{L}_\theta := \sum_{i=1}^{N}\sum_{m=1}^{M} \log\Big(f(\mathbf{x}_m^i|\mathbf{x}_{m-1}^i;\theta)\Big). \tag{8}$$

Assume the existence of an invertible symmetric positive definite "scaling matrix" matrix $\mathcal{F}_{(M,\,\theta)}$ [31] associated with the estimator; the subscripts are used to make the dependence of the scaling matrix on $M$ and $\theta$ explicit. For the "standard" case $N = 1$ in time series analysis, under some additional regularity assumptions [24, 42], one has the following limit for a *correctly specified* parametric model:

$$\mathcal{F}^{\frac{1}{2}}_{(M,\,\hat{\theta})}(\theta_M - \hat{\theta}) \overset{\mathbb{P}_{\hat{\theta}}}{\Longrightarrow} N(\mathbf{0},\mathbf{I}). \tag{9}$$

Here $\hat{\theta}$ is the true parameter vector; $\theta_M$ represents the parameters estimated with a finite time series of length $M$; $\overset{\mathbb{P}_{\hat{\theta}}}{\Longrightarrow}$ denotes convergence in distribution [42, 21] under $\mathbb{P}_{\hat{\theta}}$ (the distribution associated with the density $f(\mathbf{x};\hat{\theta})$); $N(\mathbf{0},\mathbf{I})$ denotes a normal distribution with mean zero and an identity matrix for the covariance. For a correctly specified model family, $\mathcal{F}_{(M,\,\hat{\theta})}$ can be estimated in a variety of ways [44, 31]. The appeal of MLE lies in that, asymptotically in $M$, the variance of the estimated parameters is the smallest that can be achieved by an estimator that satisfies the assumed regularity conditions [24, 42].

### 4.2 Transition Density Expansions

Here we briefly outline the key features of the recent work of Aït-Sahalia [2, 3] used in our coarse-grained computations below. The problem with using even a simple model like that given in equation 5 is that the transition density associated with the process is not known in closed form. In recent years, many attempts to approximate the transition density have appeared in the literature; some techniques depend on analytical approximations whereas others are simulation based (see, e.g. [1, 2, 5, 9, 18, 35]). We have used, with some success, the expansions found in [1, 3, 2]. High accuracy can be obtained using this method to approximate the transition density associated with a *scalar* process; the multivariate case is discussed in [3]. The basic idea behind the scalar case, presented in [1, 2], is as follows: One first transforms the process given in equation 4 into a new process [2]:

$$dY_t = \mu_Y(Y_t;\theta)dt + dW_t \tag{10}$$

$$Y \equiv \gamma(X;\theta) = \int^X \frac{du}{\sigma(u,\theta)}$$

$$\mu_Y(y;\theta) \equiv \tag{11}$$
$$\frac{\mu(\gamma^{-1}(y;\theta);\theta)}{\sigma(\gamma^{-1}(y;\theta);\theta)} - \frac{1}{2}\frac{\partial\sigma}{\partial x}(\gamma^{-1}(y;\theta);\theta)$$

An additional change of variables brings the transition density of the process closer to a standard normal density $Z \equiv \Delta^{-\frac{1}{2}}(Y - y_o)$ where $\Delta$ is the time between observations. The transformations introduced allow the use of a Hermite basis set in order to approximate the transition density of the original process via the following series:

$$p_Z(\Delta, z|y_o;\theta) \approx \tag{12}$$
$$\phi(z)\sum_{j=0}^{K}\eta_Z^{(j)}(\Delta, y_o;\theta)H_j(z)$$

$$\eta_Z^{(j)}(\Delta, y_o;\theta) \equiv \tag{13}$$
$$\frac{1}{j!}\int_{-\infty}^{\infty} H_j(z)p_Z(\Delta, z|y_o;\theta)dz :=$$
$$\frac{1}{j!}\mathbb{E}[H_j\Big(\Delta^{-\frac{1}{2}}(Y_{t+\Delta} - y_o)\Big)|Y_t = y_o;\theta]$$

In the above, $H_j$ represents the $j^{th}$ Hermite polynomial and $\phi(\cdot)$ is the standard normal density. The coefficients needed for the approximation are obtained through the conditional moments of the process $Y_t$. Aït-Sahalia outlines [2] a procedure which exploits the connection between the SDE and the associated Kolmogorov equations in order to develop a closed form expression for the $\eta_Z^{(j)}$ coefficients. The approximation is exact if $K \to \infty$ and the coefficient functions satisfy the assumptions laid out in [2]. In numerical applications one must always deal with a finite K. Problems may arise in the truncated expansion: the approximation of the density may not normalize to unity or, worse, it may become negative (see [3, 1, 5] for some possible remedies).

In the multivariate case, it becomes more difficult to introduce an analog of $Y_t$ [3]. Nonetheless, it is still possible to construct a series motivated by the methodology used in the scalar case; however, one now needs to expand in space and time, whereas the Hermite expansion yielded a series "in time only" [3]). Aït-Sahalia [3] outlines an approach which makes use of a recursion for calculating the coefficients of the expansion in the multivariate case. We

have had success in using these expansions, even in cases where convergence of the infinite series is not guaranteed by the conditions given in [1, 2]. Notice, for example, that our local models may allow a value of zero for the diffusion coefficient; using a different function class (made computationally feasible by the extension of Bakshi and Ju [5, 6]), such as sigmoidal functions for it, may help circumvent such problems. Other pathologies are discussed in [11]; estimates of the range of the parameters of interest [31, 42, 11] can enhance the algorithm performance. The comparison study [25] recommends the use of the expansions by Aït-Sahalia for a wide class of diffusion models. Beyond the estimation itself, these expansions can also be helpful in obtaining diagnostics that depend on knowledge of the transition density (such as goodness-of-fit tests [22]) and asymptotic error analysis [31].

## 5 Illustrations of Equation-Free Computation

Having estimated the parameters of a local model at a given state point opens the way to several computational possibilities. Such estimates, for example, can be used in an iterative search for zeroes of the (global, nonlinear) drift. A Newton-Raphson iteration for a (hopefully better) guess of this root involves the solution of set of linear equations for which both the matrix and the residual are available from the local linear drift. The resulting estimate of the root is then used to launch a new set of computational experiments with the "inner" SSA code, followed by a new estimation, linear equation solution, and so on to convergence. This illustrates the fundamental underpinnings of equation-free computation. Many numerical algorithms (here, root finding through Newton iteration) do not really require good closed-form global models: each iteration only requires local information (the first very few terms of a Taylor series) in order to "design" the next iteration. Traditional continuum numerical methods can thus be thought of as *protocols* for the design of a sequence of model evaluations (possibly model and Jacobian evaluations, occasionally even Hessian evaluations). In the absence of an explicit formula for the model, the same protocol can be used to design *appropriately initialized computational experiments* with a model of the system at a different level (here, the SSA simulator). Processing the results of these appropriately initialized short bursts *estimate* the quantities required for scientific computation, as opposed to *evaluating* them from a closed-form model. The so-called "coarse projective integration" is another example of the same principle. Traditional *explicit* integration routines require a call to a subroutine that *evaluates* the time-derivative of a dynamic model at a particular state. In the absence of an explicit model, short bursts of simulations of a model of the system at a different level (again, here, SSA) can be used to *estimate* these time derivatives, and, through local linear models, extrapolate the state at a later time. The fundamental assumption underpinning this entire computational framework, is that an explicit evolution equation exists, and closes, in terms of

the (known) coarse-grained observables of the fine-scale simulation (here, the concentrations of the SSA species). If this, unavailable in closed-form, equation is *deterministic*, then one only need to estimate a drift term from fine scale simulations; if, on the other hand, the coarse-grained equation is *stochastic* (fluctuations are important), then both the local drift and diffusion terms must be estimated. Certain computational tasks for stochastic *effective*, coarse-grained models require evaluations of *both* these terms (e.g. computations of stationary, equilibrium densities, or Kramers' type computations of escape times for bistable systems, see for example [29, 23]). In this paper, we perform equation-free tasks for only the drift component of the model; sometimes it may be interesting to know whether the drift component dynamics possess zeros or closed loops, as well as their parametric dependence. Also, at infinite system size (practically, for sufficiently large particle numbers) the SSA actually closes as a deterministic ODE.

### Coarse Newton-Raphson for the fixed point of the drift

In what follows we work at system sizes large enough that a diffusion approximation of the SSA output is meaningful, and -even more- the dynamics of the drift component of the diffusion are close to the kinetic ODE scheme dynamics. The neutral stability of the fixed point and the closed loops of the kinetic ODE suggest comparable features for the estimated drift, which we set out to investigate. We find the nontrivial root of the estimated drift $\mathbf{F}(\mathbf{X};\theta) = \mathbf{0}^{\dagger}$ through a coarse Newton-Raphson procedure as follows: An ensemble of $N_{path}$ SSA simulations are initialized in a neighborhood of the current guess $\mathbf{X}_0$ of the root. Each is evolved in time, and the simulations are sampled uniformly $M$ times during a time interval of length $\tau$. A local SDE model of the type (5) is estimated using the transition density expansions of Aït-Sahalia in an MLE-type scheme; the resulting model parameters are used to update the root guess through

$$\mathbf{X}_n = \mathbf{X}_{n-1} - \frac{\partial \mathbf{F}(\mathbf{X};\theta)}{\partial \mathbf{X}}^{-1} |_{\mathbf{X}=\mathbf{X}_{n-1}} \mathbf{F}(\mathbf{X}_{n-1};\theta) \approx \mathbf{X}_{n-1} - \mathbf{B}^{-1}\mathbf{A}. \qquad (14)$$

Figure 1 shows this procedure for two different values of $N_{path}$ (other parameters are noted in the caption). Newton-Raphson type procedures for isolated roots are known to converge quickly given a good initial guess; furthermore, upon convergence, the eigenvalues of the linearization of the drift are contained in the matrix $\mathbf{B}$. Estimates of these eigenvalues for different $N_{path}$ are listed, upon convergence of the root finding procedure, in Table 1. The equation-free iterates approach the deterministic ODE root (see inset); the latter is known to possess two pure imaginary eigenvalues. The estimated (from local models) eigenvalues are also characterized by a relatively small ($O(10^{-2})$) real part.

[†] We use $\mathbf{F}(\cdot;\theta)$ to denote the right hand side of a general deterministic ODE; here $\mathbf{F}(\cdot;\theta)$ is the estimated $\boldsymbol{\mu}(\cdot;\theta)$

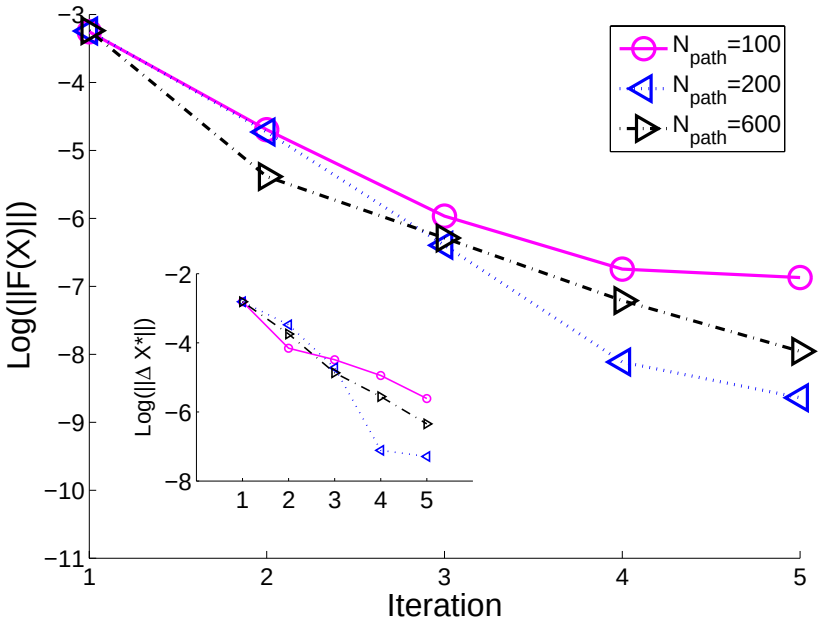



Fig. 1: Coarse NR to find stationary points. Roots of $\mathbf{F}(\mathbf{X}) \equiv \frac{\mathbf{dX}}{dt}$ using estimated (local) linear SDEs. Parameters: $N_{mol} = 1 \times 10^4$, $\tau = 5.032928126 \times 10^{-1}$, $M = 300$. $N_{path}$ values are shown in the legend and the $l^2$ distance of the current guess from the deterministic ODE root is shown in the inset. $\Delta\mathbf{X}^*$ (inset y-axis) represents the difference between the current guess and the steady state of the ODE.

Table 1: Representative real and imaginary parts of the eigenvalues of the estimated drift upon convergence to the nontrivial fixed point $\mathbf{X} \approx (0.2941, 0.41176)$; the deterministic ODE solution has a pair of pure imaginary eignenvalues.

| | Re | Im |
|---|---|---|
| $N_{path} = 100$ | $-2.53 \times 10^{-2}$ | $3.01 \times 10^{-1}$ |
| $N_{path} = 600$ | $-2.39 \times 10^{-2}$ | $2.63 \times 10^{-1}$ |

### Coarse Projective Integration for the drift

A variety of numerical integration algorithms can be implemented in our framework. Single step methods of the general form

$$\mathbf{X}_n = \mathbf{X}_{n-1} + \mathbf{\Phi}(\mathbf{X}_{n-1}, \mathbf{X}_n; \Delta t). \tag{15}$$

include the explicit and implicit Euler algorithms, (for which $\mathbf{\Phi}$ is $\Delta t\mathbf{F}(\mathbf{X}_{n-1})$ and $\Delta t\mathbf{F}(\mathbf{X}_n)$ respectively). Estimates of the drift at $\mathbf{X}_0$ can be immediately

used in a "coarse forward Euler", while the estimated $\mathbf{B}$ can be used in a root-finding procedure, along the lines illustrated above, in a "coarse backward Euler" scheme. Other schemes can be simply implemented. Here we only demonstrate (explicit) coarse forward Euler; predictor-corrector schemes (more appropriate for stiff problems) are illustrated in [10]. Representative results for our LV problem are shown in Figure 2. The deterministic ODE trajectory (dashed lines connecting points) is compared to the projective integration of the drift component of an SDE estimated locally from SSA simulation ensembles. One clearly sees the evolution of the ensemble of SSA trajectories initialized at every numerical integration point; $N_{path}$ such trajectories were evolved and observed uniformly $M$ times over a time interval $\tau$. The results were processed through the estimation scheme and the value of the drift at the original point $\mathbf{X}_0$ provided the forward Euler estimate of the "next" point through $\mathbf{X}_1 = \mathbf{X}_0 + \Delta t \mathbf{F}(\mathbf{X}_0)$. The procedure is then repeated.

Several algorithmic parameters must be carefully selected in such computations. In our case the "lifting" problem (the initialization of SSA simulations at a given value of the coarse observables) is straightforward because of the "mixed" nature of the SSA simulation; in general, the successful initialization of a fine scale code consistent with a few coarse observables can be a complicated and difficult issue, requiring, for example, preparatory constrained dynamic runs [4, 40].

Another important parameter is the length of the integrator "projective" step, $\Delta t$, which for deterministic problems is set by stability and accuracy considerations. Stability discussions for projective integration can be found in [28]; here the issue is complicated by the fact that the model is *estimated* rather than evaluated. Multiscale methods for SDEs, including error estimates, can be found in the work of [43, 16]. The total "microscopic integration time" denoted by $\tau$ and the time between observations $\equiv \delta t := \frac{\tau}{M}$ also require careful selection. If $\tau$ is too large, the simple linear model may break down as nonlinearities in the real system manifest themselves. If the assumed diffusion model is correct, there is no upper limit on $M$; yet a diffusion approximation of a different underlying process, such as the jump SSA here, will break down if the data is sampled too frequently. Similar issues have been addressed in the control literature [15]. Later on we will outline a goodness-of-fit test that can be used to guide the selection of such algorithmic parameters. In this work, short SSA trajectories in each ensemble are initialized at the same base point $\mathbf{X}_0$, or uniformly in a small neighborhood around it; we have not yet explored optimal initialization.

### Equation-Free Coarse Variational Calculations

A slight extension of the above coarse integration procedure is the implementation of equation free integration of *variational* equations. The need for these arises naturally in our example when we attempt to construct an algorithm that searches for possible closed orbits in the dynamics of the estimated drift,

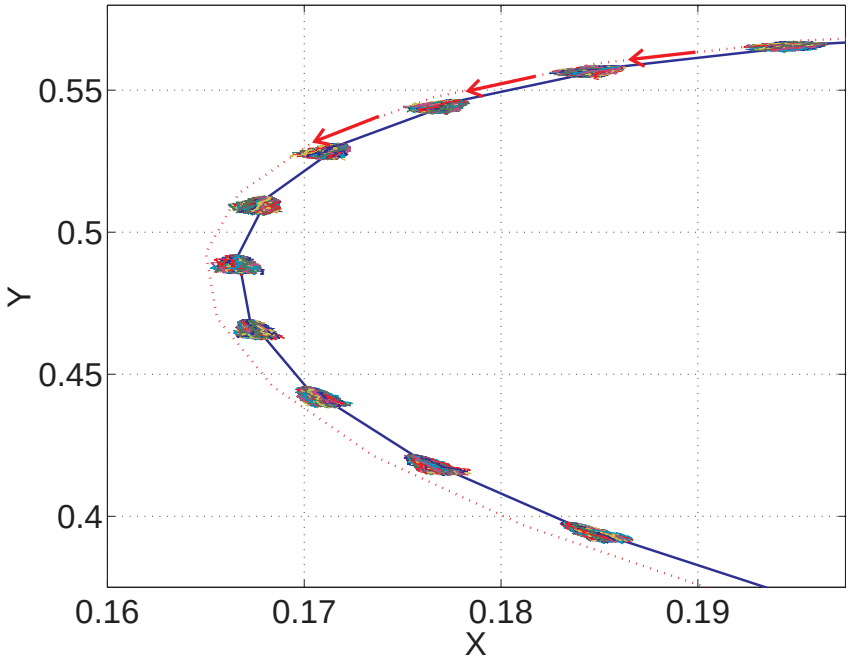



Fig. 2: Illustration of Coarse Projective Integration. $N_{path} = 600$, $N_{mol} = 1 \times 10^4$, $\Delta t = .50329$, $\tau = \frac{\Delta t}{4}$.

and attempts to converge on them. Closed orbits that are limit cycles can be found as (isolated) fixed points of an appropriate Poincaré map. In the deterministic LV problem, however, one has a one-parameter family of such orbits, and the fixed points of the Poincaré map are not isolated. Anticipating a family of such closed orbits for our estimated drift model, we *isolate* a single orbit from this one-parameter family by selecting its period (the Poincaré return time).

For a deterministic model, the initial value problem for the variational equations is

$$\begin{aligned} \frac{d\mathbf{X}}{dt} &= \mathbf{F}(\mathbf{X};\theta) \\ \mathbf{X}(t=0) &= \mathbf{X}^{\mathbf{IC}} \\ \frac{d\mathbf{V}}{dt} &= \frac{\partial \mathbf{F}(\mathbf{X};\theta)}{\partial \mathbf{X}} \cdot \mathbf{V} \\ \mathbf{V}(t=0) &= \mathbf{I}. \end{aligned} \tag{16}$$

If $\mathbf{X} \in \mathbb{R}^d$ then $\mathbf{V} \in \mathbb{R}^{d \times d}$. We use the results of integrating such variational equations to locate closed orbits as zeroes of the equation $\mathbf{G}(\mathbf{X}) \equiv \mathbf{X} - \mathbf{\Phi}_{\boldsymbol{\tau}}(\mathbf{X})$

where $\mathbf{\Phi}_\tau(\cdot)$ represents the result of integration from the (deterministic) initial condition $\mathbf{X}^{IC}$ for time $\tau$. To isolate the zeroes we seek, we select a Poincaré plane through the value $X_P = 0.3$ of the first coordinate, and the return time; we thus have one equation with one unknown, the $Y$ coordinate of the intersection of our particular closed orbit with the chosen Poincaré plane. For our coarse integration, the return time $\tau$ is typically too large to permit a single local diffusion model to accurately describe the dynamics; we therefore use the following procedure:

- Specify $\tau$ and the number $N_{grid}$ of local models we will use along the orbit, each valid for $T_f^{macro} := \frac{\tau}{N_{grid}}$.
- Simulate $N_{path}$ SSA trajectories starting at the current fixed point guess; use the data as above to estimate the first local linear model. Use its drift (and the matrix $\mathbf{B}$) to obtain the next "base point" as well as to step the variational equations for time $T_f^{macro}$.
- Repeat $N_{grid}$ times (see Figure 3).

The output of this procedure gives us the residual of the fixed point equation we wish to solve; the results of the variational integration at time $\tau$ (which, upon convergence, will give us an estimate of the monodromy matrix) are then used to compute the Jacobian of the fixed point scheme. One Newton-Raphson step for the $Y$ coordinate of the fixed point is taken, and the procedure is then repeated. Representative numerical results are shown in Figure 4. Because of the neutral dynamics, the eigenvalues of the monodromy matrix upon convergence are *both* equal to 1 (in the deterministic ODE). Table 2 shows representative eigenvalue upon convergence for different $N_{path}$ (sometimes the eigenvalues are numerically found as complex conjugates with a small imaginary part). Clearly, in addition to the algorithmic parameters involved in coarse projective integration, we should now also take into account the desired accuracy of the variational integration (quantified in part by the existence of an eigenvalue equal to unity upon convergence).

Table 2: Representative monodromy matrix eigenvalues upon convergence of the fixed point iteration for two distinct $N_{path}$ computations (see text).

| | | |
|---|---|---|
| $N_{path} = 100$ | $(0.9698, 0.1707)$ | $(0.9698, -.1707)$ |
| $N_{path} = 400$ | $(0.9159, 0.0252)$ | $(0.9159, -.0252)$ |

## 6 Discussion

We have illustrated the implementation of certain coarse-grained computations with the LV model; one focus was the coarse-grained integration of

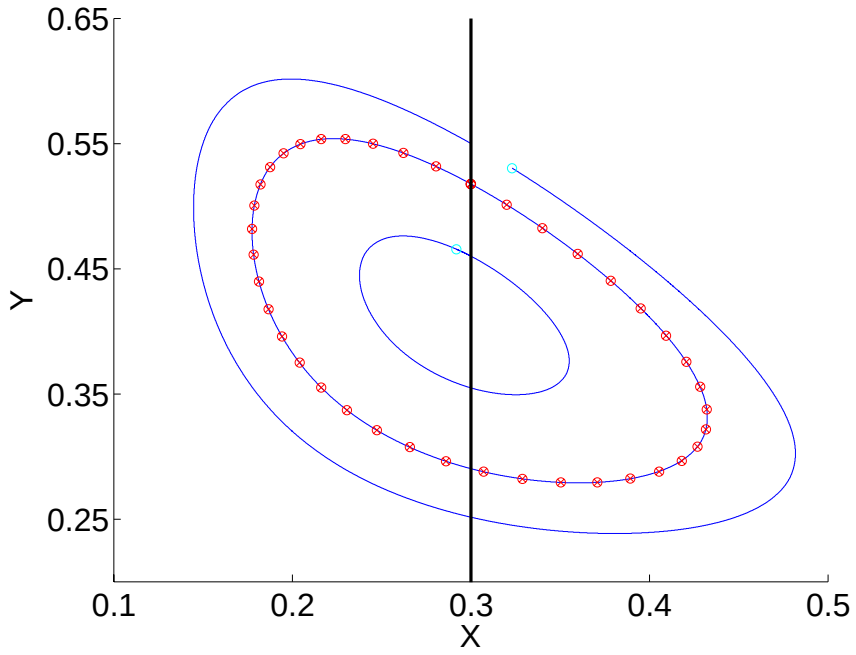



Fig. 3: Coarse closed orbit computations for the Lotka Volterra model. The deterministic model phase portrait contains an infinite number of closed orbits. Three such deterministic orbits (obtained by Runge-Kutta integration) are plotted here. To find the closed orbit with a specified period $\tau$, we use the Poincaré surface $X_P = 0.3$, shown as a solid line. The Jacobian of the coarse Newton-Raphson scheme is computed through variational integrations based on the estimated drift from ensembles of SSA simulations initialized at the $N_{grid}$ base points shown (see text).

the variational equations for the SSA-based drift estimation, as well as the modifications of the coarse Newton-Raphson iteration dictated by the neutral stability of the dynamics (the existence of infinitely many closed orbits in the ODE limit, which appears to approximately persist in our computations). The second focus was the use of Aït-Sahalia's expansions to estimate local linear SDEs from short bursts of SSA data as an intermediate step. This naturally leads to some crucial questions about the goodness-of-fit of the simple SDE models: (a) is the diffusion approximation a "good" description of the dynamics? (b) Is the linear approximation valid for the time series length chosen? and (c) How reliable is the model for making predictions/forecasts ?

One should quantitatively know how large $N_{mol}$ needs to be, for a given sampling frequency, for a diffusion model to be a statistically meaningful approximation [7, 19]. Sampling too often may be detrimental in many diffusion approximations (e.g. [15]). Local linear models (i.e. short truncations of Tay-

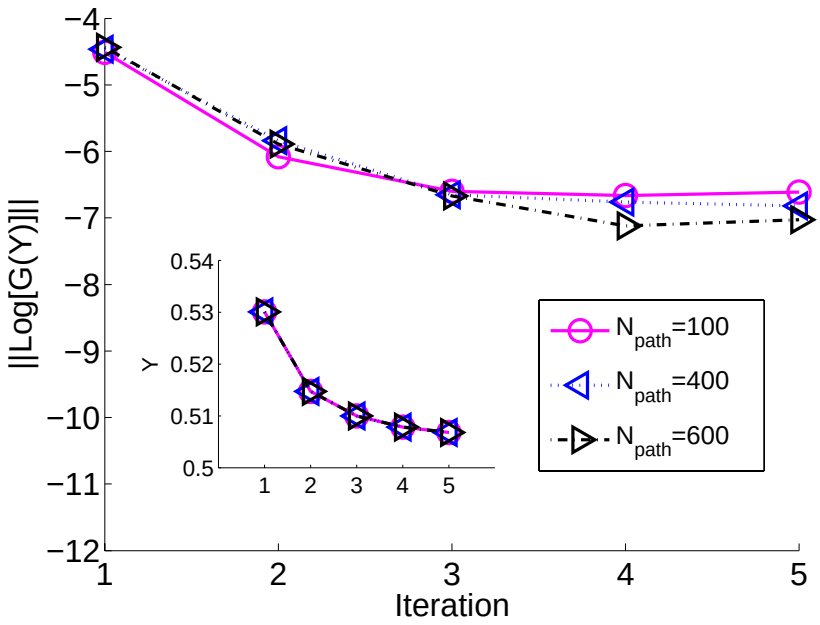



Fig. 4: Coarse Newton-Raphson for finding closed orbits of a specified period. The zeroes of $G(Y) \equiv Y - \Phi_\tau(Y)$ were calculated using a Jacobian evaluated from coarse variational integration based on SSA simulations. Parameters: $N_{mol} = 1 \times 10^4$, $\tau = 2.0131712504 \times 10^1, M = 300$, $N_{grid} = 40$ ($N_{path}$ given in the legend). The initial guess was $Y = 0.53$. For the deterministic ODE model the fixed point is $Y^{ODE} \approx 0.518$; the coarse fixed point for $N_{path} = 400$ was calculated to be $Y^{SSA} \approx 0.5075$.

lor series) are used extensively in scientific computations, but only for short time steps, whose length is determined by overall stability and accuracy considerations. Similar considerations arise in choosing the $\tau$ used for SSA data collection towards the estimation of the local linear SDE models used here; clearly, when the underlying drift is nonlinear, $\tau$ cannot be too large. A useful diagnostic tool for questions (a) and (b) applicable if one does have an accurate transition density approximation, is the probability integral transform [13, 22]. Using the data and the (assumed known) exact transition density, one creates a new random variable which, for a correctly specified model, has a known distribution. The method is applicable to both stationary and non-stationary time series; furthermore it depends on integrations of the transition density approximation rather than differentiations. Given the data, one (appropriately) estimates model parameters and then constructs the random

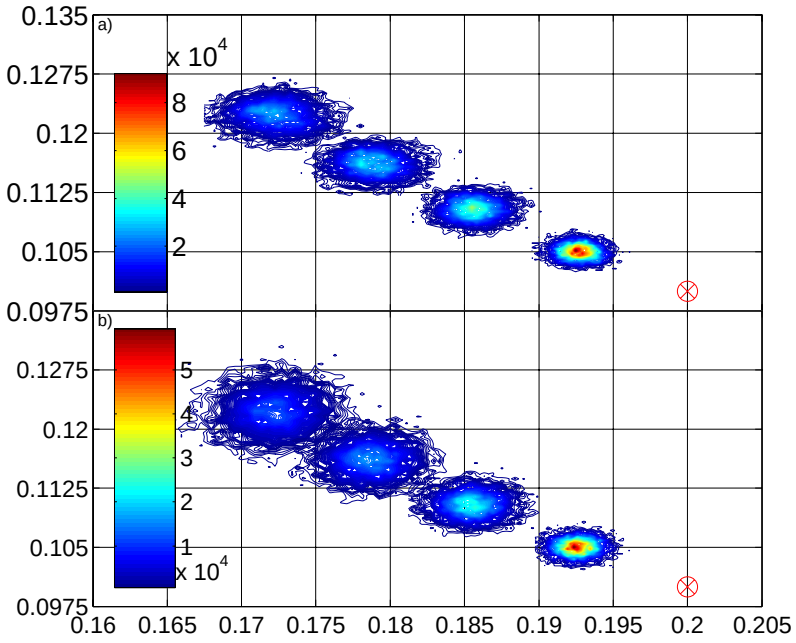



Fig. 5: Evolution of actual and model densities. The top figure shows the evolution of the SSA process, initialized as a Dirac distribution; the bottom plot shows an ensemble of numerical simulations of the ideal diffusion model using the parameters estimated from the SSA. Both distributions are plotted at $\frac{\tau}{4}, \frac{\tau}{2}, \frac{3\tau}{4}$,and $\tau$. Relevant parameters: $N_{path} = 5 \times 10^3, M = 300$, $N_{mol} = 1 \times 10^4$, $\tau = 0.50329$.

variables $Z_n$ for each observation [†] $(x_n)$. The construction below follows that in Section 3 of [13]:

$$Z_n := \int_{-\infty}^{x_n} p(x'_n|x_{n-1};\theta)dx'_n$$

$$Z_n \sim q(Z_n) \equiv \frac{d\mathbb{Q}(Z_n)}{dZ_n}$$

$$x_n \sim f(x_n|x_{n-1}) \equiv \frac{d\mathbb{F}(x_n|x_{n-1})}{dx_n}$$

[†] The method applies to both a vector and scalar process, however the construction is easiest to demonstrate in the latter case. See [22] for the multivariate extension.

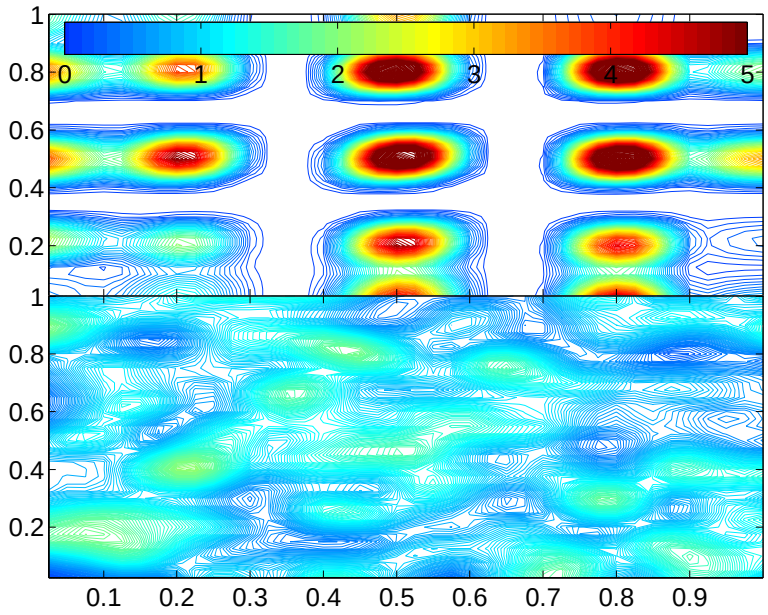



Fig. 6: Towards hypothesis testing. The function plotted corresponds to an empirical estimate of the two-dimensional density function described in [13, 22]. The data are obtained from the same ensemble of SSA simulations as in Fig 5; the top figure is for $N_{mol} = 1 \times 10^4$ and the bottom for $N_{mol} = 4 \times 10^5$. In the infinite sample limit and for a correctly specified model the density would be unity in the entire support ($[0, 1] \times [0, 1]$) of the function. The figure suggests that the observations of the larger system are closer to a diffusion model.

The symbol $\sim$ denotes that the random variable on the left of the symbol is distributed according to the density to the right. Under a correctly specified model, the $Z_n$'s are independent and uniformly distributed on $[0, 1]$, independent of the transition density [13]. In [22] a comprehensive suite of statistical tests are reviewed which exploit knowledge of the transition density and the transformation shown above. Figure 6 plots a kernel density estimate (see equation 6 on page 44 in [22]) which is based on the estimated parameters and the observed data. If the model is correctly specified, the infinite sample size density should be the product of two uniform densities. Test statistics can be created from this function (see [22] for details).

Inspection of the figures shows that, for a particular representative SSA ensemble run for $N_{mol} = 1 \times 10^4$, and a particular sampling frequency, the diffusion approximation is not acceptable; the situation appears better for $N_{mol} > 4 \times 10^5$. It is interesting to notice that, while $N_{mol} = 1 \times 10^4$ is not

large enough for the conditions of Figure 5, visual inspection of the empirical and the SDE-based density evolution might suggest otherwise. In traditional, continuum numerical algorithms issues of on-line error estimation, time-step and mesh adaptation are often built-in in modern, validated software. There is a clear necessity for incorporating, in the same spirit, hypothesis testing techniques in codes implementing the type of computations we described here; yet automating such processes appears to be a major challenge.

In our next application, we evolved an ensemble of trajectories starting from a Dirac initial distribution, and then recorded the Poincaré map for each individual trajectory over a long simulation period. Figure 7 shows the evolution of the $Y$ coordinate of these trajectories as function of the map iterate. For long times, different initial conditions in the ensemble approach some of the "extinction" fixed points of the ODE vector field (see the vertical lines in Fig. 7); once there, the system no longer changes over time. Visual inspection of the evolution of the ensemble suggests that one might try to coarse-grain the Poincaré map evolution as a model SDE; the insets in the figure show the initial evolution of the mean and the variance of the Poincaré map iterates. The smooth line in the insets, a simple least squares fit, seems to suggest a systematic evolution towards "larger" oscillations, bringing the system closer to extinction. If this evolution could be well approximated locally by a diffusion processes, approximations similar in spirit to the ones shown in this article might be used to explore features of the distribution of extinction times for the problem.

*Acknowledgement.* This work was partially supported by a Ford Foundation/NRC Fellowship to (CC) and an NSF ITR grant (IK).

## References

1. Y. Aït-Sahalia: *Closed-form likelihood expansions for multivariate diffusions*, E-print: `http://www.princeton.edu/~yacine/research.htm`, (2001)
2. Y. Aït-Sahalia: Maximum-likelihood estimation of discretely-sampled diffusions: A closed-form approximation approach. *Econometrica* **70**, 223–262 (2002)
3. Y. Aït-Sahalia, R. Kimmel: *Estimating affine multifactor term structure models using closed-form likelihood expansions*, NBER Technical Working Papers 0286, National Bureau of Economic Research, Inc, Dec. 2002. available at `http://ideas.repec.org/p/nbr/nberte/0286.html`
4. A. Amadei, A.B.M. Linssen, H.J.C. Berendsen: Essential dynamics of proteins, *Proteins* **17**, 412–425 (1993)
5. G. Bakshi, N. Ju: A refinement to Aït-Sahalia's (2002) "Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach," *Journal of Business* **78**, 2037–2052 (2005)
6. G. Bakshi, and N. Ju, H. Ou-Yang: *Estimation of continuous-time models with an applications to equity volatility dynamics* (working paper)

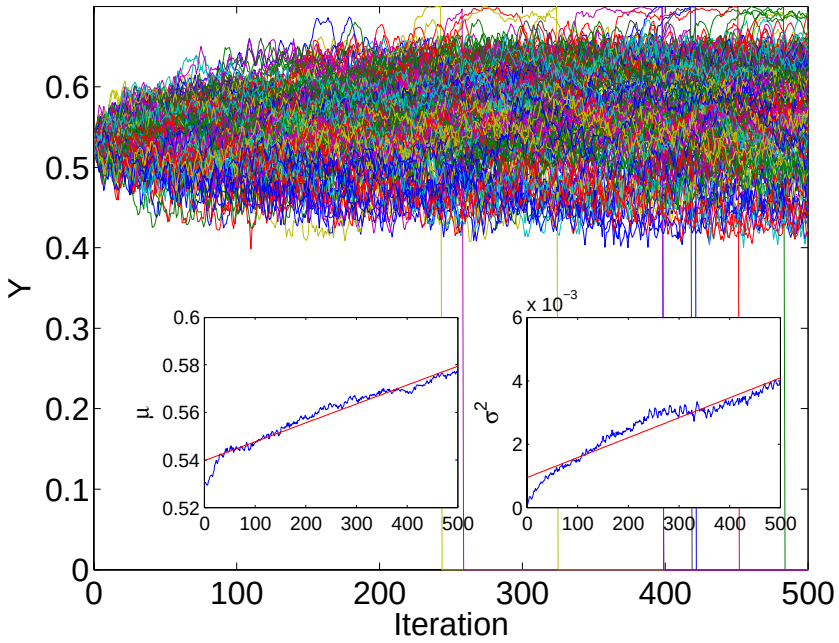



Fig. 7: SSA return map computations for the LV model. An ensemble of 200 trajectories initialized at $\mathbf{X} = (0.3, 0.53)$ are evolved, and the $Y$ coordinate of their $X_P = 0.3$ Poincaré map crossings in the negative $X$ direction is recorded. The insets (see text) suggest a systematic upward drift, bringing the system closer to extinction (this event is indicated by vertical lines, see text).


7. K. Ball, T. Kurtz, L. Popovic, G. Rempala, *Asymptotic analysis of multiscale approximations to reaction networks,* E-print: math/0508015, arXiv.org (2005)
8. M. Belkin, P. Niyogi: Laplacian eigenmaps and spectral techniques for embedding and clustering, *Advances in Neural Information Processing Systems*, vol. 4, edited by S. Becker, and Z. Ghahramani (MIT Press, Cambridge MA 2002)
9. B.M. Bibby M. Sørensen: Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, **1**, 17–39 (1995)
10. C.P. Calderon, I.G. Kevrekidis: *Estimation strategies in equation free numerical methods (in preparation)*
11. C.P. Calderon: *Fitting effective diffusion models to data associated with a "glassy potential": Estimation, classical inference procedures and some heuristics*, E-print cond-mat/0510521 on arXiv.org (submitted to SIAM MMS) (2005)
12. R.R. Coifman, S. Lafon, A.B. Lee, M. Magionni, B. Nadler, F. Warner, S.W. Zucker: Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods. *PNAS* **21**, 7432–7437 (2005)
13. F.X. Diebold, T. Gunther, A. Tay: Evaluating density forecasts with applications to financial risk management. *International Economic Review* **39**, 863–883 (1998)

14. R. Durrett: Stochastic Spatial Models. *SIAM Review* **41**, 677–718 (1999)
15. M. El-Ansary, H. Khalil: On the interplay of singular perturbations and wide-band stochastic fluctuations. *SIAM J. Control and Optimization* **24**, 83–94 (1986)
16. W. E, D. Liu, E. Vanden-Eijnden: Analysis of multiscale methods for stochastic differential equations. *Comm. on Pure and Applied Mathematics* **58**, 1544–1585 (2005)
17. M.B. Elowitz, S. Leibler: A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338 (2000)
18. A.R. Gallant, G. Tauchen, Which moments to match? *Econometric Theory* **12**, 657–681 (1996)
19. D. Gillespie: The chemical Langevin equation. *J. Chem. Phys.* **113**, 297–306 (2000)
20. D.T. Gillespie, L.R. Petzold: Improved leap-size selection for accelerated stochastic simulation. *J. Chem. Phys.* **119**, 8229–8234 (2003)
21. J.D. Hamilton: *Time Series Analysis* (Princeton University Press 1994)
22. Y. Hong, H. Li: Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. *The Review of Financial Studies* **18**, 37–84 (2005)
23. G. Hummer, I.G. Kevrekidis: Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. *J. Chem. Phys.* **118**, 10762–10773 (2003)
24. P. Jeganathan: Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* **11**, 818–887 (1995)
25. B. Jensen, R. Poulsen: Transition densities of diffusion processes: Numerical comparison of approximation techniques. *Journal of Derivatives* **9**, 18–32 (2002)
26. C. T. Kelley: *Iterative Methods for Linear and Nonlinear Equations* (SIAM, Philadelphia 1995)
27. I.G. Kevrekidis, C.W. Gear, G. Hummer: Equation-free: The computer-aided analysis of complex multiscale systems. *AIChE Jounral* **50**, 1346–1355 (2004)
28. I.G. Kevrekidis, C.W. Gear, J.M. Hyman, P.G. Kevrekidis, O. Runborg, K. Theodoropoulos: Equation-free coarse-grained multiscale computation: enabling microscopic simulators to perform system-level tasks. *Comm. Math. Sciences* **1**, 715–762 (2003)
29. D.I. Kopelevich, A.Z. Panagiotopoulos, I.G. Kevrekidis: Coarse-grained kinetic computations for rare events: Application to micelle formation. *J. Chem. Phys.* **122**, 044908–044920 (2005)
30. S. Kullback, R.A. Leibler: On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86 (1951)
31. L. Le Cam, G. L. Yang, *Asymptotics in Statistics: Some Basic Concepts* (Springer, Berlin Heidelberg New York 2000)
32. J.D. Murray: *Mathematical Biology I: An Introduction* (Springer, Berlin Heidelberg New York 2004)
33. G. Nicolis, I. Prigogine: *Self-organization in Non-equilibrium Systems* (Wiley, New York 1977)
34. G. Nicolis: *Introduction to Non-linear Science* (Cambridge University Press, Cambridge 1995)
35. A.R. Pedersen: A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian J. of Statistics*, **22**, 55–71 (1995)

36. A. Provata, G. Nicolis, F. Baras: Ocillatory Dynamics in Low Dimensional Lattices: A Lattice Lotka-Volterra Model. *J. Chem. Phys.* **110**, 8361–8368 (1999)
37. E. Schutz, N. Hartmann, Y. Kevrekidis, R. Imbihl: Microchemical engineering of catalytic reactions. *Catalysis Letters* **54**, 181–186 (1998)
38. J. Stelling, U. Sauer, Z. Szallasi, F. J. Doyle III, J. Doyle: Robustness of cellular functions. *Cell* **118**, 675–685 (2004)
39. J. B. Tenenbaum, V. De Silva, J. C. Langford: A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000)
40. G. M. Torrie, J. P. Valleau.: Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23**, 187–199 (1977)
41. J.J. Tyson, A. Csikasz-Nagy , B. Novak: The dynamics of cell cycle regulation. *BioEssays* **24**, 1095–1109 (2002)
42. A. van der Vaart: *Asymptotic Statistics* (Cambridge University Press 1998)
43. E. Vanden-Eijnden: Numerical techniques for multi-scale dynamical systems with stochastic effects. *Comm. Math. Sciences* **1**, 385–391 (2003)
44. H. White: Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25 (1982)

# Relations Between Information Theory, Robustness and Statistical Mechanics of Stochastic Uncertain Systems via Large Deviation Theory

C. D. Charalambous[1], A. Kyprianou[2] and F. Rezaei[3]

[1] Department of Electrical and Computer Engineering, University of Cyprus, 75 Kallipoleos Str, Nicosia, Cyprus, chadcha@ucy.ac.cy
[2] Department of Mechanical and Manufacturing Engineering, University of Cyprus, 75 Kallipoleos Str, Nicosia, Cyprus, akyp@ucy.ac.cy
[3] School of Information Technology and Engineering, University of Ottawa, 800 King Edward Ave., Ottawa, frezaei@site.uottawa.ca

**Summary.** Robustness issues of stochastic uncertain controlled systems are intimately related to concepts from information theory and statistical mechanics. These relations are unveiled by using the theory of large deviations through the solution of two fundamental optimization problems. In one of the optimization problems the aim is to minimize the relative entropy between a nominal measure induced by a reference stochastic system and the measures induced by the uncertain stochastic systems subject to energy constraints. In the other the supremum of an energy functional with respect to the measures induced by the nominal and uncertain systems, subjected to constraints on the relative entropy between the measures induced by the nominal and uncertain stochastic systems is sought. The solutions of these problems, by virtue of their formulation, satisfy the $H^\infty$ robustness criterion and through statistical mechanics arguments the associated optimal sensitivity can be characterized.

## 1 Introduction

This chapter analyses controlled stochastic uncertain systems in the framework of large deviations. Since the theory of large deviations is closely related to information theory and statistical mechanics a further analysis is carried out to demonstrate these connections in the context of stochastic uncertain systems. In the introduction, for completeness a brief overview of the important concepts in statistical mechanics, $H^\infty$ theory of robustness, information theory and large deviations is given. Section 2 is written in order to emphasize the close relationship between the $2^{nd}$ law of Thermodynamics and its pertaining Clausius inequality, and the Willems dissipation inequality [1],

and then to demonstrate the link between statistical mechanics and classical thermodynamics. These relations are established in Sections 4 and 5. In anticipation of what follows, Section 2 casts statistical mechanics in a variational formulation.

Section 3 gives the duality relationships on which Sections 4 and 5 prominently rely on, in order to first analyse robustness of stochastic uncertain control systems, and then characterise their optimal solutions. Finally in Section 6, as an example, the large deviation principle is applied to diffusion processes and the connection to the thermodynamic entropy is highlighted.

Statistical mechanics provides the mathematical framework, which deals with uncertainty of systems composed of large population of particles (or molecules). The uncertainty that statistical mechanics deals with emanates from the variability of the microstates that correspond to a particular macroscopic property. Specifically, for a system consisting of a population of $\mathcal{N}$ particles, the microstate is represented by the vector of positions and momenta of the particles and its thermodynamic variables the macroscopic properties. The key ingredient of statistical mechanics, which also provides the link to classical thermodynamics is the partition function. Once the partition function, a mathematical object that counts all the microstates compatible with a given macrostate, is computed all the thermodynamic properties of a system can be derived [2].

The basic mathematical notion associated with robust control is the $H^\infty$-norm of the system, a measure of how much of the exogenous disturbance signals are attenuated by the controller [3]. In order to elucidate this on physical grounds lets denote the disturbance signal that enters an uncertain plant G, Figure 1, by $w$ and let it be an element of an $L_2$ space (the space of square integrable Lebeasgue measurable functions). Then the controller's K, Figure 1, action affects the operator $T_K$ that maps $w \in L_2(\mathcal{W})$, the Hilbert space of the disturbance signals, to the plant output $z$ which belongs to $L_2(\mathcal{Z})$, the Hilbert space of the output signals. The norm of the system is the induced norm between the Hilbert spaces of this operator given by,

$$\|T_K\|_\infty = \|T_K\| = \sup_{\|w\|_{L_2(\mathcal{W})} \neq 0} \frac{\|z\|^2_{L_2(\mathcal{Z})}}{\|w\|^2_{L_2(\mathcal{W})}} = \sup_{\|w\|_{L_2(\mathcal{W})} \leq 1} \|z\|^2_{L_2(\mathcal{Z})} \tag{1}$$

such that $T_K : \mathcal{W} \to \mathcal{Z}$.

Ensuring, for a given controller $K$ that

$$\|T_K\|_\infty \leq \gamma \tag{2}$$

and interpreting the norm $\|w\|$ as the energy of the signal generated by the family of disturbances having finite energy, then the controller $K$ attenuates the family of disturbances by a level $\gamma$, known in $H^\infty$ parlance as the sensitivity level. Combination of equations (1) and (2) yield,

$$\|z\|_{L_2(\mathcal{Z})} \leq \gamma \|w\|_{L_2(\mathcal{W})}, \forall w \in L_2(\mathcal{W}) \tag{3}$$

which is equivalent to,

$$\sup_{w \in L_2(\mathcal{W})} \left( \|z\|_{L_2(\mathcal{Z})} - \gamma \|w\|_{L_2(\mathcal{W})} \right) \leq 0 \tag{4}$$

The controller should always seek to dissipate the transmission of energy from $w$ to $z$ measured by the norm $\|z\|_{L_2(\mathcal{Z})}$, over the unit ball uncertainty by an amount determined by $\gamma$. The controller should minimize the maximum dissipation leading to the following game formulation,

$$\inf_{K} \sup_{w \in L_2(\mathcal{W})} \left( \|z\|_{L_2(\mathcal{Z})} - \gamma \|w\|_{L_2(\mathcal{W})} \right) \leq 0 \tag{5}$$

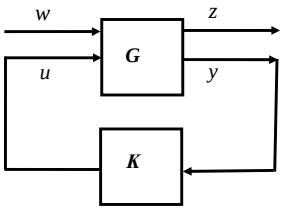



Fig. 1: Block Diagram Representation of Uncertain Systems

The link between robustness as discussed above and statistical mechanics is established in Section 5.

Information theory is the mathematical framework in which problems in science and engineering are formulated, and ultimately their solution are sought, such as of data transmission over noisy(less) channels and data compression [4]. However, its fundamental notions of entropy and relative entropy have found applications in areas such as statistics [5], statistical mechanics [6, 7], and computational complexity [8]. The entropy of the random experiment of drawing letters from a finite alphabet $A = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ of $n$ letters with probability mass function $\mathbf{p} = (p_1, \ldots p_n)$, where $p_i$ is the probability of drawing $\alpha_i$, is given by the functional,

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln p_i \tag{6}$$

To define relative entropy or Kullback-Leibler distance, assume a second similar random experiment, but with the drawing action governed by a different probability mass function $\mathbf{q}$. Then the relative entropy is defined between the two probability mass functions $\mathbf{p}$ and $\mathbf{q}$ as the functional,

$$H(\mathbf{p}|\mathbf{q}) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} \tag{7}$$

The entropy is interpreted as the amount of uncertainty of the single random experiment whereas the relative entropy characterizes the discrepancy between the two probability mass functions that govern the random action of drawing in the two experiments. The link between relative entropy optimization and robustness is established in Section 4.

An area of probability in which the concepts of entropy and relative entropy stand prominently is that of the large deviations. This theory qualifies probabilistic events as rare and, drawing from both the theories of Statistical Probability and Mathematical Analysis [9], describes the notion of typical and atypical events [10] mathematically. The whole mathematical edifice of large deviations stems from a definition that, through an action functional puts limiting bounds on the behavior of a family of probability measures indexed by either a real valued parameter $\epsilon \in \Re$ or an integer valued parameter $n$ as they tend to their limit. The milestone theorem of Large Deviations theory is the Cramer theorem. It characterizes the limiting behavior of the empirical mean $S_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_j$ of independent and identically distributed random variables distributed according to the probability measure $P$ and taking values in $\Re^d$ as follows. The empirical mean $S_n$ is a random variable of probability law $P_n$ and the cumulant generating function is defined as

$$\Lambda(\lambda) \equiv \log E\left(e^{\langle \lambda, X_i \rangle}\right) \tag{8}$$

where the expectation is taken with respect to the probability measure $P$ and $\langle ., . \rangle$ denotes the inner product in $\Re^d$. If the mean value $\overline{x}$ exists, then $S_n$ converges to $\overline{x}$ in probability as $n$ goes to $\infty$. Otherwise stated, for any closed set $F$ of $R^d$ that does not contain $\overline{x}$ then $P_n(F) \to 0$ as $n \to \infty$. The Legendre-Fenchel transform of $\Lambda(\lambda)$ is defined as,

$$\Lambda^*(x) \equiv \sup_{\lambda \in \Re^d} (\langle \lambda, x \rangle - \Lambda(\lambda)) \tag{9}$$

The Cramer's theorem states that the logarithmic rate of this convergence is bounded by an action functional given by $\Lambda^*(.)$, as follows:

- for any closed set $F$ of $R^d$

$$\limsup_{n \to \infty} \frac{1}{n} \log P_n(F) \leq - \inf_{x \in F} \Lambda^*(x) \tag{10}$$

- for any open set $G$ of $R^d$

$$\liminf_{n \to \infty} \frac{1}{n} \log P_n(G) \geq - \inf_{x \in G} \Lambda^*(x) \tag{11}$$

Sanov's theorem extends Cramer's theorem to empirical measures induced by finite sequences $\mathbf{y} = (y_1, \ldots, y_k)$ with each $y_i$ assuming values in the alphabet $A = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ with occurence probability of each letter being $p_i$. For each finite random sequence $\mathbf{y}$, the empirical measure is $L_k^{\mathbf{y}} = \frac{1}{k}(\#\alpha_1 \in \mathbf{y}, \ldots, \#\alpha_n \in \mathbf{y})$, where each $\#a_i$ counts the occurences of $a_i$ in a sequence $\mathbf{y}$. In this case each empirical measure may be identified with points of $\Re^n$ and satifies the large deviation bounds as $k$ tends to $\infty$ with bounds imposed by $\Lambda^*$ which for this case $\Lambda^*(x) = H(x|\mu)$. Sanov's theorem then in an obvious way provides the link between large deviations, statistical mechanics and information theory [11, 12]. The link between Legendre-Frenchel transform and robustness is established in Section 4.

## 2 Thermodynamics and Statistical Mechanics

### 2.1 Thermodynamics: an Overview and its Link to Statistical Mechanics

Classical equilibrium thermodynamics are to a great extend founded on two laws. The law of energy conservation, $1^{st}$ law of thermodynamics, which for general stationary systems is given by

$$Q - W = \Delta U \tag{12}$$

where $Q$ is the net heat transfer to the system, $W$ the net work done on or by the system and $\Delta U$ the change of its internal energy $U$ due to the net energy transfer to or from the system as work and heat. The second law of thermodynamics codifies the degradation of the energy quality, by irreversible processes, of a thermodynamic system which is not in equilibrium with its surrounding. It can be formulated as

$$\mathrm{d}S \leq \frac{\mathrm{d}Q}{T} \tag{13}$$

with equality being valid for reversible processes. In equation (13), S is the macroscopic thermodynamic property of entropy and $dQ$ the heat transfer from a system at temperature T. For a reversible process the first law may be recast as

$$T\Delta S - W = \Delta U \tag{14}$$

and solving for $W_{rev}$, the reversible work is given by,

$$W_{rev} = T\Delta S - \Delta U \tag{15}$$

and it is the maximum work that can be extracted from the process, otherwise known as Helmhotz free energy, $F$.

The Clausius inequality (13) can be viewed as a dissipation inequality for an irreversible proces, say from state 1 to state 2 that produces the entropy,

$$S_{gen} = S_2 - S_1 - \int_1^2 \frac{\mathrm{d}Q}{T} = \Delta S - \frac{\Delta Q}{T} \tag{16}$$

Solving for $\Delta Q$ and substituting in (12) the actual irreversible work is given by

$$W_{irr} = -\Delta U + T\Delta S - TS_{gen} \tag{17}$$

The difference between the reversible and irreversible work

$$A = W_{rev} - W_{irr} = TS_{gen} \tag{18}$$

is a measure of the irreversibility, $A$, or dissipation of the useful work and is given by the product of entropy generated and the temperature T. In order to make the connection of entropy and dissipation as described here put emphasis to the fact that the existence of the dissipation inequality implies entropy production by an irreversible process. A further important fact is when a thermodynamic system is disturbed out of its equilibrium an irreversible process comes into action to bring it back to the equilibrium maximum entropy state, by generating entropy.

Boltzman linked the microscopic properties of the particles that composed a thermodynamic system with their associate macroscopic properties by his celebrated equation,

$$S = k \ln(\Omega) \tag{19}$$

where S is the entropy, macroscopic property, and $\Omega$ the number of microstates that are compatible with the state. Starting from this equation and using combinatorial arguments that are taking into consideration the various possible states that the system can access, for brevity assume a finite number $n$ of such states, it can be shown that the entropy is given by

$$S = -k \sum_i^n p_i^* \ln p_i^* \tag{20}$$

where $k = 1.3806503 \times 10^{-23} \frac{J}{K}$ is the Boltzmann constant and $p_i^*$ is the probability of finding the system in state $i$. For a thermodynamic system in contact with a heat bath of temperature T, Gibbs computed $p_i^*$ to be,

$$p_i^* = \frac{e^{-E_i/kT}}{Q};\ Q = \sum_{j=1}^n e^{-E_j/kT} \tag{21}$$

where $Q$ is the partition function of the system. Substituting (21) in (20) the Helmhotz free energy $F$, a macroscopic thermodynamic quantity can be expressed in terms of the partition function $Q$ as,

$$F = -kT \ln Q = -kT \log \sum_{j=1}^N e^{-E_j/kT} \tag{22}$$

so this is $F = U - TS$, when $p = p^*$ and hence,

$$F = \sum_{i=1}^{n} E_i p_i^* + kT \sum_{i=1}^{n} p_i^* \ln p_i^* = U(p^*) - kTS(p^*) \tag{23}$$

The link between statistical mechanics and thermodynamics can be established by substituting the differential form of the first law of thermodynamics, (here we assume a mechanical work of the form $W = PV$, where P is the pressure and V the volume),

$$\mathrm{d}U = T\mathrm{d}S - P\mathrm{d}V \tag{24}$$

in the differential form of the Helmhotz free energy

$$\mathrm{d}F = \mathrm{d}U - T\mathrm{d}S - S\mathrm{d}T \tag{25}$$

to get

$$\mathrm{d}F = -S\mathrm{d}T - P\mathrm{d}V \tag{26}$$

Identifying the coefficients of (26) with the corresponding partial differentials,

$$S = -\left(\frac{\partial F}{\partial T}\right)_V = k \ln Q + kT \left(\frac{\partial \ln Q}{\partial T}\right)_V$$

$$P = -\left(\frac{\partial F}{\partial V}\right)_T = kT \left(\frac{\partial \ln Q}{\partial T}\right)_T \tag{27}$$

Therefore starting from the partition function, a mathematical structure that counts the number of possible microstates of a system the various thermodynamic quantities can be obtained [13]. The link between free energy and storage function is obtained in Section 6 via Large Deviations.

### 2.2 Variational Interpretation of Statistical Mechanics

The following theorem gives a variational interpretation of the basic principle of Statistical Mechanics which characterises the thermodynamical equilibrium states as those of maximum entropy [14]. This principle will be used in the subsequent sections to relate the statistical mechanics concepts to robustness concepts.

**Theorem 1.** *Let $\Sigma$ a non-empty denumerable set endowed with the discrete topology and $\mathcal{M}(\Sigma) = \left\{\pi = (\pi_1, \ldots, \pi_N), \pi_j \geq 0, \sum_{j=1}^{N} \pi_j = 1, 1 \leq j \leq N\right\}$. 1) For every measurable function $E_j : \Sigma \to \Re, 1 \leq j \leq N$, and a fixed probability vector $\mu \in \mathcal{M}(\Sigma)$*

$$\log\Big(\sum_{j=1}^{N} e^{\frac{-E_j(x)}{kT}} \mu_j(x)\Big)^{kT} =$$

$$\sup_{\nu\in\mathcal{M}(\Sigma)}\Big\{-\sum_{j=1}^{N}E_j(x)\nu_j(x)-kT\sum_{j=1}^{N}\frac{\nu_j(x)}{\mu_j(x)}\log\frac{\nu_j(x)}{\mu_j(x)}\Big\} \tag{28}$$

*Moreover, the supremum is attained at*

$$\nu_n^*(x)=\frac{e^{-E_n(x)/kT}\mu_n(x)}{\sum_{j=1}^{N}e^{-E_j(x)/kT}\mu_j(x)},\ 1\le n\le N \tag{29}$$

*2) For every measurable function $E_j:\Sigma\to\Re, 1\le j\le N$*

$$\log\Big(\sum_{j=1}^{N}e^{\frac{-E_j(x)}{kT}}\Big)^{kT}=$$

$$\sup_{\nu\in\mathcal{M}(\Sigma)}\Big\{-\sum_{j=1}^{N}E_j(x)\nu_j(x)-kT\sum_{j=1}^{N}\nu_j(x)\log\nu_j(x)\Big\} \tag{30}$$

*Moreover, the supremum is attained at*

$$\nu_n^*(x)=\frac{e^{-E_n(x)/kT}}{\sum_{j=1}^{N}e^{-E_j(x)/kT}},\ 1\le n\le N \tag{31}$$

*3) The basic principle of Statistical Mechanics, (30), (31) are the dual equations associated with the primal problem of maximizing the Entropy subject to an average energy constraint, defined by*

$$\sup_{\nu\in\mathcal{M}(\Sigma)}\Big\{-k\sum_{j=1}^{N}\nu_j(x)\log\nu_j(x)\Big\};\ \textit{subject to}\ \sum_{j=1}^{N}E_j(x)\nu_j(x)\le\gamma,\gamma\in\Re \tag{32}$$

**Proof.** 1) Follows from Theorem 2. The equivalence between 2) and 3) follows from standard primal-dual arguments of convex optimization.

Therefore, the optimal measure $\nu^*$ can be characterised as the measure closest to the nominal measure (uniform measure in this case) which satisfies the constraint on the expected values of the random variables $E_j$.

## 3 Robustness of Stochastic Uncertain Systems: General Setting

In this section, an abstract formulation of robustness of uncertain stochastic systems is introduced, which is related to the computation of induced norm. Finally, it is shown that this type of robustness problems are equivalent to the variational Theorem 1, thus establishing a connection between robustness and statistical mechanics.

### 3.1 Duality Relations

The quantities introduced in the next definition are employed extensively in subsequent sections to characterize the optimal strategies and to establish certain relations implied by the problem formulation. Applications to fully and partially observed stochastic control systems are found in [15, 16].

**Definition 1.** *Let $\nu, \mu \in \mathcal{M}(\Sigma)$ and $\ell : \Sigma \to \Re$ a measurable function.*
*1) The moment generating function of $\ell$ with respect to $\mu$ is defined by*

$$M_\mu(s) \equiv E_\mu\Big(e^{s\ell}\Big) = \int_\Sigma e^{s\ell}\, \mathrm{d}\mu \in (0, \infty], \;\; s \in \Re \tag{33}$$

*2) The cumulant generating function of $\ell$ with respect to $\mu$ is defined by*

$$\Psi_\mu(s) \equiv \log M_\mu(s) = \log \int_\Sigma e^{s\ell}\, \mathrm{d}\mu \in (-\infty, \infty], \;\; s \in \Re \tag{34}$$

*3) The Legendre-Fenchel Transform of $\Psi_\mu(s)$ is defined by*

$$\Psi^\star_\mu(x) \equiv \sup_{s \in \Re} \Big\{ sx - \Psi_\mu(s) \Big\}, \;\; x \in \Re \tag{35}$$

*$\Psi^\star_\mu(x)$ is also called the Entropy Rate Functional of $\ell$.*
*4) The relative entropy of $\nu$ with respect to $\mu$ is defined by*

$$H(\nu|\mu) \equiv \begin{cases} \int_\Sigma \log(\frac{\mathrm{d}\nu}{\mathrm{d}\mu})\, \mathrm{d}\nu \text{ if } \nu << \mu \text{ and } \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L_1(\nu) \\ \\ +\infty \qquad\qquad\qquad \text{otherwise} \end{cases} \tag{36}$$

It can be shown that $\Psi_\mu(s)$ as a function of $\ell$ is convex, $H(\nu|\mu)$ as a function of $\mu, \nu \in \mathcal{M}(\Sigma)$ is convex in both arguments, $H(\nu|\mu) \geq 0$, and $H(\nu|\mu) = 0$, if and only if $\mu = \nu$. Thus, $H(\nu|\mu)$ is a measure of discrepancy between the two measures. Moreover, $M_\mu(s), \Psi_\mu(s)$ are convex functions of $s \in \Re$. The moment generating function (33) and cumulant generating function (34), when employed in the context of stochastic control and filtering, represent the so-called risk-sensitive pay-off [17, 18, 19, 20]. For linear quadratic problems it has been long observed [17, 20] that the solution of risk-sensitive problems is equivalent to the solution of the minimax game formulation of the disturbance attenuation problem. Similarly, connections between risk-sensitive pay-off functionals and deterministic and stochastic minimax games with square integrable disturbances are also established in [21, 22, 23, 24]. As it has been discussed in the introduction, the Legendre-Fenchel Transform (35) is employed in Large Deviations Theory to identify the entropy rate functional, which describes the exponential rate of convergence to zero of rare events.
In addition, the cumulant generating function and the relative entropy are in duality with respect to a Legendre-Fenchel transform, and the following result is a variant of a theorem found in [24].

**Theorem 2.** *For a given $s \in \Re$, and $\ell : \Sigma \to \Re$ a measurable function such that $s\ell$ is bounded from below*

$$-\Psi_\mu(s) = -\log E_\mu(e^{s\ell}) = \inf_{\nu \in \mathcal{M}(\Sigma)} \Big\{ H(\nu|\mu) - \int_\Sigma s\ell \, \mathrm{d}\nu \Big\} \tag{37}$$

*Moreover, if $\ell e^{s\ell} \in L_1(\mu)$, then the infimum in (37) is attained by the tilted probability measure $\nu^*$ given by*

$$d\nu^* = \frac{e^{s\ell} \mathrm{d}\mu}{\int_\Sigma e^{s\ell} \, \mathrm{d}\mu} \tag{38}$$

**Proof.** The proof is given in [24].

### 3.2 Abstract Formulation

Let $(\Sigma, d)$ denote a complete separable metric space, and $(\Sigma, \mathcal{B}(\Sigma))$ the corresponding measurable space in which $\mathcal{B}(\Sigma)$ are identified as the Borel sets generated by open sets in $\Sigma$. Let $\mathcal{M}(\Sigma)$ denote the set of probability measures on $(\Sigma, \mathcal{B}(\Sigma))$, $\mathcal{U}_{ad}$ the set of admissible controls, and $BC(\Sigma)$ the set of continuous bounded real-valued functions, $\ell^u : \Sigma \to \Re$ for a given $u \in \mathcal{U}_{ad}$.
Here, $\mathcal{M}(\Sigma)$ denotes the set of all possible measures induced by the stochastic systems, while $\ell^u \in BC(\Sigma)$ denotes the energy function or fidelity criterion associated with a given choice of the control law $u \in \mathcal{U}_{ad}$.
It is clear that there is one to one relation between Theorem 1 and Theorem 2. In fact, Theorem 1 is a special case of Theorem 2, simply let $s \to kT, \ell^u \to -E, \mu^u \to \sum_{j=1}^N \delta(x-j)$. Therefore, any problem which is related to Theorem 2, it is also related to the Statistical Mechanics variational equations.

## 4 Robustness of Stochastic Uncertain Systems: an Energy Constraint Formulation

In this section, the optimization problem described in the introduction in the context of $H^\infty$ control is introduced.

### 4.1 Problem Statement

**Definition 2.** *Let $u \in \mathcal{U}_{ad}$, let $\ell^u \in BC(\Sigma)$, and $\mu^u \in \mathcal{M}(\Sigma)$ which is a fixed nominal measure, and $m \equiv E_{\mu^u} = \int_\Sigma \ell^u \, \mathrm{d}\mu^u$, $\gamma \in \Re$.*
*1) Find $\nu^{u,*} \in \mathcal{M}(\Sigma)$ which solves*

$$J_o(u, \nu^{u,*}) = \inf_{\{\nu^u \in \mathcal{M}(\Sigma); \int_\Sigma \ell^u \, \mathrm{d}\nu^u \le \gamma \ \}} H(\nu^u|\mu^u) \tag{39}$$

*when*
$m \equiv E_{\mu^u}(\ell^u) = \int_\Sigma \ell^u \, \mathrm{d}\mu^u > \gamma$;
*2) Find* $\nu^{u,*} \in \mathcal{M}(\Sigma)$ *which solves*

$$J_p(u, \nu^{u,*}) = \inf_{\{\nu^u \in \mathcal{M}(\Sigma); \int_\Sigma \ell^u \, \mathrm{d}\nu^u \geq \gamma\}} H(\nu^u | \mu^u) \tag{40}$$

*when*
$m \equiv E_{\mu^u}(\ell^u) = \int_\Sigma \ell^u \, \mathrm{d}\mu^u < \gamma$;

*Remark 1.* The fidelity constraints $E_{\nu^u}(\ell^u) \leq \gamma$, $E_{\nu^u}(\ell^u) \geq \gamma$ represent average energy constraints with respect to the unknown measure $\nu^u \in \mathcal{M}(\Sigma)$, such as integral quadratic constraints, tracking errors, etc., while $\gamma$ is a parameter which is in some relation with $m \equiv E_{\mu^u}(\ell^u)$, that is, either $m > \gamma$ or $m < \gamma$. In particular, as shown in subsequent sections, the case (39), with $m > \gamma$ will correspond to the optimistic scenario (emphasizing the best cases) in which the strategies are risk-seeking, while the case (40), with $m < \gamma$ will correspond to the pessimistic scenario (emphasizing the worst cases) in which the strategies are risk-averse.

The constrained problems of Definition 2 can be converted into unconstrained problems by introducing the Lagrangian and the dual functionals. To do so for every $s \in \Re$, define the Lagrangian

$$J^{s,\gamma}(u, \nu^u) \equiv H(\nu^u | \mu^u) - s\Big(E_{\nu^u}(\ell^u) - \gamma\Big) \tag{41}$$

and its associated dual functional

$$J^{s,\gamma}(u, \nu^{u,*}) = \inf_{\nu^u \in \mathcal{M}(\Sigma)} J^s(u, \nu^u) \tag{42}$$

In addition, define the quantity

$$\varphi^{s^*}(u, \gamma) \equiv \sup_{s \in \Re} J^{s,\gamma}(u, \nu^{u,*}) \tag{43}$$

The above problems have various implications in minimax games, some of which are described below.

### 4.2 Related Problems

**Disturbance Attenuation in Robustness.** For a given $u \in \mathcal{U}_{ad}$ let $L_2(\nu^u; \mathcal{H}) \equiv \Big\{\phi^u : \Sigma \to \mathcal{H}; \phi^u$ is a random variable such that $\int_\Sigma \|\phi\|^2_{\mathcal{H}} \, \mathrm{d}\nu^u < \infty\Big\}$ denotes the Hilbert space of random variables. Let $L_2(\nu^u; \mathcal{Z})$ and $L_2(\nu^u; \mathcal{D})$ denote the Hilbert spaces of tracking signals and disturbance signals, respectively. For a given $u \in \mathcal{U}_{ad}$, let $T^u : \mathcal{D} \to \mathcal{Z}$ be a bounded linear operator with induced norm defined by

$$J(u) \equiv ||T^u|| = \sup_{||d||_{L_2(\nu^u;\mathcal{D})} \neq 0} \frac{||z||^2_{L_2(\nu^u;\mathcal{Z})}}{||d||^2_{L_2(\nu^u;\mathcal{D})}},\ z = T^u d \tag{44}$$

The sub-optimal disturbance attenuation is to ensure that for all $u \in \mathcal{U}_{ad}$ that $J(u) \leq \frac{1}{s}, s > 0$, which is equivalent to

$$\begin{aligned} J^s(u) &= \sup_{d \in L_2(\nu^u;\mathcal{D})} \Big\{ s \int ||z||^2_{\mathcal{Z}} \, \mathrm{d}\nu^u - \frac{1}{2} \int ||d||^2_{\mathcal{D}} \, \mathrm{d}\nu^u \Big\} \\ &= - \inf_{d \in L_2(\nu^u;\mathcal{D})} \Big\{ \int ||d||^2_{\mathcal{D}} \, \mathrm{d}\nu^u - s \int ||z||^2_{\mathcal{Z}} \, \mathrm{d}\nu^u \Big\} \end{aligned} \tag{45}$$

and ensuring that the pay-off is non-positive.
When $\nu^u$ is absolutely continuous with respect to $\mu^u$, then it can be shown (see [25]) that $H(\nu^u|\mu^u) = \frac{1}{2} \int ||d||^2_{\mathcal{D}} \mathrm{d}\, \nu^u$. Therefore, the dual functional associated with converting the primal problem (40) into the equivalent unconstrained optimization

$$J^{s,\gamma}(u, \nu^{u,*}) = \inf_{\nu^u \in \mathcal{M}(\Sigma)} \Big\{ H(\nu^u|\mu^u) - s\Big( E_{\nu^u}\Big(\ell^u\Big) - \gamma \Big) \Big\} \tag{46}$$

is equivalent to the sub-optimal disturbance attenuation problem (45) (let $\ell^u = ||z||^2_{\mathcal{Z}}$). Moreover, larger values of $s$ imply higher attenuation and hence higher dissipation. An application of the above results to general nonlinear partially observable systems is discussed in [25].

**Legendre-Fenchel or Cramer Transform.** In the context of large deviations, the dual functionals associated with converting the primal problems (39), (40) into equivalent unconstrained optimization problems are equal to the Legendre-Fenchel or Cramer transforms of $\ell^u$ defined by

$$\begin{aligned} I(\gamma) &\equiv \sup_{s \in \Re} \Big\{ s\gamma - \log E_{\mu^u} \Big\{ e^{s\ell^u} \Big\} \Big\} \\ &= \sup_{s \in \Re} \inf_{\nu^u \in \mathcal{M}(\Sigma)} \Big\{ H(\nu^u|\mu^u) - s\Big( E_{\nu^u}\Big(\ell^u\Big) - \gamma \Big) \Big\} \end{aligned} \tag{47}$$

The Legendre-Fenchel or Cramer transform of $\ell^u$ is employed in Large Deviations Theory to identify the entropy rate functional $I(\gamma)$ associated with rare events.

**Optimistic Versus Pessimistic Optimization.** In the context of robust disturbance attenuation of uncertain systems, the measure $\mu^u \in \mathcal{M}(\Sigma)$ corresponds to the nominal measure, $\nu^u \in \mathcal{M}(\Sigma)$ corresponds to the uncertain measure, and the fidelity constraints $E_{\nu^u}\Big\{\ell^u\Big\} \leq \gamma$ and $E_{\nu^u}\Big\{\ell^u\Big\} \geq \gamma$ represent average energy constraints. It can be shown that [26]:

1. for (39) the average energy constraint with respect to all uncertain measures $\nu^u << \mu^u$, $E_{\nu^u}\Big\{\ell^u\Big\} \leq \gamma$, is below the average energy of the nominal model, $E_{\mu^u}\Big\{\ell^u\Big\} > \gamma$; hence it represents an optimistic scenario.

2. for (40) the average energy constraint with respect to all uncertain measures $\nu^u << \mu^u$, $E_{\nu^u}\left\{\ell^u\right\} \geq \gamma$ is above the average energy of the nominal model $E_{\mu^u}\left\{\ell^u\right\} < \gamma$; hence it represents a pessimistic scenario.

The parameter $s \in \Re$ is the Lagrange multiplier associated with the dual functional of the primal problems (39), (40). In particular, $s \leq 0$ corresponds to (39) while $s \geq 0$ corresponds to (40).

**Risk-Averse Versus Risk-Seeking Optimization.** In the context of risk-sensitive pay-offs, (39) corresponds to an optimistic pay-off functional (emphasizing the best cases) in which the strategies are risk-seeking, and (40) corresponds to a pessimistic pay-off functional (emphasizing the worst cases) in which the strategies are risk-averse. Moreover, in the context of uncertain stochastic systems, risk-averse strategies always imply dissipation inequalities.

### 4.3 Properties of the Optimal Solution

The next Lemma presents several properties of the unconstrained problems, including monotonicity properties of the dual functional with respect to $\gamma$ and conditions for finding the Lagrange multiplier.

**Lemma 1.** *For a given $u \in \mathcal{U}_{ad}$, assume $\ell^u \in BC(\Sigma)$. Then the following statements hold.*
*1) The dual functional $J^{s,\gamma}(u, \nu^u)$ is related to the cumulant generating function of $\ell^u$ with respect to $\mu^u \in \mathcal{M}(\Sigma)$ via*

$$J^{s,\gamma}(u, \nu^{u,*}) = s\gamma - \Psi_{\mu^u}(s), \ \gamma \in \Re, \ s \in \Re \tag{48}$$

*2) The dual functional $J^{s,\gamma}(u, \nu^{u,*})$ is concave in $s \in \Re$.*
*3) The supremum of the dual functional $J^{s,\gamma}(u, \nu^{u,*})$ over $s \in \Re$ is the Legendre-Fenchel Transform of $\Psi_{\mu^u}(s)$ and*

$$\varphi^{s^*}(u, \gamma) = \Psi^{\star}_{\mu^u}(\gamma) = \sup_{s \in \Re}\{s\gamma - \Psi_{\mu^u}(s)\} \tag{49}$$

*4) $\varphi^{s^*}(u, \gamma)$ is a convex function of $\gamma \in \Re$.*
*5) $\varphi^{s^*}(u, \gamma) \geq 0, \forall \gamma \in \Re$.*
*6) If $\ell^u \in L_1(\mu^u)$ and $E_{\mu^u}(\ell^u) = m$ then*
*a) $\varphi^{s^*}(u, m) = 0$.*
*b) $\varphi^{s^*}(u, \gamma)$ is non-decreasing for $\gamma \in [m, \infty)$, that is,*

$$\varphi^{s^*}(u, \gamma_1) \leq \varphi^{s^*}(u, \gamma_2), \ m \leq \gamma_1 \leq \gamma_2 < \infty$$

*c) $\varphi^{s^*}(u, \gamma)$ is non-increasing for $\gamma \in (-\infty, m]$, that is,*

$$\varphi^{s^*}(u, \gamma_2) \geq \varphi^{s^*}(u, \gamma_1), \ -\infty < \gamma_2 \leq \gamma_1 \leq m$$

*d) $J^{s,\gamma}(u, \nu^{u,*})$ is differentiable with respect to $s$ at point $s = s_0$ and*

$$\frac{d}{ds}J^{s,\gamma}(u,\nu^{u,*})\mid_{s=s_0}=\gamma-\frac{E_{\mu^u}\Big(\ell^u e^{s_0\ell^u}\Big)}{E_{\mu^u}\Big(e^{s_0\ell^u}\Big)}=\gamma-E_{\nu^{u,*}}\Big(\ell^u\Big) \tag{50}$$

*where*

$$\mathrm{d}\nu^{u,*}=\frac{e^{s_0\ell^u}\mathrm{d}\mu^u}{\int_\Sigma e^{s_0\ell^u}\,\mathrm{d}\mu^u} \tag{51}$$

*In addition, $J^{s,\gamma}(u,\nu^{u,*})$ is twice continuously differentiable with respect to $s$ at point $s=s_0$ and*

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}J^{s,\gamma}(u,\nu^{u,*})\mid_{s=s_0}=-\Big\{E_{\nu^{u,*}}\Big((\ell^u)^2\Big)-\Big(E_{\nu^{u,*}}(\ell^u)\Big)^2\Big\}\leq 0 \tag{52}$$

*e) Let*

$$s^*=\arg\sup_{s\in\Re}\Big\{s\gamma-\Psi^\star_{\mu^u}(s)\Big\} \tag{53}$$

*Then*

$$s^*=\arg\sup_{s\in\Re}J^{s,\gamma}(u,\nu^{u,*})=\arg\sup_{s\in\Re}\inf_{\nu^u\in\mathcal{M}(\Sigma)}\{H(\nu^u|\mu^u)-s\Big(E_{\nu^u}(\ell^u)-\gamma\Big)\Big\} \tag{54}$$

*Moreover,*

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big\{s\gamma-\Psi_{\mu^u}(s)\Big\}|_{s=s^*}=0\ \textit{implies}\ E_{\nu^{u,*}}\Big(\ell^u\Big)|_{s=s^*}=\gamma \tag{55}$$

*where $\nu^{u,*}$ is given by (51). Moreover, a necessary condition for the supremum of the dual functional $J^{s,\gamma}(u,\nu^{u,*})$ over $s\in\Re$ is that $s^*$ occurs on the boundary of the linear constraint.*

**Proof.** The proof for statements 1) to 6).a-c) is standard [27].

Next, the regions over which $J^{s,\gamma}(u,\mu^{u,*})$ is maximized, are identified, and conditions for finding the Lagrange multipliers are derived, for both problems of Definition 2.

**Theorem 3.** *Recall the problem of Definition 2,*
*1) Risk-Seeking Scenario. Consider problem (39).*
*Suppose $m\equiv E_{\mu^u}\Big\{\ell^u\Big\}=\int_\Sigma \ell^u\,\mathrm{d}\mu^u>\gamma$.*
*Then there exists a minimizing measure $\nu^{u,*}\in\mathcal{M}(\Sigma)$ which satisfies*

$$J^{s^*,\gamma}(u,\nu^{u,*})=\sup_{s\leq 0}\Big\{J^{s,\gamma}(u,\nu^{u,*})\Big\}=\varphi^{s^*}(u,\gamma)=\sup_{s\leq 0}\Big\{s\gamma-\Psi_{\mu^u}(\gamma)\Big\}$$

$$\equiv\Psi^\star_{\mu^u}(\gamma)=\inf_{\Big\{\nu^u\in\mathcal{M}(\Sigma);\ \int_\Sigma \ell^u\,\mathrm{d}\nu^u\leq\gamma\Big\}}H(\nu^u|\mu^u) \tag{56}$$

*and for some $s \leq 0$, then $\nu^{u,*} \in \mathcal{M}(\Sigma)$ is given by*

$$\mathrm{d}\nu^{u,*} = \frac{e^{s\ell^u}\mathrm{d}\mu^u}{\int_\Sigma e^{s\ell^u}\,\mathrm{d}\mu^u},\ s \leq 0 \tag{57}$$

*Moreover,the supremum over $s \leq 0$ in (56) is attained at $s^* < 0$ given by*

$$\gamma = E_{\nu^{u,*}}\Big\{\ell^u\Big\}|_{s=s^*} \leq E_{\nu^{u,*}}\Big\{\ell^u\Big\} < E_{\mu^u}\Big\{\ell^u\Big\} = m,\ \forall s \in [s^*, 0] \tag{58}$$

*2) Risk-Averse Scenario. Consider problem (40).*
*Suppose $m \equiv E_{\mu^u}\Big\{\ell^u\Big\} = \int_\Sigma \ell^u\,\mathrm{d}\mu^u < \gamma$.*
*Then there exists a minimizing measure $\nu^{u,*} \in \mathcal{M}(\Sigma)$ which satisfies*

$$J^{s^*,\gamma}(u,\nu^{u,*}) = \sup_{s\geq 0}\Big\{J^{s,\gamma}(u,\nu^{u,*})\Big\} = \varphi^{s^*}(u,\gamma) = \sup_{s\geq 0}\Big\{s\gamma - \Psi_{\mu^u}(\gamma)\Big\} \tag{59}$$

$$\equiv \Psi^\star_{\mu^u}(\gamma) = \inf_{\Big\{\nu^u\in\mathcal{M}(\Sigma);\ \int_\Sigma \ell^u\,\mathrm{d}\nu^u \geq \gamma\Big\}} H(\nu^u|\mu^u) \tag{60}$$

*and for some $s \geq 0$, then $\nu^{u,*} \in \mathcal{M}(\Sigma)$ is given by*

$$\mathrm{d}\nu^{u,*} = \frac{e^{s\ell^u}\mathrm{d}\mu^u}{\int_\Sigma e^{s\ell^u}\,\mathrm{d}\mu^u},\ s \geq 0 \tag{61}$$

*Moreover, the supremum over $s \geq 0$ in (59) is attained at $s^* > 0$ given by*

$$\gamma = E_{\nu^{u,*}}\Big\{\ell^u\Big\}|_{s=s^*} \geq E_{\nu^{u,*}}\Big\{\ell^u\Big\} > E_{\mu^u}\Big\{\ell^u\Big\} = m,\ \forall s \in [0, s^*] \tag{62}$$

**Proof.** Similar to derivations found in [13, 25].

## 5 Robustness of Stochastic Uncertain Systems: a Relative Entropy Constraint Formulation

In this section the optimization for robustness is undertaken with respect to an energy like objective functional under relative entropy constraints of the uncertain measure $\nu^u$ taken with respect to a fixed nominal measure $\mu^u$.

### 5.1 Problem Statement

**Definition 3.** *Let $u \in \mathcal{U}_{ad}$, and $\ell^u \in BC(\Sigma)$ which is a fixed nominal measure, and $R \in (0,\infty)$.*
*Find $\nu^{u,*} \in \mathcal{M}(\Sigma)$ which achieves the supremum*

$$J(u,\nu^{u,*}) = \sup_{\{\nu^u\in\mathcal{M}(\Sigma);H(\nu^u|\mu^u)\leq R\}} \int_\Sigma \ell^u \mathrm{d}\,\nu^u \tag{63}$$

*where $R \in (0,\infty)$.*

Next, for every $s \in \Re$, define the Lagrangian associated with the problem of Definition 3

$$J^{s,R}(u,\nu^u) \equiv E_{\nu^u}(\ell^u) - s\Big(H(\nu^u|\mu^u) - R\Big) \tag{64}$$

and its associated dual functional

$$J^{s,R}(u,\nu^{u,*}) = \sup_{\{\nu^u \in \mathcal{M}(\Sigma)\}} J^s(u,\nu^u) \tag{65}$$

In addition, define the quantity

$$\varphi^{s^*}(u,R) \equiv \inf_{s\geq 0} J^{s,R}(u,\nu^{u,*}) \tag{66}$$

### 5.2 Related Problems

**Disturbance Attenuation in Robustness.** For a given $u \in \mathcal{U}_{ad}$ let $L_2(\nu^u;\mathcal{H}) \equiv \Big\{\phi^u : \Sigma \to \mathcal{H}; \phi^u$is a random variable such that $\int_\Sigma \|\phi\|^2_{\mathcal{H}}\, \mathrm{d}\nu^u < \infty\Big\}$ denote the Hilbert space of random variables. Let $L_2(\nu^u;\mathcal{Z})$ and $L_2(\nu^u;\mathcal{D})$ denote the Hilbert Spaces of tracking signals and disturbance signals, respectively. For a given $u \in \mathcal{U}_{ad}$, let $T^u : \mathcal{D} \to \mathcal{Z}$ be a bounded linear operator with induced norm defined by

$$J(u,d^*) \equiv ||T^u|| = \sup_{||d||_{L_2(\nu^u;\mathcal{D})}\neq 0} \frac{||z||^2_{L_2(\nu^u;\mathcal{Z})}}{||d||^2_{L_2(\nu^u;\mathcal{D})}} = \sup_{\frac{1}{2}||d||_{L_2(\nu^u;\mathcal{D})}\leq R} ||z||^2_{L_2(\nu^u;\mathcal{Z})} \tag{67}$$

Then the optimal control $u^* \in \mathcal{U}_{ad}$ is found by minimizing the induced norm

$$J(u^*,d^*) \equiv \inf_{u\in\mathcal{U}_{ad}} ||T^u|| = \inf_{u\in\mathcal{U}_{ad}} \sup_{\frac{1}{2}||d||_{L_2(\nu^u;\mathcal{D})}\leq R} ||z||^2_{L_2(\nu^u;\mathcal{Z})} \tag{68}$$

The induced norm is equivalent to the optimal disturbance attenuation. For a given $u \in \mathcal{U}_{ad}$, the induced norm is found by defining the dual functional

$$J^{s^*}(u,d^*) = \inf_{s\geq 0} \sup_{d\in L_2(\nu^u;\mathcal{D})} \Big\{ \int ||z||^2_{\mathcal{Z}}\, \mathrm{d}\nu^u - s\Big(\frac{1}{2}\int ||d||^2_{\mathcal{D}}\, \mathrm{d}\nu^u - R\Big)\Big\} \tag{69}$$

Moreover, the optimal control $u^* \in \mathcal{U}_{ad}$ is found by minimizing the induced norm, and it is given by

$$J^{s^*}(u^*,d^*) = \inf_{u\in\mathcal{U}_{ad}} \inf_{s\geq 0} \sup_{d\in L_2(\nu^u;\mathcal{D})} \Big\{ \int ||z||^2_{\mathcal{Z}}\, \mathrm{d}\nu^u - s\Big(\frac{1}{2}\int ||d||^2_{\mathcal{D}}\, \mathrm{d}\nu^u - R\Big)\Big\} \tag{70}$$

in which $\inf_{u\in\mathcal{U}_{ad}} \inf_{s\geq 0}$ is interchanged.
When $\nu^u$ is absolutely continuous with respect to $\mu^u$, and the nominal model

is described by stochastic differential equations which are driven by Brownian motion or general Martingales, then it can be shown that $H(\nu^u|\mu^u) = \frac{1}{2}\int ||d||^2_{\mathcal{D}}\, \mathrm{d}\nu^u$. In this case, the primal problem (68) and its dual problem (70) are equivalent to the problem of Definition 3, that is,

$$J(u^*,\nu^{u,*}) = J^{s^*,R}(u^*,\nu^{u,*}) =$$
$$= \inf_{s\geq 0}\inf_{u\in\mathcal{U}_{ad}}\inf_{\nu^u\in\mathcal{M}(\Sigma)}\Big\{E_{\nu^u}\Big(\ell^u\Big) - s\Big(H(\nu^u|\mu^u) - R\Big)\Big\} = J(u^*,d^*) = J^{s^*}(u^*,d^*) \tag{71}$$

(let $\ell^u = ||z||^2_{\mathcal{Z}}$). Moreover, the smaller the values of $s$ the higher the attenuation and hence the higher dissipation of output power with respect to the input power.

**Risk-Averse Versus Risk-Seeking Optimization.** In the context of risk-sensitive pay-offs, the problem of Definition 3 corresponds to an optimistic pay-off functional (emphasizing the best cases), when the Lagrange multiplier $s \leq 0$, in which the strategies are risk-seeking, and to a pessimistic pay-off functional (emphasizing the worst cases) in which the strategies are risk-averse, when the Lagrange multiplier $s \geq 0$.

### 5.3 Properties of the Optimal Solution

**Corollary 1.** *For a given $u \in \mathcal{U}_{ad}$, and for some $s \in \Re$ such that $\frac{\ell^u}{s} \in BC(\Sigma)$, the following statements hold.*
*1) The dual functional $J^{s,R}(u,\nu^{u,*})$ is related to the cumulant generating function of $\ell^u$ with respect to $\mu^u \in \mathcal{M}(\Sigma)$ via*

$$J^{s,R}(u,\nu^{u,*}) = s \sup_{\left\{\nu^u\in\mathcal{M}(\Sigma);H(\nu^u|\mu^u)<\infty\right\}} \Big\{\frac{1}{s}\int_\Sigma \ell^u\,\mathrm{d}\nu^u - H(\nu^u|\mu^u)\Big\} + sR \tag{72}$$

$$= s\log\int_\Sigma e^{\frac{\ell^u}{s}}\,\mathrm{d}\mu^u + sR = s\Psi_{\mu^u}(\frac{1}{s}) + sR \tag{73}$$

*Moreover, if the supremum in (72) is attained at $\nu^{u,*} \in \mathcal{M}(\Sigma)$ and it is given by*

$$\mathrm{d}\nu^{u,*} = \frac{e^{\frac{\ell^u}{s}}\mathrm{d}\mu^u}{\int_\Sigma e^{\frac{\ell^u}{s}}\,\mathrm{d}\mu^u} \tag{74}$$

*In addition, "The average energy of the system" = "The Helmoltz Free Energy" $+s\times$ "The Relative Entropy of the system", that is,*

$$\int_\Sigma \ell^u\,\mathrm{d}\nu^{u,*} = s\log\int_\Sigma e^{\frac{\ell^u}{s}}\,\mathrm{d}\mu^u + sH(\nu^{u,*}|\mu),\ s\in(0,\infty) \tag{75}$$

*2) The dual functional $J^{s,R}(u,\nu^{u,*})$ is convex in $s > 0$.*

*3) The function* $\Gamma_{\mu^u}(s) \equiv s\Psi_{\mu^u}(\frac{1}{s})$ *is a non-increasing function of* $s \in (0,\infty)$*, that is,*

$$\Gamma_{\mu^u}(s_1) = s_1 \log E_{\mu^u}\Big\{e^{\frac{\ell^u}{s_1}}\Big\} \leq s_2 \log E_{\mu^u}\Big\{e^{\frac{\ell^u}{s_2}}\Big\} = \Gamma_{\mu^u}(s_2),\ 0 < s_2 \leq s_1 \quad (76)$$

*4) The infimum of the dual functional* $J^{s,R}(u,\nu^{u,*})$ *over* $s > 0$ *defined by*

$$\Phi^*_{\mu^u}(R) \equiv \varphi^{s^*}(u,R) = \inf_{s>0}\Big\{s\Psi_{\mu^u}(\frac{1}{s}) + sR\Big\} \quad (77)$$

*is a concave functional of* $R \geq 0$*.*
*5)*

$$E_{\mu^u}\Big\{\ell^u\Big\} \leq \Phi^*_{\mu^u}(R) = \varphi^{s^*}(u,R) \leq R + \log E_{\mu^u}\Big(e^{\ell^u}\Big) \quad (78)$$

*Moreover if* $\ell^u$ *is* $\nu^u$*-essentially bounded for all* $\nu^u \in \mathcal{A}$*, then the above bounds become*

$$E_{\mu^u}\Big\{\ell^u\Big\} \leq \Phi^*_{\mu^u}(R) = \varphi^{s^*}(u,R) \leq \min\Big\{R + \log E_{\mu^u}\{\ell^u\}, ||\ell^u||_\infty\Big\}$$

*6)The infimum of the functional* $J^{s,R}(u,\nu^{u,*})$ *over* $s > 0$ *is uniquely attained at*

$$H(\nu^{u,*}|\mu^u)|_{s=s^*} = R \quad (79)$$

*where* $\nu^{u,*}$ *is given by (74). That is, a necessary condition for the infimum of the dual functional* $J^{s,R}(u,\nu^{u,*})$ *over* $s > 0$ *is that* $s^*$ *occurs on the boundary of the relative entropy constraint. Moreover,*

$$\frac{\mathrm{d}}{\mathrm{d}s} s \log \int_\Sigma e^{\frac{\ell^u}{s}}\,\mathrm{d}\mu^u = \log \int_\Sigma e^{\frac{\ell^u}{s}}\,\mathrm{d}\mu^u - \frac{1}{s}E_{\nu^{u,*}}\{\ell^u\} = -H(\nu^{u,*}|\mu^u) \quad (80)$$

*7) Under the assumptions of 6), the relative entropy* $H(\nu^{u,*}|\mu^u)$ *is a non-increasing function of* $s > 0$*, that is,*

$$0 \leq H(\nu^{u,*}|\mu^u)|_{s=s_2} \leq H(\nu^{u,*}|\mu^u)|_{s=s_1} \leq H(\nu^{u,*}|\mu^u)|_{s=s^*} = R,\ 0 < s^* \leq s_1 \leq s_2 \quad (81)$$

**Proof.** Similar to derivations found in [25].

*Remark 2.* The various statements of Corollary 1 establish the various paths which connect information theory, robustness of stochastic uncertain systems, statistical mechanics and thermodynamics. An example of such connection has been discussed earlier in the context of the related Disturbance Attenuation Problem, where it is shown that Lagrange multiplier $s$ is the sensitivity reduction associated with problems in which the induced norm is less than or equal to $s$, that is, $||T^u|| \leq s$, and that $s^*$ is the optimal sensitivity reduction. These connection are elucidated further in the forecoming discussion

**Monotonicity Properties.** The statement 1 of Corollary 1 states that the dual functional is proportional to the Free Energy of $\frac{\ell^u}{s}$. Moreover, the worst

case measure is the equilibrium measure of statistical mechanics, in which the denomimator of (57) is the partition function. Furthermore, (80) is a fundamental identity in statistical mechanics. It furthermore states that the dual functional is proportional to the Free Energy of the system $\Gamma_{\mu^u}(s) \equiv s\Psi_{\mu^u}(\frac{1}{s})$. In the context of risk-sensitive pay-off's, it states that the dual functional is proportional to risk-averse optimization (pessimistic), in which $s$ is the Lagrange multiplier of the unconstrained problem. Moreover, 3) states that the risk-sensitive pay-off or Free Energy $\Gamma_{\mu^u}(s)$ is a non-increasing function of the sensitivity parameter or Lagrange multiplier $s \in (0, \infty)$.

**Characterization of Optimal Sensitivity Reduction and 2nd Law of Thermodynamics.** Corollary 1,7) states that the optimal sensitivity reduction $s^*$ corresponds to the case when the relative entropy of the worst case measure with respect to the nominal measure is exactly at the boundary of the constraint. In addition, 8) states that the relative entropy is non-increasing as a function of the sensitivity parameter $s \in [s^*, \infty)$. This monotonicity property can be used to devise a simple algorithm to compute $s^*$, by starting with an arbitrary $s \in [s^*, \infty)$ and then performing a sequence of relative entropy calculations till its value occurs at the boundary. Moreover, since the best possible dissipation of the system corresponds to the optimal sensitivity reduction, or when the induced norm is equal to $s^*$, then the non-increasing property of the Relative Entropy of the system (81) implies an increase in the Power Dissipation of the system. Thus, (81) is a statement of the 2nd Law of Thermodynamics, which states that higher dissipation (e.g., smaller dissipation factor $s$) implies higher relative entropy of the system.

**Upper and Lower Bounds of the Optimal Solution.** Corrolary 1, 6), gives lower and upper bounds for the optimal solution via the a priori information of the original problem, specifically, the nominal measure and the energy functional. The lower bound is trivial (follows from Jensen's inequality) while the upper bound is non-trivial. These bounds are important in judging the performance of sub-optimal solutions to the optimal solution.

**Convergence of Induced Norm to the $L_1$ Norm.** Corrolary 1, 4) states that the optimal solution is a concave functional of $R \geq 0$. Moreover, 5) states that the case $R = 0$ corresponds to the risk-neutral pay-off. Also, using the relations established in the context of the Related Disturbance Attenuation Problem, Corrolary 1, 5) states that as $R \to 0$, the induced norm converges to an $L_1$ norm, e.g., $\lim_{R\to 0} ||T^u|| = \sup_{\left\{\nu^u \in \mathcal{M}(\Sigma); H(\nu^u|\mu^u) \leq R\right\}} \int_\Sigma \ell^u \, \mathrm{d}\nu^u = \int_\Sigma \ell^u \, \mathrm{d}\mu^u$.

In the next theorem, the Corrolary 1 is employed to show equivalence between the unconstrained and constrained problems, which implies that all properties implied in the Corrolary 1 hold for the problem of Definition 3.

**Theorem 4.** *For a given $u \in \mathcal{U}_{ad}$, suppose $\ell^u \in BC(\Sigma)$ and there exists a $u \in \mathcal{U}_{ad}$ such that $J^{s,R}(u, \nu^{u,*}) < \infty$. Then the following statements hold.*

*1) The primal constrained problem of Definition 3 and the unconstrained dual problem are equivalent, that is,*

$$J^{s^*,R}(u,\nu^{u,*}) = \inf_{s>0} J^{s,R}(u,\nu^{u,*}) = \varphi^{s^*}(u,R) = \inf_{s>0}\Big\{sR+s\Psi_{\mu^u}(\frac{1}{s})\Big\} \equiv \Phi^*_{\mu^u}(R) \tag{82}$$

$$= \sup_{\Big\{\nu^u\in\mathcal{M}(\Sigma);\ H(\nu^u|\mu^u)\le R\Big\}} \int_\Sigma \ell^u \,\mathrm{d}\nu^u = J(u,\nu^{u,*}) \tag{83}$$

*and the results of Lemma 1 hold.*
*2)*

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}J^{s,R}(u,\nu^{u,*}) = \frac{1}{s^3}\Big(E_{\nu^{u,*}}(\ell^u)^2 - \Big(E_{\nu^{u,*}}(\ell^u)\Big)^2\Big) \ge 0,\ \forall s\in[s^*,\infty) \tag{84}$$

*3)Then optimal pay-off is equal to the average energy of the system with respect to the worst case measure $\nu^{u,*}$ given by*

$$J^{s^*,R}(u,\nu^{u,*}) = \varphi^{s^*}(u,R) = E_{\nu^{u,*}}\Big\{\ell^u\Big\}|_{s=s^*} = \frac{\int_\Sigma \ell^u e^{\frac{1}{s}\ell^u}\,\mathrm{d}\mu^u}{\int_\Sigma e^{\frac{1}{s}\ell^u}\,\mathrm{d}\mu^u}|_{s=s^*} \tag{85}$$

**Proof.** Statements 1), 2) follow from simple reformulation of the constrained optimization problem of Definition 3 to an unconstrained optimization problem and Corrolary 1. Statement 3) follows from the equation of the dual functional of the unconstrained optimization problem.

*Remark 3.* The third Statement of the Theorem 4, states that the optimal performance is given by the average energy with respect to the worst case measure. Moreover, 2) states that the second derivative of the optimal performance is proportional to the variance of the energy function with respect to the worst case measure.

## 6 The Large Deviations Principle Applied to Diffusion Processes

In this section we construct the action functional and a deterministic measure on cylinder sets of a Hilbert space. This is a consequence of the Large Deviation principle applied to Brownian motion found in [28, 27].

**Assumption** $f:\Re^n\to\Re$, $\sigma:\Re^n\to\Re^n\otimes\Re^n$ are uniformly Lipschitz continuous, $\sigma$ is bounded and $a(x)\equiv\sigma(x)\sigma'(x)$ is positive definite, that is, there exists an $k\in[1,\infty)$ such that

$$\|f(x)-f(y)\| + \|\sigma(x)-\sigma(y)\| \le k\|x-y\|,$$

$$\|\sigma(x)\| \leq k, \quad \exists \lambda > 0 \ni \sigma(x)\sigma(x) \geq \lambda I_{n\times n}.$$

The LDP associated with diffusion processes is usually applied to the space $\Big(\mathcal{X}, \mathcal{B}_{\mathcal{X}}\Big) = \Big(C_{0,T} \equiv C([0,T];\Re^n), \mathcal{B}_{0,T} \equiv \mathcal{B}(C([0,T];\Re^n))\Big)$, which is a Banach space with the uniform norm $||\cdot||_{C_{0,T}}$. The diffusion process $\{X^\epsilon(t)\} : C_{0,T} \to C_{0,T}$ is the unique solution of the stochastic Ito differential equation

$$dX^\epsilon(t) = f(X^\epsilon(t))dt + \sqrt{\epsilon}\sigma(X^\epsilon(t))\mathrm{d}w(t), \ X^\epsilon(0) = x, \tag{86}$$

where the assumption is satisfied. For a given bounded function $f$ let $\{P^\epsilon\}_{\epsilon>0}$ denote the probability measure induced by $\{X^\epsilon(t)\}$ on $\Big(C_{0,T}, \mathcal{B}_{0,T}\Big)$. Then $P^\epsilon = \mathcal{W}^\epsilon \circ X^{\epsilon,-1}$ where $\mathcal{W}^\epsilon$ is the measure induced by $\{\sqrt{\epsilon}w(t)\}$ and $X^\epsilon : C_{0,T} \to C_{0,T}$ is defined by $X^\epsilon = F^\epsilon(g)$, where $X^\epsilon$ is the unique continuous solution of $X^\epsilon(t) = x + \int_0^t f(X^\epsilon(s))ds + g^\epsilon(t)$.
Introduce the Hilbert space

$$H^1_{0,T} = H^1\Big([0,T];\Re^n\Big) \equiv \Big\{\phi \in C([0,T];\Re^n);$$

$$\phi(t) = \int_0^t \dot{\phi}(s)\,\mathrm{d}s, \int_0^T ||\dot{\phi}(s)||^2_{\Re^n}\,\mathrm{d}s < \infty\Big\}$$

which is the space of absolutely continuous functions with square-integrable derivatives. Then $\Big\{\Big(C_{0,T}, \mathcal{B}_{0,T}, P^\epsilon\Big)\Big\}_{\epsilon>0}$ satisfies the LDP, which is an application of the contraction principle; the action functional is given by

$$I^{x,f}_{H^1_{0,T}}(X) = \Big\{ \begin{array}{c} -\frac{1}{2}\int_0^T ||a^{-\frac{1}{2}}(X(s))\Big(\dot{X}(s) - f(X(s))\Big)||^2_{\Re^n}\,\mathrm{d}s, \\ -\infty, \end{array}$$

where $-\infty$ corresponds to the case when $X - x \notin H^1_{0,T}$. Equivalently,

$$I^{x,f}_{H^1_{0,T}}(X) = I^{x,0}_{H^1_{0,T}}(w) = \Big\{ \begin{array}{c} -\frac{1}{2}\int_0^T ||a^{-\frac{1}{2}}(X(s))\dot{w}(s)||^2_{\Re^n}\,\mathrm{d}s \\ -\infty \end{array}$$

where $H^{1,w}_{0,T} \equiv \Big\{w \in H^1_{0,T}; X(t) = x + \int_0^t f(X(s))\,\mathrm{d}s + \int_0^t \sigma(X(s))\dot{w}(s)\Big\}$.

### 6.1 Connections to Thermodynamic Entropy

Large Deviations to relate the Free Energy $\log \int_{C_{0,T}} e^{\int_0^T \ell(X)ds} dP^\epsilon(X)$ to the Macroscopic Thermodynamic Entropy $S$ of Section 2. Clearly, for any bounded and continuous function $\ell$, by the Laplace-Varadhan Theorem of Large Deviations [27] we have

$$S(x) \equiv \lim_{\epsilon \to 0} \epsilon \log \int_{C_{0,T}} e^{\frac{1}{\epsilon}\int_0^T \ell(X)ds}\,\mathrm{d}P^\epsilon_x(X)$$

$$= \sup_{w \in H^{1,w}_{0,T}} \Big\{ \int_0^T \Big( \ell(X(s)) - \frac{1}{2} ||a^{-\frac{1}{2}}(X(s)) \dot{w}(s)||^2_{\Re^n} \Big) \, \mathrm{d}s; \qquad (87)$$

$$\dot{x}(s) = f(x(s)) + \sigma(x(s)) \, \dot{w}(s), 0 \leq s \leq T \Big\}$$

Here $W_R(x, \dot{w}) \equiv \frac{1}{2} ||a^{-\frac{1}{2}}(x) \dot{w}||^2_{\Re^n} - \ell(x)$ is the supply of energy into the system, and $S(x)$ is the available storage, maximum extractable energy of the system, and hence a storage function [1]. It is important to notice that the dissipation inequality in [1] is expressed in terms of free energy of thermodynamics and not in terms of entropy as the Clausius inequality is expressed. The functional $S(x)$ is employed in robust control problems to establish a dissipation inequality [1].

Referring to Figure 1, the dynamic equation in (87) is identified as the plant G on which the uncertainty is imposed by the disturbance noise $\dot{w}$ identified with the input $w$ of the Figure 1. In this context, under certain conditions, equation (87) is related to equation (4).

## 7 Conclusion

This paper establishes various connections between Robustness, Information Theory, Large Deviations and Statistical Mechanics for stochastic uncertain systems. These connections are established by introducing two fundamental optimization problems. The characteristics of the optimal solutions are presented with a discussion on how these properties are connected to the optimal states of statistical mechanical systems. The monotonicity properties of the sensitivity level $s$ with respect to relative entropy are reminiscent of the monotonicity properties of the temperature with respect to entropy that one encounters in the fluctuation dissipation theory of statistical mechanics. This connection has also implications in devising algorithms of computing the optimal sensitivity $s^*$ of complicated controlled systems. Detail derivations and additional properties and connections among these fields are found in [15, 16].

## References

1. J. Willems: Dissipative dynamical systems part I: general theory. Arch. Rational Mech. Anal., 321–351, **45**,(1972)
2. L.M. Bellac, F. Mortessagne, G.G. Batroumi: *Equilibrium and Non-Equilibrium Statistical Thermodynamics* (Cambridge University Press 2004)
3. G. Zames: Feedback and Optimal Sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Trans. Aut. Cont.* **26**, 301–320 (1981)
4. T.M. Cover, J.A. Thomas: *Elements of Information Theory* (John Wiley and Sons Inc. 1991)

5. S. Kullback: *Information Theory and Statistics* (John Wiley and Sons Inc., New York 1959)
6. E.T. Jaynes: Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957)
7. E.T. Jaynes: Information theory and statistical mechanics. *Phys. Rev.* **108**, 171–190(1957)
8. A.N. Kolmogorov: Logical basis for information theory and probability. *IEEE Trans. Inform. Theory* **14**, 662–664 (1968)
9. A. Dembo, O. Zeitouni: *Large Deviations Techniques and Applications*, (Springer, Berlin Heidelberg New York 1998)
10. C.E. Shannon: A mathematical theory of communication. *Bell System Tech. J.* **27**, 379–423, 623–656 (1948)
11. R.S. Ellis: *Entropy, Large Deviations, and Statistical Mechanics.* (Springer, Berlin Heidelberg New York 1985)
12. R.S. Ellis: The theory of large deviations: from Boltzmann's 1877 calculation to equilibrium macrostatesin 2D turbulence. *Phys. D* **133**, 106–136 (1999)
13. C.D. Charalambous, F. Rezaei , A. Kyprianou: Relations between information theory, robustness and statistical mechanics of stochastic systems. In: *43rd IEEE Conference on Decision and Control, 14-17 December 2004*, 3479–3484 (2004)
14. D. Ruelle: A variational formulation of equilibrium statistical mechnaics and the Gibbs phase rule. *Commun. Math. Phys* **5**, 324–329 (1967)
15. F. Rezaei, C.D. Charalambous, A. Kyprianou: Optimization of fully observable nonlinear stochastic uncertain control systems. In: *43rd IEEE Conference on Decision and Control, 14-17 December 2004*, 2556–2560 (2004).
16. F. Rezaei, C. D. Charalambous, A. Kyprianou: Optimization of non-linear stochastic uncertain relaxed controlled systems: entropy rate functionals and robustness. In: *43rd IEEE Conference on Decision and Control, 14-17 December 2004*, 2561–2565 (2004)
17. D. Jacobson: Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Trans. Aut. Cont.* **18**, 124–131, (1973)
18. J. Speyer: An adaptive terminal guidance scheme based on an exponential cost criterion with applications to homing guidance. *IEEE Trans. Aut. Contr.* **21**, 371–375, (1976)
19. P. Whittle: Risk-sensitive linear/quadratic/gaussian control. *Adv. Applied Prob.* **13**, 764–777 (1981)
20. A. Bensoussan, J.H. van Schuppen: Optimal control of partially observable stochastic systems with an exponential-of-integral performance index. *SIAM J. Cont. Optim.* **23**, 599–613, (1985)
21. P. Whittle: A risk sensitive maximum principle. *Syst. Cont. Lett.* **15**, 183–192 (1990)
22. W.H. Fleming, W.M. McEneaney: Risk-sensitive control and differential games. In: *Stochastic Theory and Adaptive Control*, eds. T.E. Duncan and B. Pasik-Duncan, (Conference on Decision and Control 1992).
23. M. James, J. Baras, R. Elliott: Risk sensitive control and dynamic games for partially observed discrete time non-linear systems. *IEEE Trans. Aut. Cont.* **39**, 780–792, (1994)
24. P.D. Pra, L. Meneghini, W. Runggaldier: Some connections between stochastic control and dynamic games. *Math. Cont., Sig., Syst.* **9**, 303–326, (1996)

25. C.D. Charalambous, F. Rezaei, N. Ahmed: Optimization of stochastic uncertain systems: minmax games and robustness. *SIAM J Cont. and Optim.* (submitted), (2004)
26. C.D. Charalambous, F. Rezaei: Optimization of stochastic uncertain systems: large deviations and robustness. In: *Proc. 42nd IEEE Conference on Decision and Control, Hawaii, U.S.A., 2003*, vol. 4, 4260–4264 (2003)
27. J.D. Deuschel, W.D. Stroock: *Large Deviations*, vol. 137 in Pure and Applied Mathematics (Academic Press Inc., Boston 1989)
28. S.R.S. Varadhan: *Large Deviations and Applications* (SIAM, Philadelphia 1984)

# Part III

## Kinetics and Model Reduction

# Exactly Reduced Chemical Master Equations

M. R. Roussel[1,3] and R. Zhu[1,2]

[1] Department of Chemistry and Biochemistry, University of Lethbridge, Lethbridge, Alberta, Canada, T1K 3M4.
[2] Current address: Department of Chemistry, University of Calgary, Calgary, Alberta, Canada, T2N 1N4,
[3] roussel@uleth.ca, home page: http://people.uleth.ca/~roussel

**Summary.** In the small-number limit, we must abandon the description of chemical systems in terms of continuous concentration variables which evolve according to deterministic rate equations in favor of a discrete stochastic formulation. The probability distribution for the molecular populations however does obey a deterministic equation called the chemical master equation. Any desired population statistic (mean, standard deviation, etc.) can be obtained from the probability distribution. Unfortunately, the master equation consists of a huge set of differential equations, and it is thus in general impractical to use it directly. In this paper, we review the ideas underlying discrete population modeling and the chemical master equation. We then develop methods for reducing the chemical master equation to a much smaller set of differential equations by exploiting the same time-scale separation which leads to the emergence of a hierarchy of attracting manifolds in the mass-action case. Finally, we develop a method for generating an initial condition for the reduced model based on a generalization of the stationary reactant approximation.

## 1 Introduction

The dynamics of chemical systems span a huge range of time scales, from the femtosecond range of internal motions and intramolecular energy transfers, through the microsecond range which characterizes the time between molecular encounters, and finally to the much longer time scales which may in general govern reactive events [1]. Correspondingly, chemical systems can be described at a number of different levels which may or may not include processes in a given range of time scales. Figure 1 shows *some* of the methods which can be used to describe the kinetics of a chemical system, and their interrelationships. Note that the figure only shows methods which are essentially descriptive in nature. Methods which properly belong to the discipline of chemical dynamics (ab initio molecular dynamics, transition state theory, etc.) which aim to predict rates of chemical reactions from first principles [1] are not included. Molecular dynamics straddles these two worlds since it includes both fully

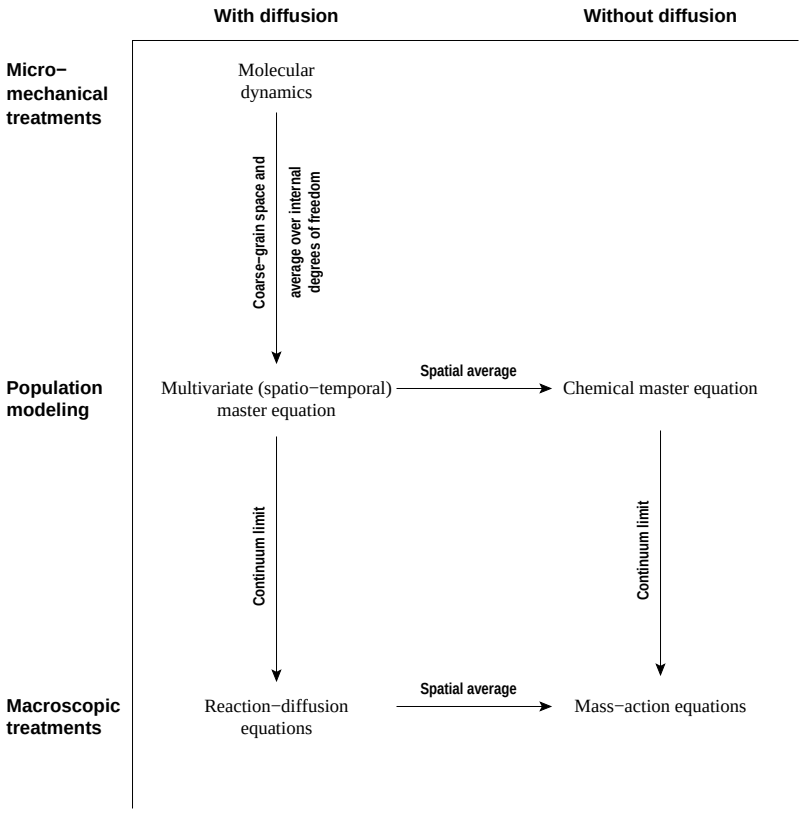



Fig. 1: Levels of description in chemical kinetics.

ab initio quantum-mechanical techniques and empirical simulation methods calibrated from experimental data, along with every conceivable variation in between [2, 3]. At the other end of the hierarchy of methods, we have macroscopic treatments which take no account of the discrete nature of molecules, let alone of their internal degrees of freedom. This paper will discuss a problem which arises in the middle ground of population modeling. In population models, we attempt to describe the evolution in time of the number of molecules (the population) of a particular type. By "type", we might just mean an identifiable chemical species, although in more detailed models the type might also include the quantum state of a molecule. Either way, having given up the mechanical level of description, reactions become random events to be described by a statistical theory.

We focus particularly on the case of well-mixed chemical systems. There are two major approaches to studying the evolution of molecular populations for well-mixed system. The first consists in writing down an evolution equa-

tion for the probability distribution of the populations. This evolution equation is known in the literature as the chemical master equation (CME) [4, 5]. The CME's major drawback is that it consists of a huge number of ordinary differential equations (ODEs) for which analytic solutions are not typically available and which are troublesome to solve numerically. The alternative is to use a simulation method to obtain sample trajectories of the underlying stochastic process [6, 7]. While stochastic simulations are a practical alternative to the solution of the master equation, the necessity to compute a large number of stochastic trajectories in order to obtain estimates of the moments of the distribution (e.g. the average and standard deviation of the number of molecules) makes this task computationally demanding as well. Accordingly, a great deal of effort has gone into methods to accelerate stochastic simulations [8, 9, 10, 11, 12, 13, 14].

Our approach is a bit different [15], though not without antecedent [16, 10, 11, 17, 12, 13, 14]. Conceptually, our work is closest to that of Janssen who developed a projection operator formalism to obtain a reduced master equation in which degrees of freedom corresponding to rapidly responding intermediates have been eliminated [16]. Shibata, following a similar approach, obtained a reduced master equation with added diffusive terms [17]. Because a typical chemical system evolves on many different time scales, even after we have eliminated the very rapid internal motions of the molecules, the master equation itself displays relaxation over a range of time scales. Accordingly, the CME is stiff, a numerical property of systems with multiple, widely separated time scales which makes their numerical integration difficult [18, 19, 20]. Relaxation on different time scales however opens up interesting possibilities for model reduction. Indeed, if we aren't interested in the relaxation on the fastest time scales of a system, elimination of the fast modes leads very directly to a reduced model. Since the fast modes in a chemical system can be very fast indeed, these modes are often not easily accessible experimentally and are thus of less interest than the observable slow evolution toward the final probability distribution. Unlike other methods based on analysis of the master equation, our methods are essentially *exact*: The reduced model we obtain simply confines the solutions to the slow manifold [21, 22, 23, 24], the hypersurface which contains the purely slow parts of the evolution to the final distribution. The accurate computation of this manifold for the CME is tricky, but not impossible. We obtain a reduced master equation which is of a similar form to the original CME, but which is both much smaller and non-stiff. The main difficulty lies in projecting the initial condition in the full probability space onto the manifold in such a way that the evolution of the reduced system tracks that of the full system after decay of transients.

In the next section, we introduce the ideas underlying stochastic population modeling and offer a simple derivation of the CME. Because it is easier to see ideas in action than to discuss them abstractly, throughout this paper we consider the competitive inhibition (CI) mechanism of enzyme kinetics as an example:

$$\mathrm{E} + \mathrm{S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{C} \underset{k_2}{\overset{k_{-2}}{\rightleftharpoons}} \mathrm{E} + \mathrm{Q} \tag{1a}$$

$$\mathrm{E} + \mathrm{X} \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} \mathrm{H} \tag{1b}$$

In this mechanism, E is an enzyme which catalyzes the conversion of the substrate (reactant) S to the product Q via the intermediate C. X is an inhibitor which forms a nonproductive complex with the enzyme. The inhibition here is competitive since the inhibitor and substrate cannot both bind the enzyme at the same time. The rate constants shown here are conventional macroscopic mass-action rate constants. The rate constants in step (1a) have been numbered in such a way as to highlight the symmetry of this reaction with respect to S and Q. Note that this numbering also has the advantage that all of the $k_i$'s are second-order rate constants, while the $k_{-i}$'s are first-order constants.

## 2 Stochastic Population Modeling and the Chemical Master Equation

In stochastic population modeling, the state of the system is described by a vector of populations, $\mathbf{N}(t) = [N_1(t), N_2(t), \ldots, N_n(t)]^{\mathrm{T}}$, where $N_i$ is the number of molecules of type $i$ at time $t$, $n$ is the number of different types of molecules considered in the model, and the superscripted T indicates the matrix transpose. In our example, the state could be described by the vector

$$\mathbf{N}(t) = [N_C(t), N_E(t), N_H(t), N_Q(t), N_S(t), N_X(t)]^{\mathrm{T}}, \tag{2}$$

where $N_C$ is the number of C molecules, $N_E$ is the number of E molecules, and so on. However, because the mechanism (1) represents a closed chemical system, the following mass conservation relations [25, 26, 27] reduce the number of independent variables to three:

$$N_S + N_C + N_Q = N_{S0}, \tag{3a}$$

$$N_E + N_C + N_H = N_{E0}, \tag{3b}$$

$$N_X + N_H = N_{X0}, \tag{3c}$$

where $N_{S0}$, $N_{E0}$ and $N_{X0}$ are, respectively, the initial numbers of molecules of S, E and X in an experiment in which separate solutions of these three substances are mixed at $t = 0$. Accordingly, if we think of these three quantities as parameters, the state of the system can be completely described by the vector

$$\mathbf{N}(t) = [N_C(t), N_S(t), N_X(t)]^{\mathrm{T}}, \tag{4}$$

among other possible triplets.

A first-order reaction is a random event which might involve, for instance, intramolecular energy transfers [1]. If we only keep track of particle numbers, the collisions which are necessary for the occurrence of second-order reactions are also random events [5]. Ternary collisions are exceedingly rare in the gas phase, and can generally be treated as a sequence of two molecular encounters in solution, but these also would be random events in a homogeneous population model. Accordingly, a statistical theory is most appropriate to the description of chemical kinetics on this level. Let $\mathcal{N}$ be the space of all possible vectors $\mathbf{N}$ satisfying the mass conservation and non-negativity constraints. Then $P(\mathcal{N}, t)$ is the probability distribution over the space $\mathcal{N}$ at time $t$ and $P(\mathbf{N}, t)$ is the probability that the system is in a particular state $\mathbf{N}$ at time $t$.

We offer here a simple derivation of the evolution equation for $P(\mathcal{N}, t)$, the chemical master equation, which brings out the essential elements of the theory but which is admittedly not very rigorous. More elaborate treatments are available elsewhere [4, 5]. The essence of Boltzmann's assumption of molecular chaos [28] is that collisions rapidly erase any dynamical memory of the initial condition in the gas phase, i.e. that the autocorrelation times of particle trajectories are short. Provided we are satisfied with a model which won't describe events which occur on shorter time scales than the mean collision time, the evolution of the probability density can then be treated as a Markov process [29]. In solution, because of solvent caging effects [1], vigorous mixing is required to maintain homogeneity and to fulfill the conditions leading to a Markov process. In the text that follows, we will use the relatively simpler language appropriate to gas-phase kinetics. However, reactions in solution are included in this formalism if we think of diffusion through the solvent as playing the memory-erasing role of collisions: A Markov process will result when we consider times which are long compared to the time scale over which diffusion and/or mixing homogenizes mesoscopic volumes.

We focus on the probability of one particular state, $P(\mathbf{N}, t)$. If the evolution can be described as a Markov process, then we should be able to write

$$P(\mathbf{N}, t + \Delta t) = F(\mathcal{N}(t), \Delta t),$$

where $F$ is a functional of the distribution at time $t$ and of $\Delta t$, provided $\Delta t$ is significantly larger than the characteristic autocorrelation time of particle trajectories. If $\Delta t$ is sufficiently small, but not so small as to violate the above constraint, and assuming that $F$ is a continuous functional of $\Delta t$, then we should be able to write

$$P(\mathbf{N}, t + \Delta t) = P(\mathbf{N}, t) + W_{\mathbf{N}}(\mathcal{N}(t))\Delta t,$$

where $W_{\mathbf{N}}$ is the instantaneous rate of change of $P(\mathbf{N}, t)$. Probability must be conserved, so states gain or lose probability by transfer from other states. In our case, these transfers are caused by chemical reactions. Let $\mathcal{R}$ be the set of chemical reactions, and $\boldsymbol{\nu}_r$ be the stoichiometry vector for reaction $r$, i.e. after a reaction of type $r$ starting from state $\mathbf{N}$, the new state is $\mathbf{N} + \boldsymbol{\nu}_r$.

For sufficiently small $\Delta t$, only one reaction of any given type is likely to have occurred, so we can write

$$P(\mathbf{N}, t + \Delta t) = P(\mathbf{N}, t) + \Delta t \sum_{r \in \mathcal{R}} \left[ a_r(\mathbf{N} - \boldsymbol{\nu}_r) P(\mathbf{N} - \boldsymbol{\nu}_r, t) - a_r(\mathbf{N}) P(\mathbf{N}, t) \right]. \tag{5}$$

In this equation $a_r(\mathbf{N})$ is the (conditional) probability per unit time that a reaction of type $r$ occurs given that the system is in state $\mathbf{N}$. These quantities are known in the literature as *reaction propensities* [6, 7]. They are assumed to be independent of time, an assumption which may break down if either very small spatial scales or times much shorter than the mean collision time are considered [30]. The former is not an issue for us since we are assuming a well-mixed system, but would be an issue in a spatio-temporal master equation [29]. The first term in the sum represents transitions to state $\mathbf{N}$ while the second term represents transitions from this state into other states. These terms can be written as products involving the instantaneous probabilities on the assumption that these probabilities do not change appreciably in a time $\Delta t$.

If we rearrange equation (5) in the obvious way and take the limit as $\Delta t \to 0$, we get

$$\begin{aligned} \frac{\mathrm{d}P(\mathbf{N}, t)}{\mathrm{d}t} &= \lim_{\Delta t \to 0^+} \frac{P(\mathbf{N}, t + \Delta t) - P(\mathbf{N}, t)}{\Delta t} \\ &= \sum_{r \in \mathcal{R}} \left[ a_r(\mathbf{N} - \boldsymbol{\nu}_r) P(\mathbf{N} - \boldsymbol{\nu}_r, t) - a_r(\mathbf{N}) P(\mathbf{N}, t) \right]. \end{aligned} \tag{6}$$

The limit taken to obtain equation (6) is one of those peculiar physicist's limits in which time intervals which are too short (shorter than the mean time between collisions) are excluded. Nevertheless, the master equation (6) gives a satisfactory account of the evolution of the probability distribution $P(\mathcal{N}, t)$ for a chemically reacting system on appropriate time scales.

The reaction propensities $a_r$ can be derived from collision theory [5]. The classical theory of reactions in solution gives similar expressions for the reaction probabilities [1, 31] from which the propensities are derived. The reaction propensity can be written in the form [6, 7]

$$a_r(\mathbf{N}) = \kappa_r h_r(\mathbf{N}),$$

where $h_r(\mathbf{N})$ is the number of different combinations of reactant molecules participating in reaction $r$ which can be formed from the set implied by the value of the state vector $\mathbf{N}$. In reactions with simple dynamics, in the limit of a large system and with suitable initial conditions, the mean value of $\mathbf{N}$ obtained by solving the master equation should agree with the solution of the corresponding mass-action rate equations. In order for this to be so, the stochastic rate constants must be related to the mass-action constants as follows:

$$\kappa_r = k_r/(N_A V)^{\phi_r - 1} \prod_{\nu_{ri}<0} (-\nu_{ri})!, \tag{7}$$

where $N_A$ is Avogadro's constant, $V$ is the system volume, $\nu_{ri}$ is the $i$th component of the stoichiometric vector of reaction $r$, and $\phi_r$ is the order of the reaction, i.e. the sum of the stoichiometric coefficients of the reactants:

$$\phi_r = -\sum_{\nu_{ri}<0} \nu_{ri}.$$

The product which appears at the end of equation (7) has to do with a difference in definition between mass-action rate constants and stochastic rate constants. In the former, the statistical factor which arises in reactions in which two or more molecules of the same species appear as reactants is absorbed into the rate constant, while in stochastic kinetics, this factor appears in the combinatorial term $h_r$. Note that this product is equal to unity if $\nu_{ri} = -1$ for all reactants. Equation (7) assumes that $k_r$ is given in molar units, which is almost always the case in solution, but not in the gas phase. In the latter case where rate constants are often given on a per molecule basis, the factor of $N_A$ is omitted.

According to equation (7), first-order stochastic and mass-action rate constants have the same value. Thus, for mechanism (1), $\kappa_{-1} = k_{-1}$, $\kappa_{-2} = k_{-2}$ and $\kappa_{-3} = k_{-3}$. Applying equation (7) to the second-order rate constants on the other hand, we get $\kappa_1 = k_1/V$, $\kappa_2 = k_2/V$ and $\kappa_3 = k_3/V$. The factors $h_r$ for our reactions are as follows: $h_1 = N_E N_S$, $h_{-1} = h_{-2} = N_C$, $h_2 = N_E N_Q$, $h_3 = N_E N_X$, and $h_{-3} = N_H$. If our state vector is given by equation (4), then $N_E$, $N_Q$ and $N_H$ are calculated using equations (3). Furthermore, the stoichiometric vectors corresponding to the full state vector (2) are given by

$$\left[\boldsymbol{\nu}_1\ \boldsymbol{\nu}_2\ \boldsymbol{\nu}_3\right] = \begin{bmatrix} 1 & 1 & 0 \\ -1 & -1 & -1 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

The stoichiometric vectors for the reverse reactions are just $\boldsymbol{\nu}_{-i} = -\boldsymbol{\nu}_i$.

Putting it all together, we get the following chemical master equation for the CI mechanism, first studied by Jachimowski and coworkers [32]:

$$\begin{aligned}\frac{\mathrm{d}P(N_S,N_C,N_X)}{\mathrm{d}t} &= \kappa_1(N_E+1)(N_S+1)P(N_C-1,N_S+1,N_X)\\ &\quad+\kappa_{-1}(N_C+1)P(N_C+1,N_S-1,N_X)\\ &\quad+\kappa_2(N_E+1)(N_Q+1)P(N_C-1,N_S,N_X)\\ &\quad+\kappa_{-2}(N_C+1)P(N_C+1,N_S,N_X)\\ &\quad+\kappa_3(N_E+1)(N_X+1)P(N_C,N_S,N_X+1)\\ &\quad+\kappa_{-3}(N_H+1)P(N_C,N_S,N_X-1)\\ &\quad-P(N_C,N_S,N_X)\left[\kappa_1N_EN_S+\kappa_{-1}N_C+\kappa_2N_EN_Q\right.\\ &\qquad\left.+\kappa_{-2}N_C+\kappa_3N_EN_X+\kappa_{-3}N_H\right].\end{aligned}$$

In this equation, as noted above, $N_E$, $N_Q$ and $N_H$ are calculated using the conservation relations (3).

Several observations can be made about the chemical master equation:

1. The master equation is *linear* and can be written in the form

$$\dot{\mathbf{P}} = \mathbf{R}\mathbf{P}, \tag{8}$$

   where $\mathbf{P}$ is the vector of probabilities of the states defined by (4), and $\mathbf{R}$ is a constant coefficient matrix. In principle, solving the CME is therefore trivial since its solutions can be written as a superposition of exponential decay modes along the eigenvectors of $\mathbf{R}$, with amplitudes determined by the initial conditions.
2. The master equation will typically be a *huge* set of differential equations. The number of independent molecular populations in a chemical mechanism is equal to the number of chemical reactions (counting the forward and reverse reactions as one reaction). The number of states is the number of different sets of populations which satisfy the conservation and non-negativity constraints. This number is clearly of order $\xi^\rho$, where $\xi$ is an extensivity parameter (e.g. the volume) and $\rho$ is the number of reactions. Even for small $\rho$, this number grows very quickly, and of course we may be interested in reaction networks, such as those in living cells, where the populations may be relatively small but where the number of reactions is very large. To illustrate this point, figure 2 shows how the number of states of the CI mechanism grows as we increase the volume at fixed concentrations. Note the very small volumes considered in this figure. To put these numbers in perspective, consider that the volume of a mitochondrion or of a small bacterium is about $10^{-15}$ L [33, 34]. As a result of the very large size of the CME, only a few special cases which can be solved analytically have been studied in detail. Numerical solutions of the CME are rarely seen, simulations of the Markov process being preferred.
3. The master equation is *sparse*. While the size of the matrix $\mathbf{R}$ clearly grows as the square of the number of states, and thus as $\xi^{2\rho}$, the number of nonzero terms in this matrix is only $2\rho+1$ times the number of states,

i.e. the number of nonzero terms also grows as $\xi^\rho$. This pattern of growth is also illustrated for the CI mechanism in figure 2. We can exploit the fact that matrix $\mathbf{R}$ is sparse by using techniques specially designed for such matrices.

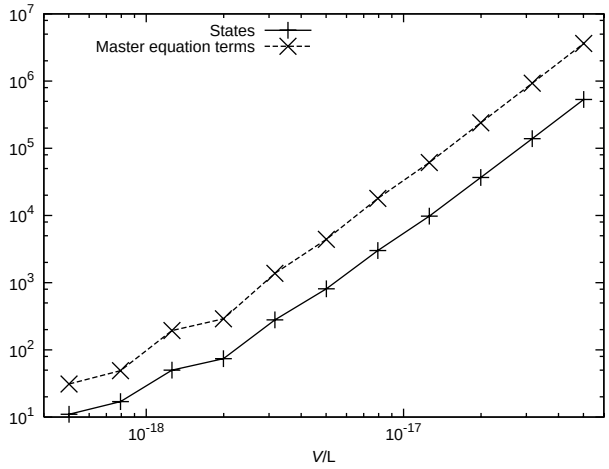



Fig. 2: Number of states and of nonzero terms in the matrix $\mathbf{R}$ as a function of volume for the CI mechanism at the following total concentrations of substrate, enzyme and inhibitor: $[\mathrm{S}]_0 = 10\,\mu$mol/L, $[\mathrm{E}]_0 = 2\,\mu$mol/L and $[\mathrm{X}]_0 = 3\,\mu$mol/L.

### 2.1 Solution Structure and Model Reduction Strategy

As mentioned above, the solutions of the linear system (8) can be written in the form

$$\mathbf{P} = \sum_{i=1}^{\sigma} a_i e^{\lambda_i t} \mathbf{e}_i,$$

where $\sigma$ is the number of states, $\lambda_i$ is an eigenvalue of $\mathbf{R}$, and $\mathbf{e}_i$ is the corresponding eigenvector. In a closed chemical system, the eigenvalues of $\mathbf{R}$ all have negative real parts, except for a zero eigenvalue associated with probability conservation. The eigenvector associated with the zero eigenvalue is the equilibrium probability distribution of the system. Modes with larger (more negative) eigenvalues are exhausted sooner than those with smaller eigenvalues (borrowing language from the literature on the computational singular perturbation method [35]). Accordingly, after some time, the solutions tend to

$$\mathbf{P} \approx \sum_{i=1}^{d} a_i e^{\lambda_i t} \mathbf{e}_i, \tag{9}$$

assuming that we have ordered the eigenvalues from smallest to largest and that only $d$ modes remain active. Another way of understanding this equation is that the solutions are eventually confined to the $d$-dimensional hyperplane defined by the $d$ leading eigenvectors of $\mathbf{R}$. This hyperplane is the $d$-dimensional *slow eigenspace* of $\mathbf{R}$. Provided there is a reasonable separation between the time scales $1/\Re(\lambda_d)$ and $1/\Re(\lambda_{d+1})$, where $\Re()$ represents the real part, and assuming that we are only interested in motion on the time scale $1/\Re(\lambda_d)$ or slower, then a satisfactory description of the dynamics can be obtained by confining the model to this eigenspace, which we denote $\mathcal{S}$. As emphasized in our earlier work, the slow eigenspace is a slow invariant manifold of the CME [15]. Note that equation (9) cannot be used directly to obtain a reduced model because of numerical instabilities which arise in very large problems.

On the $d$-dimensional slow eigenspace, we can choose $d$ independent variables such that the remaining $\sigma - d$ variables can be written as functions of the former set. Specifically, we let

$$\mathbf{P} = \begin{bmatrix} \mathbf{u} \\ \mathbf{m} \end{bmatrix}, \tag{10}$$

where $\mathbf{u}$ are the independent and $\mathbf{m}$ the dependent variables. Note that this may require a reordering of the original variables. Because the master equation is linear, we can in principle compute a matrix $\mathbf{M}$ such that

$$\mathbf{m} = \mathbf{M}\mathbf{u} \tag{11}$$

on $\mathcal{S}$. Using this relationship, we can replace the variables in the vector $\mathbf{m}$ wherever they appear in the rate equations for $\mathbf{u}$ by expressions involving only variables in $\mathbf{u}$. Thus, the original model has been reduced from one in $\sigma$ variables to one in $d$ variables.

There are several technical issues to be resolved in order to bring this program to fruition:

1. We have to choose $d$. In principle, we could choose any $d < \sigma$ we like, provided we do not split a complex-conjugate or degenerate pair of eigenvalues in so doing. It is best however if we exploit a large gap in the eigenvalue spectrum where there is a clear separation of time scales. Moreover, there is very little point in constructing a reduced model unless $d \ll \sigma$.
2. We need to choose the independent variables $\mathbf{u}$.
3. Having chosen $d$ and $\mathbf{u}$, we must compute a basis for the slow eigenspace $\mathcal{S}$, then obtain the reduced rate equations for $\mathbf{u}$.
4. Finally, we need to generate initial conditions for the reduced model such that the trajectory computed for this model shadows that of the full model after decay of transients.

At each step, we want to avoid dealing with the full model. We could for instance generate the initial conditions of the reduced model by integrating

the full model until the trajectory so generated approaches $\mathcal{S}$. In our view, the reduced model should, insofar as this is possible, be a complete model which includes, among other things, a prescription for computing the initial conditions in the context of that model alone. To do otherwise loses much of the advantage of generating a reduced model, particularly for problems such as this one where the full model is huge and awkward to handle.

In the following sections, we will demonstrate how each of these problems can be addressed. In some cases, our solutions are fully developed. In others, there is still considerable room for research. The general procedure was first described in our earlier paper [15]. However, we have made a number of improvements since this paper appeared which we describe for the first time here.

## 3 Methods and Results

### 3.1 Choosing the Dimension of $\mathcal{S}$ and the Reduced Model Variables

Ideally, we would choose $d$, the dimension of the slow eigenspace $\mathcal{S}$, based on a pair of criteria: First, we would like $d$ to include all modes up to a gap in the eigenvalue spectrum of $\mathbf{R}$. Secondly, we would like to include all modes whose time scales are observable in the context of a given experiment. The latter is beyond the scope of this paper. We thus focus on the former problem.

The qualitative appearance of the spectrum of $\mathbf{R}$ depends greatly on the structure of the reaction mechanism, less so on the values of the rate constants, and very little on either the extensivity parameter or on the initial concentrations of the various species appearing in the mechanism. This is an empirical observation, but one which is extremely useful. Among other useful consequences, this observation allows us to study how the spectrum behaves for a few small examples, i.e. at smaller values of $V$, and then to apply this knowledge directly for larger systems.

Figure 3 shows the eigenvalue spectrum of the matrix $\mathbf{R}$ for the CI mechanism. For these values of the parameters, there are 1248 states, and thus 1248 eigenvalues of $\mathbf{R}$. A first large gap occurs after the 37th eigenvalue (including the leading zero eigenvalue not shown in the figure). For the parameters of this figure, $37 = N_{S0} + 1$. Smaller gaps recur every 37th eigenvalue at first, then additional large gaps occur farther on in the spectrum. If we repeat this calculation with different parameters, the sizes of the gaps vary, but we typically find the same qualitative picture. In particular, there is almost always a prominent gap after the $(N_{S0} + 1)$st eigenvalue. We can guess why: Relaxation to the equilibrium point requires stepwise adjustments in the numbers of substrate and product molecules. Including zero, there are $N_{S0} + 1$ possible values of $N_S$ or $N_Q$. This behavior is a stochastic modeling counterpart of the formation of a one-dimensional slow manifold in the mass-action ODEs

for the CI mechanism [36]. Of course, it is possible to find parameter values where the spectrum has quite a different appearance. Figure 4 shows an example where the spectrum has no prominent gaps. The parameters used to draw this figure correspond to slow [37, 38, 39] or sluggish inhibition [36], which is closely related to the hysteretic enzyme concept [40, 41, 42]. In the deterministic counterpart of this case, relaxation to equilibrium cannot be reduced globally to motion along a one-dimensional slow manifold [36]. There is however a globally defined two-dimensional slow manifold which attracts all trajectories. Interestingly, the separation of time scales which leads to the appearance of a two-dimensional slow manifold in the deterministic system is not apparent here. In such cases, one can arbitrarily decide to keep only a certain number of modes, perhaps based on the observational time scale of a proposed experiment. It is however not that easy to identify these cases a priori, particularly since the computation of the full eigenvalue spectrum, as we have done here for illustration, becomes very difficult as the size of the matrix $\mathbf{R}$ grows. The best advice one can give at this time is to either choose $d$ based on heuristics developed from small, typical cases as we have done here, or else to use sparse matrix techniques to find a suitable $d$ which includes all time scales of interest. We have not yet pursued the treatment of this case in any detail.

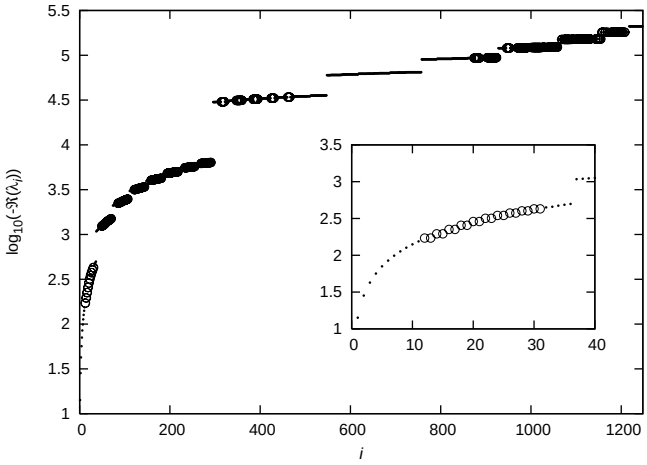



Fig. 3: Real parts of the eigenvalues of $\mathbf{R}$ plotted vs the eigenvalue number for the CI mechanism with $k_1 = 10^7\,\mathrm{M^{-1}s^{-1}}$, $k_{-1} = 10\,\mathrm{s^{-1}}$, $k_2 = 10^6\,\mathrm{M^{-1}s^{-1}}$, $k_{-2} = 3 \times 10^4\,\mathrm{s^{-1}}$, $k_3 = 10^8\,\mathrm{M^{-1}s^{-1}}$, $k_{-3} = 700\,\mathrm{s^{-1}}$, $[\mathrm{S}]_0 = 10\,\mu\mathrm{M}$, $[\mathrm{E}]_0 = 2\,\mu\mathrm{M}$, $[\mathrm{X}]_0 = 3\,\mu\mathrm{M}$, and $V = 6 \times 10^{-18}\,\mathrm{L}$. For these values of the parameters, $N_{S0} = 36$. The zero eigenvalue is omitted from the plot. Dots represent real eigenvalues while open circles are used for complex eigenvalues. The inset shows the first 40 non-zero eigenvalues at greater magnification.

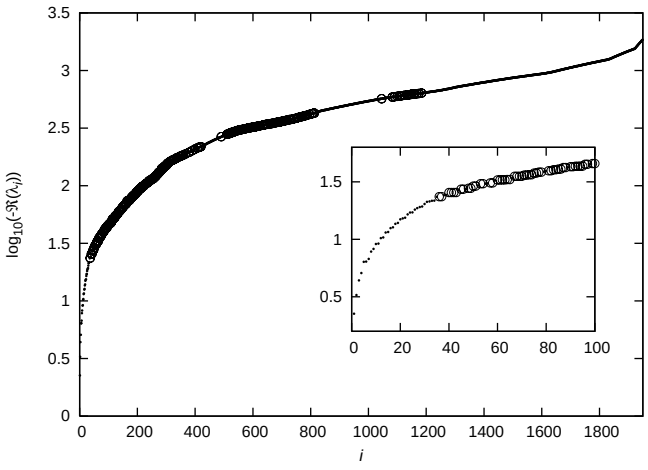



Fig. 4: Real parts of the eigenvalues of $\mathbf{R}$ plotted vs the eigenvalue number for the CI mechanism with $k_1 = 10^7\,\mathrm{M}^{-1}\mathrm{s}^{-1}$, $k_{-1} = 7\,\mathrm{s}^{-1}$, $k_2 = 4\times10^6\,\mathrm{M}^{-1}\mathrm{s}^{-1}$, $k_{-2} = 26\,\mathrm{s}^{-1}$, $k_3 = 5\times10^5\,\mathrm{M}^{-1}\mathrm{s}^{-1}$, $k_{-3} = 2\,\mathrm{s}^{-1}$, $[\mathrm{S}]_0 = 30\,\mu\mathrm{M}$, $[\mathrm{E}]_0 = 2\,\mu\mathrm{M}$, $[\mathrm{X}]_0 = 1\,\mu\mathrm{M}$, and $V = 5\times10^{-18}\,\mathrm{L}$. For these values of the parameters, $N_{S0} = 90$. The zero eigenvalue is omitted from the plot. Dots represent real eigenvalues while open circles are complex eigenvalues. The inset provides a magnified view of the first 100 eigenvalues.

Given that, most of the time, the leading eigenvalues are associated with the stepwise adjustment in the balance between substrate and product molecules, a choice of independent variables for the reduced model suggests itself immediately: We should include in $\mathbf{u}$ one probability variable for each value of $N_S$. We therefore define $\mathbf{u}$ as follows:

$$u_i = P(0, i-1, N_{X0}), \qquad i = 1, 2, \ldots, N_{S0}+1. \tag{12}$$

The choice of $N_C = 0$ and $N_X = N_{X0}$ is arbitrary. No doubt a smarter choice could be made, but we have so far not found an algorithm which does better than this simple ad hoc selection.

### 3.2 Evolution Equation on the Slow Eigenspace

Having chosen $d$, we need to compute a basis for the slow eigenspace. The leading eigenvectors are, of course, such a basis. The leading eigenvectors can, moreover, be computed efficiently using sparse matrix techniques. There is however a problem: Since we are dealing in very high-dimensional spaces, it will usually be the case that some of these eigenvectors will be nearly degenerate. We therefore can't use them directly as a basis for the slow eigenspace $\mathcal{S}$ since we would then run into problems with ill-conditioned matrices in some

steps of the calculation, nor can we easily orthonormalize them, for essentially the same reason.

A similar problem arises in the ILDM method for finding approximate slow manifolds of systems of nonlinear differential equations. The solution there is to use a Schur decomposition of the Jacobian matrix [43]. The first $n$ vectors of a Schur basis of a matrix span the same subspace as the first $n$ eigenvectors, but are orthogonal to each other. Since we want to use sparse matrix techniques which avoid a full Schur decomposition of $\mathbf{R}$, we need to compute a partial Schur basis. Moreover, since the solutions of equation (6) are real-valued, we want to avoid unnecessary complex arithmetic. We therefore prefer a real partial Schur basis. This basis can be computed using a variant of the Jacobi-Davidson method developed by van Noorden [44].

We have experimented quite a bit with van Noorden's MATLAB code (known as `rjdqr`) and, although we have not tried every conceivable combination of parameters, we can recommend the following procedures, at least as a starting point for further investigation: First, we have found that preconditioning using an incomplete LU decomposition often makes the calculation of the Schur basis unstable. We therefore do not precondition the matrix $\mathbf{R}$. Iterative algorithms like the Jacobi-Davidson method require an initial guess for the eigenvectors. We have tried a variety of physically motivated trial vectors, but finally found that what worked best was simply to start with a set of random vectors, which are then orthonormalized. Satisfactory results are however not always obtained using a random orthonormal set, a point to which we will return shortly. Internally, `rjdqr` uses a Krylov subspace iteration method to solve a correction equation. The code is written to accommodate either GMRES [45] or BiCGStab($\ell$) [46]. We have found the latter to be more stable in this application.

Let the basis of $\mathcal{S}$, computed as described above, be $\mathbf{V}$. Once we have this basis, it is a relatively straightforward exercise in linear algebra to compute the reduced model. The matrix representation of $\mathbf{V}$ has dimensions $\sigma \times d$. We define the matrix $\mathbf{V_u}$ to be the $d \times d$ submatrix of $\mathbf{V}$ whose rows correspond to the components of $\mathbf{u}$. Similarly, $\mathbf{V_m}$ is the $(\sigma - d) \times d$ submatrix of $\mathbf{V}$ whose rows correspond to the components of $\mathbf{m}$. The matrix $\mathbf{M}$, which gives the value $\mathbf{m}$ on $\mathcal{S}$ given $\mathbf{u}$ (equation (11)), is then easily seen to be

$$\mathbf{M} = \mathbf{V_m} \mathbf{V_u}^{-1}.$$

As is well known however, matrix inversion is an ill-behaved operation, particularly for large matrices. If we multiply both sides of the above equation by $\mathbf{V_u}$ on the right, we obtain $\mathbf{M}\mathbf{V_u} = \mathbf{V_m}$. After taking a transpose, we get

$$\mathbf{V_u}^{\mathrm{T}} \mathbf{M}^{\mathrm{T}} = \mathbf{V_m}^{\mathrm{T}}.$$

$\mathbf{M}$ is obtained by solving the above linear equation.

To derive the evolution equation, first define the submatrices $\mathbf{R_{u,u}}$ and $\mathbf{R_{u,m}}$ of $\mathbf{R}$ such that

$$\dot{\mathbf{u}} = \mathbf{R}_{\mathbf{u},\mathbf{u}}\mathbf{u} + \mathbf{R}_{\mathbf{u},\mathbf{m}}\mathbf{m}.$$

Given the relationship (11), this becomes

$$\dot{\mathbf{u}} = \mathbf{K}\mathbf{u},$$

with

$$\mathbf{K} = \mathbf{R}_{\mathbf{u},\mathbf{u}} + \mathbf{R}_{\mathbf{u},\mathbf{m}}\mathbf{M}. \tag{13}$$

### 3.3 Initial Conditions for the Reduced Model

We can construct a sensible ansatz for the initial condition of the reduced model based on a simple piece of physical reasoning. There are two common ways to set up an enzyme kinetic experiment. We can start with three separate solutions (one of enzyme, one of substrate and one of inhibitor) which are mixed rapidly at $t = 0$. Alternatively, we can pre-incubate the enzyme with the inhibitor and mix in a substrate solution at $t = 0$. In either case, the reaction starts with an induction period during which the enzyme-substrate complex accumulates. Since, in typical experiments, the rise time of the complex C is short, not much product will be formed during this transient period. Thus, we expect

$$F = \langle N_S \rangle + \langle N_C \rangle \tag{14}$$

to still be relatively large after decay of transients. Our approach is simply to maximize this quantity on $\mathcal{S}$ subject to the constraints $P(N_S, N_C, N_X) > 0 \ \forall \ (N_S, N_C, N_X)$ and $\sum P(N_S, N_C, N_X) = 1$. Note that this is a linear programming problem since moments of the distribution can be computed by

$$\langle N_i^k \rangle = \sum_{\mathbf{N}} N_i^k P(\mathbf{N}).$$

Moreover, each of the $P(\mathbf{N})$ can be expressed in terms of the reduced set $u$: If we order the components of $\mathbf{P}$ as in equation (10), then

$$\mathbf{P} = \mathbf{\Pi}\mathbf{u},$$

where

$$\mathbf{\Pi} = \begin{bmatrix} \mathbf{I}_d \\ \mathbf{M} \end{bmatrix},$$

and $\mathbf{I}_d$ is the $d \times d$ identity matrix. The initial condition can therefore be computed using standard linear programming software. We used the GNU Linear Programming Kit (GLPK) [47] accessed in MATLAB through the GLPKMEX interface [48]. The time required to solve the linear program is too small to be reliably measured, and is negligible in the context of the overall computation.

### 3.4 Obtaining an Accurate Reduction

If we carry out the procedure described above to obtain a reduced model using just one set of randomly generated trial vectors, we get very inconsistent results. Figure 5 shows the standard deviation of $N_S$ defined by

$$\sigma_S^2 = \sum_{\mathbf{N}} \left(N_S - \langle N_S \rangle\right)^2 P(\mathbf{N})$$

for two different computations of the reduced model, along with the exact result obtained by integrating the full model. We show this particular statistic because it seems particularly sensitive to errors in the computed Schur basis. Note that one of the two computations disagrees very badly with the full model, particularly at small times. It is also relatively easy to find instances where the linear program used to construct the initial condition gives the nonsensical result $\mathbf{u}(0) = 0$. The problem is the classical curse of dimensionality: In the very high-dimensional phase space of this problem, the subspace spanned by a set of randomly chosen trial vectors may have only a very small projection onto some of the Schur vectors defining $\mathcal{S}$. Because of the iterative method of solution, this tends to affect the accuracy of the last few vectors computed, i.e. of those vectors corresponding to the faster of the $d$ eigenmodes retained.

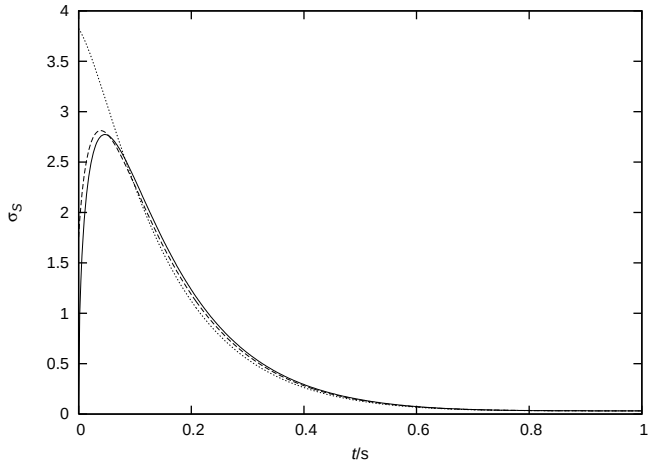



Fig. 5: Standard deviation of $N_S$ vs time for the full model (solid), and two different computations of the reduced model (dashed and dotted curves). The parameters were as in Fig. 3, except $V = 5 \times 10^{-18}$ L. The full model was integrated using MATLAB's `ode15s`, a stiff integrator, with the initial conditions $P(N_{S0}, 0, N_{X0}) = 1$ and $P(N_S, N_C, N_X) = 0$ for all other $\mathbf{N}$. The reduced model was integrated using the non-stiff solver `ode45`.

Fortunately, it is relatively easy to detect an inaccurate reduction: In the bad cases, the maximum value of $F$ found by our linear program is significantly smaller than $N_{S0}$. For instance, in figure 5 where $N_{S0} = 30$, the dotted curve corresponds to an optimum value of $F$ of 23.9, whereas the dashed curve corresponds to $F = 27.5$. The amount by which $F$ will be smaller than $N_{S0}$ in an optimal reduction will vary with the parameters of the model since this depends on the average rate of product formation during the transient phase. Nevertheless, the general principle enunciated here should hold for a wide variety of chemical systems in which reduction is feasible.

Since inaccurate reductions are not overwhelmingly common, we have adopted a simple trial-and-error strategy in which we first compute the Schur basis with a very loose tolerance, and then tighten the tolerance if the initial computation produces a reasonable value of $F$. We check again if $F$ is sufficiently large. If not, we start over with a new set of random trial vectors. Figure 6 shows the evolution of the standard deviation for several reduced models obtained by the rejection algorithm with different randomly generated trial vectors. Note that the results are now much more consistent. Interestingly, the reduced model slightly underestimates the maximum in the standard deviation. Unfortunately, this rejection algorithm involves significant overhead: On a 2.8 GHz Pentium 4 with 1 Gb of physical memory running Linux, computing the reduced model took between 83 and 1674 s, depending on how many iterations it took to hit on a good set of trial vectors. The mean and standard deviation of the computation time were 680 and 587 s, respectively, corresponding to a mean iteration count of 15 with a standard deviation of 13. Integrating the full model only took 8.7 s. There are several reasons why these statistics should not discourage us from further investigation of this method:

1. We are using an interpreted version of `rjdqr`. A compiled version would run much faster. We have also not exhausted the possibilities for optimizing the overall procedure.
2. Integration of the reduced model takes a negligible amount of time (too small to obtain reliable statistics) both because of the size reduction of the system and because of a reduction in stiffness. Among other benefits, we can use a non-stiff integrator on the reduced system as noted in the caption to figure 5. Accordingly, in studies where we want to integrate the reduced model repeatedly, e.g. to study the evolution from different initial distributions, using the reduced model might still be advantageous since the reduction only has to be done once for a given set of parameters. (In such a study, the method for generating a sensible initial distribution described in section 3.3 might only be used to test the quality of the reduction as described in section 3.4.)

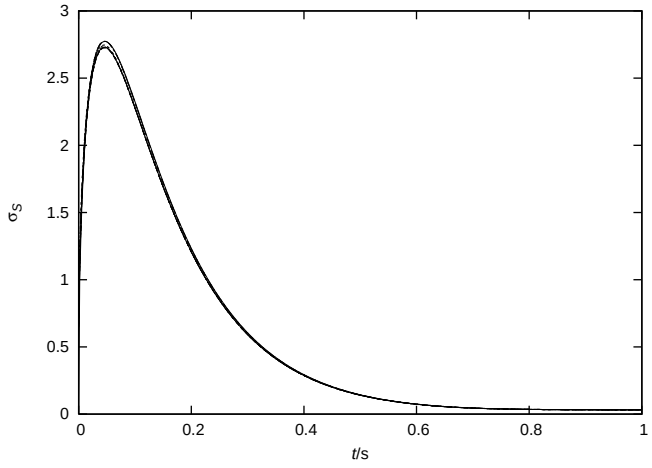



Fig. 6: Standard deviation of $N_S$ vs $t$ for the full model (solid) and for several computations of the reduced model using the rejection algorithm described in the text. The parameters were as in figure 5. The initial computation was carried out with a tolerance of $10^{-3}$ in `rjdqr`. The reduced model was rejected at this step if $F < 0.7N_{S0}$. The Schur vectors were refined to a tolerance of $10^{-8}$ with a rejection threshold at this step of $0.95N_{S0}$.

## 4 Conclusions

We have presented a method for computing a reduced stochastic model and found that in typical cases for the CI mechanism, $N_{S0}+1$ states are sufficient to faithfully reproduce the dynamics of the full model after decay of transients. Since the number of molecules of any given type scales like the extensivity parameter $\xi$, the number of variables in our reduced model also scales like $\xi$ instead of the $\xi^\rho$ scaling of the full probability space. The number of terms in the master equation (non-zero entries of the matrix $\mathbf{R}$) also scales as $\xi^\rho$, where $\rho$ is the number of chemical reactions in the model. On the other hand, the matrix $\mathbf{K}$ which governs the dynamics of the reduced model, defined by equation (13), is not sparse. Thus it will typically have a number of non-zero entries which scales as $\xi^2$. We would therefore expect the methods developed here to be particularly advantageous for models with $\rho \gg 2$. Already with $\rho = 3$, we see that the reduced model is enormously more efficiently simulated than the full master equation, again due in large part to the reduction in stiffness. The initial generation of the reduced model is the only step whose efficiency is of concern.

We have shown how a physically sensible initial condition for the reduced master equation can be generated without integrating the full model. We believe that this method can be generalized to a wide variety of cases, including ordinary mass-action chemical models. For instance, in the mass-action model

corresponding to the mechanism (1), the one-dimensional slow manifold, when it exists [36], can be parameterized by $s$, the concentration of substrate. Although we have not tested this idea, it seems likely that the value of $s$ which satisfies $s + c_{\mathcal{M}}(s) = s_0$, where $c_{\mathcal{M}}(s)$ is the equation of the manifold, will provide a good initial condition for the reduced model. This construction shows that the initial condition ansatz proposed here is in fact essentially a refined version of the reactant stationary approximation [20] corrected to satisfy mass conservation relationships.

There are still some significant technical challenges to address to make this reduction method truly practical. Currently, we choose $d$ and the set of independent variables $\mathbf{u}$ heuristically (equation (12)). This is of course not very satisfying. While it is fairly obvious how we could find $d$ adaptively, simply by computing the eigenvalues one at a time until we find a spectral gap, it is much less obvious how to choose the independent variables automatically.

The other major challenge to be resolved involves the generation of the basis for the slow eigenspace. We either need to find a clever way to generate the trial vectors which is more likely to generate a good starting point for iterative extraction of the Schur basis, or we need to use a different method altogether to generate this basis. When we do have a good basis, we get excellent results from the reduced model with our physically motivated initial condition. We are therefore encouraged to pursue this line of investigation.

*Acknowledgement.* We gratefully acknowledge Dr Tycho van Noorden of the Technische Universiteit Eindhoven, both for providing his MATLAB code for computing a real partial Schur basis, and for helpful discussions. We would also like to thank Catharine J. Roussel for useful discussions. This work was funded by an Alberta Ingenuity Fellowship to R.Z., and by a grant from the Natural Sciences and Engineering Research Council of Canada to M.R.R.

## References

1. J.I. Steinfeld, J.S. Francisco, W.L. Hase: *Chemical Kinetics and Dynamics*. Second edn. (Prentice Hall, Upper Saddle River, NJ 1999)
2. M.E. Tuckerman: *Ab Initio* molecular dynamics: Basic concepts, current trends and novel applications. *J. Phys.: Condens. Matter* **14**, R1297–R1355 (2002)
3. T. Ziegler: Tools of the trade in modeling inorganic reactions. From balls and sticks to HOMO's and LUMO's. *J. Chem. Soc., Dalton Trans.*, 642–652 (2002)
4. I. Oppenheim, K.E. Shuler, G.H. Weiss: *Stochastic Processes in Chemical Physics: The Master Equation*. (MIT Press, Cambridge, MA 1977)
5. D.T. Gillespie: A rigorous derivation of the chemical master equation. *Physica A* **188**, 404–425 (1992)
6. D.T. Gillespie: A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434 (1976)
7. D.T. Gillespie: Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977)

8. M.A. Gibson, J. Bruck: Efficient exact stochastic simulation of chemical systems with many species and many channels. *J. Phys. Chem. A* **104**, 1876–1889 (2000)
9. D.T. Gillespie: Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.* **115**, 1716–1733 (2001)
10. E.L. Haseltine, J.B. Rawlings: Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *J. Chem. Phys.* **117**, 6959–6969 (2002)
11. C.V. Rao, A.P. Arkin: Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *J. Chem. Phys.* **118**, 4999–5010 (2003)
12. Y. Cao, D.T. Gillespie, L.R. Petzold: The slow-scale stochastic simulation algorithm. *J. Chem. Phys.* **122**, 014116 (2005)
13. Y. Cao, D. Gillespie, L. Petzold: Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems. *J. Comput. Phys.* **206**, 395–411 (2005)
14. Y. Cao, D.T. Gillespie, L.R. Petzold: Avoiding negative populations in explicit Poisson tau-leaping. *J. Chem. Phys.* **123**, 054104 (2005)
15. M.R. Roussel, R. Zhu: Reducing a chemical master equation by invariant manifold methods. *J. Chem. Phys.* **121**, 8716–8730 (2004)
16. J.A.M. Janssen: The elimination of fast variables in complex chemical reactions. II. Mesoscopic level (reducible case). *J. Stat. Phys.* **57**, 171–185 (1989)
17. T. Shibata: Reducing the master equations for noisy chemical systems. *J. Chem. Phys.* **119**, 6629–6634 (2003)
18. C.F. Curtiss, J.O. Hirschfelder: Integration of stiff equations. *Proc. Natl. Acad. Sci. U.S.A.* **38**, 235–243 (1952)
19. C.W. Gear: *Numerical Initial Value Problems in Ordinary Differential Equations* (Prentice-Hall, Englewood Cliffs, N.J. 1971)
20. G.M. Côme: Mechanistic modelling of homogeneous reactors: A numerical method. *Computers & Chem. Eng.* **3**, 603–609 (1979)
21. S.J. Fraser: The steady state and equilibrium approximations: A geometrical picture. *J. Chem. Phys.* **88**, 4732–4738 (1988)
22. A.H. Nguyen, S.J. Fraser: Geometrical picture of reaction in enzyme kinetics. *J. Chem. Phys.* **91**, 186–193 (1989)
23. M.R. Roussel, S.J. Fraser: On the geometry of transient relaxation. *J. Chem. Phys.* **94**, 7106–7113 (1991)
24. M.R. Roussel, S.J. Fraser: Invariant manifold methods for metabolic model reduction. *Chaos* **11**, 196–206 (2001)
25. M.H. Holmes, J. Bell: The application of symbolic computing to chemical kinetic reaction schemes. *J. Comp. Chem.* **12**, 1223–1231 (1991)
26. S. Schuster, T. Höfer: Determining all extreme semi-positive conservation relations in chemical reaction systems: A test criterion for conservativity. *J. Chem. Soc., Faraday Trans.* **87**, 2561–2566 (1991)
27. R.I. Ben-Aïm, V. Viossat: A geometric representation of species concentrations in chemical kinetics. *New J. Chem.* **25**, 864–868 (2001)
28. J. Jeans: *An Introduction to the Kinetic Theory of Gases* (Cambridge University Press, Cambridge 1952)
29. C.W. Gardiner: *Handbook of Stochastic Methods*, Second edn. (Springer, Berlin Heidelberg New York 1985)
30. H.X. Zhou: Theory and simulation of the influence of diffusion in enzyme-catalyzed reactions. *J. Phys. Chem. B* **101**, 6642–6651 (1997)

31. S.A. Rice: *Diffusion-Limited Reactions,* vol. 25 of Comprehensive Chemical Kinetics (Elsevier, Amsterdam 1985)
32. C.J. Jachimowski, D.A. McQuarrie, M.E. Russell: A stochastic approach to enzyme-substrate reactions. *Biochemistry* **3**, 1732–1736 (1964)
33. R.Y. Stanier, E.A. Adelberg, J.L. Ingraham: *General Microbiology*, Fourth edn. (Macmillan, London 1977)
34. B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson: *Molecular Biology of the Cell*, Third edn. (Garland, New York 1994)
35. S.H. Lam: Using CSP to understand complex chemical kinetics. *Combust. Sci. Technol.* **89**, 375–404 (1993)
36. M.R. Roussel, S.J. Fraser: Global analysis of enzyme inhibition kinetics. *J. Phys. Chem.* **97**, 8316–8327 (1993); **98**, E5174 (1994).
37. J.W. Williams, J.F. Morrison: The kinetics of reversible tight-binding inhibition. *Meth. Enzymol.* **63**, 437–467 (1979)
38. J.F. Morrison, C.T. Walsh: The behavior and significance of slow-binding enzyme inhibitors. *Adv. Enzymol. Related Areas Mol. Biol.* **61**, 201–301 (1988)
39. S.E. Szedlacsek, R.G. Duggleby: Kinetics of slow and tight-binding inhibitors. *Meth. Enzymol.* **249**, 144–180 (1995)
40. C. Frieden: Kinetic aspects of regulation of metabolic processes — The hysteretic enzyme concept. *J. Biol. Chem.* **245**, 5788–5799 (1970)
41. C. Frieden: Slow transitions and hysteretic behavior in enzymes. *Annu. Rev. Biochem.* **48**, 471–489 (1979)
42. K.E. Neet, G.R. Ainslie, Jr.: Hysteretic enzymes. *Meth. Enzymol.* **64**, 192–226 (1980)
43. U. Maas, S.B. Pope: Simplifying chemical kinetics: Intrinsic low-dimensional manifolds in composition space. *Combust. Flame* **88**, 239–264 (1992)
44. T. van Noorden: *Computing a partial real ordered generalized Schur form using the Jacobi-Davidson method.* Preprint 1307, Utrecht University (2004)
45. Y. Saad, M.H. Schultz: GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.* **7**, 856–869 (1986)
46. G.L.G. Sleijpen, D.R. Fokkema: BiCGstab($\ell$) for linear equations involving matrices with complex spectrum. *Electron. Trans. Numer. Anal.* **1**, 11–32 (1993)
47. A. Makhorin: GNU Linear Programming Kit (GLPK), version 4.8. `http://www.gnu.org/software/glpk/glpk.html` (2005)
48. N. Giorgetti: GLPKMEX - a Matlab MEX interface for the GLPK library, version 0.7. `http://www.dii.unisi.it/~giorgetti/downloads.html#glpkmex` (2005)

# Model Reduction in Kinetic Theory

H. Struchtrup

Department of Mechanical Engineering, University of Victoria, PO Box STN CSC 3055, Victoria BC V8W 3P6, Canada, struchtr@me.uvic.ca

**Summary.** Methods to derive macroscopic transport equations for rarefied gases from the Boltzmann equations are presented. Featured methods include the Chapman-Enskog expansion, Grad's moment method, and the author's *order of magnitude method*. The resulting macroscopic equations are compared and discussed by means of simple problems, including linear stability, shock wave structures, and Couette flow.

## 1 Introduction

The most important scaling parameter to characterize processes in rarefied gases is the Knudsen number Kn, defined as the ratio between the mean free path of a particle and a relevant reference length scale (e.g. channel width, wavelength, etc.). For a complex flow the local Knudsen number might differ by several orders of magnitude between locations. Thus, rarefied gas flows are multiscale problems.

Processes in rarefied gases are well described by the Boltzmann equation [1, 2, 3], a non-linear integro-differential equation that describes the evolution of the particle distribution function $f$ in phase space, i.e. on the microscopic level. The numerical solution of the Boltzmann equation, either directly [4] or via the Direct Simulation Monte Carlo (DSMC) method [5], is very time expensive.

If the Knudsen number is small, the Boltzmann equation can be reduced to simpler models, which allow faster solutions,

If Kn $< 0.01$ (say), the equations of ordinary hydrodynamics—the laws of Navier-Stokes and Fourier (NSF)—can be derived from the Boltzmann equation. The NSF equations are macroscopic equations for mass density $\rho$, velocity $v_i$ and temperature $T$, and thus pose a mathematically less complex problem than the Boltzmann equation.

Macroscopic equations for rarefied gas flows at Knudsen numbers above 0.01 are highly desirable, since they promise to replace the Boltzmann equation with simpler equations that still capture the relevant physics. Several

methods are available to derive the desired higher order equations, and all of these suggest different sets of equations.

Naturally, the complexity of the subject forbids a detailed discussion of this rather complex topic on the space available. Thus, this contribution aims mainly at presenting the main ideas of the most important methods, and to point out the relations and differences between the various sets of equations. The reader searching for greater detail is referred to the cited literature, in particular to the author's textbook [3].

## 2 Basic Kinetic Theory

We shall consider mon-atomic ideal gases exclusively. The basic quantity in kinetic theory is the particle distribution function $f\left(\mathbf{x},t,\mathbf{c}\right)$; $\mathbf{x}$ and $t$ are the space and time variables, respectively, and $\mathbf{c}$ denotes the microscopic velocities of particles. The distribution function is defined such that $f\left(\mathbf{x},t,\mathbf{c}\right)d\mathbf{c}d\mathbf{x}$ gives the number of gas particles in the phase space cell $d\mathbf{c}d\mathbf{x}$ at time $t$.

Macroscopic quantities are obtained by taking suitable averages (moments) of the phase density. The basic hydrodynamic variables are obtained according to

$$\rho = m\int f d\mathbf{c}\ ,\ \rho v_i = m\int c_i f d\mathbf{c}\ ,\ \rho u = \frac{3}{2}\rho\theta = \frac{m}{2}\int C^2 f d\mathbf{c}\ . \tag{1}$$

Here, $\theta = \frac{k}{m}T$ is the temperature in energy units (that will be used from now on instead of $T$), $m$ is the mass of a particle, $k$ denotes Boltzmann's constant, and $C_i = c_i - v_i$ is the peculiar velocity. $u$ denotes the specific internal energy, and $(1)_3$ must be considered as the definition of temperature.

The phase density is obtained as the solution of the Boltzmann equation,

$$\frac{\partial f}{\partial t} + c_i\frac{\partial f}{\partial x_i} = \mathcal{S}\left(f,f\right)\ . \tag{2}$$

Here, $\mathcal{S}$ is the collision term which describes the change of $f$ due to collisions among particles. The full expression for $\mathcal{S}$ can be found in the literature [1, 2, 3], here we only list its most important properties:

1. Mass, momentum and energy are conserved in a collision,

$$m\int \mathcal{S}d\mathbf{c} = 0\ \ ,\ \ m\int c_i\mathcal{S}d\mathbf{c} = 0\ \ ,\ \ \frac{m}{2}\int C^2\mathcal{S}d\mathbf{c} = 0\ . \tag{3}$$

2. The production of entropy is always non-negative (H-theorem),

$$\Sigma = -k\int \ln f\mathcal{S}d\mathbf{c} \geqslant 0\ . \tag{4}$$

3. In equilibrium the phase density is a Maxwellian distribution, i.e.

$$\mathcal{S} = 0 \Longrightarrow f = f_M = \frac{\rho}{m}\frac{1}{\sqrt{2\pi\theta}^3}\exp\left[-\frac{C^2}{2\theta}\right]\ . \tag{5}$$

The Boltzmann collision term $\mathcal{S}$ is a complex non-linear integral expression in $f$ that depends also on the interaction potential between the particles. Its mathematical treatment becomes particularly simple for particles interacting with a repulsive fifth-order power potential, the so-called Maxwell molecules. More realistic potentials, e.g. general power laws, hard sphere molecules, or Lennard-Jones potentials introduce higher complexity.

Simplified expressions for $\mathcal{S}$ that capture its basic properties are often used, the most popular of these is the BGK model [6] where

$$\mathcal{S}_{BGK} = \nu \left( f_M - f \right) \; ; \tag{6}$$

$\nu$ is the average collision frequency for a particle.

Multiplication of the Boltzmann equation (2) with $\left\{ m, mc_i, \frac{m}{2} C^2 \right\}$ and subsequent integration over the microscopic velocity yields the conservation laws for mass, momentum and internal energy,

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \frac{\partial \rho v_k}{\partial x_k} &= 0 \; , \\ \rho \frac{\partial v_i}{\partial t} + \rho v_k \frac{\partial v_i}{\partial x_k} + \frac{\partial p}{\partial x_i} + \frac{\partial \sigma_{ik}}{\partial x_k} &= 0 \; , \\ \frac{3}{2} \rho \frac{\partial \theta}{\partial t} + \frac{3}{2} \rho v_k \frac{\partial \theta}{\partial x_k} + \frac{\partial q_k}{\partial x_k} &= - \left( p \delta_{ij} + \sigma_{ij} \right) \frac{\partial v_i}{\partial x_j} \; . \end{aligned} \tag{7}$$

Here, the pressure $p$ obeys the ideal gas law, $p = \rho\theta$, and

$$q_i = \frac{m}{2} \int C^2 C_i f d\mathbf{c} \tag{8}$$

denotes the heat flux vector. The pressure tensor is defined as

$$p\delta_{ij} + \sigma_{ij} = m \int C_i C_j f d\mathbf{c} \quad \text{with} \quad \sigma_{ij} = m \int C_{\langle i} C_{j \rangle} f d\mathbf{c} \tag{9}$$

where $\sigma_{ij}$ denotes the stress, that is the symmetric and tracefree part of the pressure tenor, with $\sigma_{ii} = 0$, $\sigma_{ij} = \sigma_{ji}$. Indices in angular brackets denote symmetric trace-free tensors (see [3], Appendix A.2).

The conservation laws (7) together with the definitions for stress and heat flux (9, 8) are exact, that is they are valid for any solution $f$ of the Boltzmann equation. Mathematically, the five conservation laws do not form a closed set of equations for the hydrodynamic variables $\{\rho, v_i, \theta\}$, since they contain stress and heat flux as well.

The idea of macroscopic continuum approximations is to close the set of equations by deriving additional macroscopic equations for $\sigma_{ij}$ and $q_i$ from the Boltzmann equation by means of rational approximation procedures. Various methods available to this end, and the corresponding additional equations for $\sigma_{ij}$ and $q_i$, will be discussed in the sequel.

For completeness we mention that multiplication of the Boltzmann equation with $-k \ln \frac{f}{y}$ ($y$ is a constant) yields the balance of entropy, which has a non-negative production (4) [1, 2, 3].

## 3 Chapman-Enskog Method

The best known approach to derive macroscopic transport equations from the Boltzmann equation is the Chapman-Enskog (CE) method [1, 2, 3, 7, 8]. The CE method is based on the dimensionless form of the Boltzmann equation which contains the Knudsen number as a scaling parameter for the collision term,

$$\frac{\partial f}{\partial t} + c_i \frac{\partial f}{\partial x_i} = \frac{1}{\mathrm{Kn}} \mathcal{S}\left(f, f\right) \ . \tag{10}$$

In the limit $\mathrm{Kn} \to 0$, the collision term must vanish, and it follows from the properties of $\mathcal{S}$ that the corresponding phase density is the local Maxwellian (5), $f^{(0)} = f_M$. Evaluation of $\sigma_{ij}$ and $q_i$ with the Maxwellian gives zero stress and heat flux,

$$\sigma_{ij}^{(0)} = q_i^{(0)} = 0 \ . \tag{11}$$

Insertion of this into the conservation laws (7) yields the well known Euler equations.

The idea of the CE expansion method is to add corrections to the local equilibrium distribution by adding terms of higher orders in the Knudsen number,

$$f = f^{(0)} + \mathrm{Kn} f^{(1)} + \mathrm{Kn}^2 f^{(2)} + \cdots \ . \tag{12}$$

An important condition on the expansion (12) is that the hydrodynamic variables $\{\rho, v_i, \theta\}$ are the same at any level of expansion, so that

$$\rho \left\{1, v_i, \frac{3}{2}\theta\right\} = m \int \left\{1, c_i, \frac{C^2}{2}\right\} f^{(0)} d\mathbf{c} \ , \ 0 = \int \left\{1, c_i, \frac{C^2}{2}\right\} f^{(\alpha)} d\mathbf{c} \ \ (\alpha \geq 1) \, .$$

These compatibility conditions guarantee that only the equations for the non-equilibrium variables $\sigma_{ij}$ and $q_i$ change with increasing degree of approximation,

$$\sigma_{ij} = \sigma_{ij}^{(0)} + \mathrm{Kn}\sigma_{ij}^{(1)} + \mathrm{Kn}^2 \sigma_{ij}^{(2)} + \cdots \ , \quad q_i = q_i^{(0)} + \mathrm{Kn} q_i^{(1)} + \mathrm{Kn}^2 q_i^{(2)} + \cdots \ . \tag{13}$$

The expansion parameters $f^{(\alpha)}$ are determined successively by plugging the series (12) into the Boltzmann equation, and equating terms with the same factors in powers of the Knudsen number. This leads to an iterative structure, where the correction at order $\alpha$ is a function of (derivatives of) the lower order corrections, $f^{(\alpha)} = \mathcal{F}\left(f^{(\beta)}, 0 \leq \beta < \alpha\right)$, see e.g. [7, 3]. All correction terms depend only on the hydrodynamic variables and their gradients,† since the zeroth order term—the Maxwellian—depends only on the hydrodynamic variables $\{\rho, v_i, \theta\}$. Stress and heat flux are computed from the approximation (12) by accounting for terms up to a certain order, and the resulting expressions will relate $\sigma_{ij}$ and $q_i$ to the hydrodynamic variables and their gradients.

† Time derivatives are replaced by means of the conservation laws [7, 3].

Obviously, to zeroth order the expansion yields the Euler equations (11). The first order correction gives the laws of Navier-Stokes and Fourier,

$$\sigma_{ij}^{(1)} = -2\mu \frac{\partial v_{\langle i}}{\partial v_{j\rangle}} \quad , \quad q_i^{(1)} = -\kappa \frac{\partial \theta}{\partial x_i} \; . \tag{14}$$

The most important success of the CE method is that it gives accurate expressions for viscosity $\mu$ and heat conductivity $\kappa$, which relate these to the microscopic interaction potential and the hydrodynamic variables. In particular one finds, in accordance with experiments, that the viscosity depends only on temperature, and not on density. For power potentials the CE method yields

$$\mu = \mu_0 \left( \frac{\theta}{\theta_0} \right)^{\omega} \tag{15}$$

with $\omega = 1/2$ for hard spheres and $\omega = 1$ for Maxwell molecules; experiments indicate $\omega \simeq 0.8$ for argon [5]. Heat conductivity and viscosity are related through the Prandtl number,†

$$\mathrm{Pr} = \frac{5}{2} \frac{\mu}{\kappa} \simeq \frac{2}{3} \; .$$

The value of Pr varies only slightly (less than 1%) with the molecule model, and measured values are close to 0.66 [2, 3].

The second order contributions are the Burnett equations [9, 2, 7, 3],

$$\begin{aligned} \sigma_{ij}^{(2)} = \frac{\mu^2}{p} \Bigg[ & \varpi_1 \frac{\partial v_k}{\partial x_k} S_{ij} - \varpi_2 \left( \frac{\partial}{\partial x_{\langle i}} \left( \frac{1}{\rho} \frac{\partial p}{\partial x_{j\rangle}} \right) + \frac{\partial v_k}{\partial x_{\langle i}} \frac{\partial v_{j\rangle}}{\partial x_k} + 2 \frac{\partial v_k}{\partial x_{\langle i}} S_{j\rangle k} \right) \\ & + \varpi_3 \frac{\partial^2 \theta}{\partial x_{\langle i} \partial x_{j\rangle}} + \varpi_4 \frac{\partial \theta}{\partial x_{\langle i}} \frac{\partial \ln p}{\partial x_{j\rangle}} + \varpi_5 \frac{1}{\theta} \frac{\partial \theta}{\partial x_{\langle i}} \frac{\partial \theta}{\partial x_{j\rangle}} + \varpi_6 S_{k\langle i} S_{j\rangle k} \Bigg] \; , \end{aligned} \tag{16}$$

$$\begin{aligned} q_i^{(2)} = \frac{\mu^2}{\rho} \Bigg[ & \theta_1 \frac{\partial v_k}{\partial x_k} \frac{\partial \ln \theta}{\partial x_i} - \theta_2 \left( \frac{2}{3} \frac{\partial^2 v_k}{\partial x_k \partial x_i} + \frac{2}{3} \frac{\partial v_k}{\partial x_k} \frac{\partial \ln \theta}{\partial x_i} + 2 \frac{\partial v_k}{\partial x_i} \frac{\partial \ln \theta}{\partial x_k} \right) \\ & + \theta_3 S_{ik} \frac{\partial \ln p}{\partial x_k} + \theta_4 \frac{\partial S_{ik}}{\partial x_k} + 3\theta_5 S_{ik} \frac{\partial \ln \theta}{\partial x_k} \Bigg] \; . \end{aligned} \tag{17}$$

The Burnett coefficients $\varpi_\alpha$, $\theta_\alpha$ depend on the molecule type, and for power potentials with exponent $\gamma$ some values are given in Table 1 [10].

The third order expansion yields the super-Burnett equations. Their computation is extremely cumbersome, and the full three-dimensional non-linear super-Burnett equations were never derived. One only finds the linearized equations in 3-D [11, 12, 13, 3], and the non-linear equations for one-dimensional geometry [11, 14, 15, 3].

† Our defintion of the Prandtl number differs from the usual one by a factor $\frac{k}{m}$ due to the use of $\theta$ instead of $T$.

| $\gamma$ | $\omega$ | $\varpi_1$ | $\varpi_2$ | $\varpi_3$ | $\varpi_4$ | $\varpi_5$ | $\varpi_6$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 1 | 10/3 | 2 | 3 | 0 | 3 | 8 | 75/8 | 45/8 | $-3$ | 3 | 39/4 |
| 7.66 | 0.8 | 3.600 | 2.004 | 2.761 | 0.254 | 1.784 | 7.748 | 10.160 | 5.656 | $-3.014$ | 2.761 | 9.019 |
| $\infty$ | 0.5 | 4.056 | 2.028 | 2.418 | 0.681 | 0.219 | 7.424 | 11.644 | 5.822 | $-3.09$ | 2.418 | 8.286 |

Table 1: Burnett coefficients for power potentials ($\gamma = 5$ for Maxwell molecules, $\gamma = 7.66$ for argon, $\gamma = \infty$ for for hard sphere molecules) [10].

The equations of Navier-Stokes and Fourier cease to be accurate for Knudsen numbers above $\sim 0.01$, and one would expect that Burnett and super-Burnett equations are valid for larger Knudsen numbers. Unfortunately, however, the higher order equations become linearly unstable for processes involving small wavelengths, or high frequencies [12, 15], and they lead to unphysical oscillations in steady state processes [16], and thus cannot be used in numerical simulations .

There is no clear argument why the Chapman-Enskog expansion leads to unstable equations. It seems that a first order Chapman-Enskog expansion leads generally to stable equations, while higher order expansions generally yield unstable equations, although exceptions apply, e.g. see [17, 18].

Zhong et al. suggested the "augmented Burnett equations" where some terms of super-Burnett order (but not the actual super-Burnett terms) are added to the Burnett equations to stabilize these [19, 20]. The augmented Burnett equations still are unstable in space [15, 3], and they lack a rational derivation from the Boltzmann equation [15].

For reference in subsequent sections we print the distribution function for the NSF equations (with $\xi = C/\sqrt{\theta}$)

$$f_{|CE} = f_M \left[ 1 + \frac{\sigma_{ik}^{(1)}}{2p} A\left(\xi^2\right) \frac{C_{\langle i} C_{k\rangle}}{\theta} + \frac{2}{5} \frac{q_k^{(1)}}{p\theta} B\left(\xi^2\right) C_k \left( \frac{C^2}{2\theta} - \frac{5}{2} \right) \right] . \quad (18)$$

The dimensionless functions $A$ and $B$ result from the approximate inversion of the Boltzmann collision term [2, 7, 3] and thus depend on the interaction potential. For Maxwell molecules they are constants, $A = B = 1$, while for all other interaction potentials they are polynomials; fourth order polynomials give an excellent approximation [2, 10]. Setting $A = B = 1$ leads to small deviations between theory and measurement of viscosity and heat conductivity.

## 4 Grad Moment Method

The Chapman-Enskog method leads to expressions for stress and heat flux that contain higher derivatives of the hydrodynamic variables. Grad suggested a quite different approach, in which the number of variables is extended beyond the 5 hydrodynamic variables $\rho, v_i, \theta$, by adding stress $\sigma_{ij}$, heat flux $q_i$

and other moments to the list of variables [21, 22]. The corresponding transport equations for the additional variables are obtained as moments of the Boltzmann equation and are first order partial differential equations for the moments. They do not form a closed set for the variables and require a closure argument. For this Grad suggests to approximate the phase density by an expansion in Hermite polynomials about the equilibrium distribution (the local Maxwellian), where the expansion coefficients are related to the moments.

A crucial point in the method is the question which and how many moments are needed to describe a process. The answer depends on the particular process, but experience shows that the number of moments must be increased with increasing Knudsen number [23, 24, 25, 26, 27, 28]. Grad's method does not provide an argument that links the Knudsen number to the set of moments to be considered as variables.

The best known set of Grad-type moment equations is the 13 moment system, which has the variables $\{\rho, v_i, \theta, \sigma_{ij}, q_i\}$. The corresponding moment equations are obtained by multiplying the Boltzmann equation with $\left\{m, mc_i, \frac{m}{2}C^2, mC_{\langle i}C_{j\rangle}, \frac{m}{2}C^2C_i\right\}$. This gives the conservation laws (7) plus additional moment equations for stress and heat flux,

$$\frac{\partial \sigma_{ij}}{\partial t} + v_k \frac{\partial \sigma_{ij}}{\partial x_k} + \sigma_{ij}\frac{\partial v_k}{\partial x_k} + \frac{4}{5}\frac{\partial q_{\langle i}}{\partial x_{j\rangle}} + 2p\frac{\partial v_{\langle i}}{\partial x_{j\rangle}} + 2\sigma_{k\langle i}\frac{\partial v_{j\rangle}}{\partial x_k} + \frac{\partial m_{ijk}}{\partial x_k} = \mathcal{P}_{ij} \ , \quad (19)$$

$$\frac{\partial q_i}{\partial t} + v_k \frac{\partial q_i}{\partial x_k} + \frac{5}{2}p\frac{\partial \theta}{\partial x_i} + \frac{5}{2}\sigma_{ik}\frac{\partial \theta}{\partial x_k} + \theta\frac{\partial \sigma_{ik}}{\partial x_k} - \theta\sigma_{ik}\frac{\partial \ln \rho}{\partial x_k} + \frac{7}{5}q_k\frac{\partial v_i}{\partial x_k} + \frac{2}{5}q_k\frac{\partial v_k}{\partial x_i} + \frac{7}{5}q_i\frac{\partial v_k}{\partial x_k} + \frac{1}{2}\frac{\partial R_{ik}}{\partial x_k} + \frac{1}{6}\frac{\partial \Delta}{\partial x_i} + m_{ijk}\frac{\partial v_j}{\partial x_k} - \frac{\sigma_{ij}}{\rho}\frac{\partial \sigma_{jk}}{\partial x_k} = \mathcal{P}_i \ . \quad (20)$$

Equations (19, 20) contain additional moments of the distribution function, which are defined as

$$\Delta = m\int C^4 \left(f - f_M\right) d\mathbf{c} \ , \ R_{ij} = m\int \left(C^2 - 7\theta\right) C_{\langle i}C_{j\rangle} f d\mathbf{c} \ ,$$
$$m_{ijk} = m\int C_{\langle i}C_jC_{k\rangle} f d\mathbf{c} \ . \quad (21)$$

The terms on the right hand sides are the moments of the Boltzmann collision term,

$$\mathcal{P}_{ij} = m\int C_{\langle i}C_{j\rangle}\mathcal{S} d\mathbf{c} \quad , \quad \mathcal{P}_i = \frac{m}{2}\int C^2C_i\mathcal{S} d\mathbf{c} \ . \quad (22)$$

Obviously, the set of equations can be closed by finding expressions for $\Delta, R_{ij}, m_{ijk}, \mathcal{P}_{ij}, \mathcal{P}_i$ that relate these to the basic 13 variables $\{\rho, v_i, \theta, \sigma_{ij}, q_i\}$. To this end, the Grad method provides the distribution [21, 22, 3]

$$f_{|13} = f_M\left[1 + \frac{\sigma_{ik}}{2p}\frac{C_{\langle i}C_{k\rangle}}{\theta} + \frac{2}{5}\frac{q_k}{p\theta}C_k\left(\frac{C^2}{2\theta} - \frac{5}{2}\right)\right] \ . \quad (23)$$

This function recovers the basic 13 variables, and allows to compute the unknowns (21, 22) as

$$\Delta = R_{ij} = m_{ijk} = 0 \ , \ \mathcal{P}_{ij} = -\frac{p}{\mu}\sigma_{ij} \ , \ \mathcal{P}_i = -\frac{2}{3}\frac{p}{\mu}q_i \ . \tag{24}$$

Insertion of (23) into (19, 20) gives, together with (7), the closed set of equations for the 13 variables.

By comparing the distribution functions (18) and (23) it becomes evident that they are quite similar. However, there are two differences: (a) the CE phase density contains only the first approximations to stress and heat flux, $\sigma_{ij}^{(1)}$ and $q_i^{(1)}$, while the Grad distribution contains both as independent variables, $\sigma_{ij}$ and $q_i$. (b) In the Grad function, the coefficients $A$ and $B$ assume the values for Maxwell molecules, $A = B = 1$. From the last point one will infer that the Grad 13 moment equations will be best suited for Maxwell molecules, while for other molecule types they can only be an approximation.

The Grad 13 equations have two major drawbacks: (a) The equations are symmetric hyperbolic for most values of the variables, and this leads to shock structures with discontinuities (sub-shocks) for Mach numbers above 1.65 [23, 26]. (b) Since Grad's method is not linked to the Knudsen number, the range of applicability for the equations is unclear.

These problems remain for Grad-type equations with more variables, which give smooth shocks up to higher but not too high Mach numbers [26, 27]. The 13 moment equations do not describe Knudsen boundary layers [29, 30, 24], increasing the number of moments allows to compute these [31, 24, 32].

For some problems, in particular for large Mach or Knudsen numbers, one has to face hundreds of moment equations, but the relation between moment number and Knudsen or Mach number is not clear. Computations for hundreds of moments are only manageable for simple geometries and problems [33, 23, 24], and were never performed in two or three dimensions. Indeed, the goal of a macroscopic set of equations must be to have a simplification compared to the Boltzmann equation, and using hundreds of moments does not achieve this goal.

## 5 Combining the Chapman-Enskog and Grad Methods

In most of the available literature, the two classical methods—Grad moment method and Chapman-Enskog expansion—are treated as being completely unrelated. However, using a method akin to the Maxwellian iteration of Truesdell and Ikenberry [34, 35], Reinecke and Kremer extract the Burnett equations from Grad-type moment systems [10, 36]. Which set of moments one has to use for this purpose depends on the interaction potential. For Maxwell molecules it is sufficient to consider Grad's set of 13 moments.

In [30] it was shown that this iteration method is equivalent to the CE expansion of the moment equations. In the original CE method one first expands, and then integrates the resulting distribution function to compute its moments. In the Reinecke-Kremer-Grad method, the order of integration and expansion is exchanged.

For Maxwell molecules the Burnett equations result from the second order CE expansion of the 13 moments set, while the super-Burnett equations result from the 3rd order CE expansion of the 26 moment set (which adds $\Delta, R_{ij}, m_{ijk}$ to the list of variables) [30, 13, 15, 3].

While the Reinecke-Kremer-Grad method does not give new results, it allows an easier access to higher order CE expansions, in particular the Burnett equations. The method does not solve the stability problems of the Burnett equations

M. Torrilhon and the present author used a different way to combine the two methods by assuming different time scales for the 13 basic variables of the theory on one side, and all higher moments on the other [13, 15, 3]. This allows to perform a CE expansion around a non-equilibrium state which is defined through the 13 variables. This method, which appeared first as a side note in Grad's contribution to the Encyclopedia of Physics [22], gives a regularizing correction to the Grad13 equations, the regularized 13 moment equations (R13 Eqs.).

The same idea was used by Karlin et al. [37] for the linearized Boltzmann equation. They compute an approximation to the distribution function, which is used to derive a set of 13 linear equations for the 13 moments. Their equations are the linearized form of the R13 equations.

The R13 equations are not hyperbolic, give smooth shock structures for all Mach numbers, and they are stable. Therefore, this combination of the CE and Grad methods yields a marked improvement over the original methods. The R13 equations will be shown and discussed later, in Sec. 6, which presents an alternative method of derivation.

The Grad distributions, e.g. the 13 moment phase density $f_{|13}$ (23), define non-equilibrium manifolds in phase space [37, 38]. The Maxwellians form a subset on these non-equilibrium manifolds (they are the appropriate Grad distribution for the 5 moments case, i.e. the Euler equations). The Grad closure restricts the phase space so that the gas cannot access all states in phase space, but only those on the Grad non-equilibrium manifold. This strong restriction is inherent to Grad's closure, and has no physical foundation, since Grad distributions cannot be extracted from the Boltzmann equation. This is different, of course, for the Maxwellians, which are those phase densities that give a zero collision term. With no argument from physics to support the Grad distributions, it seems to be daring to restrict the gas on the Grad non-equilibrium manifolds. One way to relax the Grad assumption—at least somewhat—is to allow states in the *vicinity* of the Grad manifolds. This stands in analogy to the relation between Euler and NSF equations, which describe the equilibrium manifold, and its vicinity.

It must be emphasized that there is no evidence in physics to support the existence of pseudo-equilibrium manifolds for the gas. In particular there is no guideline for choosing the relevant moments, or the pseudo-equilibrium distribution, which could, e.g., be a Grad distribution with any number of moments. A somewhat popular alternative are distribution functions that result from maximizing entropy or extended thermodynamics, see [23, 39, 40, 41, 42, 43] as well as [3] for details on, and problems associated with, this approach.

## 6 Order of Magnitude Method

The weak point in the Grad method is that no statement is made to connect Knudsen numbers and relevant moments. As a result, the derivation of the R13 equations as outlined above required the assumption of different time scales for the basic 13 moments, and higher moments. While this assumption leads to a set of equations with desired behavior, it is difficult to justify, since the characteristic times of all moments are, in fact, of the same order.

An alternative approach to the problem was presented by Struchtrup in [44, 45, 3], partly based on earlier work by Müller et al. [46].

The *order of magnitude method* considers not the Boltzmann equation itself, but its infinite system of moment equations for symmetric and trace-free moments

$$u^{(a)}_{\langle i_i \cdots i_n \rangle} = m \int C^{2a} C_{\langle i_1} \cdots C_{i_n \rangle} f d\mathbf{c} \quad (a, n = 0, 1, 2, \ldots) \,. \tag{25}$$

Here, due to space restrictions, we cannot present the method in detail, but only describe its main steps; in particular we shall not show the general moment equations for the moments (25).

The method of finding the proper equations with *order of accuracy* $\lambda_0$ in the Knudsen number consists of the following three steps:

1. Determination of the *order of magnitude* $\lambda$ of the moments.
2. Construction of a moment set with minimum number of moments at any order $\lambda$.
3. Deletion of all terms in all equations that would lead only to contributions of orders $\lambda > \lambda_0$ in the conservation laws for energy and momentum.

Step 1 is based on a Chapman-Enskog expansion where a moment $\phi$ is expanded according to

$$\phi = \phi_0 + \mathrm{Kn}\phi_1 + \mathrm{Kn}^2\phi_2 + \mathrm{Kn}^3\phi_3 + \cdots \,,$$

and the leading order of $\phi$ is determined by inserting this ansatz into the complete set of moment equations. A moment is said to be of leading order $\lambda$ if $\phi_\beta = 0$ for all $\beta < \lambda$. This first step agrees with the ideas of [46], where,

however, the authors do not perform a Chapman-Enskog expansion, but a Maxwellian iteration [35].

In Step 2, new variables are introduced by linear combination of the moments originally chosen. The new variables are constructed such that the number of moments at a given order $\lambda$ is minimal. This step does not only simplify the later discussion, but gives an unambiguous set of moments at order $\lambda$. This ensures that the final result will be independent of the initial choice of moments. Note that, while the basic set of moments (25) makes it easy to identify the order of magnitude (in Step 1), any alternative complete set of moments could have been chosen to arrive at the same new variables after Step 2.

Step 3 follows from the definition of the order of accuracy $\lambda_0$: A set of equations is said to be accurate of order $\lambda_0$, when stress $\sigma_{ij}$ and heat flux $q_i$ are known within the order $\mathcal{O}\left(\mathrm{Kn}^{\lambda_0}\right)$. The evaluation of this condition is based on the fact that all moment equations are strongly coupled. This implies that each term in any of the moment equations has some influence on all other equations, in particular on the conservation laws. A theory of order $\lambda_0$ will consider only those terms in all equations whose leading order of *influence* in the conservation laws is $\lambda \leq \lambda_0$. Luckily, in order to evaluate this condition, it suffices to start with the conservation laws, and step by step, order by order, add the relevant terms that are required

The accounting for the order of accuracy is the main difference between the order of magnitude approach and Consistently Ordered Extended Thermodynamics (COET) [46], which assumes that *all* terms in *all* moment equations that are of leading order $\lambda \leq \lambda_0$ or smaller must be retained. The order of magnitude approach leads to smaller systems of equations for a given order, and can be performed for the full three dimensional and time dependent equations, while [46] presents the equations only for one-dimensional steady state processes.

The order of magnitude method was applied to the special cases of Maxwell molecules and the BGK model in [44, 3], and it was shown that it yields the Euler equations at zeroth order, the Navier-Stokes-Fourier equations at first order, and Grad's 13 moment equations (with omission of the non-linear term $\frac{\sigma_{ij}}{\rho}\frac{\partial \sigma_{jk}}{\partial x_k}$) at second order. The regularized 13 moment equations (R13) are obtained as the third order approximation, they consists of the conservation laws (7) and the balance laws for stress (19) and heat flux (20) which now are closed by the expressions†

† There are some differences between the original R13 equations of [13] and the equations presented here, which result from the order of magnitude method. The original equations contain some higher (4th) order terms, and were derived for the linearized collision operator, see [3] for details and discussion.

$$\Delta = -\frac{\sigma_{ij}\sigma_{ij}}{\rho} - 12\frac{\mu}{p}\left[\theta\frac{\partial q_k}{\partial x_k} + \theta\sigma_{kl}\frac{\partial v_k}{\partial x_l} + \frac{5}{2}q_k\frac{\partial\theta}{\partial x_k} - q_k\theta\frac{\partial \ln\rho}{\partial x_k}\right] ,$$

$$R_{ij} = -\frac{4}{7}\frac{1}{\rho}\sigma_{k\langle i}\sigma_{j\rangle k} - \frac{24}{5}\frac{\mu}{p}\left[\theta\frac{\partial q_{\langle i}}{\partial x_{j\rangle}} + q_{\langle i}\frac{\partial\theta}{\partial x_{j\rangle}} - \theta q_{\langle i}\frac{\partial\ln\rho}{\partial x_{j\rangle}} + \frac{10}{7}\theta\sigma_{k\langle i}\frac{\partial v_{\langle j\rangle}}{\partial x_{k\rangle}}\right]$$

$$m_{ijk} = -2\frac{\mu}{p}\left[\theta\frac{\partial\sigma_{\langle ij}}{\partial x_{k\rangle}} - \sigma_{\langle ij}\theta\frac{\partial\ln\rho}{\partial x_{k\rangle}} + \frac{4}{5}q_{\langle i}\frac{\partial v_{j}}{\partial x_{k\rangle}}\right] \quad . \tag{26}$$

The moments of the collision operator $(24)_{2,3}$ are exact for Maxwell molecules, and remain unchanged, $\mathcal{P}_{ij} = -\frac{p}{\mu}\sigma_{ij}$, $\mathcal{P}_i = -\frac{2}{3}\frac{p}{\mu}q_i$.

A closer inspection of the regularized equations (26) shows that the terms added to the original Grad 13 moment equations are of super-Burnett order.

For general, i.e. non-Maxwellian, molecule types the order of magnitude method was performed to second order in [45, 3]; the derivation of the third order equations would be far more involved than for Maxwell molecules. Again the equations at zeroth and first order are the Euler and NSF equations (with exact viscosity, heat conductivity and Prandtl number). The second order equations are a generalization of Grad's 13 moment equations,†

$$\begin{aligned}\frac{D\sigma_{ij}}{Dt} + \sigma_{ij}\frac{\partial v_k}{\partial x_k} + 2\sigma_{k\langle i}\frac{\partial v_{j\rangle}}{\partial x_k} + \frac{4}{5}\Pr\frac{\varpi_3}{\varpi_2}\left(\frac{\partial q_{\langle i}}{\partial x_{j\rangle}} - \omega q_{\langle i}\frac{\partial\ln\theta}{\partial x_{j\rangle}}\right)\\ + \frac{4}{5}\Pr\frac{\varpi_4}{\varpi_2}q_{\langle i}\frac{\partial\ln p}{\partial x_{j\rangle}} + \frac{4}{5}\Pr\frac{\varpi_5}{\varpi_2}q_{\langle i}\frac{\partial\ln\theta}{\partial x_{j\rangle}} + \left(\frac{\varpi_6}{\varpi_2} - 4\right)\sigma_{k\langle i}S_{j\rangle k}\\ = -\frac{2}{\varpi_2}\frac{p}{\mu}\left[\sigma_{ij} + 2\mu\frac{\partial v_{\langle i}}{\partial x_{j\rangle}}\right] ,\end{aligned} \tag{27}$$

$$\begin{aligned}\frac{Dq_i}{Dt} + q_k\frac{\partial v_i}{\partial x_k} + \frac{5}{3}q_i\frac{\partial v_k}{\partial x_k} - \frac{5}{2}\frac{1}{\Pr}\sigma_{ik}\frac{\partial\theta}{\partial x_k} + \frac{5}{4}\frac{1}{\Pr}\frac{\theta_3}{\theta_2}\theta\sigma_{ik}\frac{\partial\ln p}{\partial x_k}\\ + \frac{5}{4}\frac{1}{\Pr}\frac{\theta_4}{\theta_2}\theta\left(\frac{\partial\sigma_{ik}}{\partial x_k} - \omega\sigma_{ik}\frac{\partial\ln\theta}{\partial x_k}\right) + \frac{5}{2}\frac{1}{\Pr}\frac{3}{2}\frac{\theta_5}{\theta_2}\sigma_{ik}\frac{\partial\theta}{\partial x_k}\\ = -\frac{1}{\theta_2}\frac{5}{2}\frac{1}{\Pr}\frac{p}{\mu}\left[q_i + \frac{5}{2}\frac{\mu}{\Pr}\frac{\partial\theta}{\partial x_i}\right] .\end{aligned} \tag{28}$$

Here, the coefficients $\varpi_\alpha$, $\theta_\alpha$ are the Burnett coefficients of Table 1 and $\omega$ is the viscosity exponent of (15).

Jin and Slemrod [47, 48] proposed an alternative regularization by constructing a set of equations that (a) gives the Burnett equations in a second order CE expansion, and (b) gives a positive entropy production for all values of the variables. Up to second order their equations agree with the generalized Grad 13 equations (27, 28) to which they add terms of super-Burnett order that were designed to achieve their goal (b). These higher order terms cannot be justified from the Boltzmann equation [3].

† Recall that Grad's 13 moment equations are only suitable for Maxwell molecules.

We summarize as follows: The order of magnitude method reproduces the established results of the CE expansion at zeroth (Euler) and first (NSF) order. Moreover it provides a new link between the Knudsen number and Grad's 13 moment equations which turn out to be of second order in the Knudsen number, together with a generalization of these for non-Maxwellian molecules. Finally, the method provides a rational derivation of the R13 equations that does not require artificial assumptions.

## 7 Relations Between the Various Sets of Equations

The derivation of macroscopic equations from the Boltzmann equation is simplest for the special case of Maxwell molecules. Accordingly, theories of higher orders in the Knudsen number, like the super-Burnett and the R13 equations, are—at present—only available for Maxwell molecules.

The Chapman-Enskog expansion of increasing order gives the Euler, Navier-Stokes-Fourier, Burnett, and super-Burnett equations.

The augmented Burnett equations contain terms of super-Burnett order, which are added ad hoc, and cannot be derived from the Boltzmann equation.

Grad-type moment equations can be constructed for arbitrary moment sets, but the 13 moment system and the 26 moment system are particularly interesting, since they are equations of orders $\mathcal{O}\left(\mathrm{Kn}^2\right)$ and $\mathcal{O}\left(\mathrm{Kn}^4\right)$, respectively. They can be obtained also from the order of magnitude approach which gives the R13 equations as the proper equations at order $\mathcal{O}\left(\mathrm{Kn}^3\right)$, and the NSF and Euler equations at orders $\mathcal{O}\left(\mathrm{Kn}^1\right)$ and $\mathcal{O}\left(\mathrm{Kn}^0\right)$, respectively.

Jin and Slemrod's equations are accurate to order $\mathcal{O}\left(\mathrm{Kn}^2\right)$, but contain terms of super-Burnett order, $\mathcal{O}\left(\mathrm{Kn}^3\right)$, which cannot be derived from the Boltzmann equation.

A Chapman-Enskog expansion of higher order moment equations can be performed by means of CE expansions for stress and heat flux (see (13)); the results agree with those of the CE expansion of the Boltzmann equation.

The relations between the various sets of equations are depicted in Table 2, in which an arrow between two sets of equations indicates that one set can be derived from the other (e.g. the Burnett eqs. from the Grad13 eqs. by means of a CE expansion). Note that at a given order the equations derived from the CE method and from the order of magnitude approach are quite different, due to the marked differences in methodology. Indeed, the CE based equations (e.g. the super-Burnett equations, at third order), contain less information than their counterparts (e.g. the R13 equations, also at third order), since the former can be derived from the latter, but not vice versa.

For other types of interaction potentials, accurate sets of equations are only available to order $\mathcal{O}\left(\varepsilon^2\right)$, namely the generalized 13 moment equations which were obtained from the order of magnitude method, and the Burnett equations from the CE method. The Euler and NSF equations form the proper

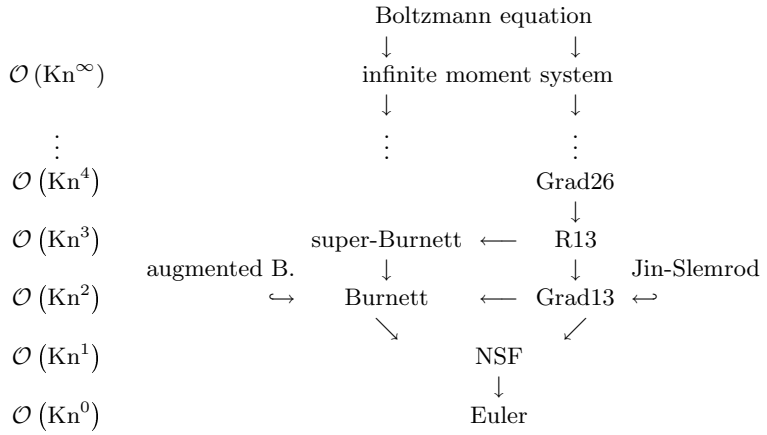



Table 2: The hierarchy of macroscopic equations for Maxwell molecules [3].

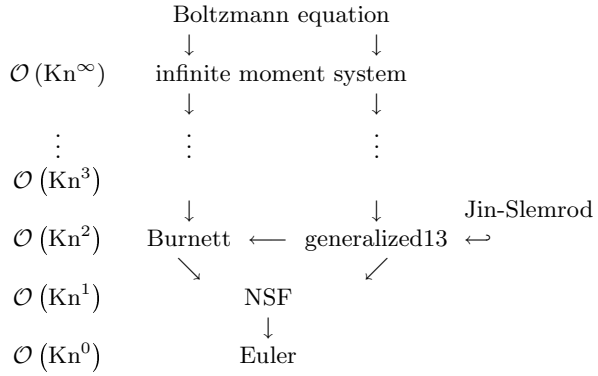



Table 3: The hierarchy of macroscopic equations for molecules with arbitrary interaction potentials [3].

equations at zeroth and first order. The Jin-Slemrod equations are available as well. Table 3 shows the known equations as well as their order of accuracy, and their relations.

## 8 Applications

The previous sections gave an overview over several methods to derive macroscopic equations for rarefied gas flows; see Tables 2 and 3. We now turn the attention to some applications in order to discuss the behavior and quality of

the equations. As before, space restrictions forbid to go into detail, and the interested reader is referred to the cited literature.

**Linear stability:** Bobylev [12] has shown that the Burnett and super-Burnett equations are unstable in transient problems. This failure is the most important reason to discard the Burnett and super-Burnett equations (and thus the CE expansion), and to strive for alternative methods.

For one-dimensional processes the linearized transport equations of the previous sections can be written as

$$\frac{\partial u_A}{\partial t} + \mathcal{A}_{AB}\frac{\partial u_B}{\partial x} + \mathcal{B}_{AB}\frac{\partial^2 u_B}{\partial x^2} + \cdots = \mathcal{C}_{AB}u_B \ , \tag{29}$$

with constant matrices $\mathcal{A}_{AB}, \mathcal{B}_{AB}, \mathcal{C}_{AB}, \ldots$ For the solution, we assume plane waves of the form

$$u_A = \tilde{u}_A \exp\left[i\left(\Omega t - kx\right)\right]$$

where $\tilde{u}_A$ is the complex amplitude of the wave, $\Omega$ is its frequency, and $k$ is its wave number. The equations (29) can then be written as

$$\mathcal{G}_{AB}\left(\Omega, k\right)\tilde{u}_B = 0 \text{ where } \mathcal{G}_{AB}\left(\Omega, k\right) = i\Omega\delta_{AB} - ik\mathcal{A}_{AB} - k^2\mathcal{B}_{AB} + \cdots - \mathcal{C}_{AB}$$

and nontrivial solutions require $\det\left[\mathcal{G}_{AB}\left(\Omega, k\right)\right] = 0$; the resulting relation between $\Omega$ and $k$ is the dispersion relation.

If a disturbance in space is considered, the wave number $k$ is real, the frequency is complex, $\Omega = \Omega_r\left(k\right) + i\Omega_i\left(k\right)$, and stability requires $\Omega_i\left(k\right) \geq 0$.

If a disturbance in time at a given location is considered, the frequency $\Omega$ is real, while the wave number is complex, $k = k_r\left(\Omega\right) + ik_i\left(\Omega\right)$. Then, for a wave traveling in positive $x$-direction ($k_r > 0$), the damping must be negative ($k_i < 0$), while for a wave traveling in negative $x$-direction ($k_r < 0$), the damping must be positive ($k_i > 0$). Thus, if $k\left(\Omega\right)$ is plotted in the complex plane with $\Omega$ as parameter, the curves should not touch the upper right nor the lower left quadrant.

Thus, in order to test the stability of a given set of equations, one has to test for stability in time and space. However, for the Burnett and super-Burnett equations, most authors only consider stability in time, and ignore stability in space [12, 19].

Figure 1, taken from [13, 15], considers the stability against local disturbances of frequency $\Omega$. The figure shows the solutions of the dispersion relation for the different sets of equations; the dots mark the points where $\Omega = 0$. Grad 13 equations, and NSF equations give two different modes each, and none of the solutions violates the condition of stability. The R13 equations have 3 modes, all of them are stable. This is different for the Burnett (3 modes), super-Burnett (4 modes), and augmented Burnett (4 modes) equations: the Burnett equations have one unstable mode, and super-Burnett and augmented Burnett have two unstable modes.

The test for stability in time [15] shows that NSF, augmented Burnett, Grad 13, and R13 equations are stable for all wavelengths, while Burnett and super-Burnett equations are unstable.

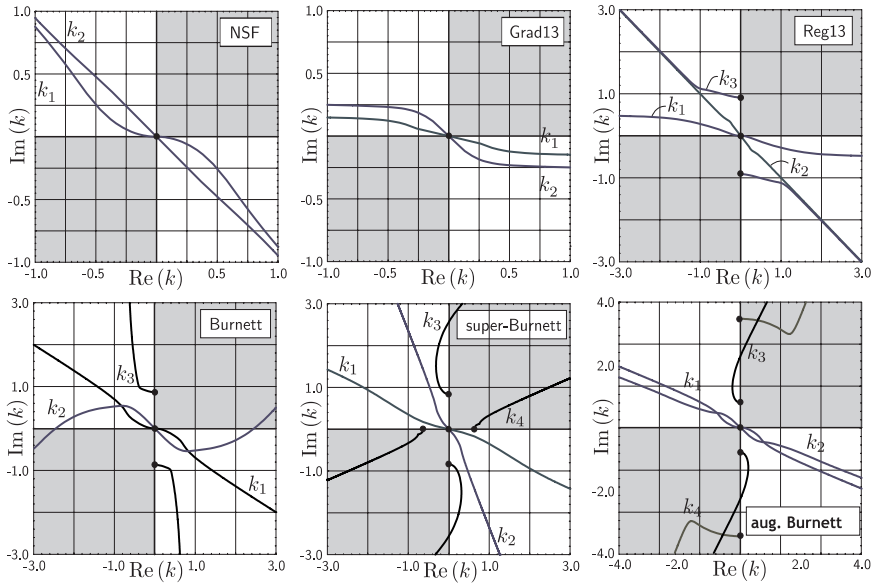



Fig. 1: The solutions $k\,(\Omega)$ of the dispersion relation in the complex plane with $\Omega$ as parameter for Navier-Stokes-Fourier, Burnett, Super-Burnett, augmented Burnett, Grad 13, and R13 equations [13, 15]. The dots denote the points where $\Omega = 0$.

All Burnett type equations, including the augmented Burnett equations, fail the tests for stability. The NSF, Grad-type equations and the R13 equations are stable for all frequencies and for disturbances of any wavelength.

**Shock structures:** The computation of shock structures is a standard test for macroscopic equations designed to describe rarefied gas flows, and we present some of the results of [15].

A one-dimensional steady shock structure connects two equilibrium states, where the values of density $\rho_0$, $\rho_1$, velocity $v_0$, $v_1$, and temperature $\theta_0$, $\theta_1$ in the two equilibrium states are related through the Rankine-Hugoniot relations [3]. The relevant parameter for the shock is the inflow Mach number $M_0 = v_0 \Big/ \sqrt{\frac{5}{3}\theta_0}$.

We compare the shock structures obtained from macroscopic equations for rarefied flows to DSMC results obtained with Bird's code [5], and plot results in dimensionless form [51, 15, 3]. Figure 2 shows the density and heat flux profiles of a $M_0 = 2$ shock calculated with the NSF and Grad13 equations as well as with the Burnett and super-Burnett equations for Maxwell molecules. The NSF result simply mismatches the profile, while the Grad13 solution shows a strong subshock. The Burnett and super-Burnett solutions are spoiled by oscillations in the downstream part of the shock, which arise due to the spatial instabilities. Thus, for the computation of shock structures the Burnett equations and super-Burnett-equations have to be rejected.

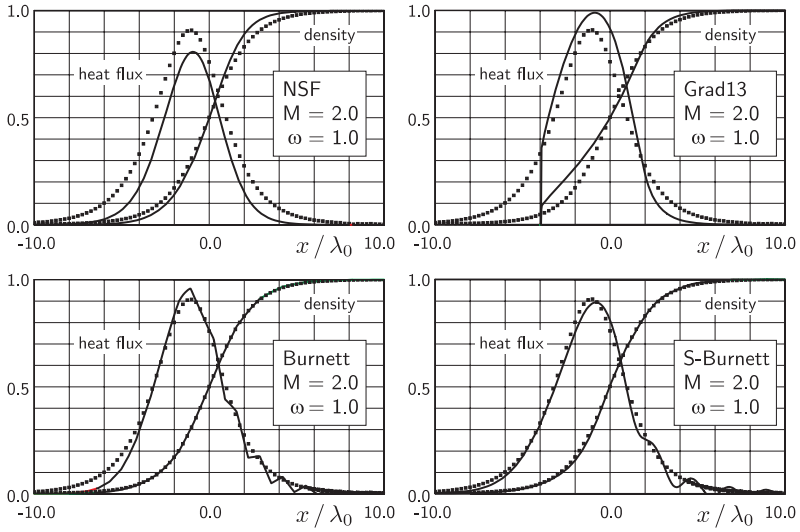



Fig. 2: Shock structure solutions for the NSF equations, Grad 13 equations, and Burnett and super-Burnett equations, for Maxwell molecules at Mach number $M_0 = 2$ (solid lines). Both Burnett results exhibit non-physical oscillations in the downstream region. The squares represent the DSMC solution.

In [52] DSMC results for velocity and temperature are used to compute stress $\sigma$ and heat flux $q$ from the Burnett equations. Comparison with the actual DSMC results for $\sigma$ and $q$ shows considerable improvement over the NSF equations. Thus, the Burnett equations contain the proper physics of the shock, but are useless, since their mathematical structure does not allow to compute a stable solution. Fiscko and Chapman [14] deleted one linear term from the Burnett and super-Burnett equations to obtain stable shock solutions in reasonable agreement to DSMC simulations. Obviously, the mathematical properties of the equations are changed by deleting terms ad hoc, and thus it is not surprising that they obtained stable behavior.

R13 equations and augmented Burnett equations give good results for a wider range of Mach numbers. Figure 3 compares shock structures for Maxwell molecules at $M_0 = 2$ and $M_0 = 4$ for R13 and augmented Burnett with DSMC results. For $M_0 = 2$ the density profiles exhibit no visible differences and both models match the DSMC results very well. The shape of the heat flux is captured very well by the R13 equations, while the augmented Burnett equations do not reproduce the maximum value and the upstream relaxation. The deviations from the DSMC solutions are more pronounced for $M_0 = 4$, where the R13 results begin to deviate in the upstream part. In the tail of the augmented Burnett profiles small oscillations are present, due to instability in space. This happens since the solution was obtained from a boundary value problem; no stability problems arise when the augmented Burnett equations

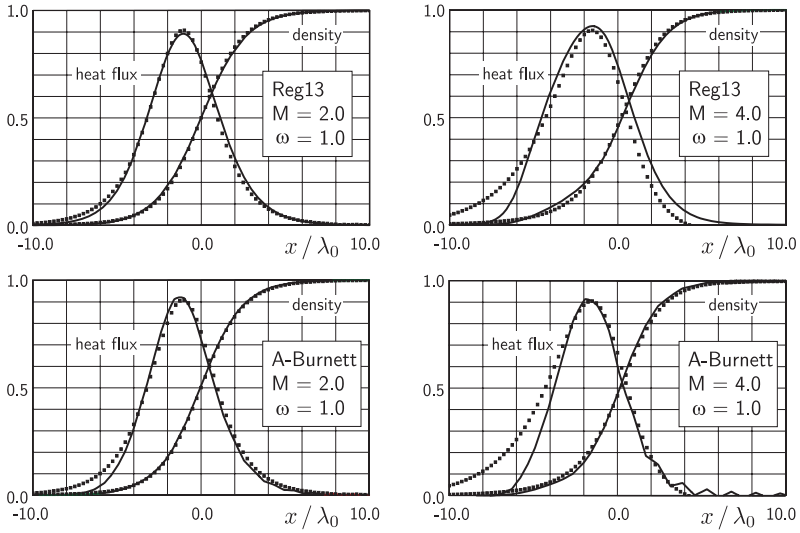



Fig. 3: Shock structures in a gas of Maxwell molecules with Mach numbers $M_0 = 2$ and $M_0 = 4$. The upper row shows the solution of the R13 equations, while the lower row shows the results of the augmented Burnett equations. The squares correspond to the DSMC solution.

are solved by by time stepping into steady state [19, 20]. Altogether, the results of the R13 system for Maxwell molecules agree better with DSMC results than the solutions of the augmented Burnett equations. For higher Mach numbers both deviate somewhat from DSMC results.

A shock is often characterized by the shock thickness, defined as [49, 50]

$$\delta = \frac{\rho_1 - \rho_0}{\max\left(\frac{\partial \rho}{\partial x}\right)} \,. \tag{30}$$

Figure 4 compares thickness results for the R13 system to measurements in argon ($\omega = 0.8$) [49, 50]. The computed shock thickness yields a striking agreement with the experimental data. The results of the augmented Burnett equations with $\omega = 0.8$ lead to a similar agreement, while the NSF results lie far off. The good agreement of the shock thickness for high Mach numbers should not be overemphasized, since the single parameter $\delta$ cannot reflect the complete profile, so that the agreement with shock thickness measurements does not imply a reliable description of the complete profile. Nevertheless, the information that $\delta$ does reflect—a mean thickness—is predicted by the R13 equations accurately even for high Mach numbers.

**Couette flow:** The biggest obstacle for any higher order model for rarefied gas flows is to find proper boundary conditions. This is a difficult problem, and no conclusive answers can be given at present. We shall not discuss the problems, but only present some Couette flow calculations that were obtained

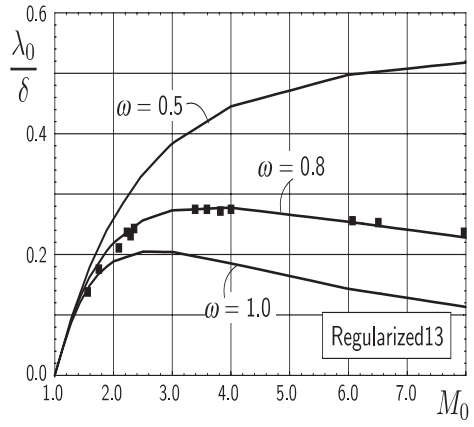



Fig. 4: Comparison of the inverse shock thickness for the R13 equations with measurements for Argon (squares, $\omega \approx 0.8$). The curve of the augmented Burnett equations with $\omega = 0.8$ shows a similar agreement.

by partial fitting of boundary conditions. Couette flow describes the motion of a gas between two infinite parallel plates at distance $L$, moving against each other with the speed $v_W^L$, details can be found in [3].

For constructing the solution it is assumed that the non-equilibrium quantities can be split into bulk (B) and Knudsen layer (L) contributions, $\phi = \phi_B + \phi_L$, where the Knudsen layer contributions vanish in some distance from the wall.

The bulk equations follow from a Chapman-Enskog expansion of the steady state equations in Couette geometry; only the second order equations are considered, which read ($y = x_2/L$ is the dimensionless space variable)

$$\sigma_{12} = -\frac{\mu P_0}{\rho\theta L}\frac{dv}{dy}\ ,\ q_2 = -\frac{15}{4}\frac{\mu P_0}{\rho\theta L}\frac{d\theta}{dy}\ ,\ \sigma_{22} = -\frac{6}{5}\frac{\sigma_{12}\sigma_{12}}{P_0}\ ,\ q_1 = \frac{7}{2}\frac{\sigma_{12}q_2}{P_0}\ . \tag{31}$$

The first two equations, for $\sigma_{12}$ and $q_2$, are the laws of Navier-Stokes and Fourier multiplied with the factor $P_0/\rho\theta = P_0/p$. When these are used with the conservation laws (7), it suffices to prescribe the jump and slip boundary conditions [3]

$$\begin{aligned} v^\alpha - v_W^\alpha &= \frac{-\frac{2-\chi}{\chi}\alpha_1\sqrt{\frac{\pi}{2}}\sqrt{\theta}\sigma_{12}n^\alpha - \frac{1}{5}\alpha_2 q_1}{\rho\theta + \frac{1}{2}\sigma_{22}}\ , \\ \theta^\alpha - \theta_W^\alpha &= -\frac{\frac{2-\chi}{2\chi}\beta_1\sqrt{\frac{\pi}{2}}\sqrt{\theta}q_2 n^\alpha + \frac{1}{4}\theta\sigma_{22}}{\rho\theta + \frac{1}{2}\sigma_{22}} + \frac{V^2}{4}\ , \end{aligned} \tag{32}$$

with correction factors $\alpha_1, \alpha_2, \beta_1$ close to unity [1, 3]. The constant $P_0$ follows from the prescribed mass between the plates.

More interesting are the equations for the normal stress, $\sigma_{22}$, and the heat flux parallel to the wall, $q_1$. Both vanish in the NSF theory, and thus their non-zero values describe pure rarefaction effects. In particular it must be noted that there is no temperature gradient in the $x$-direction: $q_1$ is a heat flux that is not driven by a temperature gradient.

Depending on the set of equations used, $\sigma_{ij}$ and $q_i$ can have linear Knudsen layer contributions as well. The CE expansion, that gave the bulk solution, discards these linear parts [3], but they can be obtained from the linearized equations [16].

The superposition of bulk solution and Knudsen layers for the R13 equations gives

$$v = v_{|B} - \frac{2}{5} q_{1|L} \ , \ \ \theta = \theta_{|B} - \frac{2}{5} \sigma_{22|L} \ , \ \ \sigma_{12} = \sigma_{12|B} \ , \ \ \sigma_{22} = \sigma_{22|B} + \sigma_{22|L} \ ,$$
$$p = P_0 - \sigma_{22|L} \ , \ \ \rho = \frac{p}{\theta} \ , \ \ q_1 = q_{1|B} + q_{1|L} \ , \ \ q_2 = q_{2|B} \ . \tag{33}$$

with the Knudsen layer terms

$$q_{1|L} = A \sinh\left[\sqrt{\tfrac{5}{9}} \tfrac{y-\frac{1}{2}}{\mathrm{Kn}}\right] \quad , \ \ \sigma_{22|L} = D \cosh\left[\sqrt{\tfrac{5}{6}} \tfrac{y-\frac{1}{2}}{\mathrm{Kn}}\right] \ . \tag{34}$$

The constants of integration $A$ and $D$, which should be computed from boundary conditions for stress and heat flux, were fitted to DSMC simulations. Figure 5 compares results of DSMC calculations, NSF equations with jump and slip boundary conditions, and the R13 equations for $\mathrm{Kn} = 0.1$. R13 matches the DSMC simulations quite well; the most visible differences lie in the bulk values for $\sigma_{12}$ and $\sigma_{22}$. The temperature maximum is reproduced very well, while some differences can be observed at the boundaries. NSF, on the other hand, cannot neither describe Knudsen boundary layers nor the rarefaction effects described by $\sigma_{22}$ and $q_1$.

NSF and Grad 13 equations do not give linear Knudsen layers at all. The Burnett and super-Burnett equations cannot describe Knudsen layers of the type (34) but give periodic solutions of the form $A \sin\left[\lambda \frac{x-1/2}{\mathrm{Kn}}\right]$, $B \cos\left[\lambda \frac{x-1/2}{\mathrm{Kn}}\right]$. The augmented Burnett equations give expressions of the type (34), but the signs for the heat flux parallel to the flow does not match the DSMC simulations [16].

## 9 Conclusions and Outlook

Several methods to derive macroscopic equations for rarefied gases, and the resulting equations were presented, including the classical Chapman-Enskog and Grad methods, and the new order of magnitude method. Our interest was focussed on equations for Knudsen numbers above 0.01, i.e. beyond the validity of the Navier-Stokes equations.

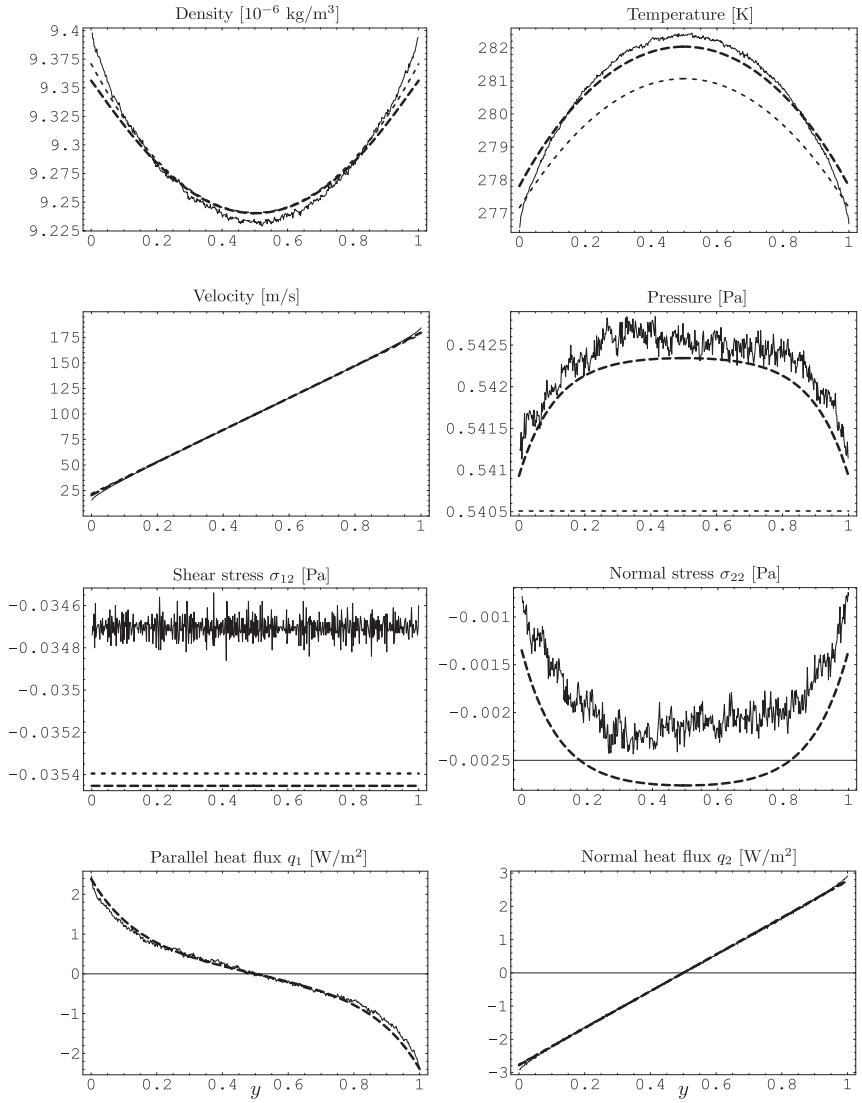



Fig. 5: Couette flow at Kn = 0.1,with $v_W^L = 200\frac{\mathrm{m}}{\mathrm{s}}$. Continuous line: DSMC, finely dashed line (- - - - ): NSF, dashed line ($-\,-\,-\,-$ ): superposition of bulk solution and linear Knudsen layer solution. Recall that NSF implies $q_1 = \sigma_{22} = 0$ (curves not shown).

The CE method suggests the Burnett and super-Burnett equations which, however, are unstable, lead to spurious oscillations in linear steady state processes, and fail to accurately describe shock structures. Efforts to augment the equations were only partially successful.

The Grad method leads to stable equations that can describe Knudsen layers when more than 13 variables are considered. Due to the hyperbolicity of the equations, they lead to unphysical subshocks in shock structure problems.

The order of magnitude method suggest the Grad 13 moment equations at second order and the R13 equations at third order. The latter are superior to the competing sets of equations for several reasons: (a) they contain the Burnett and Super-Burnett equations as can be seen by means of a CE expansion in the Knudsen number, (b) they are linearly stable for all wavelengths and frequencies, (c) they show phase speeds and damping coefficients that match experiments better than those for the NSF equations, or the Grad13 system, (d) they exhibit Knudsen boundary layers, and (e) they lead to smooth shock structures for all Mach numbers.

While the R13 equations have many desirable features, a number of difficult problems must be solved before the R13 equations (or any other model above the NSF equations) can be used as a reliable tool. (a) Reliable boundary conditions must be developed. (b) Industrially relevant gases are diatomic (air!) or polyatomic, and higher order equations for these and for mixtures must be derived. (c) The multiscale character of rarefied flows requires advanced numerical methods that, based on a well chosen local Knudsen number, use the most efficient set of equations in a flow region; this requires the interplay of solvers for NSF, R13 and Boltzmann equations, and reliable switching and transition conditions. (d) Currently, only the Jin-Slemrod equations are accompanied by a proper entropy inequality, and it is desirable to find equivalents for the other higher order models.

These problems are under investigation, and we hope to be able to present solutions in the future.

*Acknowledgement.* This research was supported by the Natural Sciences and Engineering Research Council (NSERC).

## References

1. C. Cercignani: *Theory and application of the Boltzmann Equation.* Scottish Academic Press, Edinburgh 1975
2. S. Chapman, T.G. Cowling: *The mathematical Theory of Non-Uniform Gases* (Cambridge University Press 1970)
3. H. Struchtrup: *Macroscopic Transport Equations for Rarefied Gas Flows—Approximation Methods in Kinetic Theory*, Interaction of Mechanics and Mathematics Series (Springer, Heidelberg 2005)

4. T. Ohwada: Heat flow and temperature and density distributions in a rarefied gas between parallel plates with different temperatures. Finite difference analysis of the nonlinear Boltzmann equation for hard sphere molecules. *Phys. Fluids* **8**, 2153–2160 (1996)
5. G. Bird: *Molecular gas dynamics and the direct simulation of gas flows* (Clarendon Press, Oxford 1994)
6. P.L. Bhatnagar, E.P. Gross, M. Krook: A Model for collision processes in gases. I. Small Amplitude Processes in Charged and Neutral One-Component Systems. *Phys. Rev.* **94**, 511–525 (1954)
7. M.N. Kogan: *Rarefied Gas Dynamics* (Plenum Press, New York 1969)
8. J.H. Ferziger, H.G. Kaper: *Mathematical theory of transport processes in gases* (North-Holland, Amsterdam 1972)
9. D. Burnett: The distribution of molecular velocities and the mean motion in a non-uniform gas. *Proc. Lond. Math. Soc.* **40**, 382–435 (1936)
10. S. Reinecke, G.M. Kremer: Method of Moments of Grad. *Phys. Rev. A* **42**, 815–820 (1990)
11. M.Sh. Shavaliyev: Super-Burnett Corrections to the Stress Tensor and the Heat Flux in a Gas of Maxwellian Molecules. *J. Appl. Maths. Mechs.* **57**, 573–576 (1993)
12. A.V. Bobylev: The Chapman-Enskog and Grad methods for solving the Boltzmann equation. *Sov. Phys. Dokl.* **27**, 29–31 (1982)
13. H. Struchtrup, M. Torrilhon: Regularization of Grad's 13-moment-equations: Derivation and Linear Analysis. *Phys. Fluids* **15**, 2668–2680 (2003)
14. K.A. Fiscko, D.R. Chapman: Comparison of Burnett, Super-Burnett and Monte Carlo Solutions for Hypersonic Shock Structure. In: *Proceedings of the 16th Symposium on Rarefied Gasdynamics* (AIAA, Washington 1989), 374–395
15. M. Torrilhon, H. Struchtrup: Regularized 13-Moment-Equations: Shock Structure Calculations and Comparison to Burnett Models. *J. Fluid Mech.* **513**, 171–198 (2004)
16. H. Struchtrup: Failures of the Burnett and Super-Burnett equations in steady state processes. *Cont. Mech. Thermodyn.* **17**, 43–50 (2005)
17. I.V. Karlin, A.N. Gorban: Hydrodynamics from Grad's equations: What can we learn from exact solutions? *Ann. Phys. - Berlin* **11**, 783–833 (2002)
18. Y. Zheng, H. Struchtrup: Burnett equations for the ellipsoidal statistical BGK Model. *Cont. Mech. Thermodyn.* **16**, 97–108 (2004)
19. X. Zhong, R.W. MacCormack, D.R. Chapman: Stabilization of the Burnett Equations and Applications to High-Altitude Hypersonic Flows. *AIAA 91-0770* (1991)
20. X. Zhong, R.W. MacCormack, D.R. Chapman: Stabilization of the Burnett Equations and Applications to Hypersonic Flows. *AIAA Journal* **31**, 1036 (1993)
21. H. Grad: On the Kinetic Theory of Rarefied Gases. *Comm. Pure Appl. Math.* **2**, 325 (1949)
22. H. Grad: Principles of the Kinetic Theory of Gases. In: *Handbuch der Physik*, vol. 12, ed. by S. Flügge (Springer, Berlin 1958)
23. I. Müller, T. Ruggeri: *Rational Extended Thermodynamics*, Springer Tracts in Natural Philosophy, vol. 37 (Springer, New York 1998)
24. H. Struchtrup: Heat Transfer in the Transition Regime: Solution of Boundary Value Problems for Grad's Moment Equations via Kinetic Schemes. *Phys. Rev. E* **65**, 041204 (2002)

25. H. Struchtrup: An Extended Moment Method in Radiative Transfer: The Matrices of Mean Absorption and Scattering Coefficients. *Ann. Phys.* **257**, 111–135 (1997)
26. W. Weiss: Continuous shock structure in extended Thermodynamics. *Phys. Rev. E* **52**, 5760 (1995)
27. J.D. Au: Nichtlineare Probleme und Lösungen in der Erweiterten Thermodynamik. Dissertation, Technical University Berlin 2000
28. J.D. Au, M. Torrilhon, W. Weiss: The Shock Tube Study in Extended Thermodynamics, *Phys. Fluids* **13**, 2423–2432, (2001)
29. H. Struchtrup: Kinetic schemes and boundary conditions for moment equations. *ZAMP* **51**, 346–365 (2000)
30. H. Struchtrup: Some remarks on the equations of Burnett and Grad. IMA Volume **135** "Transport in Transition Regimes," (Springer, New York 2003)
31. D. Reitebuch, W. Weiss: Application of High Moment Theory to the Plane Couette Flow. *Cont. Mech. Thermodyn.* **11**, 227 (1999)
32. H. Struchtrup: Grad's Moment Equations for Microscale Flows. In: *Symposium on Rarefied Gasdynamics 23*, AIP Conference Proceedings **663**, 792–799 (2003)
33. W. Weiss: Zur Hierarchie der Erweiterten Thermodynamik. Ph.D. thesis, Technical University Berlin (1990)
34. E. Ikenberry, C. Truesdell: On the pressures and the flux of energy in a gas according to Maxwell's kinetic theory I. *J. of Rat. Mech. Anal.* **5**, 1–54 (1956)
35. C. Truesdell, R.G. Muncaster: *Fundamentals of Maxwell's kinetic theory of a simple monatomic gas* (Academic Press, New Yorck 1980)
36. S. Reinecke, G.M. Kremer: Burnett's equations from a (13+9N)-field theory. *Cont Mech. Thermodyn.* **8**, 121–130 (1996)
37. I.V. Karlin, A.N. Gorban, G. Dukek, T.F. Nonnenmacher: Dynamic correction to moment approximations, *Phys. Rev. E* **57** , 1668–1672 (1998)
38. A.N. Gorban, I.V. Karlin: *Invariant Manifolds for Physical and Chemical Kinetics*, Lecture Notes in Physics, vol. 660, (Springer, Berlin 2005)
39. W. Dreyer: Maximization of the Entropy in Non-equilibrium. *J. Phys. A: Math. Gen.* **20**, 6505 (1987)
40. C.D. Levermore: Moment Closure Hierarchies for Kinetic Theories. *J. Stat. Phys.* **83**, 1021–1065 (1996)
41. M. Junk: Domain of definition of Levermore's five-moment system. *J. Stat. Phys.* **93**, 1143–1167 (1998)
42. W. Dreyer, M. Junk, M. Kunik: On the approximation of the Fokker-Planck equation by moment systems. *Nonlinearity* **14**, 881–906 (2001)
43. M. Junk: Maximum entropy moment problems and extended Euler equations. *Transport in Transition Regimes, IMA Vol. Math. Appl.* **135**, 189–198 (Springer, New York 2003)
44. H. Struchtrup: Stable transport equations for rarefied gases at high orders in the Knudsen number. *Phys. Fluids* **16** , 3921–3934 (2004)
45. H. Struchtrup: Derivation of 13 moment equations for rarefied gas flow to second order accuracy for arbitrary interaction potentials. *Multiscale Model. Simul.* **3**, 211–243 (2004)
46. I. Müller, D. Reitebuch, W. Weiss: Extended Thermodynamics – Consistent in Order of Magnitude, *Cont. Mech. Thermodyn.* **15**, 113-146 (2003)
47. S. Jin, M. Slemrod: Regularization of the Burnett equations via relaxation. *J. Stat. Phys.* **103**, 1009–1033 (2001)

48. S. Jin, L. Pareschi, M. Slemrod: A Relaxation Scheme for Solving the Boltzmann Equation Based on the Chapman-Enskog Expansion. *Acta Mathematicas Applicatae Sinica* (English Series) **18**, 37–62 (2002)
49. B. Schmidt: Electron Beam Density Measurements in Shock Waves in Argon. *J. Fluid Mech.* **39**, 361 (1969)
50. H. Alsmeyer: Density Profiles in Argon and Nitrogen Shock Waves Measured by the Absorbtion of an Electron Beam, *J. Fluid Mech.* **74**, 497 (1976)
51. G.C. Pham-Van-Diep, D.A. Erwin, E.P. Muntz, Testing Continuum Descriptions of Low-Mach-Number Shock Structures. *J. Fluid Mech.* **232**, 403 (1991)
52. E. Salomons, M. Mareschal, Usefulness of the Burnett Description of Strong Shock Waves. *Phys. Rev. Lett.* **69**, 269–272 (1992)

# Novel Trajectory Based Concepts for Model and Complexity Reduction in (Bio)Chemical Kinetics

D. Lebiedz, V. Reinhardt, J. Kammerer

Interdisciplinary Center for Scientific Computing, University of Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany,
lebiedz@iwr.uni-heidelberg.de

**Summary.** Based on increasing availability of high-accuracy data from high-throughput experimental techniques, detailed kinetic models for complex reaction mechanisms come more and more into applications. They are for instance used in computer simulations aimed at optimization of technical process operation or for virtual experiments in a systems biology approach to cellular biochemistry. Since high-dimensional models from large-scale mechanisms are difficult to handle in both computationally expensive spatiotemporal simulations and interpretation of system functions, sound mathematical methods for model and complexity reduction are important. Here, model reduction aims at reducing the degrees of freedom necessary for a sufficiently accurate description of the system dynamics whereas complexity reduction is supposed to help in providing functional insight into the dynamic structure and functional properties of complex reaction networks which is particularly important in biology. In this article we review recent developments from our group in both areas which rely on trajectory based concepts. First, we review a concept which is related to maximal relaxation of chemical forces under given constraints in terms of least-square minimal entropy production of single reaction steps and present applications for model reduction of chemical reaction mechanisms. Second, we discuss a sensitivity approach to phase flow analysis which can be exploited for complexity reduction in biochemical networks by identifying some aspects of the dynamic coupling structure of system components.

## 1 Introduction

Kinetic modeling of complex (bio)chemical reaction systems can be described as a systematic mapping of reality to mathematical equations describing the dynamical behavior of the system under investigation. This process necessarily fades out most microscopic details associated with the reaction process but rather tries to capture the central features giving rise to the macroscopic

behavior. Describing microscopically highly complicated systems with billions of degrees of freedom by only a few characteristic macroscopic variables is at the core of the connection between quantum physics and phenomenological thermodynamics as a classical multi-scale problem. Statistical physics aims at bridging the gap between atomic properties and thermodynamic variables like temperature, pressure and energy.

In principle, each modeling attempt of natural phenomena is related to such multi-scale problems and it is the art of modeling to find reasonable levels of description for the aspects to be investigated. In the so called top-down approach, one starts with an abstract system level description and refines the corresponding model according to experimental observation. The bottom-up approach starts with first-principle physical and chemical laws, often on a very fine or intermediate scale and incorporates all detailed knowledge that it available into the model. The latter approach is more rigorous and if successful provides a much deeper insight into the system but generally leads to large-scale models with a huge number of degrees of freedom. It is mostly impossible to use such high-dimensional models in spatiotemporal simulations of the system dynamics on larger scales or for gaining insight into dynamical mechanisms without further analysis.

Here, model reduction techniques come into play, which aim at filtering out the essential degrees of freedom from a bottom-up constructed first-principle model with respect to the system properties of interest. In case of chemical kinetics this property of interest is often the transient reaction dynamics itself. Many well known model reduction methods exploit intrinsic multiple time scales to construct low dimensional approximations of full reaction mechanisms that describe the "long-term" dynamics accurately. This means to resolve on a larger time scale and neglect fast scales in a reasonable sense by introducing generalized approximations similar to quasi-steady-state or partial equilibrium. Many realizations of this approach have been described, a comprehensive overview can be found in [7, 17] and in [12] the most common model reduction techniques and their underlying concepts are mentioned and corresponding references are given.

Here, we present a novel approach that is, however, also related to multiple time scales. We do not consider the time scales themselves but rather regard a dissipative chemical reaction system as a set of reaction steps that gradually relax to partial equilibrium. In thermodynamic equilibrium finally all reaction steps are fully relaxed (principle of detailed balance, see [11]), but on the way to equilibrium, they relax with a tendency proportional to their driving force, the chemical affinity. By artificially relaxing the reaction system maximally under given constraints (the current system state characterized by a few selected degrees of freedom), it is possible to compute a low-dimensional approximation to the kinetic mechanism, a reduced model. We will point out these ideas and present example applications in section 2.

Another problem with large-scale kinetic models derived from first-principles is that numerical simulations in high-dimensional phase space yield

only restricted insight into the dynamic mechanism, component coupling relations and functional properties of the system itself. However, this insight is highly important when studying biochemical signaling and metabolic networks in cells [10]. Modern systems biology approaches use mathematical modeling to understand such systems and their functions and methods for analyzing large-scale models are helpful and necessary in that context [9, 22]. Here, we present a numerical algorithm that reduces the complexity of biochemical networks by providing insight into some aspects of the dynamic coupling structure of system components. The method analyzes the relative behavior of trajectory bundles in phase space and is related to the problem of finding locally reduced models that capture the essential system dynamics with some desired accuracy while enslaving (relaxing) the remaining dynamical modes. Details will be explained in section 3 and biochemical example applications will be discussed.

## 2 Model Reduction: Constrained Relaxation of Chemical Forces and Minimal Entropy Production Trajectories

While many common model reduction techniques in chemical kinetics make explicit use of the time-scale separation concept, the approach presented here, which was first introduced in [12], is based on finding a criterion for the maximal relaxation of chemical forces (reaction affinities) under given constraints in terms of some fixed species concentrations. The latter have to be chosen *a priori* as representatives (called progress variables) of a reduced model.

From a thermodynamic point of view this criterion is related to a generalized concept for the distance of a chemical system from its attractor which is given by the thermodynamic equilibrium under isolated conditions. This fact can be used for a model reduction approach in the way that a given number of initial values is specified and the remaining initial values and a special trajectory (called minimal entropy production trajectory, MEPT, in the following) which converges towards equilibrium are computed such that the affinities of single reaction steps are minimized in a least squares weighting along this trajectory. This can be formulated in the sense that the sum of least square deviations from zero of the entropy production contribution of each single reaction step along this particular trajectory is as small as possible under the initial value constraints while approaching equilibrium. The latter assumption can also be interpreted as the demand that all thermodynamic (chemical) forces and dynamic modes of the system are and remain maximally relaxed under the given constraints. Hence the entropy production plays formally the role of a kinetic potential in a dissipative system with its gradient as the driving force towards equilibrium. In the following, we will explain the use of this criterion for model reduction and show results computed for two example systems.

### 2.1 Method and Numerical Realization

In non-equilibrium thermodynamics, it is common to express the entropy change of a system by $\mathrm{d}S = \mathrm{d}S_i + \mathrm{d}S_e$, where $\mathrm{d}S_e$ describes the entropy exchange between the system and its environment by flows of heat and matter, implying $\mathrm{d}S_e = 0$ for isolated systems. According to the second law of thermodynamics, $\mathrm{d}S_i \geq 0$ holds for any spontaneous process [6]. In the following, we assume the chemical system under consideration to be isolated, hence we have $\mathrm{d}S = \mathrm{d}S_i \geq 0$.

For an elementary reaction step with the reaction rates $r_{\rightarrow}, r_{\leftarrow}$ for forward and backward reactions respectively, entropy production $\frac{\mathrm{d}S}{\mathrm{d}t}$ can be computed by

$$\frac{\mathrm{d}S_i}{\mathrm{d}t} = R(r_{\rightarrow} - r_{\leftarrow}) \ln\left(\frac{r_{\rightarrow}}{r_{\leftarrow}}\right), \tag{1}$$

where $R$ denotes the gas constant.

The additivity of entropy production for several elementary reaction steps allows to compute the total entropy production from purely kinetic information for arbitrary reaction systems as long as the detailed elementary step mechanism is known and kinetic data are available. It follows again from the second law of thermodynamics that the entropy production of an isolated system is a Lyapunov function. i.e. it is positive definite, monotonic decreasing and approaching zero at thermodynamic equilibrium. Exploiting these properties, the task of model reduction can be formulated using entropy production as a measure for the relaxation of single reaction steps in the following sense: Fix some initial concentrations chosen as representatives of a low-dimensional approximation of the full kinetic model (called progress variables) and minimize with respect to the remaining variable initial concentration values the square deviation of entropy production from zero in each single reaction step. Mathematically this can be formulated as a variational boundary value problem:

$$\min_{c_i, i \in I_{\mathrm{free}}} \int_0^T \sum_{j=1}^{n_{\mathrm{reac}}} \left(\frac{\mathrm{d}S_j}{\mathrm{d}t}\right)^2 \mathrm{d}t \tag{2}$$

subject to

$$\begin{aligned} \tfrac{\mathrm{d}c_i}{dt} &= f_i(c) \qquad \forall i \in I_{fixed} \cup I_{free} \\ c_i(0) &= \text{const}, \; i \in I_{\mathrm{fixed}} \\ |c_i(T) - c_{i_{\mathrm{eq}}}| &\leq \varepsilon, \; i \in I_{\mathrm{fixed}}, \qquad T \text{ free}, \end{aligned} \tag{3}$$

and subject to possible chemical conservation relations.

The index sets $I_{\mathrm{free}}$ and $I_{\mathrm{fixed}}$ contain the indices of free initial variables and fixed initial variables (progress variables) respectively and $n_{ref}$ is the number of reactions in the elementary step mechanism. The integration in (2) comprises the sum of the square deviations of the entropy production of each single reaction step from their minimal value zero. The goal is the minimization of this integral with respect to the free initial concentrations. As the system dynamics get infinitely slow when approaching equilibrium,

the equilibrium point $c_{i_{\mathrm{eq}}}$, $i \in I_{\mathrm{fixed}}$ is approximated in the above formulation within a surrounding of radius $\epsilon$. The end time $T$ is free in the problem formulation as it is *a priori* unknown.

Elaborate mathematical optimization techniques exist for the numerical solution of variational boundary problems as formulated in (2)-(3). The numerical solutions for the examples in the next subsection have been computed using `MUSCOD-II` [14] developed for optimal control of large scale dynamical systems modeled by ordinary differential or differential algebraic equations. `MUSCOD-II` is based on multiple shooting applied in the context of sequential quadratic programming (SQP) methods as introduced by Bock [3] for numerical optimization. The multiple shooting approach is robust and efficient and has particular advantages for the problem of computing MEPTs that will be discussed in section 2.3 in more detail.

### 2.2 Example Applications and Discussion

*Three component model system*

To illustrate the above MEPT method, it was tested with a simple kinetic model system involving three chemical species:

$$A + A \leftrightarrow B \leftrightarrow C \tag{4}$$

The composition of the isolated reaction system can be completely described in terms of concentration values $\Phi_i$ for each chemical species involved. Due to mass conservation, $\Phi_A + 2\Phi_B + 2\Phi_C$ holds for system (4). To keep the notation as simple as possible, $2c_A = \Phi_A$, $c_B = \Phi_B$ and $c_C = \Phi_C$ are defined and constant factors arising by the stoichiometry are included in the velocity constants.

With these simplifications the kinetic equations in dimensionless variables are

$$\begin{aligned} \tfrac{\mathrm{d}c_A(t)}{\mathrm{d}t} &= -k_1 c_A^2 + k_{-1} c_B \\ \tfrac{\mathrm{d}c_B(t)}{\mathrm{d}t} &= k_1 c_A^2 - k_{-1} c_B - k_2 c_B + k_{-1} c \\ \tfrac{\mathrm{d}c_C(t)}{\mathrm{d}t} &= k_2 c_B - k_{-2} c_C \end{aligned} \tag{5}$$

with the mass conservation relation

$$c_A(t) + c_B(t) + c_C(t) \equiv 1.0. \tag{6}$$

Fig. 1 shows the results obtained when computing a minimal entropy production trajectory (MEPT) for this system using $k_1 = 1, k_{-1} = 10^{-5}, k_2 = 0.01$, and $k_{-2} = 10^{-5}$ as rate constants. Here, $c_C$ has been chosen as the only progress variable with fixed value $c_C(0) = c_0 = 0.1$.

For comparison with a widely applied model reduction technique, we computed a reduced description of (5) using the ILDM method introduced by Maas and Pope [15] which can be analytically treated in this simple test case:

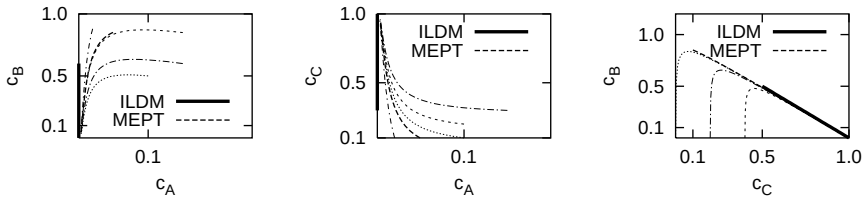



Fig. 1: Comparison of analytically computed ILDM with MEPT results for 3-component system (5) according to (2)-(3) together with a selection of bundling trajectories

$$c_A = 0, c_C = 1 - c_B \tag{7}$$

This ILDM is included in the figures as a bold line.

The results demonstrate that the minimal entropy production trajectories are not identical to the ILDMs especially farther away from equilibrium. This is a consequence of the fact that the ILDM approach assumes fast time scales to be fully relaxed. This assumption is only an approximation as long as the system is not in equilibrium. Opposed to that, the minimal entropy production method leads to maximally relaxed thermodynamic forces under given constraints but does not enforce mode relaxation by *a priori* assumption. Especially far from equilibrium the latter approach obviously describes the real situation much more accurately.

*Simplified hydrogen combustion mechanism*

We present further results for the following six component hydrogen combustion mechanism which describes the kinetics of this system semi-quantitatively.

$$\begin{array}{rcll}
\mathrm{H_2} & \overset{k_1,k_{-1}}{\longleftrightarrow} & 2\mathrm{H} \quad , & k_1 = 2.0, \quad k_{-1} = 216.0 \\
\mathrm{O_2} & \overset{k_2,k_{-2}}{\longleftrightarrow} & 2\mathrm{O} \quad , & k_2 = 1.0, \quad k_{-2} = 337.5 \\
\mathrm{H_2O} & \overset{k_3,k_{-3}}{\longleftrightarrow} & \mathrm{H} + \mathrm{OH} \, , & k_3 = 1.0, \quad k_{-3} = 1400.0 \\
\mathrm{H_2} + \mathrm{O} & \overset{k_4,k_{-4}}{\longleftrightarrow} & \mathrm{H} + \mathrm{OH} \, , & k_4 = 1000.0, \, k_{-4} = 10800.0 \\
\mathrm{O_2} + \mathrm{H} & \overset{k_5,k_{-5}}{\longleftrightarrow} & \mathrm{O} + \mathrm{OH} \, , & k_5 = 1000.0, \, k_{-5} = 33750.0 \\
\mathrm{H_2} + \mathrm{O} & \overset{k_6,k_{-6}}{\longleftrightarrow} & \mathrm{H_2O} \quad , & k_6 = 100.0, \quad k_{-6} = 0.7714
\end{array} \tag{8}$$

The kinetic model of this mechanism is given by

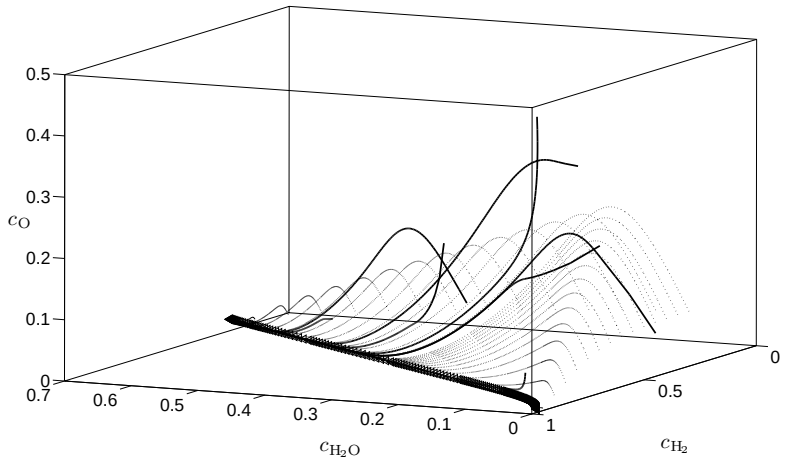



Fig. 2: Illustration of the MEPT method for the $H_2$ combustion mechanism (8). The triangles depict the one-dimensional MEPT with $H_2O$ as progress variable, the dotted lines represent two-dimensional MEPTs ($H_2O$ and $H_2$ as progress variables). The solid lines depict arbitrary trajectories.

$$\begin{aligned}
\frac{\mathrm{d}c_{\mathrm{H}_2}}{\mathrm{d}t} &= -k_1 c_{\mathrm{H}_2} + k_{-1} c_{\mathrm{H}}^2 - k_4 c_{\mathrm{H}_2} c_O + k_{-4} c_{\mathrm{H}} c_{\mathrm{OH}} - k_6 c_{\mathrm{H}_2} c_{\mathrm{O}} + k_{-6} c_{\mathrm{H2O}} \\
\frac{\mathrm{d}c_{\mathrm{H}}}{\mathrm{d}t} &= 2k_1 c_{\mathrm{H}_2} - 2k_{-1} c_{\mathrm{H}}^2 + k_3 c_{\mathrm{H}_2\mathrm{O}} + k_{-3} c_{\mathrm{H}} c_{\mathrm{OH}} + k_4 c_{\mathrm{H}_2} c_{\mathrm{O}} - k_{-4} c_{\mathrm{H}} c_{\mathrm{OH}} \\
&\quad -k_5 c_{\mathrm{O}_2} c_{\mathrm{H}} + k_{-5} c_{\mathrm{O}} c_{\mathrm{OH}} \\
\frac{\mathrm{d}c_{\mathrm{O}_2}}{\mathrm{d}t} &= -k_2 c_{\mathrm{O}_2} + k_{-2} c_{\mathrm{O}}^2 - k_5 c_{\mathrm{H}} c_{\mathrm{O}_2} + k_{-5} c_{\mathrm{O}} c_{\mathrm{OH}} \\
\frac{\mathrm{d}c_{\mathrm{O}}}{\mathrm{d}t} &= 2k_2 c_{\mathrm{O}_2} - 2k_{-2} c_{\mathrm{O}}^2 - k_4 c_{\mathrm{H}_2} c_{\mathrm{O}} + k_{-4} c_{\mathrm{H}} c_{\mathrm{OH}} + k_5 c_{\mathrm{H}} c_{\mathrm{O}_2} - k_{-5} c_{\mathrm{O}} c_{\mathrm{OH}} \\
&\quad +k_{-4} c_{\mathrm{H}} c_{\mathrm{OH}} - k_4 c_{\mathrm{H}_2} c_{\mathrm{O}} - k_6 c_{\mathrm{H}_2} c_{\mathrm{O}} + k_{-6} c_{\mathrm{H}_2\mathrm{O}} \\
\frac{\mathrm{d}c_{\mathrm{H}_2\mathrm{O}}}{\mathrm{d}t} &= -k_3 c_{\mathrm{H}_2\mathrm{O}} + k_{-3} c_{\mathrm{H}} c_{\mathrm{OH}} + k_6 c_{\mathrm{H}_2} c_{\mathrm{O}} - k_{-6} c_{\mathrm{H}_2\mathrm{O}} \\
\frac{\mathrm{d}c_{\mathrm{OH}}}{\mathrm{d}t} &= k_3 c_{\mathrm{H}_2\mathrm{O}} - k_{-3} c_{\mathrm{H}} c_{\mathrm{OH}} + k_4 c_{\mathrm{H}_2} c_{\mathrm{O}} - k_{-4} c_{\mathrm{H}} c_{\mathrm{OH}} + k_5 c_{\mathrm{H}} c_{\mathrm{O}_2} \\
&\quad -k_{-5} c_{\mathrm{O}} c_{\mathrm{OH}}
\end{aligned} \tag{9}$$

Together with the two conservation equations

$$2c_{\mathrm{H}_2} + 2c_{\mathrm{H}_2\mathrm{O}} + c_{\mathrm{H}} + c_{\mathrm{OH}} = C_1 \tag{10}$$

$$2c_{\mathrm{O}_2} + c_{\mathrm{H}_2\mathrm{O}} + c_{\mathrm{O}} + c_{\mathrm{OH}} = C_2 \tag{11}$$

we have a system with four degrees of freedom.

By fixing one initial condition we can compute a single trajectory with maximally relaxed chemical forces in the sense discussed above as we did for the last 3-component example mechanism. The MEPT for a fixed initial concentration of $H_2O$ ($10^{-4}$)) and constants $C_1 = 2.0, C_2 = 1.0$ is depicted with triangles in Fig. 2.

However, the formulation (2)-(3) leaves the freedom to choose an arbitrary number of progress variables (variables with fixed initial conditions) as long

as there are still degrees of freedom left in the system. Hence, to illustrate the principal applicability of the MEPT method for model reduction to higher dimensions than one, we computed families of MEPTs using both $H_2O$ and $H_2$ as progress variables. We first varied the initial concentration of $H_2$ from 0.3 to 0.95 with the initial concentration of $H_2O$ set to $10^{-4}$. Then we varied the initial concentration of $H_2O$ from 0.05 to 0.65 with the initial concentration of $H_2$ fixed to 0.3.

We call the so computed trajectories "two-dimensional" MEPTs because two progress variables are used (dotted lines in Fig. 2). These MEPTs form a smooth two-dimensional surface and all of them relax to the MEPT with one progress variable ("one-dimensional" MEPT). Trajectories from arbitrary initial conditions (solid lines in Fig. 2) first relax to the two-dimensional surface, then to the one-dimensional MEPT and finally to equilibrium.

### 2.3 Outlook

Entropy production $\frac{\mathrm{d}S_i}{\mathrm{d}t}$ remains well defined in open systems kept away from equilibrium by boundary flows of energy and matter if at least the concept of local equilibrium is valid [6]. This concept assumes that the non-equilibrium system can be described using the macroscopic variables known from equilibrium thermodynamics as functions of time and/or space. Assuming stability of this local equilibrium, in the so called linear regime (valid for systems not to far from equilibrium) the second law of thermodynamics leads to the result that the entropy production of a spontaneous process is still positive definite, monotonic decreasing and approaching a minimum at a non-equilibrium steady states (Prigogine's minimum entropy production principle [16]). Thus, in this case the entropy production can still be used as a Lyapunov function and the MEPT model reduction concept as described above can be applied. However, also in many far from equilibrium situations, a quite general evolution criterion exists for dissipative systems [16, 11]. The latter describes the monotonous evolution of a dynamical system towards its attractor and could be exploited for a model reduction approach analogous to the MEPT. Thus, the whole concept can in principle be extended towards application for chemical systems that do not reach equilibrium or steady state but for example limit cycle attractors.

In various applications the ultimate aim of model reduction is to incorporate the reduced model into extensive spatiotemporal simulations including physical transport processes which would be prohibitively expensive numerically if a detailed reaction mechanism containing multiple time scales was used. There are two possibilities for realization of this aim via the MEPT method. The first computes a low dimensional approximation *in situ*, meaning online during numerical integration of the spatiotemporal system (mostly partial differential equations, PDE). The second computes reduced models *a priori* and offline and tabulates them appropriately as a function of some chosen degrees of freedom for parameterization (progress variables) of the

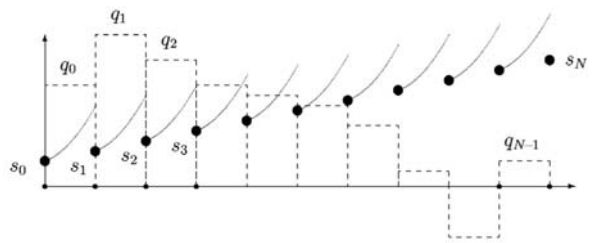



Fig. 3: Multiple shooting discretization and decoupled piecewise trajectory integration for given initial values at the multiple shooting nodes $s_i$ and optionally some input control functions $q_i$.

low-dimensional approximation in the range of values required. In both cases huge numbers of MEPT families have to be computed which are parameterized by different fixed initial values that vary continously. This demands efficient numerical solution of the corresponding optimization problem and multiple shooting (see Fig. 2.3) in combination with an initial value embedding approach is a promising strategy here.

Direct numerical treatment of optimization for dynamical processes with differential equations as constraints require a projection of the original optimization problem in function space to a finite dimensional approximation. This can be achieved by numerically integrating the differential equations and decoupling the optimization itself from the numerical integration in an outer loop (sequential method). A different approach uses a full discretization of the differential equations and treats the resulting algebraic equations as constraints in a nonlinear optimization problem (simultaneous method). While the sequential method is easy to implement, the simultaneous approach is generally more robust and efficient, however gives rise to a large number of additional variables (at the discretization nodes). A powerful compromise between both concepts is the multiple shooting idea.

The multiple shooting method integrates the state trajectories only on small initially decoupled time subintervals of the full time horizon (see Fig. 2.3). A BDF-type (backward differentiation formulae) stiff integrator [1] is used for this in `MUSCOD-II`. As a consequence one has to solve $n$ initial value problems instead of one. In the case of dynamical instabilities like strongly diverging state trajectories this makes the accurate numerical computation of derivative information required for widely used sequential quadratic programming (SQP) optimization much more stable or in some cases possible at all. However, at first sight this improvement seems to be again at the expense of additional unknown variables corresponding the state trajectory values at the multiple shooting nodes as for the discretization nodes in the simultaneous approach. But the mathematical structure of the multiple shooting discretization

can be successfully exploited numerically and the corresponding optimization problem can be solved with roughly the same effort as in the single shooting (sequential) approach [14]. Furthermore, continuity of the state trajectories for the final optimal solution in the multiple shooting nodes has to be assured which can be easily done be adding additional equality constraints to the optimization problem. These require that the end point for the preceding multiple shooting interval $s_i$ matches with the initial value $s_{i+1}$ of the next (see Fig. 2.3).

The multiple shooting approach allows in particular the incorporation of *a priori* information about the optimal solution (trajectory) which can be used to set initial conditions for the state variables at the multiple shooting nodes. This freedom is on the one hand a often significant help on the way towards a global optimum even though local optimization algorithms are used. On the other hand, it can be used for embedding a solution already available into a family of neighboring problems in the sense of parametric optimization where parameters slightly change from one optimization problem to the next. This has been demonstrated by Diehl et al. for nonlinear model predictive control (NMPC) applications [4]. It can be transferred to the problem of computing families of MEPTs in applications both for the in situ approach and the offline tabulation. Here, in particular, purely one initial value of neighbored optimization problems is different which can be efficiently exploited through initial value embedding. The embedding strategy computes a linear approximation of the new solution with slightly shifted initial value and initializes the whole optimization problem very efficiently with this linear extrapolation. This generally assures fast convergence to the new solution. The application of these ideas to the computation of MEPT families will be highly beneficial for large-scale mechanisms and higher-dimensional MEPTs.

## 3 Complexity Reduction of Biochemical Reaction Networks

We present an algorithm for complexity analysis of biochemical systems [13] which identifies the degree to which chemical species decouple from the whole network on phase space trajectory pieces in the sense that small perturbations of these species locally do not affect the system dynamics essentially. This method is based on the framework of sensitivity analysis along nominal state trajectories. The singular value decomposition of the phase flow sensitivity matrix allows identification of "enslaved" dynamical modes as those with the smallest sensitivity coefficients and thus the piecewise reduction of the complete model by forced relaxation of these modes. Additionally the algorithm is able to determine local conservation relations corresponding to unity singular values. A subsequent system decomposition into relaxed, nonconstant and constant subspaces provides insight into the system's complexity through the dynamic phase space structure. The coupling analysis is realized

by computation of the contribution of each biochemical species to each of the three subspaces.

### 3.1 Method and Algorithm

The general idea of the complexity analysis by piecewise model decomposition is based on a discrete representation of a continuous time finite-dimensional dynamical system. The whole simulation interval is partitioned into $m$ time windows $[0,T],[T,2T],...,[(m-1)T,mT]$ with length $T$. The idea underlying the complexity analysis is related to an approximation of the full ordinary differential equation (ODE) model consisting of kinetic rate equations for biochemical species by a differential algebraic equation (DAE) model with algebraic (relaxed) and differential (active) variables (modes) $z(t)$ and $y(t)$ on each of the $m$ pieces of a nominal trajectory starting at given initial values.

$$\dot{x}(t) = \frac{\mathrm{d}x}{\mathrm{d}t} = F(x)(\mathrm{ODE}) \longrightarrow \dot{y}(t) = \frac{\mathrm{d}y}{\mathrm{d}t} = f(y,z),\ 0 = g(y,z)(\mathrm{DAE}) \quad (12)$$

The algebraic part $0 = g(y,z)$ of the DAE system representing mode enslavement (relaxation) defines a manifold on which the remaining active modes $y(t)$ evolve. For its computation we transform on each time interval $[(k-1)T,kT], k = 1,...,m$ the original ODE system by vectors $u_1,u_2,...,u_n$ representing contracting and expanding directions between nominal and slightly perturbed trajectories in phase space and relax the strongly contracting modes to the manifold (enslavement). In order to find these directions we use sensitivities (numerical derivatives) characterizing the response of the species concentrations at the end-point of a small trajectory piece after perturbations of the initial values.

$$\delta x(kT) = W(kT) \cdot \delta x((k-1)T), W(kT) := \frac{\partial x(kT)}{\partial x((k-1)T)}, k = 1,...,m \quad (13)$$

$\delta x((k-1)T)$ and $\delta x(kT)$ denote initial and final perturbations of the nominal values $x((k-1)T), x(kT)$ on the time horizon $[(k-1)T,kT]$. The sensitivity matrix $W(kT)$ describes the propagation of the initial perturbations with the phase flow.

The calculation of sensitivity matrices can be carried out via different approaches. The external numerical differentiation method uses the final values of nominal and neighbored (starting from the slightly perturbed initial value) trajectories for approximative evaluation of the sensitivity matrices by the finite difference quotient. For the first time window ($k = 1$) the sensitivity matrix $W(T)$ can be computed by

$$\frac{\partial x(T)}{\partial x(0)} = \frac{x(T,x(0)+\delta x(0)) - x(T,x(0))}{\delta x(0)} \quad (14)$$

In order to obtain high numerical accuracy for sensitivities one has to guarantee high accuracy for nominal and perturbed trajectories, which leads to an

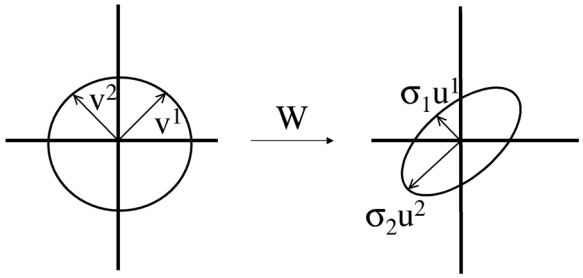



Fig. 4: Geometric interpretation of the singular value decomposition of a $2 \times 2$ matrix $W$ as defined in equation (13), mapping of a sphere to a hyperellipse.

enormous rise in computing time. Moreover, different step sizes of the numerical integrator (while computing nominal and perturbed trajectories) may lead to severe numerical instabilities [5]. To avoid these difficulties we use in our algorithm the internal numerical differentiation (IND) technique proposed by Bock [2]. This technique freezes the adaptively generated time grid for numerical integration of the nominal trajectory. This means that the nominal and perturbed trajectories are computed with the same step size and order by the numerical integrator. IND is implemented in the robust integrator `DAESOL` [1] based on a BDF-method (backward differentiation formula). `DAESOL` is well suited for the solution of initial value and boundary value problems of stiff ordinary differential (ODE) and differential-algebraic equations (DAE) along with sensitivity analysis in the form required here.

For identification of large and small sensitivity modes we use the well known singular value theorem in linear algebra [21] assuring for every regular matrix the existence of a singular value decomposition (SVD) with uniquely determined singular values. SVD decomposes a given matrix $W$ into a product of three matrices

$$W = U \cdot \Sigma \cdot V^T, \ \Sigma = diag(\sigma_i), i = 1, ..., n \tag{15}$$

where $U$ is an orthonormal matrix containing the left singular vectors (as columns), $V$ is an analogous orthonormal matrix of the right singular vectors and $\Sigma$ is a diagonal matrix containing the singular values $\sigma_i$. The diagonal elements of $\Sigma$ can be arranged in descending order such that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n > 0$. The geometrical interpretation of the SVD is illustrated in Fig. 4. The column vectors $u^j$ of $U$ correspond to the semi-axes of the final perturbation hyperellipse with the lengths $\sigma_i$ whereas the right singular vectors $v^j$ of $V$ are the pre-images of these principal semi-axes under the linear map $W$ in the sphere of equidistant initial perturbations [21] .

The geometrical visualization of the SVD applied to the sensitivity matrix $W(kT)$ in (13) reveals that the unit space of the initial perturbations

$\delta x((k-1)T)$ is mapped to a hyperellipse. The column vectors of the matrix $U$ represent directions in the phase space in which contractions and expansions of the initial perturbations with the factor $\sigma_i$ occur. Obviously the shortest semi-axes (with the smallest singular values) belong to the strongly contracting directions. In order to identify the corresponding dynamical modes automatically based on the SVD we transform the original ODE system by the matrix $U^T$.

$$U^T \frac{\mathrm{d}x}{\mathrm{d}t} = U^T F \tag{16}$$

This transformation rotates the original axes of the phase space representing biochemical species into directions of the left singular vectors $(u^j)$.

In the second step we iteratively relax the strongly contracting dynamical modes. This is equivalent to setting the projection of the ODE vector field onto these directions to zero. In this way we obtain a DAE system in transformed coordinates $y(t), z(t)$

$$(\dot{y}, \dot{z}) = U^T \frac{\mathrm{d}x}{\mathrm{d}t} = U^T F = (f, g) = (f, 0) \tag{17}$$

In order to determine the differential dimension (number of differential variables) of an accurately approximating DAE system (17) correctly (i.e. error-controlled maximal relaxation on each time interval $[(k-1)T, kT], k = 1, ..., m)$ we introduce a straightforward relaxation criterion. In each iteration we check the relative error in the remaining differential variables (active modes) at the end point of the current trajectory piece caused by relaxation of the enslaved modes.

$$\frac{|y_i^\star(kT) - y_i(kT)|}{|y_i(kT)|} \leq TOL, i = 1, ..., n_{red} \tag{18}$$

$y_i^\star(kT)$ denotes the solution of the DAE approximation (17) with differential dimension $n_{red}$ and $y_i(kT)$ is the solution of the transformed full ODE system (16). If this error-criterion is fulfilled with a user defined tolerance $TOL$ the reduction of the differential variables by one is accepted and we try to relax the next strongest contracting mode. The iterative procedure stops when the relative error criterion is violated. Consistent initial values for the DAE system (17) are computed within `DAESOL` by a homotopy-like continuation method [1] starting from the current point in phase space. This corresponds to an instantaneous relaxation of the modes with small singular values.

In addition to the local relaxation of modes, the discrete dynamical systems approach based on singular value decomposition of piecewise phase flow sensitivity matrices allows identification of constant dynamical modes (local conservation relations). Modes corresponding to phase space directions in which perturbations do not alter with the phase flow are characterized by singular values equal to 1. After the identification of the maximal number of relaxed modes (smallest possible number of differential equations on the interval $[(k-1)T, kT]$) by the algorithm described above we try to fix the transformed differential variable $y_m$ corresponding to the smallest difference

$|1-\sigma_{y_m}|$. For this purpose we evaluate the solution of the DAE approximation (12) on this interval while fixing $y_m$ to its value $y_m((k-1)T)$ and check again the error criterion (18). If this criterion is fulfilled the variable $y_m$ is eliminated from the DAE system by fixing it to its initial value on the current trajectory piece. Then we repeat the described procedure with the dynamical mode corresponding to the next smallest deviation $|1-\sigma|$ of its singular value from 1. Finally we continue the whole algorithmic procedure of relaxing and fixing dynamical modes on the next time interval $[kT,(k+1)T])$.

The results of the algorithm described above can be exploited for a network coupling analysis of biochemical species. For this purpose we compute the projections of the original coordinate axes unit vectors (representing the biochemical species) onto the constant, non-constant and relaxed subspaces spanned by the corresponding orthogonal column vectors $u^j$ from the SVD. These projections of the original species $i$ are given by

$$p_i^{const} := \sum_{j\in\mathcal{C}} u_i^j \cdot u^j,\ p_i^{nonconst} := \sum_{j\in\mathcal{A}} u_i^j \cdot u^j,\ p_i^{rel} := \sum_{j\in\mathcal{F}} u_i^j \cdot u^j,\ i=1,...,n \tag{19}$$

where $\mathcal{C}$ is a set of indices corresponding to the constant directions $u^j$. $\mathcal{A}$ and $\mathcal{F}$ are defined as $\mathcal{A} := \{1,...n_{red}\} \setminus C$ and $\mathcal{F} := \{n_{red}+1,...,n\}$ respectively. $u_i^j$ denote the i-th component of the left singular vector $u^j$. Due to the orthogonality of the vectors $u^j$ the contributions of the species to these subspaces can be determined as follows

$$r_i^{space} := \frac{\|p_i^{space}\|}{\|p_i^{const}\| + \|p_i^{nonconst}\| + \|p_i^{rel}\|}, i=1,...,n \tag{20}$$

where $space \in \{const, nonconst, rel\}$. The contributions of the original species to the whole non-relaxed (active) subspace including constant and non-constant directions $u^j$ is given by

$$r_i^{act} = r_i^{const} + r_i^{nonconst}, i=1,...,n \tag{21}$$

### 3.2 Application Results and Discussion

*Single enzyme Michaelis-Menten system*

In order to test the complexity reduction algorithm described in the previous section, we analyze the coupling behavior of biochemical species from two reaction systems displaying qualitatively different kinds of dynamical behavior. The first one is the well known irreversible Michaelis-Menten kinetics for a single enzyme system

$$E + S \overset{k_{\pm 1}}{\longleftrightarrow} ES \xrightarrow{k_2} E + P \tag{22}$$

described by two ordinary differential equations

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = -k_1(E_T - x_2)x_1 + k_{-1}x_2$$
$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = k_1(E_T - x_2)x_1 - k_2x_2 - k_{-1}x_2 \tag{23}$$

with $x_1 = [S], x_2 = [ES]$ (substrate and enzyme-substrate-complex concentrations respectively),$[E_T] = [E] + [ES] = 100.0$ (total enzyme concentration), rate coefficients $k_1 = 1.0, k_{-1} = 1.0, k_2 = 0.5$ and initial conditions $x_1(0) = 100.0, x_2(0) = 0$. According to Fig. 5 the dynamics of Michaelis-

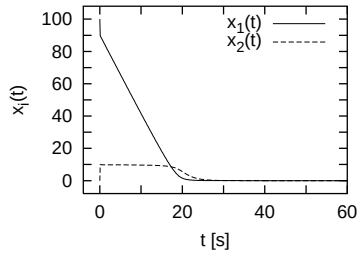



Fig. 5: Numerical simulation of the Michaelis-Menten model (23).

Menten system exhibit 3 phases. After a very short initiation phase with formation of enzyme-substrate-complex the second phase (plateau between 0 and 20s) with nearly constant enzyme-substrate-complex and decreasing substrate concentration can be distinguished. In the third phase (after 20s) a decay of the enzyme-substrate-complex is observed whereas the concentration of the substrate is nearly zero because it has been completely consumed. The relative contributions of each species to the active subspace are depicted in Figure 6 and reflect the stage dynamics described above. For instance in the second phase, where the dynamics of the system is dominated by the substrate decay, the contribution of $x_2$ is 0% (fully relaxed) and the whole active dynamics is governed by of $x_1$ (contribution of 100%). Thus, the computational results of our algorithm (Fig. 6) confirm the validity range of the well-known quasi-steady-state assumption commonly applied for the enzyme-substrate-complex in Michaelis-Menten type kinetics to describe the whole system with a single rate equation. The dimension of the constant subspace is in this case always zero.

*Peroxidase-oxidase (PO) oscillator*

To demonstrate that our algorithm can provide valuable insight into more complex biochemical reaction networks we analyze the peroxidase-oxidase (PO) oscillator [18, 8], a kinetic model for the enzymatic reduction of oxygen to water by $NADH$:

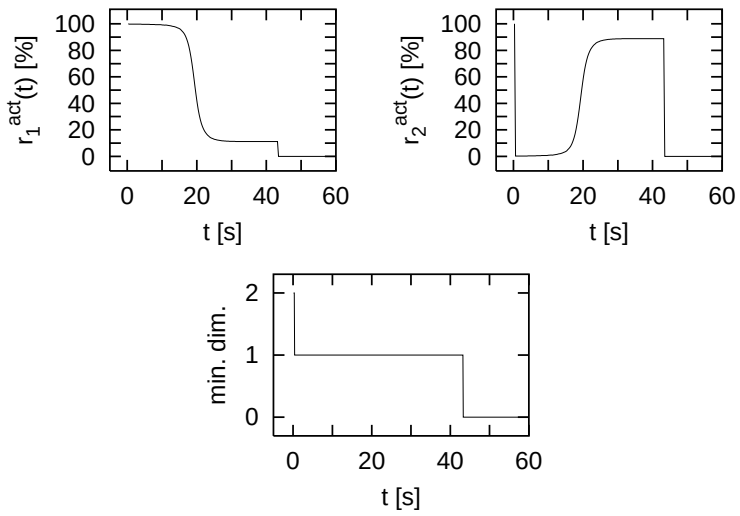



Fig. 6: Michaelis-Menten model: *Top*: Relative contribution $r_i^{act}(t)$ of each chemical species $x_i, i = 1, 2$ in % to the subspace of active dynamical modes according to equation (21). *Bottom*: Minimum dimension (dimension of active subspace) for the model (23) as a function of time computed by the presented algorithm with an error tolerance $TOL = 10^{-2}$ in (18) and interval length $T = 0.3$ (see (13)).

$$NADH + \frac{1}{2}O_2 + H^+ \longrightarrow NAD^+ + H_2O \tag{24}$$

To provide an internal standard for component decoupling, we extend the PO model artificially by a loosely coupled hypothetic second enzyme $Enz_{act/inact}$ which has an active and an inactive state, is activated by superoxide ions and not involved in the oscillatory mechanism. Based on a detailed reaction mechanism consisting of 10 differential equations, Figure 7 shows numerical simulation results for four selected chemical species reflecting the rich dynamical behavior of the PO system (relaxation oscillations followed by regular oscillations and finally a steady state).

Figure 8 shows the dimension of the active subspace (number of differential variables) over time corresponding to the degrees of freedom necessary to describe the complete dynamics of the full system accurately. Obviously, on large parts of the simulation interval the phase space dynamics of the original ODE system can be represented accurately by a reduced system with dimension $\leq 4$. However, within a small region of each period in the relaxation oscillation regime ($0 \leq t \leq 2800$), the full mechanism is required. Figure 9 shows the results of the dynamic coupling analysis for the biochemical species $NADH, NAD^{\cdot}, H_2O_2$ and $Enz_{act}$. The contribution of $NADH$ to the active subspace is relatively large. Thus this species couples to the essential dynam-

| reaction[a] | rate expression | constant |
|---|---|---|
| (1) $NADH + O_2 + H^+ \longrightarrow NAD^+ + H_2O_2$ | $k_1[NADH][O_2]$ | 3.0 [b] |
| (2) $H_2O_2 + Per^{3+} \longrightarrow coI$ | $k_2[H_2O_2][Per^{3+}]$ | $1.8 \times 10^7$ [b] |
| (3) $coI + NADH \longrightarrow coII + NAD^{\cdot}$ | $k_3[coI][NADH]$ | $4.0 \times 10^5$ [b] |
| (4) $coII + NADH \longrightarrow Per^{3+} + NAD^{\cdot}$ | $k_4[coII][NADH]$ | $2.6 \times 10^5$ [b] |
| (5) $NAD^{\cdot} + O_2 \longrightarrow NAD^+ + O_2^-$ | $k_5[NAD^{\cdot}][O_2]$ | $2.0 \times 10^7$ [b] |
| (6) $O_2^- + Per^{3+} \longrightarrow coIII$ | $k_6[O_2^-][Per^{3+}]$ | $1.7 \times 10^6$ [b] |
| (7) $2O_2^- + 2H^+ \longrightarrow H_2O_2 + O_2$ | $k_7[O_2^-]^2$ | $2.0 \times 10^7$ [b] |
| (8) $coIII + NAD^{\cdot} \longrightarrow coI + NAD^+$ | $k_8[coIII][NAD^{\cdot}]$ | $11.0 \times 10^7$ [b] |
| (9) $2NAD^{\cdot} \longrightarrow NAD_2$ | $k_9[NAD^{\cdot}]^2$ | $5.6 \times 10^7$ [b] |
| (10) $Per^{3+} + NAD^{\cdot} \longrightarrow Per^{2+} + NAD^+$ | $k_{10}[Per^{3+}][NAD^{\cdot}]$ | $1.8 \times 10^6$ [b] |
| (11) $Per^{2+} + O_2 \longrightarrow coIII$ | $k_{11}[Per^{2+}][O_2]$ | $1.0 \times 10^5$ [b] |
| (12) $\longrightarrow NADH$ | $k_{12}$ | 0.132 |
| (13) $O_2(gas) \longrightarrow O_2(liquid)$ | $k_{13}[O_2]_{eq}$ | $4.4 \times 10^{-3}$ [d,e] |
| $(-13)$ $O_2(liquid) \longrightarrow O_2(gas)$ | $k_{-13}[O_2]$ | $4.4 \times 10^{-3}$ [d] |
| (14) $Enz_{inact} + O_2^- \longrightarrow Enz_{act}$ | $\frac{k_{14}[O_2^-]^5}{(K_f^5 + [O_2^-]^5)}$ | 0.005 [b] ($k_{14}$) |
| | | 0.4 [cf] ($K_f$) |
| (15) $Enz_{act} \longrightarrow Enz_{inact}$ | $k_{15}[Enz_{act}]$ | 1.6 [d] |

Table 1: Detailed model of the peroxidase–oxidase reaction coupled to the activation of an enzyme *Enz* ([a] $Per^{3+}$ and $Per^{2+}$ indicate iron(III) and iron(II) peroxidase respectively. coI, coII and coIII indicate the enzyme intermediates compound I, compound II and compound III. [b] In $M^{-1}\ s^{-1}$. [c] In $M$. [d] In $s^{-1}$. [e] The value of $[O_2]_{eq}$ is $12\mu M$. [f] The amount of $Enz_{inact}$ is assumed to be large compared to $Enz_{act}$ and therefore considered to be constant: total concentration included in rate constant $k_{14}$). The initial condition values are 12.0 $\mu M$ for $O_2$ and 1.5 $\mu M$ for $Per^{3+}$, all other initial concentrations are zero.

ics of the full reaction network in a sense that small perturbations of $NADH$ cause a significant disturbance of the system dynamics. In contrast, the other three species seem to decouple from the network due to mode relaxation. The activated enzyme $Enz_{act}$, our internal standard that has been added with a kinetic law that assures decoupling, in fact decouples on the intervals where it is not active (compare with Figure 7). The decoupling of the species $H_2O_2$ is in accordance with recent results obtained from a different approach which is also related to model reduction and described in [20]. At first sight, the decoupling of $NAD^{\cdot}$ radicals seems to be in conflict with the fact that they play a significant role in the autocatalysis cycle within the PO reaction [19]. But numerical simulations of the PO mechanism starting from a slightly perturbed initial value of $NAD^{\cdot}$ variable show that perturbations of this species indeed do not change the systems dynamics. The perturbed system levels off very fast to the unperturbed dynamics.

Fig. 10 shows the dimension analysis of the constant subspace for the PO reaction system. The maximal dimension of this subspace does not exceed 3.

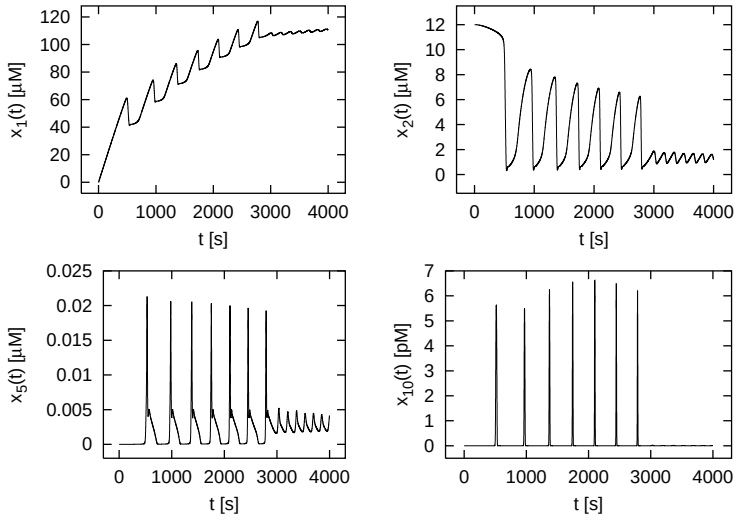



Fig. 7: PO model (see Table 1): Numerical simulation of the species $x_i, i = 1 : NADH, i = 2 : O_2, i = 5 : NAD^{\cdot}, i = 10 : Enz_{act}$ using `DAESOL` [1].

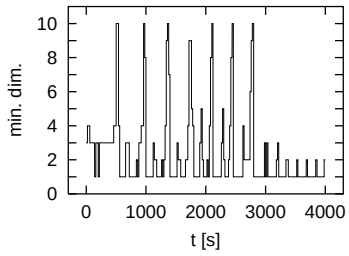



Fig. 8: Minimum dimension (dimension of active subspace) for the PO model as a function of time for two different dynamical regimes (relaxation oscillations: $0 \leq t \leq 2800s$ and harmonic oscillations: $t \geq 2800s$, see also Fig. 7), $TOL = 10^{-2}$ in (18) and interval length $T = 20.0$ (see (13)).

It reflects the qualitative change in the system dynamics from transient relaxation oscillations to regular oscillations with small amplitude (at approximately $t = 3000$). In the latter regime at least one mode is constant during a full oscillation period. The contributions of the species $NADH, O_2$ and $NAD^{\cdot}$ to the constant subspace are depicted in Figure 11. The $NAD^{\cdot}$ radical does not contribute to the constant subspace due to its decoupling from the active subspace (compare Fig. 9). High contribution ($\approx 70\%$) of the species $NADH$ in each oscillation period is caused by its only slowly increasing concentration (Fig. 7) during larger oscillatory changes in most other species. The same

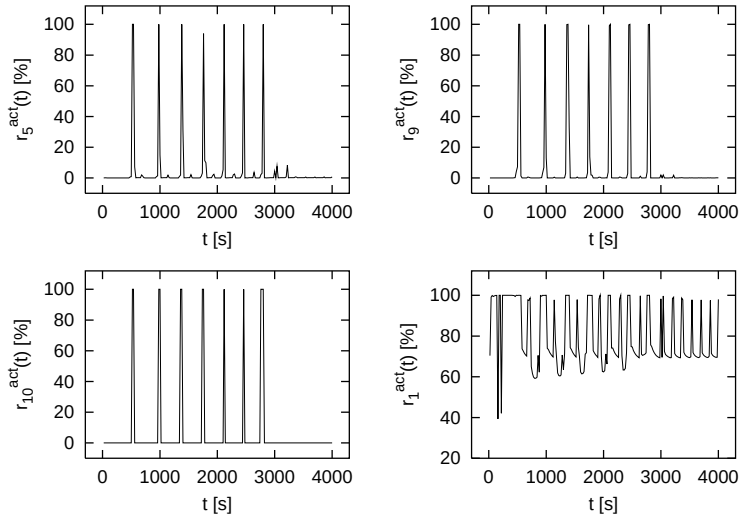



Fig. 9: PO model: Relative contribution $r_i^{act}(t)$ of the chemical species $x_i, i = 5 : NAD^{\cdot}, i = 9 : H_2O_2, i = 10 : Enz_{act}, i = 1 : NADH$ in % to the subspace of active dynamical modes (including constant and non-constant directions) according to equation (21).

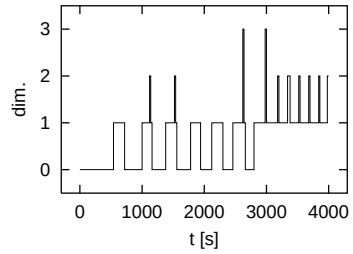



Fig. 10: PO model: Dimension of the constant subspace, $TOL = 10^{-2}$ in (18) and interval length $T = 20.0$ (see (13)).

argument holds partly also for the spieces $O_2$. However, its contribution to the constant subspace is smaller ($\approx 30\%$). Conservation relations (constant modes) that are valid over the full time horizon are not found because chemical mass conservation laws have been already eliminated during kinetic modeling of the reaction mechanism.

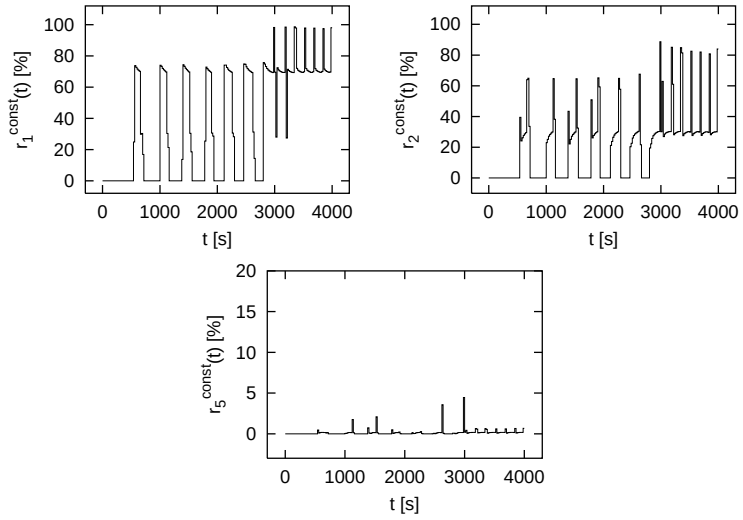



Fig. 11: PO model: Relative contribution $r_i^{const}(t)$ of the chemical species $x_i, i = 1 : NADH, i = 2 : O_2, i = 5 : NAD^{\cdot}$ in % to the subspace of constant dynamical modes according to equation (20).

### 3.3 Outlook

Although the approach to complexity reduction presented here provides a useful analysis tool for network coupling in biochemical kinetics, there is further need to identify network modules in the sense of subnetworks that are tightly coupled within themselves and loosely coupled to the rest of the network. Modules and redundancy are central themes in biochemical networks and closely connected to their physiological functions. A sensitivity related analysis similar to the one discussed here may be promising but probably has to take additionally into account sensitivities of sensitivities (second derivatives) and the topology of the reaction network.

*Acknowledgement.* The authors would like to thank Prof. J. Warnatz (IWR, Heidelberg) for financial and scientific support, Prof. H.G. Bock (IWR, Heidelberg) for providing the `MUSCOD-II` and `DAESOL` code and Dr. J. Zobeley and Dr. U. Kummer (EML Research, Heidelberg) for collaboration.

## References

1. I. Bauer, F. Finocchi, W.J. Duschl, H.-P. Gail, J.P. Schlöder: Simulation of chemical reactions and dust destruction in protoplanetary accretion discs. *Astron. Astrophys.* **317**, 273–289 (1997)

2. H.G. Bock: Numerical treatment of inverse problems in chemical reaction kinetics. In: *Modelling of Chemical Reaction Systems (Springer Series in Chemical Physics)*, ed. by K.H. Ebert, P. Deuflhard, and W. Jäger, chapter 8, 102–125 (Springer, Heidelberg 1981)
3. H.G. Bock, K.J. Plitt: A multiple shooting algorithm for direct solution of constrained optimal control problems. In: *Proceedings of the ninth IFAC world congress, Budapest*, 242–247 (Pergamon Press 1984)
4. M. Diehl, H.G. Bock, J.P. Schlöder: A real-time iteration scheme for nonlinear optimization in optimal feedback control. *SIAM J. Control Optim.* **43** (5), 1714–1736 (2005)
5. C.W. Gear, T. Vu: Smooth numerical solution of ordinary differential equations. In: *Numerical Treatment of Inverse Problems in Differential and Integral Equations*, ed. by P. Deuflhard E. Hairer (Birkäuser, Boston 1983)
6. P. Glansdorff, I. Prigogine: *Thermodynamic Theory of Structure, Stability and Fluctuations* (Wiley, New York 1971)
7. A.N. Gorban, I.V. Karlin, A.Y. Zinovyev: Constructive methods of invariant manifolds for kinetic problems. *Phys. Rep.* **396**, 197–403 (2004)
8. M.J.B. Hauser, U. Kummer, A.Z. Larsen, L.F. Olsen: Oscillatory dynamics protect enzymes and possibly cells against toxic substances. *Faraday Discuss.* **120**, 215–227 (2001)
9. R. Heinrich, S. Schuster: *The Regulation of Cellular Systems* (Chapman & Hall, New York 1996)
10. H. Kitano: Computational systems biology. *Nature* **420**, 206–210 (2002)
11. D. Kondepudi, I. Prigogine: *Modern Thermodynamics* (John Wiley & Sons, Baffins Lane, England 1936)
12. D. Lebiedz: Computing minimal entropy production trajectories: An approach to model reduction in chemical kinetics. *J. Chem. Phys.* **120** (15), 6890–6897 (2004)
13. D. Lebiedz, J. Kammerer, U. Brandt-Pollmann: An automatic network coupling analysis for dynamical systems based on detailed kinetic models. *Phys. Rev. E* **72**, 041911–1–8 (2005)
14. D.B. Leineweber, I. Bauer, H.G. Bock, J.P. Schlöder: An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization - part I: theoretical aspects. *Comput. Chem. Eng.* **27**, 157 (2003)
15. U. Maas, S.B. Pope: Simplifying chemical kinetics: Intrinsic low-dimensional manifolds in composition space. *Combust. Flame* **88**, 239–264 (1992)
16. G. Nicolis, I. Prigogine: *Self-organization in Nonequilibrium Systems* (Wiley, New York 1977)
17. M.S. Okino, M.L. Mavrovouniotis: Simplification of mathematical models of chemical reaction systems. *Chem. Rev.* **98**, 391–408 (1998)
18. A. Scheeline, D.L. Olson, E.P. Williksen, G.A. Horras, M.L. Klein, R. Larter: The peroxidase-oxidase oscillator and its constituent chemistries. *Chem. Rev.* **97**, 739–756 (1997)
19. I. Schreiber, Y.-F. Fen, J. Ross: Categorization of some oscillatory enzymatic reaction. *J. Phys. Chem.* **100**, 8556–8566 (1996)
20. R. Straube, D. Flockerzi, S.C. Müller, M.J.B. Hauser: Reduction of chemical reaction networks using quasi-integrals. *J. Phys. Chem. A* **109**, 441–450 (2005)
21. L.N. Trefethen, D. Bau: *Numerical Linear Algebra* (SIAM, Philadelpia 1997)

22. J. Zobeley, D. Lebiedz, J. Kammerer, A. Ishmurzin, U. Kummer: A new time-dependent complexity reduction method for biochemical systems. *Trans. Comput. Syst. Biol.* **1**, 90–110 (2005)

# Dynamics of the Plasma Sheath

M. Slemrod*

Department of Mathematics, University of Wisconsin–Madison, Madison, Wisconsin 53706, USA, slemrod@math.wisc.edu

**Summary.** The motion of the interface separating the sheath boundary layer and quasi-neutral plasma is formulated in the terms of level set motion. The equations for the motion are derived and given as a system of partial differential equations. When restricted to the case of planar, cylindrical, and spherical symmetries, these equations become ordinary differential equations.

## 1 Introduction

In a recent sequence of papers with Ha [4] and Feldman & Ha [3]. I have generalized an earlier paper of K.-U. Riemann and Th. Daube [8] for the propagation of a sheath interface in a plasma of ions and electrons. The purpose of this article is to present an elementary exposition of the underlying issues, both physical and mathematical, that arise is this most natural of multi-scale problems. However while the presentation is elementary, there is an underlying theme to the work. Namely for multi-scale systems, one can do worse than avoid transition layers completely and replace the transition layers by propagating sharp interfaces described as level sets [6]. The reason is simple: what one loses in accuracy of resolution of small scales one gains in both the simplicity of the level set formalism and the underlying geometric understanding.

This paper is divided into four sections after this Introduction. Section 2 describes the underlying physical problem within the context of a classical gas dynamics. Section 3 reformulates the problem of sheath formation for the plasma problem. Section 4 formulates an approach to the location of the plasma sheath interface for planar, cylindrical, and spherical symmetries. Finally Section 5 generalizes Section 4's results to the general non-symmetric case and then recovers the dynamics of the sheath interface in the symmetric cases as an application of the general theory.

* This research was sponsored by the U.S. National Science Foundation under grants DMS-0203858 and DMS-0243722.

## 2 The Euler Equations with Planar, Radical, and Spherical symmetry

Consider a gas with constant temperature satisfying the ideal gas law

$$p(\rho) = c^2\rho \tag{1}$$

where $p$ is the pressure, $\rho$ is the density, and $c$ is the speed of sound (a constant). The equations of motion are given by the balance of mass and momentum:

$$\rho_t + (\rho u)_r = -\frac{\alpha}{r}(\rho u), \tag{2}$$

$$(\rho u)_t + (\rho u^2 + p)_r = -\frac{\alpha}{r}(\rho u^2) + \rho\beta(u)u. \tag{3}$$

where $\alpha = 0, 1, 2$ depending on whether we have planar, cylindrical, or spherical symmetry. Here $u$ is the velocity of gas and $\beta(u) \geq 0$ is the friction coefficient.

To keep things simple for now lets consider only steady motion where $\rho_t = (\rho u)_t = 0$. In this case (1), (2), (3) imply

$$u_r = \frac{\alpha c^2 u}{r(u^2 - c^2)} + \frac{\beta(u)u^2}{u^2 - c^2}. \tag{4}$$

The implication of (4) is obvious. In the cases $\alpha = 0$: $\beta(u) > 0$ will imply a singularity in the velocity as the gas attempts the transition from subsonic flow $u^2 < c^2$ to supersonic flow $u^2 > c^2$; $\alpha = 1, 2$: $\beta(u) \geq 0$ will imply the singularity.

## 3 Collisional and Collisionless Plasmas

Consider a plasma consisting of ions and electrons [5]. The density of the ions is given by $n_i$, the velocity of ions is $u$, the density of the electrons is given by $u_e$, the electric potential is $-\varphi$. Under suitable scaling of independent and dependent variables the equations of motion for a plasma of cold ions (with ion pressure identically zero) is given by conservation of mass, momentum, and Poisson equations:

$$n_{i_t} + (n_i u)_r = -\frac{\alpha}{r}(n_i - u), \tag{5}$$

$$(n_i u)_t + (n_i u^2)_r = n_i\varphi_r + n_i\beta(u)u, \tag{6}$$

$$\epsilon^2 \frac{1}{r^\alpha}\frac{\partial}{\partial r}\left(r^\alpha \frac{\partial \phi}{\partial r}\right) = n_i - n_e. \tag{7}$$

Again $\alpha = 0, 1, 2$ represents the cases of planar, cylindrical and spherical symmetry.

We assume the electron mass is much smaller than the ion mass. An asymptotic analysis of the conservation of momentum for the electrons then yields the classic formula for the electron density $n_e = e^{-\varphi}$ (Boltzmann's relation). Hence (7) simplifies to

$$\epsilon^2 \frac{1}{r^\alpha} \frac{\partial}{\partial r} \left( r^\alpha \frac{\partial \varphi}{\partial r} \right) = n_i - e^{-\varphi}. \tag{8}$$

Note we have not placed any additional source terms (ionization) in (5) but continued to place the friction term in (6). The case with $\beta(u) > 0$ is called a *collisional plasma*, if $\beta(u) \equiv 0$ it is a *collisionless plasma*.

The parameter $\epsilon > 0$ is called the Debye length and captures any length scale in our plasma. The formal limit $\epsilon \to 0$ is called the quasi-neutral limit and provides a tempting simplification to our problem.

In the quasi-neutral limit (8) implies $n_i = e^{-\varphi}$ and hence $\varphi = -\ln n_i$. Substitution of this choice of $\varphi$ into (5), (6) yields the equations of motion for our quasi-neutral plasma

$$n_{i_t} + (n_i u)_r = -\frac{\alpha}{r}(n, u), \tag{9}$$

$$(n_i u)_t + (n_i u^2 + n_i)_r = -\frac{\alpha}{r}(u_i u^2) + n_i \beta(u) u, \tag{10}$$

which are just the isothermal Euler equations (2), (3) with sound speed $c = 1$.

Now we see the difficulty encountered in plasma physics: imposition of a large potential difference across a plasma of ions and electrons will cause the ions to attempt to pass from slow subsonic flow to fast supersonic flow. But the analysis of Section 2 shows that for steady motion (9), (10) must cause a singularity to form when $u^2 = c^2$ so that $|u_r| \to \infty$ (the Bohm criterion). Hence we can only conclude that as we make the transition from subsonic to supersonic flow, the quasi-neutral limit is no longer valid and smaller $\epsilon$-scale becomes relevant. In fact asymptotic analysis [1], [7] of steady motion shows that in a bounded domain a boundary layer of order $\epsilon$ occurs which is separated form the quass-neutral regime by a transition region of order $\epsilon^{4/5}$. The boundary layer is called the plasma sheath.

## 4 Dynamics of the Plasma Sheath

One way to study dynamics of the plasma sheath is to rescale in space and time. It we set $\overline{t} = \frac{t}{\epsilon}$, $\overline{r} = \frac{r}{\epsilon}$ then in the overbar variables (5), (6), (8) become

$$n_{i_{\overline{t}}} + (n_i u)_{\overline{r}} = -\frac{\alpha}{\overline{r}}(n_i u). \tag{11}$$

$$(n_i u)_{\overline{t}} + (n_i u^2)_{\overline{r}} = n_i \varphi_{\overline{r}} + \epsilon n_i \beta(u) u, \tag{12}$$

$$\frac{1}{\overline{r}^{\alpha}}\frac{\partial}{\partial \overline{r}}\left(\overline{r}^{\alpha}\frac{\partial \varphi}{\partial \overline{r}}\right) = n_i - e^{-\varphi}. \tag{13}$$

Thus we see that on the small $\epsilon$-scale the friction term is negligible and may be neglected. (Of course the same would be true for any ionization terms as well.) So within the sheath boundary layer (11)–(13) pervade and the friction term is negligible. In fact since $\varphi$ is very large, we can neglect $e^{-\varphi}$ as well which simplifies (11)–(13) even further.

On the other hand in the far field we expect the quasi-neutral system (9)–(10) to be valid which is the $\overline{r}$, $\overline{t}$ independent variables becomes

$$n_{i_{\overline{t}}} + (n_i u)_{\overline{r}} - \frac{\alpha}{\overline{r}}(n_i u), \tag{14}$$

$$(n_i u)_{\overline{t}} + (n_i u^2 + n_i)_{\overline{r}} = -\frac{\alpha}{\overline{r}}(n_i u^2) + \epsilon n_i \beta(u) u. \tag{15}$$

Again $\epsilon$ multiplies the friction term. This reflects that friction has a small effect on the small order $\epsilon$ time scale. Now we see that finally we have two descriptions of the plasma. First in the boundary layer we have the boundary layer dynamics from (11)–(13):

$$n_{i_{\overline{t}}} + (n_i u)_{\overline{r}} = -\frac{\alpha}{\overline{r}}(n_i u), \tag{16}$$

$$(n_i u)_{\overline{t}} + (n_i u^2)_{\overline{r}} = n_i \varphi_{\overline{r}}, \tag{17}$$

$$\frac{1}{\overline{r}^{\alpha}}\frac{\partial}{\partial \overline{r}}\left(\overline{r}^{\alpha}\frac{\partial \varphi}{\partial \overline{r}}\right) = n_i \tag{18}$$

On the other hand in the far field we have quasi-neutral dynamics

$$n_{i_{\overline{t}}} + (n_i u)_{\overline{r}} = \frac{-\alpha}{\overline{r}}(n_i u), \tag{19}$$

$$(n_i u)_{\overline{t}} + (n_i u^2 + n_i)_{\overline{r}} = -\frac{\alpha}{\overline{r}}(n_i u^2). \tag{20}$$

The issue now is where to switch our two descriptions: (16)–(18) (the boundary layer dynamics) and (19)–(20) (the far field quasi-neutral dynamics).

One answer is just use the Bohm criterion and switch when $u^2 = c^2$. For dynamic problems this is not enough and the Bohm criterion must be supplemented by a second criterion. Following Riemann and Daube [8] Feldman, Ha, and Slemrod [3], [4] have used a second criterion based on the electric potential. Simply put matching asymptotic expansions in the transition layer yields

$$\varphi_r \sim \epsilon^{-\gamma},$$

where $0 < \gamma < 1$. Hence since $\overline{r} = \frac{r}{\epsilon}$ we see

$$\varphi_{\overline{r}} \sim \epsilon^{1-\gamma}$$

and $\varphi_{\overline{r}}$ is small is the transition region. We combine this observation with Bohm criterion and define the *plasma sheath interface* as the curve $\overline{r}_s(t)$ where $u^2(\overline{r}_s(t), t) = c^2$ (Bohm) and $\varphi_{\overline{r}}(\overline{r}_s(t), t) = 0$.

## 5 Generalization to Non-Symmetric Case

In Section 1–4 I have given a short description how that plasma-sheath system with a sharp interface is formulated in the canonical symmetries. But once the ideas are in place the generalization to the general non-symmetric case follows logically. This has been given in the paper of Feldman, Ha,. and Slemrod [3].

First we record the easy parts: The boundary layer sheath system and the far field quasi-neutral equations. In the interior sheath region we have

$$n_{\overline{t}} + \nabla_{\overline{\mathbf{x}}} \cdot (n\mathbf{u}) = 0, \tag{21}$$

$$\mathbf{u}_{\overline{t}} + (\mathbf{u}, \nabla_{\overline{\mathbf{x}}})\mathbf{u} = \nabla_{\overline{\mathbf{x}}}\varphi, \tag{22}$$

$$\Delta_{\overline{\mathbf{x}}}\phi = n, \tag{23}$$

where $\Delta_{\overline{x}}$ denotes the Laplacian with respect to $\overline{\mathbf{x}} = \frac{\mathbf{x}}{\epsilon}$ .

On the other hand in the far field quasi-neutral region we have

$$n_{\overline{t}} + \nabla_{\overline{\mathbf{x}}} \cdot (n\mathbf{u}) = 0 \tag{24}$$

$$\mathbf{u}_{\overline{t}} + (\mathbf{u} \cdot \nabla_{\overline{\mathbf{x}}})\mathbf{u} + \nabla_{\overline{\mathbf{x}}}(\ln n) = 0. \tag{25}$$

The main issue is how to generalize the definition of the interface between the two regions. The definition we have chosen is the obvious one.

**Definition** A *plasma sheath interface* $S(\overline{t})$ separating a quasi-neutral region and an interior sheath region is the level set of the normal component of the ion velocity and electric fields, i.e.

$$S(\overline{t}) = \{\ \overline{\mathbf{x}} \in \mathbf{R}^3;\ \mathbf{u} \cdot \boldsymbol{\nu}(\overline{\mathbf{x}}, \overline{t}) = -1 \ \text{ (or } +1 \text{ depending on the direction of flow of ions)}, \ \nabla_{\overline{\mathbf{x}}}\varphi \cdot \boldsymbol{\nu}(\overline{\mathbf{x}}, \overline{t}) = 0\ \},\ \ \overline{t} > 0,$$

where $\boldsymbol{\nu}$ is the exterior unit normal to the interface.

We note the $\pm 1$ just represents the fact that the sound speed in our formulation has been rescaled to 1. In the presence of the canonical symmetries, the definition of course reduces to the definition given in Section 4 for the symmetric cases. More importantly the level set formation gives us an immediate method of finding the dynamics of the interface. First to simplify matters we will henceforth drop the overbars on $t, \mathbf{x}$. Next we define normal time derivative of a function $f(\mathbf{x}, t)$ by

$$\frac{\delta f}{\delta t} = \partial_t f + V\boldsymbol{\nu} \cdot \nabla_{\mathbf{x}} f \tag{26}$$

where $\boldsymbol{\nu} = V\boldsymbol{\nu}$ is the velocity of the normal component interface. Hence for an interface described by a level set function $\psi(\mathbf{x}, t) = 0$, $\nabla_{\mathbf{x}}\psi$ is trivially in the normal direction to the interface and

$$\frac{\delta \psi}{\delta t} = 0. \tag{27}$$

Furthermore elementary differential geometry tells us

$$\frac{\delta \boldsymbol{\nu}}{\delta t} = \nabla_s V \tag{28}$$

when $\nabla_s$ denotes the surface gradient on the interface. Finally we differentiate the level set identities $\mathbf{u}\cdot\boldsymbol{\nu}(\mathbf{x},t) = -1$, $\nabla_{\mathbf{x}}\varphi\cdot\boldsymbol{\nu}(\mathbf{x},t) = 0$ along the propagating interface, i.e. take $\frac{\delta}{\delta t}$ of these equations. We then recover for "normal" flow, i.e. flows for which $\mathbf{u}^T = \mathbf{0}$ ($\mathbf{u}^T$ the tangential component of velocity), the following system

$$\frac{\delta \psi}{\delta t} = 0, \tag{29}$$

$$\frac{\delta n}{\delta t} = n\nabla_{\mathbf{x}}\cdot\boldsymbol{\nu} \tag{30}$$

$$(V+1) + \frac{\mathbf{h}\cdot\boldsymbol{\nu}}{n} = -\frac{1}{n}\nabla_s(V\nabla_s \ln n). \tag{31}$$

Hence $\mathbf{h}$ is an additional quantity, the ion current and (31) gives us a formula for computing $V$. (The computations are given in [3]).

In itself (29)–(31) is of course very neat. Moreover we can make an immediate observation. Since $\nabla_{\mathbf{x}}\cdot\boldsymbol{\nu}$ is twice the mean curvature $K_m$ of the interface (31) says normal velocity $V$ is inversely proportional to ion density $n$ and taking $\frac{\delta}{\delta t}$ of both sides of (31) we see $\frac{\delta V}{\delta t}$ measuring acceleration of the interface is proportional to $\frac{\delta n}{\delta t}$. If we combine this observation with (30) we see the acceleration of interface is proportional to the mean curvature of the interface (plus other terms). Thus the theory shows we have a curvature driven interface, but it is the acceleration that is driven by curvature.

Before ending this section, it is perhaps worthwhile to see what (29)–(31) gives in the cases of planar, cylindrical, and spherical symmetry, i.e. $\alpha = 0, 1, 2$. First consider planar solutions with the following ansatz: $\psi(\mathbf{x},t) = x_1 - s(t)$, $n(\mathbf{x},t) = n(x_1,t)$ and $\mathbf{h} = (h,0,0)$. In this case the surface gradients vanish and $\nabla\cdot\boldsymbol{\nu} = 0$ as well. Hence

$$\frac{\delta\psi}{\delta t} = 0, \frac{\delta n}{\delta t} = 0, \;\; V = -1 - \frac{h}{n_0}. \tag{32}$$

Since $\frac{\delta}{\delta t} = \partial_t + V\cup\nabla_{\mathbf{x}}$ this implies

$$-\dot{s} + V = 0, \;\; \partial_t n + \dot{s}\partial_{x_1} n = 0, \;\; V = -1 - \frac{h}{n}. \tag{33}$$

It is then easy to see $n(x,t) = n_0$ (a constant) and $\dot{s} = -1 - \frac{h}{n_0}$. Hence the interface is given and $x_1 = s(t)$ where $\dot{s}$ satisfies $\dot{s} = -1 - \frac{h}{n_0}$.

Next we consider the spherically symmetric solutions with $\psi(x,t) = r - s(t)$. Then we see that (29)–(31) yield

$$V = \dot{s}(t), n_t + Vn_r = \frac{2n}{r},$$

$$\dot{s} + 1 + \frac{\hat{h}}{n} = 0$$

where $\hat{h}$ is now the radial component of ion current. We combine the first two equations to get

$$n_t + \dot{s} \cdot n_r = \frac{2n}{r}, \quad \dot{s} + 1 + \frac{\hat{h}}{n} = 0. \tag{34}$$

Now define $\hat{s}(\alpha, t)$ by the characteristic curve (particle path) issued from $\alpha$ corresponding to the first equation in (34). Then

$$\frac{\mathrm{d}\hat{s}}{\mathrm{d}t}(\alpha, t) = \dot{s}(t), \quad \hat{s}(\alpha, 0) = \alpha,$$

and

$$\frac{\mathrm{d}n(\hat{s}(t), t)}{\mathrm{d}t} = \frac{2n(\hat{s}(t), t)}{\hat{s}(t)}, \quad \dot{\hat{s}}(t) + 1 + \frac{\hat{h}(\hat{s}(t), t)}{n(\hat{s}(t), t)} = 0. \tag{35}$$

Then the second equation in (35) when differentiated with respect to $t$ yields

$$\ddot{\hat{s}}(t) + 2\frac{(\dot{\hat{s}}(t) + 1)}{\hat{s}(t)} - \frac{(\dot{\hat{s}}(t) + 1)(\partial_r \hat{h}(\hat{s}(t), t)\ \dot{\hat{s}}(t) + \partial_t \hat{h}(\hat{s}(t), t))}{n} = 0$$

which combined with the ansatz $\hat{h}(t, r) = \frac{h(t)}{r^2}$ gives

$$\ddot{\hat{s}}(t) + \frac{2(\dot{\hat{s}}(t) + 1)^2}{\hat{s}(t)} - \frac{\dot{h}(t)(\dot{\hat{s}}(t) + 1)}{h(t)} = 0 \tag{36}$$

and this a second order ordinary differential equation for the interface $r = \hat{s}(t)$. The cylindrical ($\alpha = 1$) case is an analogous computation.

## References


1. R.N. Franklin, J.R. Ockendon: Asymptotic matching of plasma and sheath in an active low pressure discharge. *J. Plasma Phys.* **4**, 371–385 (1970)
2. V.A. Godyak, N. Sternberg: Dynamic model of the electrode sheaths in symmetrically driven rf discharges. *Phys. Rev. A* **42**, 2299–2312 (1990)
3. M. Feldman, S.-Y. Ha, M. Slemrod: A geometric level-set formulation of a plasma-sheath interface. *Arch. Rational Mech. Anal.* **178**, 81–123 (2005)
4. S.-Y. Ha, M. Slemrod: Global existence of plasma-ion sheaths and their dynamics. *Comm. Math. Phys.* **238**, 149–186 (2003)
5. M.A. Lieberman, A.J. Lichtenberg: *Principles of Plasma Discharges and Materials Processing* ( New York, Wiley Interscience 1994)
6. S. Osher, R.P. Fedkiw: *Level Set Methods and Dynamic Implicit Surfaces*, Appl. Math., vol. 153, (Springer, New York 2003)
7. K.-U. Riemann: The Bohm criterion and sheath formation. *J. Physics D: Appl. Phys.* **24**, 493–518 (1991)
8. K.-U. Riemann, Th. Daube: Analytical model of the relaxation of a collisionless ion matrix sheath. *J. Appl. Phys.* **86**, 1202–1207 (1999)
9. N. Sternberg, V.A. Godyak: Solving the mathematical model of the electrode sheath in symmetrically driven RF discharges. *J. Comput. Phys.* **111**, 347–353 (1994)

# Part IV

## Mesoscale and Multiscale Modeling

# Construction of Stochastic PDEs and Predictive Control of Surface Roughness in Thin Film Deposition

D. Ni and P. D. Christofides

Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA 90095, USA, pdc@seas.ucla.edu

**Summary.** In this work, we develop a systematic method for the construction of linear stochastic partial differential equation (PDE) models for feedback control of surface roughness in thin film deposition using kinetic Monte-Carlo simulations. The method is applied to a representative deposition process and is successfully validated through simulations.

## 1 Introduction

With the advancement of thin film technology, thin films of advanced materials are used in a very wide range of applications, e.g., microelectronic devices, optics, micro-electro-mechanical systems (MEMS) and biomedical products. Various deposition methods have been developed and widely used to prepare thin films such as physical vapor deposition (PVD) and chemical vapor deposition (CVD). However, the dependence of thin film properties, such as uniformity, composition and microstructure, on the deposition conditions is a severe constraint on reproducing thin film's performance. Thus, real-time feedback control of thin film deposition becomes increasingly important in order to meet the stringent requirements on the quality of thin films and reduce thin film variability.

Earlier research efforts on feedback control of thin film deposition processes focused on deposition spatial uniformity control and thin film composition control (the reader may refer to [17, 16, 42, 1, 5, 8, 32] for representative results employing a variety of control approaches). More recently, motivated by the growing industrial demands, there have been significant research efforts focusing on modelling and control of thin film growth in order to obtain thin films with well-defined microstructure. In a thin film growth process, the film is directly formed from microscopic random processes (e.g., molecule adsorption, desorption, migration and surface reaction). Precise control of film properties requires models that describe these microscopic processes and directly account

for their stochastic nature. Examples of such models include: 1) kinetic Monte-Carlo (kMC) methods [15, 12, 18, 36], and 2) stochastic partial differential equations (PDEs) [11, 45].

Kinetic Monte-Carlo methods can be readily developed and can describe the microscopic growth processes to atomistic details with multiple species and both short-range and long-range interactions. Methodologies for estimation-based feedback control and model-predictive control using kinetic Monte-Carlo models have recently been developed in [22, 23] and [30], respectively. Furthermore, feedback control using kMC models has been successfully applied to control surface roughness in a $GaAs$ deposition process using experimentally determined model parameters [24]. Since kinetic Monte-Carlo simulations provide realizations of a stochastic process which are consistent with the master equation which describes the evolution of the microscopic probability distribution, a method to construct reduced-order approximations of the master equation was also reported in [13].

However, the fact that kMC models are not available in closed-form makes very difficult to use them for system-level analysis and the design and implementation of real-time model-based feedback control systems. Motivated by this, an approach was reported in [40, 3] to identify linear deterministic models from outputs of kinetic Monte-Carlo simulators (the reader may also refer to [38] for another approach to obtain input/output models from kMC simulations) and design controllers using linear control theory to control macroscopic variables which are low statistical moments of the microscopic distributions (e.g., surface coverage, which is the zeroth moment of adspecies distribution on a lattice). However, to control higher statistical moments of the microscopic distributions, such as the surface roughness (the second moment of height distribution on a lattice), or even the microscopic configuration (such as the surface morphology), linear deterministic models may not be sufficient, because the effect of the stochastic nature of the microscopic processes becomes very significant and must be addressed both in the model construction and controller design.

Stochastic PDE models, on the other hand, which are available in closed-form, have been developed to describe the evolution of the height profile for surfaces in certain physical and chemical processes such as epitaxial growth [45] and ion sputtering [21]; in these works, the stochastic PDE models have been derived directly on the basis of the microscopic process rules through a procedure that computes the limit of the surface height of the discrete deposition process as the size of each lattice site goes to zero (see also [28, 35, 44]). More recently, Lou and Christofides [26] presented a method for feedback control of surface roughness in a thin film growth process whose surface height fluctuation can be described by the Edwards-Wilkinson (EW) equation [11], a second-order stochastic parabolic PDE (see also [25] for results on control of surface roughness in a sputtering process using the stochastic Kuramote-Sivashinsky equation). In these works, the feedback controllers were designed, using pole-placement techniques, based on the stochastic PDE

models (Edwards-Wilkinson (EW) equation [11] and stochastic Kuramote-Sivashinsky equation [25] constructed directly from the microscopic process rules) and successfully applied to the kMC model of the deposition process regulating the surface roughness to desired values. However, the construction of stochastic PDE models for thin film growth processes directly based on microscopic process rules [28, 35, 44] is, in general, a very difficult task. This bottleneck has prohibited the development of stochastic PDE models, and subsequently the design of model-based feedback control systems, for realistic deposition processes which are, in general, highly complex.

In this work, we develop a systematic method for the construction of linear stochastic PDE models for feedback control of surface roughness in thin film deposition. A thin film deposition process including molecule adsorption and surface migration is used to illustrate the application of the method. We initially reformulate a general linear stochastic PDE into a system of infinite stochastic ordinary differential equations (ODEs), and then we use a kMC simulation of the deposition process to generate surface snapshots to determine the eigenspectrum and the covariance of the stochastic ODE system. Finally, a linear stochastic PDE model is determined by least-square fitting the prederivative coefficients to match the eigenspectrum of the stochastic PDE system to the identified stochastic ODE system and the least-square-optimal form of the stochastic PDE model with model parameters expressed as functions of the process parameters is determined. Furthermore, an optimization-based feedback controller is designed using the constructed model and applied to the kMC simulation of the deposition process to control the surface roughness.

## 2 Preliminaries

### 2.1 Thin Film Growth Process

To illustrate the application of the proposed model construction methodology, we consider throughout the manuscript a thin film growth process of deposition from vapor phase, in which, the formation of the thin film is governed by two microscopic processes that occur on the surface as shown in Fig.1, i.e., the adsorption of vapor phase molecules on the surface and the migration of surface molecules. The processes of molecules adsorption and migration are very common in thin film growth processes.

More specifically, we consider a single species growth on a 1-dimensional lattice. The adsorption rate which depends on the vapor phase concentration is considered uniform over the spatial domain and constant (i.e., fixed growth rate) during each deposition, however, it could vary for different deposition runs. All surface sites are available for adsorption for all time and the adsorption rate for each surface site is given by:

$$w_a = W \tag{1}$$

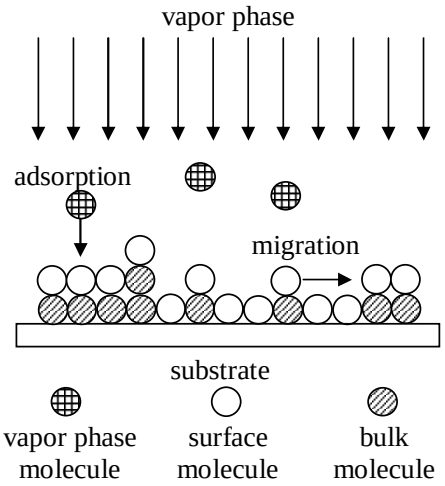



Fig. 1: The thin film growth process.

where $W$ is the growth rate in $ML/s$ (monolayers per second).

The migration rate of each surface molecule depends on its local environment. Under the consideration of only first nearest-neighbor interactions, the migration rate of surface molecules from a surface site with $n$ first nearest-neighbors is given by:

$$w_m(n) = k_{m0} e^{-\frac{E_s + nE_n}{k_B T}} \tag{2}$$

where $E_s$ is the energy barrier associated with migration due to surface effects, $E_n$ is the energy barrier associated with migration due to nearest neighbor interactions, $k_{m0}$ is the frequency constant associated with migration, $k_B$ is the Boltzmann's constant and $T$ is the substrate temperature. The values of migration energy barriers and frequency constant used in this study are taken from the literature [39] for a molecular beam epitaxy $GaAs$ process and are as follows: $E_s = 1.58\ eV$, $E_n = 0.28\ eV$ and $k_{m0} = 2k_BT/h$, where $h$ is the Planck's constant.

A kinetic Monte-Carlo simulation code following the algorithm reported in [43] is used to simulate the deposition process and obtain surface snapshots. The simulation lattice size, i.e., the total number of surface sites is denoted as $k_{max}$. Periodic boundary conditions are used in the kMC simulation to satisfy the mass balance of the migration of the surface molecules. While we focus here on a one-dimensional lattice, the model construction method can be extended to two-dimensional lattices - see [31] for details.

### 2.2 Stochastic PDE Model

As we discussed in the introduction, although there exist many first principles-based simulation codes for simulating microscopic processes, most of them are

computationally very expensive. Therefore, closed-form stochastic PDE models are favored for applications in which computation efficiency is essential, such as, for the purpose of model-based real-time feedback control.

Without any *a priori* knowledge of the deposition process, we assume that there exists a one-dimensional linear stochastic PDE of the following general form that can adequately describe the evolution of the surface of the thin film during the deposition:

$$\frac{\partial h}{\partial t} = c + c_0 h + c_1 \frac{\partial h}{\partial x} + c_2 \frac{\partial^2 h}{\partial x^2} + \cdots + c_w \frac{\partial^w h}{\partial x^w} + \xi(x,t) \tag{3}$$

where $x \in [0, \pi]$ is the spatial coordinate, $t$ is the time, $h(x,t)$ is the height of the surface at position $x$ and time $t$, and $\xi(x,t)$ is a Gaussian noise with zero mean and covariance:

$$\langle \xi(x,t)\xi(x',t') \rangle = \varsigma^2 \delta(x-x')\delta(t-t') \tag{4}$$

where $\delta(\cdot)$ is the Dirac function. Furthermore, the pre-derivative coefficients $c$ and $c_j$ in Eq.3 and the parameter $\varsigma^2$ in Eq.4 depend on the process parameters (gas flow rates, substrate temperature, etc.) $p_i(t)$:

$$\begin{aligned} c &= C[p_1(t), p_2(t), \cdots, p_d(t)] \\ c_j &= C_j[p_1(t), p_2(t), \cdots, p_d(t)] \quad j = 0, \cdots, w \\ \varsigma^2 &= C_\xi[p_1(t), p_2(t), \cdots, p_d(t)] \end{aligned} \tag{5}$$

where $C(\cdot)$, $C_j(\cdot)$ and $C_\xi(\cdot)$ are nonlinear functions to be determined.

The stochastic PDE of Eq.3 is subjected to the following periodic boundary conditions:

$$\frac{\partial^j h}{\partial x^j}(0,t) = \frac{\partial^j h}{\partial x^j}(\pi,t) \quad j = 0, \cdots, w-1 \tag{6}$$

and the initial condition:

$$h(x,0) = h_0(x) \tag{7}$$

*Remark 1.* In this work, we assume that a linear stochastic PDE model adequately describes the process dynamics, however, for the cases in which the nonlinear dynamics are significant, nonlinear stochastic PDE models would be needed. Also, we note that we use a scalar function, $h(\cdot)$, to represent the height profile of the thin film surface in the model. In general, $h(\cdot)$ can be a vector function and be used to represent any appropriate microscopic description of the thin film (such as the defect locations, grain boundaries, etc); in such a case, several stochastic PDEs should be considered simultaneously.

To study the dynamics of Eq.3, we initially consider the eigenvalue problem of the linear operator of Eq.3, which takes the form:

$$A\phi_n(x) = c_0\phi_n(x) + c_1\frac{\mathrm{d}\phi_n(x)}{\mathrm{d}x} + c_2\frac{\mathrm{d}^2\phi_n(x)}{\mathrm{d}x^2} + \cdots + c_w\frac{\mathrm{d}^w\phi_n(x)}{\mathrm{d}x^w} = \lambda_n\phi_n(x)$$
$$\frac{\mathrm{d}^j\phi_n}{\mathrm{d}x^j}(0) = \frac{\mathrm{d}^j\phi_n}{\mathrm{d}x^j}(\pi) \quad j = 0, \cdots, w-1 \quad n = 1, \cdots, \infty \tag{8}$$

where $\lambda_n$ denotes an eigenvalue and $\phi_n$ denotes an eigenfunction. A direct computation of the solution of the above eigenvalue problem yields:

$$\lambda_n = c_0 + I2nc_1 + (I2n)^2c_2 + \cdots + (I2n)^wc_w$$
$$\phi_n(x) = \sqrt{\frac{1}{\pi}}e^{I2nx} \quad n = 0, \pm1, \cdots, \pm\infty \tag{9}$$

where $\lambda_n$ denotes the $n$th eigenvalue, $\phi_n(x)$ denotes the $n$th eigenfunction and $I = \sqrt{-1}$.

To present the method that we use for parameter identification of the stochastic PDE of Eq.3, we first derive an infinite stochastic ODE representation of Eq.3 using modal decomposition and parameterize the infinite stochastic ODE system using kMC simulation. We first expand the solution of Eq.3 in an infinite series in terms of the eigenfunctions of the operator of Eq.8 as follows (i.e., the Fourier expansion in the complex form):

$$h(x,t) = \sum_{n=-\infty}^{\infty} z_n(t)\phi_n(x) \tag{10}$$

where $z_n(t)$ are time-varying coefficients. Substituting the above expansion for the solution, $h(x,t)$, into Eq.3 and taking the inner product, the following system of infinite stochastic ODEs is obtained:

$$\frac{\mathrm{d}z_n}{\mathrm{d}t} = \lambda_n z_n + c_{zn} + \xi_n(t) \quad n = 0, \pm1, \cdots, \pm\infty \tag{11}$$

and the initial conditions:

$$z_n(0) = z_{n0} \quad n = 0, \pm1, \ldots, \pm\infty \tag{12}$$

where $c_{zn} = c\int_0^\pi \phi_n^*(x)\mathrm{d}x$ (apparently $c_{z0} = c\sqrt{\pi}$ and $c_{zn} = 0 \ \forall \ n \neq 0$), $\xi_n(t) = \int_0^\pi \xi(x,t)\phi_n^*(x)\mathrm{d}x$ and $z_{n0} = \int_0^\pi h_0(x)\phi_n^*(x)\mathrm{d}x$. $\phi_n^*(x)$ is the complex conjugate of $\phi_n(x)$, the superscript star is used to denote complex conjugate in the remainder of this manuscript.

The covariances of $\xi_n(t)$ can be computed by using the following result [4].

*Result 1:* If (1) $f(x)$ is a deterministic function, (2) $\eta(x)$ is a random variable with $\langle\eta(x)\rangle = 0$ and covariance $\langle\eta(x)\eta(x')\rangle = \sigma^2\delta(x-x')$, and (3) $\epsilon = \int_a^b f(x)\eta(x)\mathrm{d}x$, then $\epsilon$ is a random number with $\langle\epsilon\rangle = 0$ and covariance $\langle\epsilon^2\rangle = \sigma^2\int_a^b f(x)f^*(x)\mathrm{d}x$.

Using the above result, we obtain $\langle \xi_n(t) \rangle = 0$ and $\langle \xi_n(t) \xi_n^*(t') \rangle = \varsigma^2 \delta(t-t')$. We note that $\xi_n(t)$ is a complex Gaussian random variable and the probability distribution function of the Gaussian distribution, $P(\xi_n, t)$, on the complex plane with zero mean and covariance $\varsigma^2 \delta(t-t')$ is defined as follows:

$$P(\xi_n, t) = \frac{1}{\sqrt{2\pi}\varsigma\delta(t-t')} e^{\frac{\xi_n \xi_n^*}{2\varsigma^2 \delta(t-t')}} \tag{13}$$

To parameterize this system of infinite stochastic ODEs, we first derive the analytic expressions for the statistical moments of the stochastic ODE states, including the expected value and covariance. By comparing the analytical expression to the statistical moments obtained by multiple kMC simulations, the parameters of the stochastic ODE system (i.e., $\lambda_n$ and $\varsigma$) can be determined.

The analytic solution of Eq.11 is obtained as follows to derive the expressions for the statistical moments of the stochastic ODE states:

$$z_n(t) = e^{\lambda_n t} z_{n0} + \frac{(e^{\lambda_n t} - 1) c_{zn}}{\lambda_n} + \int_0^t e^{\lambda_n (t-\mu)} \xi_n(\mu) \mathrm{d}\mu \tag{14}$$

Using Result 1, Eq.14 can be further simplified as follows:

$$z_n(t) = e^{\lambda_n t} z_{n0} + \frac{(e^{\lambda_n t} - 1) c_{zn}}{\lambda_n} + \theta_n(t) \tag{15}$$

where $\theta_n(t)$ is a complex random variable of normal distribution with zero mean and covariance $\langle \theta_n(t) \theta_n^*(t) \rangle = \varsigma^2 \dfrac{e^{(\lambda_n + \lambda_n^*)t} - 1}{\lambda_n + \lambda_n^*}$. Therefore, the expected value (the first stochastic moment) and the covariance (the second stochastic moment) of state $z_n$ can be expressed as follows:

$$\begin{aligned} &\langle z_n(t) \rangle = e^{\lambda_n t} z_{n0} + \frac{(e^{\lambda_n t} - 1) c_{zn}}{\lambda_n} \\ &\langle z_n(t) z_n^*(t) \rangle = \varsigma^2 \frac{e^{(\lambda_n + \lambda_n^*)t} - 1}{\lambda_n + \lambda_n^*} + \langle z_n(t) \rangle \langle z_n(t) \rangle^* \\ &n = 0, \pm 1, \cdots, \pm \infty \end{aligned} \tag{16}$$

Eq.16 holds for any initial condition $z_{n0}$. Since we are able to choose any initial thin film surface for simulation, we choose $z_{n0} = 0$ (i.e., the initial surface is flat, $h(x, 0) = 0$) to simplify our calculations. In this case, Eq.16 can be further simplified as follows (note that $c_{zn} = 0, \forall n \neq 0$):

$$\begin{aligned} &\langle z_n(t) \rangle = 0 \\ &\langle z_n(t) z_n^*(t) \rangle = \varsigma^2 \frac{e^{(\lambda_n + \lambda_n^*)t} - 1}{\lambda_n + \lambda_n^*} = \varsigma^2 \frac{e^{2Re(\lambda_n)t} - 1}{2Re(\lambda_n)} \\ &n = \pm 1, \cdots, \pm \infty \end{aligned} \tag{17}$$

where $Re(\lambda_n)$ denote the real part of $\lambda_n$, and for $z_0(t)$, it follows from Eq.16 with $\lambda_0 = 0$ that,

$$\begin{aligned} \langle z_0(t) \rangle &= \lim_{\lambda_0 \to 0} \frac{(e^{\lambda_0 t} - 1) c_{z0}}{\lambda_0} = t c_{z0} = t\sqrt{\pi} c \\ \langle z_0^2(t) \rangle &= \varsigma^2 t + t^2 \pi c^2 \end{aligned} \tag{18}$$

It can be seen in Eq.17 that the statistical moments of each stochastic ODE state depend only on the real part of the corresponding eigenvalue, and therefore, to determine the imaginary part of the eigenvalue we need to construct an extra equation. We note that $\lambda_n$ would be a complex number if the linear operator $A$ is not self-adjoint, i.e., when odd-partial-derivatives are present in the stochastic PDE (see Eq.9).

Therefore, we rewrite Eq.14 by separating the real part and the imaginary part of $z_n(t)$ as follows with initial condition $z_{n0} = 0$:

$$\begin{aligned} z_n(t) = &\frac{1}{2} \int_0^t [e^{\lambda_n (t-\mu)} + e^{\lambda_n^* (t-\mu)}] \xi_n(\mu) \mathrm{d}\mu \\ &+ \frac{1}{2} \int_0^t [e^{\lambda_n (t-\mu)} - e^{\lambda_n^* (t-\mu)}] \xi_n(\mu) \mathrm{d}\mu \\ &n = \pm 1, \cdots, \pm\infty \end{aligned} \tag{19}$$

Accordingly, the real part of $z_n(t)$ can be expressed as follows:

$$\begin{aligned} Re[z_n(t)] = &\frac{1}{2} \int_0^t [e^{\lambda_n (t-\mu)} + e^{\lambda_n^* (t-\mu)}] \xi_n(\mu) \mathrm{d}\mu \\ &n = \pm 1, \cdots, \pm\infty \end{aligned} \tag{20}$$

where $Re[z_n(t)]$ denotes the real part of $z_n(t)$. By using result 1, we have,

$$\begin{aligned} \langle Re[z_n(t)] \rangle &= 0 \\ \langle Re[z_n(t)]^2 \rangle &= \varsigma^2 [\frac{\lambda_n^* e^{2\lambda_n t} + \lambda_n e^{2\lambda_n^* t} - (\lambda_n + \lambda_n^*)}{8\lambda_n \lambda_n^*} + \frac{e^{(\lambda_n + \lambda_n^*)t} - 1}{2(\lambda_n + \lambda_n^*)}] \\ &= \varsigma^2 \{ \frac{Re(\lambda_n) e^{2Re(\lambda_n)} cos(2Im(\lambda_n) t)}{4[Re(\lambda_n)^2 + Im(\lambda_n)^2]} \\ &\quad + \frac{Im(\lambda_n) e^{2Re(\lambda_n)} sin(2Im(\lambda_n) t)}{4[Re(\lambda_n)^2 + Im(\lambda_n)^2]} \\ &\quad - \frac{Re(\lambda_n)}{4[Re(\lambda_n)^2 + Im(\lambda_n)^2]} + \frac{e^{2Re(\lambda_n)t} - 1}{4Re(\lambda_n)} \} \\ &\quad n = \pm 1, \cdots, \pm\infty \end{aligned} \tag{21}$$

where $Im(\lambda_n)$ denotes the imaginary part of $\lambda_n$. Thus, we can use Eq.17 to first determine the real part of the eigenvalue, and then use the Eq.21 to determine its imaginary part. We note that it is not recommended to determine both parts of the eigenvalue using only Eq.21, since in that case, the nonlinear least-square problem involved in the eigenvalue determination would be much more difficult to solve.

*Remark 2.* Eq.17, Eq.18 and Eq.21 show the analytical relation that relates the linear operator and the Gaussian noise in Eq.3 to the statistical moments of the states of Eq.11 which can be obtained through multiple experimental measurements or first principle simulations, and therefore, reveal a viable path to systematically construct a linear stochastic PDE of the form of Eq.3 that describes the dynamics of the microscopic processes directly from experimental or simulation data.

## 3 Model Construction

Based on the results shown in the previous section, we propose a systematic procedure to construct a linear stochastic PDE for the deposition process described in Section 2.1. In this work, we use a kinetic Monte-Carlo code to simulate the deposition process and generate surface snapshots. The proposed procedure includes the following steps: First, we design a set of simulation experiments that cover the complete range of process operation; second, we run multiple simulations for each simulation experiment to obtain the trajectories of the first and second statistical moments of the states (i.e., Fourier coefficients) computed from the surface snapshots; third, we compute the eigenvalues of the linear operator and covariance of the Gaussian noise based on the trajectories of the statistical moments of the states for each simulation experiment, and determine the model parameters of the stochastic PDE (i.e., the pre-derivative coefficients and the order of the stochastic PDE); finally, we investigate the dependence of the model parameters of the stochastic PDE on the process parameters and determine the least-square-optimal form of the stochastic PDE model with model parameters expressed as functions of the process parameters.

### 3.1 Eigenvalues and Covariance

Because there are only two process parameters considered in the deposition process studied in this work, the growth rate $W$ and the substrate temperature $T$, the simulation experiment design is straightforward. Specifically, different $W$ values and $T$ values are evenly selected from the range of process operation of interest and simulation experiments are executed with every selected $W$ value for each selected $T$ value. Therefore, we start our demonstration of the model construction methodology with the identification of the eigenvalues and

covariance. Also, we note that the trajectories of the statistical moments for each simulation experiment are computed based on 100 simulation runs taking place with the same process parameters.

In the previous section we have shown that for a deposition process with a flat initial surface, the covariance of each state $\langle z_n(t)z_n^*(t)\rangle$ should be able to be predicted by Eq.17, therefore, we can fit $\varsigma^2$ and $Re(\lambda_n)$ in Eq.17 for the profile of $\langle z_n(t)z_n^*(t)\rangle$. In order to obtain the profile of $\langle z_n(t)z_n^*(t)\rangle$, we need to generate snapshots of the thin film surface during each deposition simulation and compute the values of $z_n(t)$. Since the lattice consists of discrete sites, we let $h(kL,t)$ be the height profile of the surface at time $t$ with lattice constant $L$ ($k$ denotes the coordinate of a specific surface site), and compute $z_n(t)$ as follows:

$$z_n(t) = \int_0^{\pi} h(x,t)\phi_n^*(x)\mathrm{d}x = \sum_{k=0}^{k_{max}} h(kL,t)\int_{kL}^{(k+1)L} \phi_n^*(x)\mathrm{d}x \qquad (22)$$

where $k_{max}L = \pi$ (i.e., the lattice is mapped to the domain [0,$\pi$]). Substituting Eq.9 into Eq.22, we can derive the following expression for $z_n(t)$:

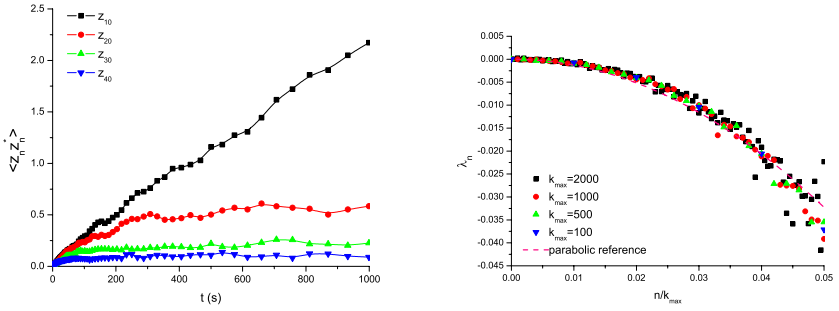
$$z_n(t) = \sum_{k=0}^{k_{max}} \frac{h(kL,t)e^{-I2kLn}}{I2\sqrt{\pi}n}(1-e^{-I2Ln}) \quad n = \pm 1, \cdots, \pm\infty \qquad (23)$$

and for $z_0(t)$, we have,

$$z_0(t) = \sum_{k=0}^{k_{max}} h(kL,t)\frac{L}{\sqrt{\pi}} = \sqrt{\pi}t\frac{\sum_{k=0}^{k_{max}} h(kL,t)}{k_{max}t} = t\sqrt{\pi}W \qquad (24)$$

To capture the dynamics of both the fast states and slow states simultaneously in the same simulation run with few surface snapshots, the snapshots are generated in a variable-time-step fashion in which the intervals between two snapshots are increased with time. This procedure is motivated by the fact that the dynamics of the fast states can be detected only at the beginning of each simulation run, and therefore, the evolving surface should be sampled more frequently in the beginning than the remainder to cope with the small time scale of these fast states. Fig.2(a) shows the typical covariance profiles of different states in a growth process. It can be seen that despite the very different time scales of the states, our method can still generate very smooth profiles for both the fast states (such as $z_{40}$, whose time scale is less than 50 $s$) and the slow states (such as $z_{10}$, whose time scale is larger than 1000 $s$).

Fig.2(b) shows the eigenvalues identified from thin film depositions occurring under the same operating conditions but simulated with different lattice size (we note that the identified eigenvalues are considered real since the imaginary part of the eigenvalues identified turned out to be very small). It can be




(a) Covariance profiles. (b) Eigenspectra.

Fig. 2: (a) Covariance profiles of $z_{10}$, $z_{20}$, $z_{30}$ and $z_{40}$; (b) Eigenvalue spectra of the infinite stochastic ODE systems identified from the kMC simulation of the deposition process with different lattice size: $k_{max} = 100$, $k_{max} = 500$, $k_{max} = 1000$ and $k_{max} = 2000$.

seen that the identified spectra are very close to each other when $n$ is rescaled with the corresponding lattice size. This is expected, since, $\phi_n(x)$ is a basis of the domain of operator $A$, and is a complex function of the frequency $n$, accordingly, $n/k_{max}$ is the length scale of the surface fluctuation described by $\phi_n(x)$ when a lattice of size $k_{max}$ is mapped to the domain of $[0, \pi]$ (we note that, for the same reason, the covariance values should be scaled with the inverse of the lattice size, $1/k_{max}$, in order to carry out a meaningful comparison).

It can also be seen in Fig.2(b) that the eigenspectra are very close to the parabolic reference curve, which implies that a second-order stochastic PDE system of the following form would be able to describe the evolution of the surface height of this deposition process:

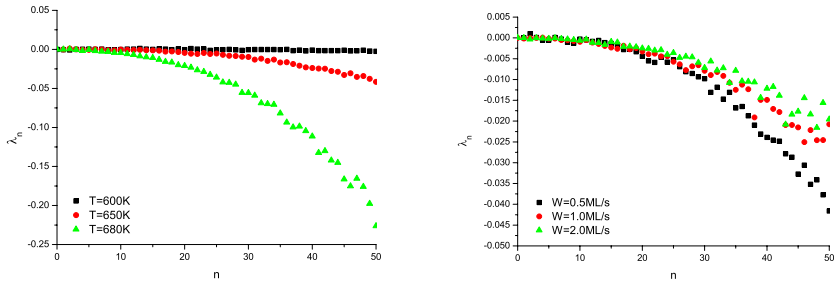
$$\frac{\partial h}{\partial t} = c + c_2 \frac{\partial^2 h}{\partial x^2} + \xi(x, t) \tag{25}$$

in which $c$, $c_2$ and the covariance of the Gaussian noise $\xi$, $\varsigma$, all depend on the microscopic processes and operating conditions. At this point, it is important to point out that Eq.25 constructed following the proposed model construction procedure from kMC data is a second-order stochastic PDE of the Edwards-Wilkinson type [11]. This is expected because of the similarity of the microscopic rules considered in the deposition process of our work and in the work of Edwards and Wilkinson [11]. However, the path followed by Edwards and Wilkinson for the construction of their stochastic PDE and the path followed by the proposed model construction procedure for the construction of Eq.25 are completely different.

For a deposition process that is similar to the one we considered here, Edwards and Wilkinson derived the well-known EW equation from a limiting procedure [11]. The EW equation is also a second-order linear stochastic PDE like the one we obtained here. The fact that our model constructed via a very different approach, coincides with the EW equation provides a good validation of our methodology.

### 3.2 Dependence on the Process Parameters

We proceed now with the derivation of the parameters of the stochastic PDE of Eq.25. From Eq.18 and Eq.24, we can see that $c = W$ for all cases. However, $c_2$ and $\varsigma^2$ identified for different deposition settings can be very different, therefore, we need to investigate their dependence on the deposition parameters to obtain their analytical expressions. $c_2$ and $\varsigma^2$ are evaluated for assorted deposition conditions and a lattice size of 1000 (i.e., $k_{max} = 1000$) is used for all the simulation runs in our study.



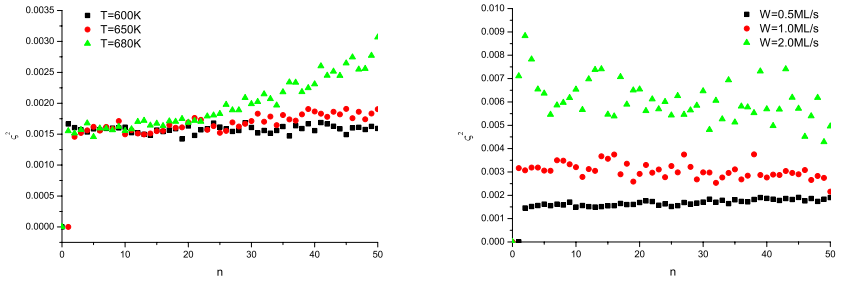

(a) Eigenspectra identified with a growth rate $W = 0.5ML/s$ for different substrate temperatures: $T = 600K$, $T = 650K$ and $T = 680K$.

(b) Eigenspectra identified with a substrate temperature $T = 650K$ for different growth rates: $W = 0.5ML/s$, $W = 1.0ML/s$ and $W = 2.0ML/s$.

Fig. 3: Eigenspectra identified from simulated deposition processes.

Fig.3(a) shows the eigenspectra identified from depositions with the same growth rate ($W = 0.5\ ML/s$) for different substrate temperatures. It can be seen that the magnitude of the eigenvalues decreases faster with increasing $n$ at higher substrate temperature. This implies that a higher substrate temperature corresponds to a larger $c_2$ in the stochastic PDE model and vice versa.

Fig.4(a) shows the covariance spectra identified from depositions with the same growth rate ($W = 0.5\ ML/s$) for different substrate temperature. Although it follows from Eq.13 that the covariance of the stochastic noise should




(a) Spectra identified with a growth rate $W = 0.5ML/s$ for different substrate temperatures: $T = 600K$, $T = 650K$ and $T = 680K$.

(b) Spectra identified with a substrate temperature $T = 650K$ for different growth rates: $W = 0.5ML/s$, $W = 1.0ML/s$ and $W = 2.0ML/s$.

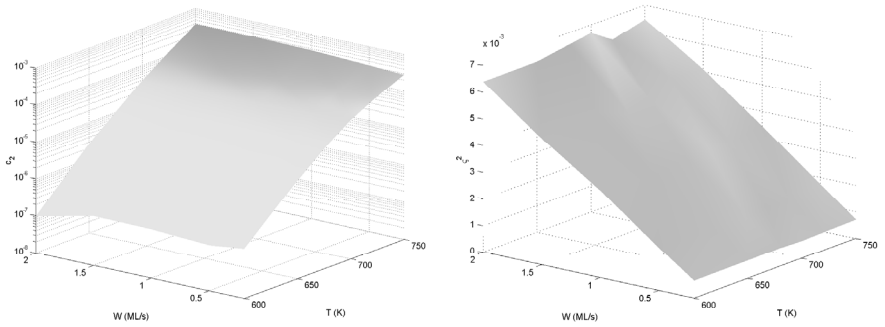
Fig. 4: Covariance spectra identified from simulated deposition processes.

be the same for all states, it is not so for high-order states in the high substrate temperature regime (e.g., $T = 680\ K$). However, because these high order states correspond to the surface fluctuations of small length scales, and at the same time, such small length scale surface fluctuations are almost negligible in the high substrate temperature regime due to the significant surface diffusion, the contribution from these high-order states at high substrate temperature becomes very small. Therefore, given that such discrepancy would not significantly affect the accuracy of the model, we compute $\varsigma^2$ only based on the low-order states. From the covariance of the low-order states shown in Fig.4(a), we may also consider $\varsigma^2$ to be independent of substrate temperature.

Fig.3(b) shows the eigenspectra identified from depositions occurring under the same substrate temperature ($T = 650K$) and different thin film growth rates. It can be seen that, at this substrate temperature, the eigenvalues die out a bit slower with increasing growth rate, which implies that a higher growth rate corresponds to a smaller $c_2$ in the stochastic PDE model and vice versa.

Fig.4(b) shows the covariance spectra identified from depositions occurring under the same substrate temperature ($T = 650K$) and different thin film growth rates. It can be seen that a higher growth rate corresponds to a larger covariance value.

To derive explicit expressions for $c_2$ and $\varsigma^2$ as functions of $T$ and $W$, we evaluate these values for different $T$ and $W$ and the results are shown in Fig.5(a) and Fig.5(b). From Fig.5(a), we can see that $ln\, c_2$ has a quasi-linear relationship with both $T$ and $W$, and thus, the following expression can be obtained for $c_2$ as a function of $T$ and $W$ through least square fitting:




(a) Profile of $c_2$. (b) Profile of $\varsigma^2$.

Fig. 5: Profile of stochastic PDE parameters as functions of substrate temperature $T$ and thin film growth rate $W$.

$$\begin{aligned} c_2(T,W) &= e^{-45.8176 + 0.0511T - 0.1620W} \\ &= \frac{e^{-32.002 + 0.0511T - 0.1620W}}{k_{max}^2} \end{aligned} \tag{26}$$

From Fig.5(b) we can see that $\varsigma^2$ depends almost linearly on both $T$ and $W$, and thus, the following expression can be obtained for $\varsigma^2$ as a function of $T$ and $W$ through least square fitting as well:

$$\varsigma^2(T,W) = 5.137 \times 10^{-8} T + 3.2003 \times 10^{-3} W \approx \frac{\pi W}{k_{max}} \tag{27}$$

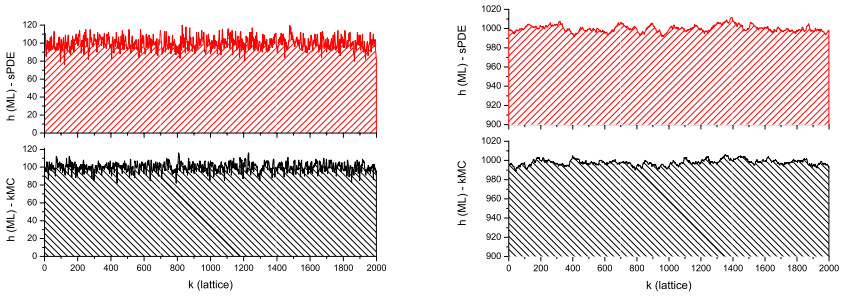
Therefore, the linear stochastic PDE model identified for the deposition process is as follows:

$$\begin{aligned} &\frac{\partial h}{\partial t} = W + (\frac{e^{-32.002 + 0.0511T - 0.1620W}}{k_{max}^2})\frac{\partial^2 h}{\partial x^2} + \xi(x,t) \\ &\frac{\partial h}{\partial x}(0,t) = \frac{\partial h}{\partial x}(\pi,t), \quad h(0,t) = h(\pi,t), \quad h(x,0) = h_0(x) \end{aligned} \tag{28}$$

where $\langle \xi(x,t)\xi(x',t') \rangle = \dfrac{5.137 \times 10^{-5} T + 3.2003 W}{k_{max}} \delta(x-x')\delta(t-t')$.

### 3.3 Validation of Stochastic PDE Model

We now proceed with the validation of the stochastic PDE model of the thin film deposition process (Eq.28). Validation experiments are conducted for a number of deposition conditions which have not been used for the model construction. We generate surface profiles using both the stochastic PDE model and the kinetic Monte-Carlo code. Fig.6(a) shows the surface profile at the end




(a) A $1000s$ deposition with substrate temperature $T = 550K$ and thin film growth rate $W = 0.1$ $ML/s$.

(b) A $400s$ deposition with substrate temperature $T = 700K$ and thin film growth rate $W = 2.5$ $ML/s$.

Fig. 6: Final thin film surface profiles generated by kMC simulation and stochastic PDE model ($k_{max} = 2000$).

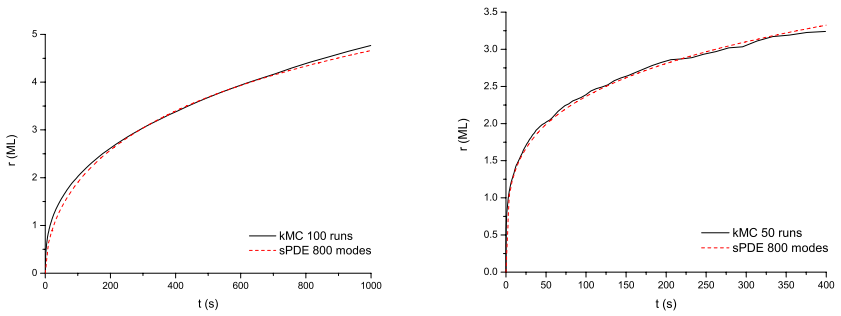
of a deposition with substrate temperature $T = 550$ $K$, thin film growth rate $W = 0.1$ $ML/s$, deposition duration of 1000 $s$ and lattice size $k_{max} = 2000$; Fig.6(b) shows the surface profile at the end of a deposition with substrate temperature $T = 700$ $K$, thin film growth rate $W = 2.5$ $ML/s$, deposition duration of 400 $s$ and lattice size $k_{max} = 2000$; we can see that both at low and high substrate temperatures, and for different growth rates, the linear stochastic PDE model constructed for the deposition process is very consistent with the kinetic Monte-Carlo simulation.

We also generate expected surface roughness profiles using both the stochastic PDE model and the kinetic Monte-Carlo simulation (average of 100 runs) for the deposition process. For simplicity, the surface roughness is evaluated in a root-mean-square fashion as follows:

$$r(t) = \sqrt{\frac{1}{\pi}\int_0^{\pi} [h(x,t) - \bar{h}(t)]^2 \mathrm{d}x} \tag{29}$$

where $\bar{h}(t) = \dfrac{1}{\pi}\displaystyle\int_0^{\pi} h(x,t)\mathrm{d}x$ is the average surface height. We note that for more detailed description of the surface morphology, the height-height correlation function may be used to evaluate the surface roughness [41].

Fig.7(a) shows the expected roughness profile of a deposition with substrate temperature $T = 550$ $K$ and thin film growth rate $W = 0.1$ $ML/s$; Fig.7(b) shows the roughness profile of a deposition with substrate temperature $T = 700$ $K$ and thin film growth rate $W = 2.5$ $ML/s$; we can see that the linear stochastic PDE model constructed for the deposition process is also very consistent with the kinetic Monte-Carlo simulation in terms of surface




(a) A $1000s$ deposition with substrate temperature $T = 550K$ and thin film growth rate $W = 0.1$ $ML/s$.

(b) A $400s$ deposition with substrate temperature $T = 700K$ and thin film growth rate $W = 2.5$ $ML/s$.

Fig. 7: Expected surface roughness profiles generated by kMC simulation and stochastic PDE model ($k_{max} = 2000$).

roughness, at both low and high substrate temperatures, for different growth rates.

## 4 Predictive Control

In this section, we design a model-based state feedback controller based on the stochastic PDE model of Eq.28 to control the thin film surface roughness of the deposition process. The difficulty of obtaining in-situ surface measurements in real-time had been one of the obstacles for implementing feedback control on thin film processes. Recently, researchers made possible to use some of the intrusive scanning probe based techniques such as the scanning tunneling microscopy (STM) [33] and atomic force microscopy (AFM) [27] in-situ, to observe in real-time the growth of the thin film. More recently, it was reported in [37] that a non-intrusive grazing incidence small angle x-ray scattering (GISAXS) method was successfully used to monitor the thin film growth in-situ in real-time; the method was capable of sampling large surface areas with sampling frequency up to 10 $Hz$ and a subnanometer resolution. Such advancements in surface metrology indeed open up the possibility for implementing feedback control systems which rely on real-time surface state measurements and possibly on state estimation algorithms [19] . On the other hand, for the cases in which state measurements are not available directly, state estimators could be used to implement output feedback control based on available measurements such as thickness and surface roughness.

### 4.1 Surface Roughness

We first proceed with the analysis of the dynamics of the surface roughness based on the stochastic PDE model constructed for the thin film deposition process. The surface roughness, $r(t)$, is defined by Eq.29. According to Eq.22, we have $\bar{h}(t) = z_0(t)\phi_0$. Therefore, $r(t)$ can be rewritten in terms of $z_n$ as follows:

$$\begin{aligned} r(t) &= \sqrt{\frac{1}{\pi}\int_0^\pi (h(x,t) - \bar{h}(t))^2 \mathrm{d}x} \\ &= \sqrt{\frac{1}{\pi}\int_0^\pi \sum_{n=-\infty, n\neq 0}^{\infty} z_n(t)\phi_n(x)\phi_n^*(x)z_n^*(t)\mathrm{d}x} \\ &= \sqrt{\frac{1}{\pi}\sum_{n=-\infty, n\neq 0}^{\infty} z_n(t)z_n^*(t)} \end{aligned} \tag{30}$$

and the expected roughness can be computed as follows:

$$\langle r(t)\rangle = \sqrt{\frac{1}{\pi}\sum_{n=-\infty, n\neq 0}^{\infty} \langle z_n(t)z_n^*(t)\rangle} \tag{31}$$

In order to design a model-based feedback controller to control the surface roughness, we first derive the analytical expression for the trajectory of $\langle r(t)\rangle$. Substituting Eq.16 into Eq.31 we obtain the following expression for $\langle r(t)\rangle$ in terms of the eigenvalues of the infinite stochastic ODE system:

$$\begin{aligned} \langle r(t)\rangle &= \sqrt{\frac{1}{\pi}\sum_{n=-\infty, n\neq 0}^{\infty} [\varsigma^2 \frac{e^{(\lambda_n + \lambda_n^*)t} - 1}{\lambda_n + \lambda_n^*} + e^{(\lambda_n + \lambda_n^*)t} z_{n0} z_{n0}^*]} \\ &= \sqrt{\frac{1}{\pi}\sum_{n=-\infty, n\neq 0}^{\infty} [\varsigma^2 \frac{e^{2Re(\lambda_n)t} - 1}{2Re(\lambda_n)} + e^{2Re(\lambda_n)t} z_{n0} z_{n0}^*]} \end{aligned} \tag{32}$$

Specifically, for the stochastic PDE model of Eq.25, $\lambda_n = -4c_2 n^2$, thus, Eq.32 can be rewritten as follows:

$$\begin{aligned} \langle r(t)\rangle &= \sqrt{\frac{1}{\pi}\sum_{n=-\infty, n\neq 0}^{\infty} (\varsigma^2 \frac{e^{-8c_2 n^2 t} - 1}{-8c_2 n^2} + e^{-8c_2 n^2 t} z_{n0} z_{n0}^*)} \\ &= \sqrt{\frac{2}{\pi}\sum_{n=1}^{\infty} (\varsigma^2 \frac{e^{-8c_2 n^2 t} - 1}{-8c_2 n^2} + e^{-8c_2 n^2 t} z_{n0} z_{n0}^*)} \end{aligned} \tag{33}$$

In order to compute an estimate of the excepted surface roughness at a future time $t$, we need to compute the infinite sum in Eq.33. However, such an

infinite summation cannot be computed directly, instead, a finite summation needs to be used to approximately compute this infinite sum. It can shown by using standard theory of infinite summation [20] that, if the following $m$th order approximation (only the first $m$th states are included in the summation) is used,

$$\begin{aligned}\hat{r}(t)^2 = \frac{2}{\pi}\sum_{n=1}^{m}(\varsigma^2\frac{e^{-8n^2c_2t}-1}{-8n^2c_2} + e^{-8n^2c_2t}z_{n0}z_{n0}^*) \\ +\frac{1}{2\pi}[e^{-8c_2(m+1)^2t}(\pi r_0^2 - \sum_{n=1}^{m} z_{n0}z_{n0}^*) + \frac{\varsigma^2}{2^{m+2}c_2}]\end{aligned} \tag{34}$$

where $r_0$ is the initial roughness value, the approximation error would be subject to the following bound:

$$|\langle r(t)\rangle - \hat{r}(t)| \leq \sqrt{\frac{1}{2\pi}[e^{-8c_2(m+1)^2t}(\pi r_0^2 - \sum_{n=1}^{m} z_{n0}z_{n0}^*) + \frac{\varsigma^2}{2^{m+2}c_2}]} \tag{35}$$

We note that the approximation error decreases with increasing $m$.

### 4.2 Predictive Control Design

We now proceed with the design of the feedback controller. Since the thin film deposition is a batch process, the control objective is to control the final surface roughness of the thin film to a desired level at the end of each deposition run. Therefore, we use an optimization-based control problem formulation (the reader may refer to [10, 7, 9, 34, 29, 6, 2] for more information on optimization-based control formulations and control of PDEs). The substrate temperature, $T$, is chosen to be the manipulated variable, while the thin film growth rate $W$ is kept constant during each deposition. Furthermore, since the process is stochastic, the controlled variable is the expected value of the final surface roughness, $\langle r(t_{dep})\rangle$, where $t_{dep}$ is the total deposition time.

Fig.8 shows the block diagram of the closed-loop system. When a real-time surface profile measurement is obtained, the states of the infinite stochastic ODE system, $z_n$, are computed. Then, a substrate temperature $T$ is computed based on states $z_n$ and the stochastic PDE model and applied to the deposition process. The substrate is held at this temperature for the rest of the deposition until a different value is assigned by the controller. The value of $T$ is determined at each time $t$ by solving, in real-time, the following optimization problem:

$$\min_{T} J = (r_{set}^2 - \langle r_{final}\rangle^2)^2 \tag{36}$$

subject to

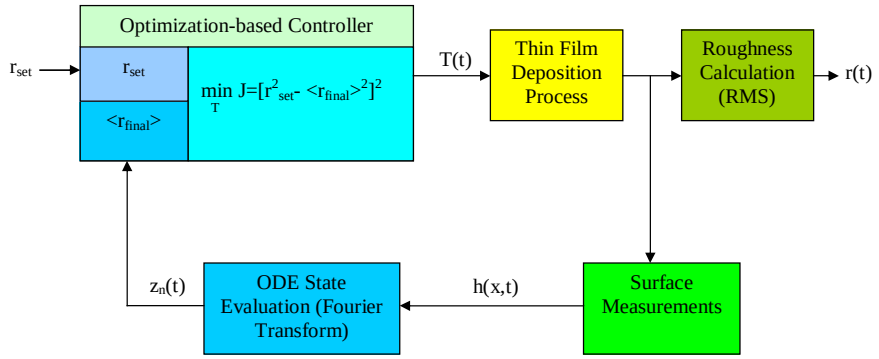



Fig. 8: Block diagram of the closed-loop system.

$$\langle r_{final}\rangle^2 = \frac{2}{\pi}\sum_{n=1}^{m}[\varsigma^2\frac{e^{-8n^2c_2(t_{dep}-t)}-1}{-8n^2c_2} + e^{-8n^2c_2(t_{dep}-t)}z_n(t)z_n^*(t)]$$
$$+\frac{1}{2\pi}\{e^{-8c_2(m+1)^2(t_{dep}-t)}[\pi r^2(t) - \sum_{n=1}^{m} z_n(t)z_n^*(t)] + \frac{\varsigma^2}{2^{m+2}c_2}\} \quad (37)$$

$$c_2 = \frac{e^{-32.002+0.0511T-0.1620W}}{k_{max}^2} \quad (38)$$

$$\varsigma^2 = \frac{\pi W}{k_{max}} \quad (39)$$

$$T_{min} \le T \le T_{max} \quad (40)$$

where $T_{min}$ and $T_{max}$ are the lowest and highest substrate temperature, respectively. We note that $J$ corresponds to the difference between the square of the desired final surface roughness $r_{set}$ and the square of the estimated final surface roughness $\langle r_{final}\rangle$ computed based on the current states $z_n$. We choose to minimize the difference of the squares of the surface roughness, i.e., the mean square of the surface height, to simplify the calculation.

The first equality constraint Eq.37 (essentially the same as Eq.34) states that the estimate of the final surface roughness, $r_{final}$, is computed based on current states $z_n(t)$ under the assumption that a substrate temperature $T$ will be used and kept constant in the rest of the deposition. The second and third equality constraints are, in fact, Eqs.26 and 27 of the stochastic PDE model of the deposition process, and since the growth rate $W$ is fixed during each deposition, the third constraint can be removed by substituting the actual value of $\varsigma^2$ into Eq.37. The optimization problem can then be solved analytically by quadratic programming using a linear approximation of Eq.37 (a standard sequential quadratic programming code can be used to solve the original nonlinear optimization problem efficiently, however, it is not used here for simplicity).

To solve the above optimization problem, our initial step is to reduce it to a quadratic programming problem with only linear constraints. To do this, we remove the second and fourth constraints (Egs.38 and 40) by first finding the optimal $c_2$ that minimizes $J$ and then computing the corresponding optimal $T$ using the equality constraint of Eq.38. In addition, we linearize the first constraint Eq.37 with respect to $c_2$ around an initial guess $\tilde{c}_2$ (we note that when $\tilde{c}_2$ is chosen close enough to the optimal $c_2$, the solution of the linearized problem should be close the solution of the original problem). The value of $\tilde{c}_2$ is computed based on the substrate temperature currently been used in Eq.38 (at $t = 0$, the $\tilde{c}_2$ is computed based on the initial substrate temperature). Therefore, the original optimization problem is reduced to:

$$\min_{c_2} J = (r_{set}^2 - \langle r_{final} \rangle^2)^2 \tag{41}$$

subject to

$$\langle r_{final} \rangle^2 = \langle r_{final}(\tilde{c}_2) \rangle^2 + (c_2 - \tilde{c}_2) \frac{\partial \langle r_{final}(\tilde{c}_2) \rangle^2}{\partial \tilde{c}_2} \tag{42}$$

$$c_{2,min} \leq c_2 \leq c_{2,max} \tag{43}$$

where $c_{2,min}$ and $c_{2,max}$ are the lower bound and upper bound of $c_2$ respectively. The second constraint is added due to the fact that $c_2$ can only take values within the corresponding range specified by Eqs.38 and 40, and $c_{2,min}$ and $c_{2,max}$ are determined as follows:

$$\begin{aligned} c_{2,min} &= \frac{e^{-32.002 + 0.0511 T_{min} - 0.1620 W}}{k_{max}^2} \\ c_{2,max} &= \frac{e^{-32.002 + 0.0511 T_{max} - 0.1620 W}}{k_{max}^2} \end{aligned} \tag{44}$$

A standard procedure based on the active set method [14] is used to solve the optimization problem of Eq.41. First, we drop the inequality constraint Eq.43, and a direct computation of the above problem by substituting the equality constraint into the objective function yields:

$$\bar{c}_2 = \tilde{c}_2 + \frac{r_{set}^2 - \langle r_{final}(\tilde{c}_2) \rangle^2}{\frac{\partial \langle r_{final}(\tilde{c}_2) \rangle^2}{\partial \tilde{c}_2}} \tag{45}$$

where $\bar{c}_2$ is the optimal value of $c_2$ without the inequality constraint Eq.43. Then, we check whether the inequality constraint is violated by $\bar{c}_2$, if the inequality constraint is inactive (i.e., the constraint is not violated), $\bar{c}_2$ is considered to be the optimal value for the linearized optimization problem. On the other hand, if the inequality constraint is active (i.e., the constraint is violated), the optimization problem is resolved accounting for the active

constraint (which serves as another equality constraint). In such case, $c_2$ can only take the value of $c_{2,min}$ (when the lower bound is violated by $\bar{c}_2$) or $c_{2,max}$ (when the upper bound is violated by $\bar{c}_2$), hence, the optimal value is just the only feasible value $c_{2,min}$ or $c_{2,max}$.

However, since Eq.42 is the linearization of Eq.37, $\bar{c}_2$ might only be a suboptimal value for the original problem. To this end, we can use this suboptimal $\bar{c}_2$ as a new guess and repeat the linearzation procedure until $\bar{c}_2$ converges to the optimal value (the convergence is guaranteed if the original problem is convex), but for the sake of simplicity, such iterative procedure is not adopted in this work. Once the optimal $c_2$ is determined, by substituting $c_2$ into Eq.38, the optimal $T$ can be obtained and used as the output of the controller.

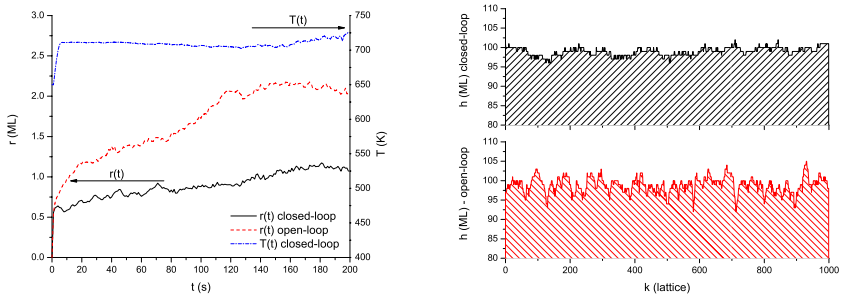
*Remark 3.* Since Eq.37, is in fact, a finite approximation of the predicted final surface roughness, to achieve a control precision $\epsilon$, $m$ should be chosen large enough for each optimization computation so that the approximation error is less than $\epsilon$ (see the Section 4.1 for detailed discussion). However, to achieve the same control precision, the minimum $m$ needed may vary depending on the specific surface configuration (i.e., current states $z_n$). On one hand, when the length scale of the surface fluctuation is very small, the magnitude of the high-order states becomes significant, hence, $m$ need to be relatively large so that these high-order states are included in the calculation. On the other hand, when the length scale of the surface fluctuation is relatively large, the contribution from the high-order states becomes negligible compared to the low-order states, hence, a relatively small $m$ should be good enough for precise calculation. Therefore, in our implementation, the desired control precision is achieved by adding more states to the finite-dimensional system until the approximation error (computed based on Eq.35) is small enough (we note that the approximation error depends on the actual values of the states as shown in Eq.35), rather than by specifying the number of states that should be evaluated from the surface snapshot before hand. However, a limit on the maximum number of states to be used is imposed to guarantee that the computation time of the controller does not prevent real-time implementation (control precision may be reduced as a trade off against the computation time).

*Remark 4.* Since the control action is computed using closed-form expressions, the computation cost is proportional to the number of states used $m$ but independent of the optimization horizon $t_{dep} - t$; however, to evaluate the values of the $m$ states, an additional computation time on the order of $k_{max}m$ is needed for each surface measurement. Nevertheless, even for a lattice size that corresponds to the largest physical dimension of the sampling area that can be achieved by common surface measurement techniques (i.e., a few microns), such computation can still be completed with seconds using currently available computing power. On the other hand, such task is almost impossible to achieve using a kMC code, whose computation cost is on the order of $k_{max}^2(t_{dep}-t)$ for merely a single run. Furthermore, we note that the evaluation

of each state is independent of other states, and therefore, can be executed in parallel, while the kMC code, being a serial calculation, is unsuitable for parallel processing.

### 4.3 Closed-Loop Simulations

A kMC code with a lattice size $k_{max} = 1000$ is used to simulate the thin film deposition process, and the substrate temperature is restricted within 300 $K$ to 900 $K$. The measurement interval, as well as the control interval, is set to be 1 $s$. We limit the maximum number of states to be used (in our case, to $m = 500$) to guarantee the maximum possible computation time for each control action is within certain requirement, however, for most of the time the number of states needed by the controller is much smaller.




(a) Surface roughness and substrate temperature profiles.

(b) Final thin film surface profile.

Fig. 9: Simulation results of a 1000 $s$ closed-loop deposition process with thin film growth rate $W = 0.5\ ML/s$ and final roughness setpoint $r_{set} = 1.0\ ML$.

Fig.9(a) shows the surface roughness and substrate temperature profiles of a closed-loop deposition process with thin film growth rate $W = 0.5\ ML/s$ and of an open-loop deposition with the same growth rate and a fixed substrate temperature $T = 650\ K$. The control objective is to drive the final surface roughness of the thin film to 1.0 $ML$ (monolayers) at the end of the 1000 $s$ deposition. It can be seen that the final surface roughness is controlled at the desired level while an open-loop deposition with the same initial deposition condition would lead to a 100% higher final surface roughness as shown in Fig.9(a) (a comparison between the surfaces of the thin films deposited with closed-loop and open-loop deposition is shown in Fig.9(b)).

Fig.10 shows the final surface roughness histogram of the thin films deposited using 100 different closed-loop depositions with final surface roughness

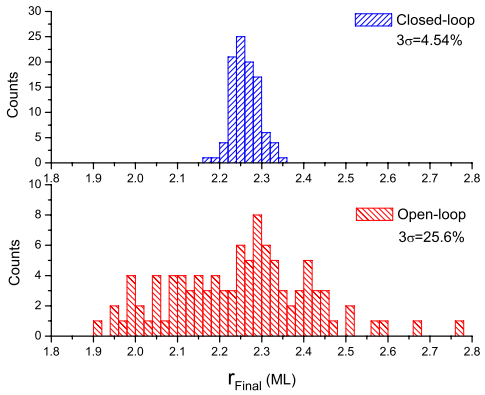
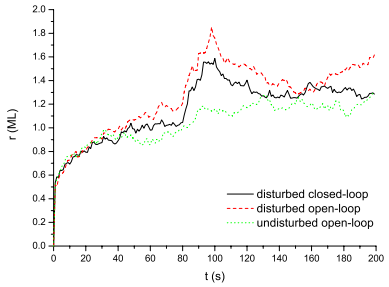



Fig. 10: Histogram of final surface roughness of 100 closed-loop and 100 open-loop thin film depositions targeted at the same surface roughness level.

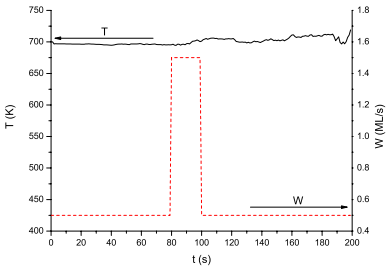
setpoint of 2.25 $ML$ and 100 different open-loop depositions. It can be seen that the average surface roughness of the thin films deposited by the open-loop depositions is very close to the average surface roughness of the thin films deposited by the closed-loop deposition and the well-designed recipe-based open-loop depositions, however, the variance among the thin films from different open-loop deposition runs is over 400% higher than that of closed-loop deposition runs even though no process disturbance is considered in the simulations. This is due to the fact that the stochastic nature of the microscopic processes of the film growth cannot be handled effectively without having a real-time feedback controller that can compensate for the stochastic deviation from the expectation. As a result, if the tolerance on the thin film surface roughness to fabricate a certain device is $\pm 0.1$ $ML$, over half of the thin films prepared by the recipe-based deposition would be disqualified. Therefore, introducing real-time feedback control system that directly aiming at the material and electrical properties of the thin films is one of the most effective, if not the only, solution to reduce cost and meet the ever increasing film quality requirements demanded by the devices which are already down to the nanometer regime.

To study the robustness of the closed-loop deposition with respect to process disturbance, open-loop and closed-loop depositions are simulated in which same process disturbance are introduced during all depositions. Particularly, for the 200 $s$ deposition, a step change in the adsorption rate is introduced at $t = 80$ $s$, and $W$ is change from 0.5 $ML$ to 1.5 $ML$; the adsorption rate $W$ remains at 1.5 $ML$ for 20 $s$ and then drop back to 0.5 $ML$ immediately (such square-wave changes in the adsorption rate may be caused by the spikes in the gas delivery system of the CVD reactor). The roughness set-point of the roughness controller is 1.3 $ML$, and the substrate temperature of all the open-loop depositions is kept constant at 700 $K$ so that the expected

final roughness of the deposited films would be 1.3 $ML$ if no disturbance is present.




(a) Profiles of surface roughness: undisturbed open-loop deposition (dotted green line), disturbed open-loop deposition (dashed red line) and disturbed closed-loop deposition (solid black line).




(b) Profiles of the process disturbance $W$, and the manipulated substrate temperature $T$ (in disturbed closed-loop deposition).

Fig. 11: Simulation results of a disturbed deposition process with nominal thin film growth rate $W = 0.5\ ML/s$ and final roughness setpoint $r_{set} = 1.3\ ML$.

Fig. 11(a) shows the typical surface roughness profile of the undisturbed open-loop deposition (dotted green line), and the surface roughness profile of the process under disturbance, of the open-loop deposition (dashed red line) and closed-loop deposition (solid black line). Fig. 11(b) shows the profiles of the disturbance variable $W$ (it is not used to compute the control action, i.e., the roughness controller is unaware of the abnormal adsorption rate) and the manipulated substrate temperature $T$ (in closed-loop deposition). It can be seen that when disturbance is introduced, the final surface roughness of the film deposited with open-loop operation is much higher than the desired level (in this case more than 20 %), while the final surface roughness of the film deposited under feedback control is still kept at the desired level (i.e., the controller is able to bring the final surface roughness down to the desired level after the disturbance when the thin film surface is unexpectedly roughened).

Fig.10 shows the histograms of the final surface roughness of thin films deposited with open-loop (with and without process disturbance) and closed-loop (with process disturbance) operations, and each histogram include 100 different simulation runs. It can be seen that the average surface roughness of the thin films deposited by the open-loop depositions is shifted up for more than 5% with the presence of process disturbance. Despite of that, the average surface roughness of the thin films deposited under feedback control is very close to the desired level even with the presence of process disturbance, and

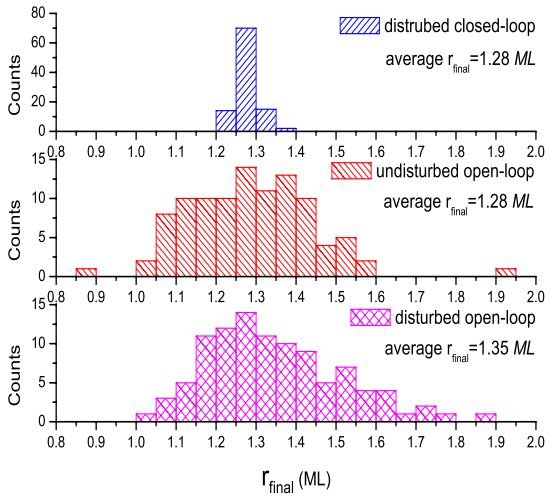



Fig. 12: Histograms of final surface roughness of thin films deposited with open-loop (with and without process disturbance) and closed-loop (with process disturbance) operations.

the variance is reduced as it is the case in the previous simulation where no disturbance was introduced.

## 5 Conclusions

In this work, we presented a systematic method for the construction of linear stochastic PDE models using data obtained from kMC simulations. A thin film deposition process including molecule adsorption and surface migration was used to illustrate the application of the method. Open-loop simulation results demonstrated the accuracy of the constructed linear stochastic PDE model for the thin film deposition process. Furthermore, an optimization-based feedback controller was designed using the constructed stochastic PDE model and closed-loop system simulation results demonstrated that the controller is capable of controlling the surface roughness of the thin film to the desired level, reduce film roughness variability and reject the effect of disturbances.

*Acknowledgement.* Financial support for this work from the NSF (ITR), CTS-0325246, is gratefully acknowledged.

## References


1. A. Armaou, P.D. Christofides: Plasma-enhanced chemical vapor deposition: Modeling and control. *Chem. Eng. Sci.* **54**, 3305–3314 (1999)
2. A. Armaou, P.D. Christofides: Dynamic optimization of dissipative PDE systems using nonlinear order reduction. *Chem. Eng. Sci.* **57**, 5083–5114 (2002)
3. A. Armaou, C.I. Siettos, I.G. Kevrekidis: Time-steppers and 'coarse' control of distributed microscopic processes. *Int. J. Robust Nonlin. Control* **14**, 89–111 (2004)
4. K.J. Åström: *Introduction to Stochastic Control Theory* (Academic Press, New York 1970)
5. J. Baker, P.D. Christofides: Finite dimensional approximation and control of nonlinear parabolic PDE systems. *Int. J. Contr.* **73**, 439–456 (2000)
6. E. Bendersky, P.D. Christofides: Optimization of transport-reaction processes using nonlinear model reduction. *Chem. Eng. Sci.* **55**, 4349–4366 (2000)
7. Y.M. Cho, P. Gyugyi: Control of rapid thermal processing: A system theoretic approach. *IEEE Trans. Control Sys. Technol.* **5**, 644–653 (1997)
8. P.D. Christofides: *Nonlinear and Robust Control of Partial Differential Equation Systems: Methods and Applications to Transport-Reaction Processes* (Birkhäuser, Boston 2001)
9. P.D. Christofides, P. Daoutidis: Finite-dimensional control of parabolic PDE systems using approximate inertial manifolds. *J. Math. Anal. & Appl.* **216**, 398–420 (1997)
10. J.W. Eaton, J.B. Rawlings: Feedback-control of chemical processes using online optimization techniques. *Comp. Chem. Eng.* **14**, 469–479 (1990)
11. S.F. Edwards, D.R. Wilkinson: The surface statistics of a granular aggregate. *Proc. R. Soc. Lond. A* **381**, 17–31 (1982)
12. K.A. Fichthorn, W.H. Weinberg: Theoretical foundations of dynamic Monte Carlo simulations. *J. Chem. Phys.* **95**, 1090–1096 (1991)
13. M.A. Gallivan, R.M. Murray: Reduction and identification methods for Markovian control systems, with application to thin film deposition. *Int. J. Robust Nonlin. Control* **14**, 113–132 (2004)
14. P.E. Gill, W. Murray, M.A. Saunders, M.H. Wright: Procedures for optimization problems with a mixture of bounds and general linear constraints. *ACM Trans. Math. Software* **10**, 282–298 (1984)
15. D.T. Gillespie: A general method for numerical simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.* **22**, 403–434 (1976)
16. D.W. Greve, T.J. Knight, X. Cheng, B.H. Krogh, M.A. Gibson, J. LaBrosse: Process control based on quadrupole mass spectrometry. *J. Vac. Sci. Technol. B* **14**, 489–493 (1996)
17. B. Johs, D. Doerr, S. Pittal, I.B. Bhat, S. Dakshinamurthy: Real-time monitoring and control during movpe growth of cdte using multiwavelength ellipsometry. *Thin Solid Films* **233**, 293–296 (1993)
18. M.A. Katsoulakis, A.J. Majda, D.G. Vlachos: Coarse-grained stochastic processes and Monte Carlo simulations in lattice systems. *J. Comp. Phys.* **186**, 250–278 (2003)
19. N. Kazantzis, C. Kravaris: Nonlinear observer design using Lyapunov's auxiliary theorem. *Syst. & Contr. Lett.* **34**, 241–247 (1999)
20. E. Kreyszig: *Advanced Engineering Mathematics*, sixth edition (John Wiley & Sons, 1988)

21. K.B. Lauritsen, R. Cuerno, H.A. Makse: Noisy Kuramote-Sivashinsky equation for an erosion model. *Phys. Rev. E* **54**, 3577–3580 (2003)
22. Y. Lou, P.D. Christofides: Estimation and control of surface roughness in thin film growth using kinetic Monte-Carlo models. *Chem. Eng. Sci.* **58**, 3115–3129 (2003)
23. Y. Lou, P.D. Christofides: Feedback control of growth rate and surface roughness in thin film growth. *AIChE J.* **49**, 2099–2113 (2003)
24. Y. Lou, P.D. Christofides: Feedback control of surface roughness of GaAs (001) thin films using kinetic Monte-Carlo models. *Comp. & Chem. Eng.* **29**, 225–241 (2004)
25. Y. Lou, P.D. Christofides: Feedback control of surface roughness in sputtering processes using the stochastic Kuramoto-Sivashinsky equation. *Comp. & Chem. Eng.* **29**, 741–759 (2005)
26. Y. Lou, P.D. Christofides: Feedback control of surface roughness using stochastic PDEs. *AIChE J.* **51**, 345–352 (2005)
27. G.Z. Mao, L. Lobo, R. Scaringe, M.D. Ward: Nanoscale visualization of crystal habit modification by atomic force microscopy. *Chem. Mater.* **9**, 773–783 (1997)
28. M. Marsili, A. Maritan, F. Toigo, J.R. Banavar: Stochastic growth equations and reparametrization invariance. *Rev. Mod. Phys.* **68**, 963–983 (1996)
29. M. Metzger, R. Backofen: Optimal temperature profiles for annealing of $GaAs$-crystals. *J. Crystal Growth* **220**, 6–15 (2000)
30. D. Ni, P.D. Christofides: Dynamics and control of thin film surface microstructure in a complex deposition process. *Chem. Eng. Sci.* **60**, 1603–1617 (2005)
31. D. Ni, P.D. Christofides: Multivariable predictive control of thin film deposition using a stochastic pde model. *Ind. & Eng. Chem. Res.* **44**, 2416–2427 (2005)
32. D. Ni, Y. Lou, P.D. Christofides, L. Sha, S. Lao, J.P. Chang: Real-time carbon content control for PECVD $ZrO_2$ thin film growth. *IEEE Trans. Semiconduct. Manufact.* **17**, 221–230 (2004)
33. Y. Nitta, M. Shibata, K. Fujita, M. Ichikawa: Nanometer-scale $Ge$ selective growth on $Si(001)$ using ultrathin $SiO_2$ film. *Surf. Sci.* **462**, 587–593 (2000)
34. H.M. Park, T.Y. Yoon, O.Y. Kim: Optimal control of rapid thermal processing systems by empirical reduction of modes. *Ind. Eng. Chem. Res.* **38**, 3964–3975 (1999)
35. S. Park, D. Kim, J. Park: Derivation of continuum stochastic equations for discrete growth models. *Phys. Rev. E* **65**, 015102(R) (2002)
36. S. Raimondeau, P. Aghalayam, A. B. Mhadeshwar, D.G. Vlachos: Parameter optimization of molecular models: Application to surface kinetics. *Ind. Eng. Chem. Res.* **42**, 1174–1183 (2003)
37. G. Renaud, R. Lazzari, C. Revenant, A. Barbier, M. Noblet, O. Ulrich, F. Leroy, J. Jupille, Y. Borensztein, C.R. Henry, J.P. Deville, F. Scheurer, and J. Mane-Mane. Real-time monitoring of growing nanoparticles. *Science* **300**, 1416–1419 (2003)
38. E. Rusli, T.O. Drews, D.L. Ma, R.C. Alkire, R.D. Braatz: Nonlinear feedback control of a coupled kinetic monte carlo–finite difference simulation. In: *Proceed. IFAC Symp. ADCHEM*, pages 597–602 (2003)
39. T. Shitara, D. D. Vvedensky, M.R. Wilby, J. Zhang, J.H. Neave, B.A. Joyce: Step-density variations and reflection high-energy electron-diffraction intensity oscillations during epitaxial growth on vicinal GaAs(001). *Physical Review B* **46**, 6815–6824 (1992)

40. C.I. Siettos, A. Armaou, A.G. Makeev, I.G. Kevrekidis: Microscopic/stochastic timesteppers and "coarse" control: a KMC example. *AIChE J.* **49**, 1922–1926 (2003)
41. M.E. Taylor, H.A. Atwater: Monte Carlo simulations of epitaxial growth: comparison of pulsed laser deposition and molecular beam epitaxy. *Appl. Surf. Sci.* **127**, 159–163 (1998)
42. A. Theodoropoulou, R.A. Adomaitis, E. Zafiriou: Inverse model based real-time control for temperature uniformity of RTCVD. *IEEE Trans. Semiconduct. Manufact.* **12**, 87–101 (1999)
43. D.G. Vlachos: Multiscale integration hybrid algorithms for homogeneous-heterogeneous reactors. *AIChE J.* **43**, 3031–3041 (1997)
44. D.D. Vvedensky: Edwards-Wilkinson eqaution from lattice transition rules. *Phys. Rev. E* **67**, 025102(R) (2003)
45. D.D. Vvedensky, A. Zangwill, C.N. Luse, M.R. Wilby: Stochastic equations of motion for epitaxial growth. *Phys. Rev. E* **48**, 852–862 (1993)

# Lattice Boltzmann Method and Kinetic Theory

S. Ansumali[1], S. S. Chikatamarla[2], C. E. Frouzakis[2], I. V. Karlin[2], and I. G. Kevrekidis[3]

[1] School of Chemical and Biomedical Engineering, Nanyang Technological University, 639798 Singapore
[2] Aerothermochemistry and Combustion Systems Laboratory, ETH Zurich, 8092 Zürich, Switzerland
[3] Department of Chemical Engineering, Princeton University, NJ 08544-5263, USA

**Summary.** One of the classical questions of non-equilibrium thermodynamics is the validity of various closure approximations in nontrivial flows. We study this question for a lid-driven cavity flow using a minimal molecular model derived from the Boltzmann equation. In this nontrivial flow, we quantify the model as a superset of the Grad moment approximation and visualize the quality of the Chapman-Enskog and Grad closure approximations. It is found that the Grad closure approximation is strikingly more robust than the Chapman-Enskog approximation at all Knudsen numbers studied. Grad's approximation is used to formulate a novel outflow boundary condition for lattice Boltzmann simulations.

## 1 Introduction

The overwhelming majority of fluid flows of physical and engineering interest are slow, i. e., characteristic flow speed $u$ is small compared to the speed of sound $c_\mathrm{s}$. This is quantified by the Mach number, $\mathrm{Ma} \sim u/c_\mathrm{s}$, which typically varies from $10^{-3} - 10^{-2}$ in hydrodynamic flows (turbines, reactors etc) to $10^{-4}$ in flows at a micrometer scale. The simplest characterization of the degree of molecularity is then the Knudsen number $\mathrm{Kn} \sim \lambda/H$, the ratio of the mean free path $\lambda$ and the characteristic scale $H$ of variation of hydrodynamic fields (density, momentum, and energy). When $\mathrm{Kn} < 10^{-3}$, one considers the hydrodynamic limit where molecularity reduces to a set of transport coefficients (viscosity, thermal conductivity etc). If, in addition, the Mach number is also small, one obtains the incompressible hydrodynamics with the ordering $\mathrm{Kn} \ll \mathrm{Ma} \ll 1$, and the flow can be characterized solely by the ratio $\mathrm{Re} \sim \mathrm{Ma}/\mathrm{Kn}$ (one of the definitions of the Reynolds number).

In recent years, the lattice Boltzmann method (LBM) has drawn considerable attention as a simulation method for flows at low Mach numbers. LBM was originally introduced as a derivative of lattice-gas models [22, 13, 32, 7] to simulate incompressible Navier-Stokes equations. LBM offers fully discrete

(in space-time-velocity) kinetic models for populations of fictitious particles with the velocities represented by links of a regular Bravais lattice (with possibly several sub-lattices). LBM operates on a highly efficient "stream-along-links-and-collide-at-nodes" schedule making the method almost ideally compliant with parallel architectures (for a general reference on LBM see, e. g. [35, 15, 36]). At present, LBM can be regarded as an established method for hydrodynamic simulations [14, 2]. Later, in a series of works [33, 4, 7, 8, 5], lattice Boltzmann equation has been derived from the continuous kinetic theory. Owing to their outstanding computational features and established relations to the continuous kinetic theory there is increasing interest in applying lattice Boltzmann models also to micro-flow simulation [10, 4, 31, 39, 6, 5].

In this paper we use the lattice Boltzmann models in order to address one of the central issues of non-equilibrium statistical physics, namely, how the system with many degrees of freedom reduces to a system with a smaller number of degrees of freedom. The study of this question was pioneered in the framework of the Boltzmann kinetic equations by the works of Hilbert, Enskog, Chapman and Grad [12, 20]. This sometimes is referred to as the closure problem. Over the years, several directions of research grew from this question, in particular, equation-free multi-scale computations [38, 27] and the method of invariant manifolds [17, 19].

However, till now there is a limited access to validation of various assumptions behind the closure approximations, especially in nontrivial flows. Therefore, numerical studies which can shed light on this question are required. Such a study is presented in this paper. We study the simplest kinetic equation pertinent to a micro-flow in a lid-driven cavity (the isothermal two-dimensional nine-velocity model [7]). We use this model in order to validate the relevance and accuracy of the two classical closure approximations, the Chapman-Enskog closure leading to the Navier-Stokes approximation of hydrodynamics, and the Grad closure approximation. Aside from the obvious relevance of this study to the general question of validity of closure approximations, it should have important practical consequences for such issues as grid refinement, boundary conditions etc.

The paper is organized as follows: For the sake of completeness, the kinetic model is briefly presented in section 2. In section 3, we show the relation of our model to the well-known Grad moment system derived from the Boltzmann kinetic equation [20]. We compare analytically the dispersion relation for the present model and the Grad moment system. This comparison reveals that the kinetic model of section 2 is a superset of Grad's moment system, rather than just a superset of the Navier-Stokes equations. In section 4, a brief description of the lattice Boltzmann method is given. In section 5, a parametric numerical study of the flow in a micro-cavity is presented. Results are also compared to direct simulation Monte Carlo data. Section 6 is the main focus of this study. In this section, the reduced description of the model kinetic equation is investigated, and a visual representation quantifying various closure assumptions is achieved. The major finding of this section is that the Grad

closure approximation is much more accurate than the Chapman-Enskog closure for all values of the Knudsen number, even in the incompressible limit of the flow. This indicates that Grad's distributions can be used for various grid-saving simulation strategies. A specific example is considered in section 7 where Grad's distribution function is used in order to establish outflow boundary condition for the simulation of open flows. The new outflow condition is validated with a three-dimensional simulation of a backward-facing step flow. We conclude in section 8 with a spectral analysis of the steady state flow (which can be regarded as yet another closure approximation) and some suggestions for further research.

## 2 Minimal Kinetic Model

We consider a two-dimensional discrete velocity model with the following set of nine discrete velocities:

$$\begin{aligned} c_x &= [0, 1, 0, -1, 0, 1 - 1, -1, 1]\,, \\ c_y &= [0, 0, 1, 0, -1, 1, 1, -1, -1]\,. \end{aligned} \tag{1}$$

The local hydrodynamic fields are defined in terms of the discrete population, $f_i$, as:

$$\sum_{i=1}^{9} f_i\{1,\, c_{x\,i},\, c_{y\,i}\} = \{\rho,\, j_x,\, j_y\}, \tag{2}$$

where $\rho$ is the local mass density, and $j_\alpha$ is the local momentum density of the model. The populations $f_i \equiv f(\mathbf{x}, \mathbf{c}_i, t)$ are functions of the discrete velocity $\mathbf{c}_i$, position $\mathbf{x}$ and time $t$. We consider the following kinetic equation for the populations (the Bhatnagar-Gross-Krook single relaxation time model):

$$\partial_t f_i + \mathbf{c}_i \cdot \partial_{\mathbf{x}} f_i = -\frac{1}{\tau}\left(f_i - f_i^{\mathrm{eq}}(f)\right), \tag{3}$$

where $\tau$ is the relaxation time, and $f_i^{\mathrm{eq}}$ is the local equilibrium [7]:

$$\begin{aligned} f_i^{\mathrm{eq}} = \rho W_i &\left(2 - \sqrt{1 + 3{u_x}^2}\right)\left(2 - \sqrt{1 + 3{u_y}^2}\right) \\ &\times \left(\frac{2u_x + \sqrt{1 + 3{u_x}^2}}{1 - u_x}\right)^{c_{xi}} \left(\frac{2u_y + \sqrt{1 + 3{u_y}^2}}{1 - u_y}\right)^{c_{yi}}, \end{aligned} \tag{4}$$

with $u_\alpha = j_\alpha/\rho$, and the weights $W_i$ are

$$W = \left[\frac{16}{36}, \frac{4}{36}, \frac{4}{36}, \frac{4}{36}, \frac{4}{36}, \frac{1}{36}, \frac{1}{36}, \frac{1}{36}, \frac{1}{36}\right]. \tag{5}$$

The local equilibrium distribution $f_i^{\mathrm{eq}}$ is the minimizer of the discrete $H$ function [25]:

$$H = \sum_{i=1}^{9} f_i \ln\left(\frac{f_i}{W_i}\right), \tag{6}$$

under the constraints of the local hydrodynamic fields (2). Note the important factorization over spatial components of the equilibrium (4). This is similar to the familiar property of the local Maxwell distribution, and it distinguishes (4) among other discrete-velocity equilibria. In the hydrodynamic regime, the model recovers the Navier-Stokes equation with viscosity coefficient $\mu = p\tau$, where $p = \rho c_s^2$ is the pressure; $c_s^2 = 1/3$ is the speed of sound in this model.

The kinetic model just described was derived upon discretization of the velocity set from continuous kinetic theory in Ref. [7] (see also an earlier relevant study [33], and an extension to a weakly compressible case [5]). It has been recently shown by several groups that this model compares well with analytical results of kinetic theory in simple flow geometries (channel flows), as well as with molecular dynamics simulations for small but finite Knudsen numbers [4, 8, 6, 31, 37, 39, 29, 41]. We use it here as a realistic kinetic theory at low Mach and Knudsen numbers in order to access the quality of various closure approximations. In the next section, we shall make a first step in quantifying this model as a superset of the Grad moment system of continuous kinetic theory.

## 3 Grad's Moment System and the Kinetic Model: Linear Case

### 3.1 The Moment System

It proves useful to represent the discrete velocity model (3) in the form of a moment system. In this section, in order to derive some analytical results, we shall consider the linearized version of the model. While any linearly independent set of variables can be used to write a moment system equivalent to (3), we choose the following nine non-dimensional moments as independent variables:

$$M = \left[\frac{\rho}{\rho_0}, \frac{j_x}{\rho_0 c_s}, \frac{j_y}{\rho_0 c_s}, \frac{P}{\rho_0 c_s^2}, \frac{N}{\rho_0 c_s^2}, \frac{P_{xy}}{\rho_0 c_s^2}, \frac{q_x}{2\rho_0 c_s^3}, \frac{q_y}{2\rho_0 c_s^3}, \frac{\psi}{2\rho_0 c_s^4}\right], \tag{7}$$

where

$$\psi = R_{yyyy} + R_{xxxx} - 2R_{xxyy} \tag{8}$$

is a scalar obtained from the $4^{th}$-order moments

$$R_{\alpha\beta\gamma\theta} = \sum_{i=1}^{9} f_i c_{\alpha i} c_{\beta i} c_{\gamma i} c_{\theta i}, \tag{9}$$

and

$$N = \sum_{i=1}^{9} f_i (c_{xi}^2 - c_{yi}^2)/2 \equiv (P_{xx} - P_{yy})/2 \tag{10}$$

is the difference of the normal stresses. Furthermore,

$$P = \sum_{i=1}^{9} f_i c_i^2,$$

is the trace of the pressure tensor, and

$$q_\alpha = \sum_{i=1}^{9} f_i c_{\alpha i} c_i^2,$$

is the energy flux obtained by contraction of the third-order moment,

$$Q_{\alpha\beta\gamma} = \sum_{i=1}^{9} f_i c_{\alpha i} c_{\beta i} c_{\gamma i}.$$

Time and space are made non-dimensional in such a way that for a fixed system size $L$ they are measured in the units of mean free time and mean free path, $\mathbf{x}' = \mathbf{x}/(L\mathrm{Kn}), t' = t/\tau$, where

$$\mathrm{Kn} = \tau c_\mathrm{s}/L$$

is the Knudsen number. The linearized equations for the moments $M$ (7) read (from now on we use the same notation for the non-dimensional variables):

$$\begin{aligned}
\partial_t \rho + \partial_x j_x + \partial_y j_y &= 0, \\
\partial_t j_x + \partial_x (P + N) + \partial_y P_{xy} &= 0, \\
\partial_t j_y + \partial_x P_{xy} + \partial_y (P - N) &= 0, \\
\partial_t P + \partial_x q_x + \partial_y q_y &= (\rho - P), \\
\partial_t N + \partial_x (q_x - Q_{xyy}) - \partial_y (q_y - Q_{yxx}) &= -N, \\
\partial_t P_{xy} + \partial_x Q_{yxx} + \partial_y Q_{yyx} &= -P_{xy}, \\
\partial_t q_x + \partial_x R_{xx\alpha\alpha} + \partial_y R_{xy\alpha\alpha} &= (2j_x - q_x), \\
\partial_t q_y + \partial_x R_{xy\alpha\alpha} + \partial_y R_{yy\alpha\alpha} &= (2j_y - q_y), \\
\partial_t \psi + \partial_x (j_x - q_x) + \partial_y (j_y - q_y) &= (2\rho - \psi).
\end{aligned} \tag{11}$$

By construction of the discrete velocities (1), the following algebraic relations are satisfied:

$$
\begin{aligned}
Q_{xyy} &= 2q_x - 3j_x, \\
Q_{yxx} &= 2q_y - 3j_y, \\
R_{xy\alpha\alpha} &= 3P_{xy}, \\
R_{xx\alpha\alpha} &= 3\left(P + \frac{1}{2}N\right) - \frac{1}{2}\psi, \\
R_{yy\alpha\alpha} &= 3\left(P - \frac{1}{2}N\right) - \frac{1}{2}\psi.
\end{aligned} \tag{12}
$$

Apart from the lack of conservation of energy and linearity of the advection, equation (11) is similar to Grad's two-dimensional 8-moment system (see [20] for the original derivation of Grad's moment systems, and [19] for a modern discussion and extensions). It should be reminded here, that the variables used in the $D$-dimensional Grad's system are density, $D$ components of the momentum flux, $D(D+1)/2$ components of the pressure tensor and $D$ components of the energy flux. The number of fields in Grad's system is 8 for $D = 2$ and 13 for $D = 3$. However, in the present case a particular component of the $4^{th}$-order moment is also included as a variable. In other words, Grad's non-linear closure for the $4^{th}$-order moment is replaced by an evolution equation with a linear advection term. We note here that while the formulation of boundary conditions for Grad's moment system remains an open problem, the boundary conditions for the extended moment system are well established through its discrete-velocity representation (3) [4]. The moment system (11) reveals the meaning of the densities appearing in model: The dimensionless density is the dimensionless pressure of the real fluid in the low Mach number limit, while the momentum flux density should be identified with the velocity in the incompressible limit. With this identification, we shall compare the moment system (11) with Grad's system.

### 3.2 One-Dimensional Grad's Moment System

Since energy is not conserved by model (3), the comparison will be with another Grad moment system which (for $D = 3$) is usually referred to as the 10-moment system. The variables used in this $D$-dimensional Grad's system are density, $D$ components of the momentum flux, and $D(D+1)/2$ components of the pressure tensor, resulting in 6 and 10 variables for $D = 2$ and $D = 3$, respectively. For one-dimensional flows, the linearized Grad's 10-moment system can be written as:

$$
\begin{aligned}
\partial_t p + \gamma \partial_x u_x &= 0, \\
\partial_t u_x + \partial_x P_{xx} &= 0, \\
\partial_t P_{xx} + 3\partial_x u_x &= -\left(P_{xx} - p\right),
\end{aligned} \tag{13}
$$

where $\gamma$ is the ratio of the specific heats of the fluid, and $\gamma = (D+2)/D$ for a $D$-dimensional dilute gas. This model can be described in terms of its dispersion

relation, which upon substitution of the solution in the form $\sim \exp(\omega t + ikx)$ reads:

$$\omega^3 + \omega^2 + 3k^2\omega + \gamma k^2 = 0. \tag{14}$$

The low wave-number asymptotic represents the large-scale dynamics (hydrodynamic scale of Kn $\ll 1$), while the high-wave number limit represent the molecular scales quantified by Kn $\gg 1$. The low wave number (Kn $\ll 1$) asymptotic, $\omega_{\rm l}$, and the large wave number (Kn $\gg 1$) asymptotic, $\omega_{\rm h}$, are:

$$\omega_{\rm l} = \left\{ \frac{(-3+\gamma)}{2} k^2 \pm i\sqrt{\gamma} k,\ -1 - (-3+\gamma)k^2 \right\},$$
$$\omega_{\rm h} = \left\{ \frac{(-3+\gamma)}{6} \pm i\sqrt{3} k,\ -\frac{\gamma}{3} \right\}.$$

The two complex conjugate modes (acoustic modes) of the $O(k^2)$ dynamics, are given by the first two roots of $\omega_{\rm l}$, and represent the hydrodynamic limit (the Navier-Stokes approximation) of the model. The third root in this limit is real and negative, corresponding to the relaxation behavior of the non-hydrodynamic variable (stress): the dominant contribution (equal to $-1$) is the relaxation rate towards the equilibrium value, while the next-order correction suggests slaving of viscous forces, which amounts to the constitutive relation for stress ($(-3+\gamma)/2k^2$ ). Furthermore, the $k^2$ dependence of the relaxation term justifies the assumption of scale separation (the higher the wave-number, the faster the relaxation). The real part of the high wave-number solution $\omega_{\rm h}$ is independent of $k$, which shows that the relaxation at very high Knudsen number is the same for all wavenumbers (so-called "Rosenau saturation" [18, 34]). Thus, the assumption of scale-separation is not valid for high Knudsen number dynamics.

### 3.3 Dispersion Relation for the Moment System

The dispersion relation for the one-dimensional version of the moment system (11) (i.e. neglecting all derivatives in the $y$-direction) reads:

$$(\omega^3+\omega^2+3k^2\omega+k^2)(\omega^3+2\omega^2+(3k^2+1)\omega+k^2)(1+\omega)((1+\omega^2)+2k^2) = 0. \tag{15}$$

The real parts of the roots of this equation (attenuation rates Re$[\omega(k)]$) are plotted in Fig. 1 as functions of the wave vector $k$. It is clear that for one-dimensional flows, the dynamics of three of the moments ($\rho$, $j_x$, and $P$) are decoupled from the rest of the variables, and follows of the dynamics of the one-dimensional Grad's moment system (13) with $\gamma = 1$.

The similarity between Grad's moment system and the present model is an important fingerprint of the kinetic nature of the latter. Grad [20] already mentioned that moment systems are particularly well suited for low Mach number flows. Qualitatively, this is explained as follows: when expansion in the Mach number around the no-flow state is addressed, the first nonlinear

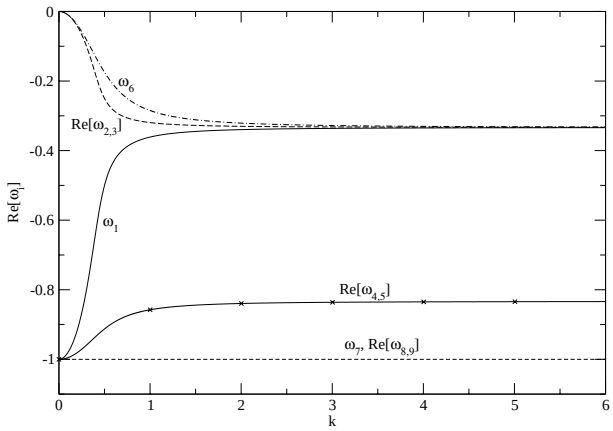



Fig. 1: Real part of the solutions of the dispersion relation (equation (15)). Roots $\omega_{2,3}$ and $\omega_1$ correspond to Grad's subsystem (13). The real-valued root $\omega_6$ and the complex conjugate roots $\omega_{2,3}$ are extended hydrodynamic modes.

terms in the advection are of order $u^2/c_s^2 \sim \mathrm{Ma}^2$. On the other hand, the same order in Ma terms in the relaxation contribute $u^2/(\tau c_s^2) \sim \mathrm{Ma}^2/\mathrm{Kn}$. Thus, if Knudsen number is also small, nonlinear terms in the advection can be neglected while the nonlinearity in the relaxation should be kept. That is why the model (3) - linear in the advection and nonlinear in the relaxation - belongs to the same domain of validity as Grad's moment systems for subsonic flows. Note that in the case of two-dimensional flows, the agreement between the present model and Grad's system is only qualitative. The present moment system is isotropic only up to $O(k^2)$. Thus, the dispersion relation of the model (3) is expected to match the one of Grad's system only up to the same order. In the hydrodynamic and slip-flow regime addressed below, this order of isotropy is sufficient. In the presence of boundaries and/or non-linearities, it is necessary to resort to numerics. In the next section we shall give details on the lattice Boltzmann and entropic lattice Boltzmann discretization of the model (3).

## 4 Lattice Boltzmann Method

The lattice Boltzmann method [22, 32, 7] is the second-order accurate implicit scheme for the kinetic equation (3). Let us briefly derive it here. After integrating (3) over the time $\delta t$, applying the trapezoidal rule in order to evaluate the BGK collision term (second-order accuracy in $\delta t$), and using the transformation [21, 1],

$$g_i(f) = f - \frac{\delta t}{2} Q_i(f), \tag{16}$$

where

$$Q_i = -\frac{1}{\tau}(f_i - f_i^{\mathrm{eq}}), \tag{17}$$

is the short-hand notation for the BGK collision term, we derive the discrete-time scheme for (3):

$$g_i(\mathbf{x} + \mathbf{c}_i\delta t, t + \delta t) = g_i(\mathbf{x}, t) + \frac{2\delta t}{2\,\tau + \delta t}\left[g_i^{\mathrm{eq}}(\mathbf{x}, t) - g_i(\mathbf{x}, t)\right]. \tag{18}$$

Furthermore, fixing the grid points in such a way that if $\mathbf{x}$ is a grid point then also $\mathbf{x} \pm \mathbf{c}_i\delta t$ are the grid points, equation (18) becomes the fully discrete second-order accurate lattice Boltzmann scheme. Note that this implicit second-order scheme for the populations $f_i$ can be interpreted as an explicit first-order scheme for the variables $g_i$ (16) obtained from a kinetic equation of the form (3) with a renormalized relaxation time $\tau'$,

$$\tau' = \tau + \frac{\delta t}{2}. \tag{19}$$

The time stepping in the second-order accurate entropic lattice Boltzmann method [26, 25, 11, 7] is done on the populations in such a way that the collision update respects the monotonicity constraint on the $H$ function:

$$f_i(\mathbf{x} + \mathbf{c}_i\delta t, t + \delta t) = f_i(\mathbf{x}, t) + \frac{\alpha\delta t}{2\,\tau + \delta t}\left[f_i^{\mathrm{eq}}(\mathbf{x}, t) - f_i(\mathbf{x}, t)\right], \tag{20}$$

where $\alpha$ replaces the factor 2 in (18), and is obtained by solving the entropy estimate,

$$H\left(\boldsymbol{f}\right) = H\left(f + \alpha Q(f)\right). \tag{21}$$

Close to the local equilibrium, $\alpha$ is equal to 2. The local adjustments of the relaxation time (via the parameter $\alpha$), as dictated by compliance with the $H$ theorem, guarantee positivity of the distribution function also for the case of discrete time steps, thereby ensuring the non-linear stability of the numerical scheme. While this is important for other applications such as flows at high Reynolds numbers, the distinction between the two schemes is not important in the present study. What will be important below is the discrete-time transform of the populations (16) which enables to interpret certain constructions of populations in the continuous kinetic theory (primarily, Grad's distributions) also for second-order accurate fully discrete schemes.

## 5 Flow in a Lid-Driven Micro-Cavity

The two-dimensional flow in a lid-driven cavity was simulated over a range of Knudsen numbers defined as Kn = Ma/Re. In the simulations, the Mach number was fixed at Ma = 0.01 and the Reynolds number, Re, was varied.

Initially, the fluid in the cavity is at rest and the upper wall of the domain is impulsively set to motion with $u_{\rm lid} = c_{\rm s}{\rm Ma}$. Diffusive boundary conditions are imposed on the walls [4], and the domain was discretized using 151 points in each spatial direction. Time integration is continued till the steady state is reached.

### 5.1 Validation with DSMC Simulation of the Micro-Cavity

In the hydrodynamic regime, the model was validated using results available from continuum simulations [3]. For higher Kn $\sim 0.1$, we compared our results with the DSMC simulation of [23]. Good agreement between the DSMC simulation and the ELBM results can be seen in Fig. 2. It can be concluded, that even for small but finite Knudsen number, the present model provides a semi-quantitative agreement, as far as the flow profile is concerned. We remind here again, that the dimensionless density in the present model corresponds to the dimensionless pressure of a real fluid so that, for quantitative comparison, the density of ELBM model should be compared with the pressure computed from DSMC.

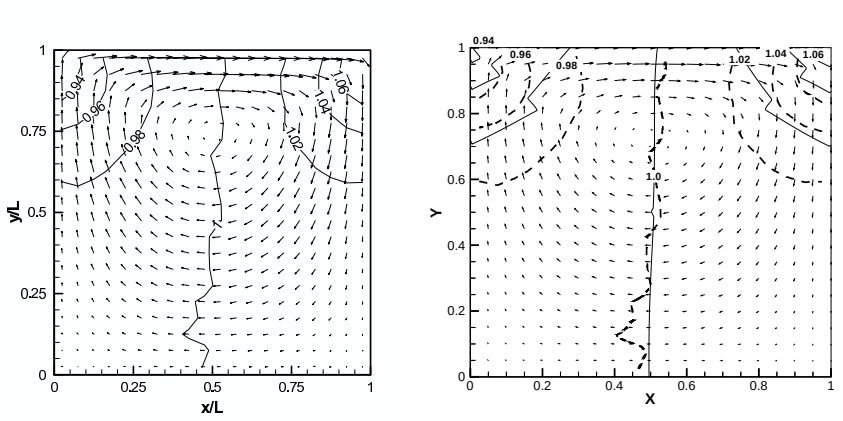


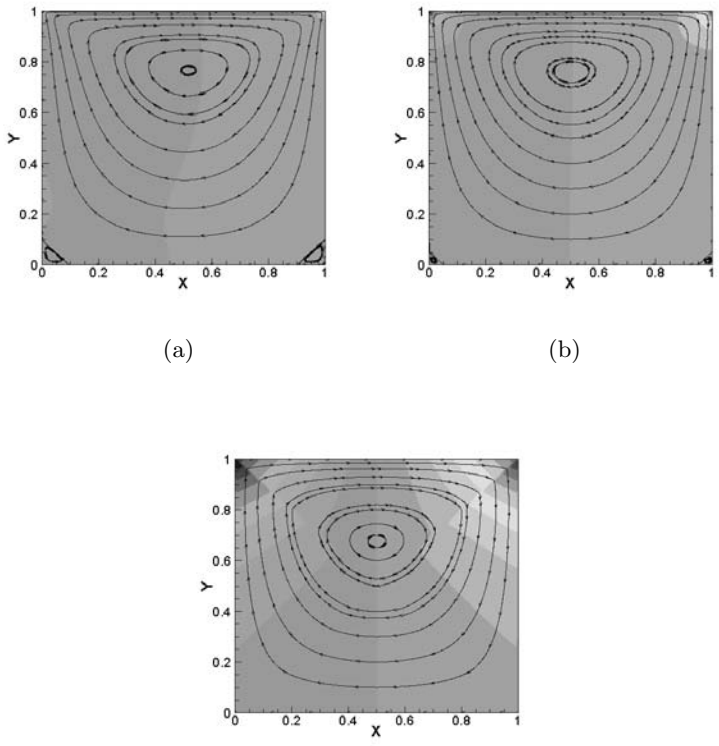

Fig. 2: Flow in a micro-cavity for Kn = 0.1 and Ma = 0.14: DSMC simulation [23] (left) , velocity vector plot and density isolines from ELBM (solid lines) with the DSMC density isolines (dashed lines) superimposed (right).

### 5.2 Parametric Study of the Flow in the Micro-Cavity

Fig. 3 shows the dimensionless density profiles with the streamlines superimposed for Kn $= 0.001, 0.01, 0.1$. For Kn $= 0.001$ (Re $= 10$), the behavior

expected from continuum simulations with a large central vortex and two smaller recirculation zones close to the lower corners can be observed. As the Knudsen number is increased, the lower corner vortices shrink and eventually disappear and the streamlines tend to align themselves with the walls.

The density profiles, as a function of Kn, demonstrate that the assumption of incompressibility is well justified only in the continuum regime, where the density is essentially constant away from the corners. This observation is consistent with the conjecture that incompressibility requires smallness of the Mach as well as of the Knudsen number. In hydrodynamic theory, the density waves decay exponentially fast (with the rate of relaxation proportional to Kn) leading effectively to incompressibility. Thus, it is expected that the onset of incompressibility will be delayed as the Knudsen number increases.




(a) (b)

(c)

Fig. 3: Density isocontours for (a) Kn = 0.001, (b)Kn = 0.01 , and (c) Kn = 0.1 (the variation of the density is $0.995 \leq \rho \leq 1.005$). Superimposed are the streamlines.

## 6 Reduced Description of the Flow

The data from the direct simulation of the present kinetic model were used to validate various closure approximations of kinetic theory in the presence of kinetic boundary layers in a non-trivial flow. In this section, we will present such an analysis for two widely used closure methods, the Navier-Stokes approximation of the Chapman-Enskog expansion and Grad's moment closure.

### 6.1 The Navier-Stokes Approximation

The Chapman-Enskog analysis [12] of the model kinetic equation leads to a closure relation for the non-equilibrium part of the pressure tensor as (the Navier-Stokes approximation):

$$\sigma_{xy} = -\tau c_{\mathrm{s}}^2(\partial_y j_x + \partial_x j_y). \tag{22}$$

Fig. 4 shows a scatter plot of the $xy$ component of the non-equilibrium part of the pressure tensor $P_{xy} - P_{xy}^{\mathrm{eq}}$, versus that computed from the Navier-Stokes approximation of the Chapman-Enskog expansion (22). The upper row is the scatter plot for all points in the computational domain, while the lower row is the scatter plot obtained after removal of the boundary layers close to the four walls of the cavity, corresponding to approximately 10 mean-free paths. In all plots, the dashed straight line of slope equal to one corresponds to the Navier-Stokes closure. These plots clearly reveal that the Navier-Stokes description is valid away from the walls in the continuum as well as in the slip-flow regime. On the other hand, it fails to represent hydrodynamics in the kinetic boundary layer, even at very low Knudsen numbers.

### 6.2 Grad's Approximation

In contrast to the Chapman-Enskog method, the Grad method has an advantage that the approximations are local in space. As the analysis of section 3 suggests, the dynamics of the density, momentum and pressure tensor are almost decoupled from the rest of the moments, at least away from the boundaries. This motivates the Grad-like approximation for the populations,

$$f_i^{\mathrm{Grad}} = W_i \left[\rho + \frac{j_\alpha c_{i\,\alpha}}{c_{\mathrm{s}}^2} + \frac{1}{2\,c_{\mathrm{s}}^4}\left(P_{\alpha\beta} - \delta_{\alpha\beta}\rho c_{\mathrm{s}}^2\right)\left(c_{i\,\alpha}c_{i\,\beta} - c_{\mathrm{s}}^2\delta_{\alpha\beta}\right)\right]. \tag{23}$$

The set of populations parameterized by the values of the density, momentum and pressure tensor (23) is a sub-manifold in the phase space of the system (3), and can be derived in a standard way using quasi-equilibrium procedures [19].

Various moments (7) can now be evaluated on the functions (23) analytically. In Fig. 5, the scatter plot of the computed energy flux $q_x$ and the Grad's

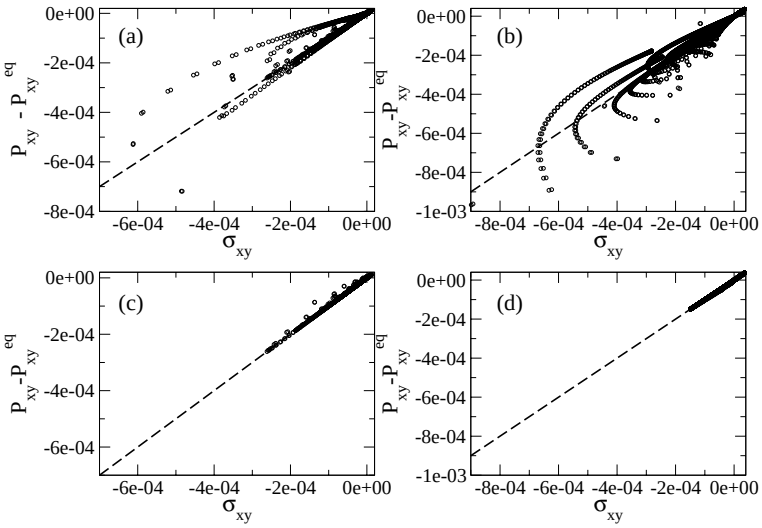



Fig. 4: Scatter plot of the non-equilibrium part of the off-diagonal component of the pressure tensor $P_{xy} - P_{xy}^{\mathrm{eq}}$ and corresponding value computed from Navier-Stokes approximation $\sigma_{xy}$ (22) for all points in the domain ((a) and (b)), and after the removal of the boundary layer corresponding to approximately 10 mean-free path ((c) and (d)). Fig. (a,c) correspond to Kn = 0.001, while Fig. (b,d) correspond to Kn = 0.01. Navier-Stokes behavior is indicated by the straight line of slope equal to one.

closure $q_x^{\mathrm{Grad}}$ is presented. Same as in Fig. 4, the off-closure points in Fig. 5 are associated with the boundary layers. The comparison of the quality with which the closure relations are fulfilled in Fig. 4 and Fig. 5 clearly indicates the advantage of that a Grad's closure. It is quite revealing that even in the case of small Knudsen numbers where one expects the Navier-Stokes closure to be good, the quality of the Grad's closure is much better. The general conclusion from the present visualization is that for slow flows Grad's closure is superior to the Navier-Stokes closure.

## 7 Application: Outflow Condition in Lattice Boltzmann Simulations

Above, we have demonstrated with a specific example (lid-driven cavity flow) that the Grad approximation contains most of the dynamics of the system. This finding is quite remarkable because it suggests that Grad's distribution function (23) can be used for extrapolation of the so-called "missing data" in the lattice Boltzmann simulations. Here we shall give an example how this can be used in imposing the outflow boundary conditions, following Ref. [16].

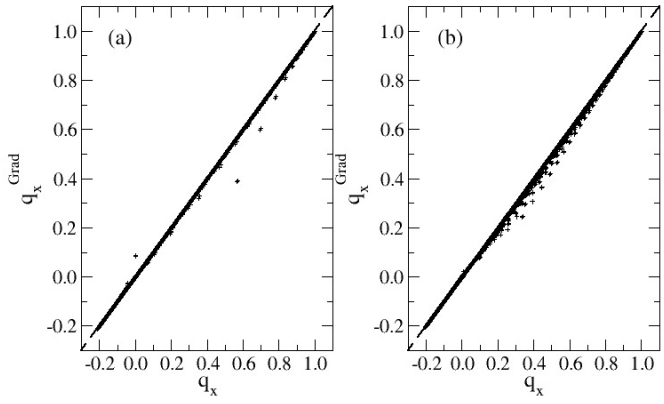



Fig. 5: Scatter plot of the computed energy flux, $q_x$, versus Grad's closure, $q_x^{\mathrm{Grad}}$: (a) Kn = 0.001, (b) Kn = 0.01.

Although the field of applications of LBM has increased considerably during the last decade, there remain outstanding issues (stability, boundary conditions, grid-refinement etc) which so far hindered a wider acceptance of LBM for computational fluid dynamics applications. One of these issues, namely numerical stability of simulations of flows at large Reynolds numbers, has been solved in the framework of the entropic formulation of LBM (see section 4). This solution was essentially based on the choice of a time step that does not violate the entropy growth condition (a physically relevant condition prescribed by the second law of thermodynamics). Furthermore, the boundary conditions at solid walls were derived from the continuous kinetic theory [4]. However, other major difficulties that are not related to the sub-grid instability at high Reynolds numbers, still persist. Such difficulties are here referred to as "missing data", and are typical in situations where off-lattice structures are present (open boundary conditions, curved solid wall boundaries, grid refinement etc.). It is common to these problems that some populations of the links at certain nodes are not available. It is best to illustrate this with an example.

A typical problem of this kind is the specification of the outlet boundary condition in duct-like flows with large aspect ratio. Such flows are most common in engineering and medical applications such as wind tunnels, blood vesicles etc. In Fig. 6, we show a situation at the outlet node. Since there are no lattice nodes beyond the outlet, populations of the three discrete velocities pointing into the fluid are not known and need to be fixed by additional considerations. At present, there is no established way to cope with this problem

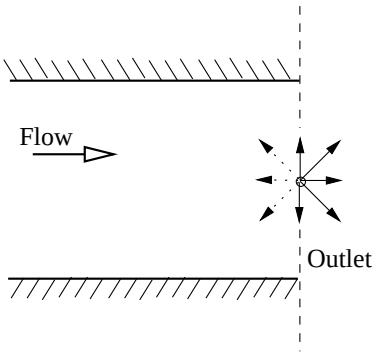



Fig. 6: Situation at the outlet node. Data are missing for the populations of the velocities pointing into the fluid (dash).

[14]. Because specification of pressure at the outlet has no significance in long pipes, one relies on interpolation schemes (see, e.g. [40]). Interpolation often becomes a major source of inaccuracy in the simulations.

We shall use Grad's approximations for the populations (23) in order to extrapolate the missing populations. Namely, we impose the outlet condition (see Fig. 6) by the following rule: At time step $n+1$, the populations of the links at the outlet pointing into the fluid are assigned the values (23), whereas the values of the moments $\rho$, $\mathbf{j}$, and $P_{\alpha\beta}$ are taken from the previous time step $n$ at the same nodes. Initially, all links are at equilibrium.

The three-dimensional backwards-facing step flow was used to validate the outlet boundary condition. The standard $D3Q15$ lattice Boltzmann model with the polynomial equilibrium [32] was used. Geometry of the setup was chosen to model the experiment of Armaly et al [9]: The channel length ($X$) was $20S$, where $S$ is the backwards facing step height, the channel width ($Y$) was $2S$. The step heights was $S = 10$ (lattice units), the step length was $2S$. The ratio of the span width ($Z$) to the step height was equal $36:1$ (that is, the span width was $36S$ lattice units). The total number of grid points was about $1.5 \times 10^6$. Kinetic boundary conditions [4] were applied on the wall nodes. The inflow was a fully developed velocity profile in a duct flow (simulated separately in the duct with the dimension $15S \times S \times 36S$). The inflow velocity maximum ranged between $10^{-2}$ to $4 \times 10^{-2}$ while the kinematic viscosity was fixed at $\nu = 10^{-3}$. The outlet condition (23) was applied both in the backwards-facing step channel and in the the auxiliary duct simulations. All simulations were done on a single-processor facility (PC) till steady state was reached in whole domain, a single run time ranged between one to several hours depending on the Reynolds number, $\mathrm{Re} = (2US)/\nu$, where $U$ is cross-section averaged inlet velocity.

Before reporting the results, it should be pointed out that the same three-dimensional lattice Boltzmann model with the outlet boundary conditions

based on a simple second-order interpolation formula for the missing populations (see, e. g. [40]) failed at Reynolds number Re $< 50$. The reason for such a poor performance is the errors which start at the outlet and propagate upstream.

The range of Reynolds number covered in our simulation with the new outlet was $100 <$ Re $< 392$. In Fig. 7, snapshots of the velocity on the mid-plane at Re $= 270$ are shown in the full computation domain, including the outlet. It is visible in Fig. 7 that the velocity profile stays smooth during the whole simulation. In Fig. 8, the primary flow reattachment length (the distance at which the velocity field on the bottom wall becomes directed towards the outlet) is compared with the results of the simulations of the incompressible Navier-Stokes equation by various numerical techniques [24, 28], with the recent two-dimensional lattice Boltzmann simulation on a non-uniform grid [40], as well as with the experimental data of Armaly et al [9], and was found to be in excellent agreement. We stress that the accuracy and stability achieved with the new outlet boundary condition allowed us to use a small step size of only ten grid points, much less than it would be required with different boundary conditions.

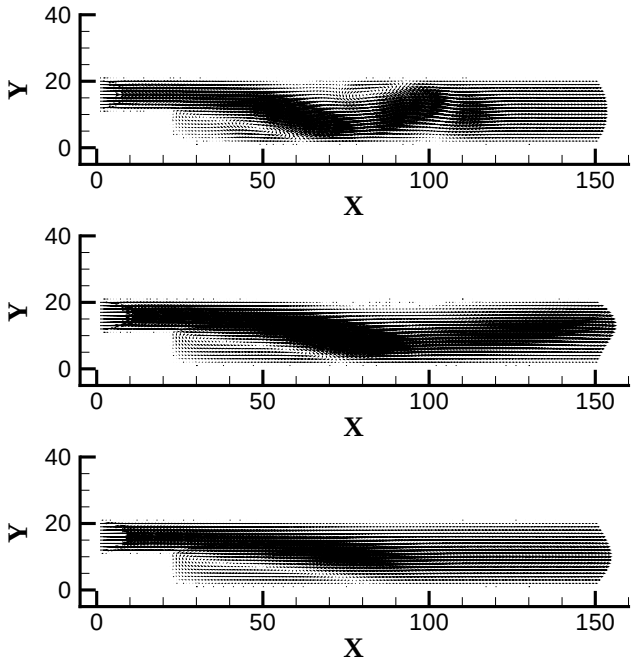



Fig. 7: Snapshots of the velocity field on the mid-plane at Re $= 270$ at $9 \times 10^3$, $18 \times 10^3$, and $40 \times 10^3$ time steps in lattice units (from top to bottom).

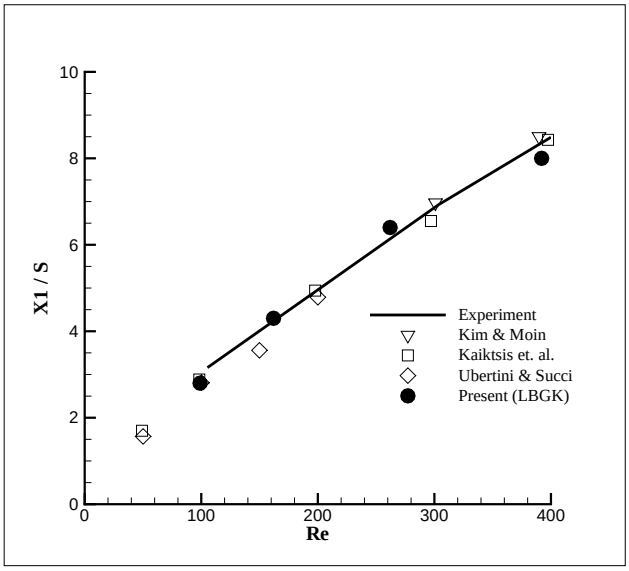



Fig. 8: Primary reattachment length $X_1$ normalized by the step height $S$. Comparison of the present simulation with the experiment of Armaly et al [9], and simulations of Kaiktsis et al [24], Kim and Moin [28], and Ubertini and Succi [40].

## 8 Discussion

We considered a specific example of a kinetic model in order to compare various theories of non-equilibrium thermodynamics in a nontrivial flow situation. Our major finding is that the minimal kinetic model can be quantified as a superset of the Grad's moment systems, and hence the populations stay close to the low-dimensional manifolds described by discrete-time Grad's populations (23). This is at variance with a viewpoint that lattice Boltzmann method is a superset of just the Navier-Stokes equations. For the case of a driven cavity flow, different closure approximations were tested against the direct simulation data. Grad's closure for the minimal model was found to perform better than the Navier-Stokes approximation in the whole range of Knudsen number. Thanks to its simplicity, the Grad approximation *within* the lattice Boltzmann models can be used in the situations where a part of the information about populations is missing in order to reconstruct the unavailable data. Such situations are quite frequent in lattice Boltzmann simulations, for example, in the case of in- and outflow boundary conditions, grid refinement etc. The fact that Grad's sub-manifold contains almost all of the dynamics can be used then in order to extrapolate populations on the missing links of the lattices via the explicit formula (23).

Whereas we have explored two classical closures of kinetic theory, we conclude this paper with a validation of yet another closure based on a spectral

decomposition. To that end, the ELBM code was coupled with ARPACK [30] in order to compute the leading eigenvalues and the corresponding eigenvectors of the Jacobian field of the corresponding map at the steady state. In all cases, the eigenvalues are within the unit circle (Fig. 9(a)). The leading eigenvalue is always equal to one (reflecting mass conservation), and the corresponding eigenvector captures most of the structure of the steady state. As the Knudsen number decreases, eigenvalues tend to get clustered close to the unit circle. This happens because when the Knudsen number is small the incompressibility assumption is a good approximation, and mass is also conserved locally. The very close similarity between Fig. 9(b) and Fig. 4(a), reveals that states perturbed away from the steady state along the leading eigenvector are also described well by the Navier-Stokes closure.

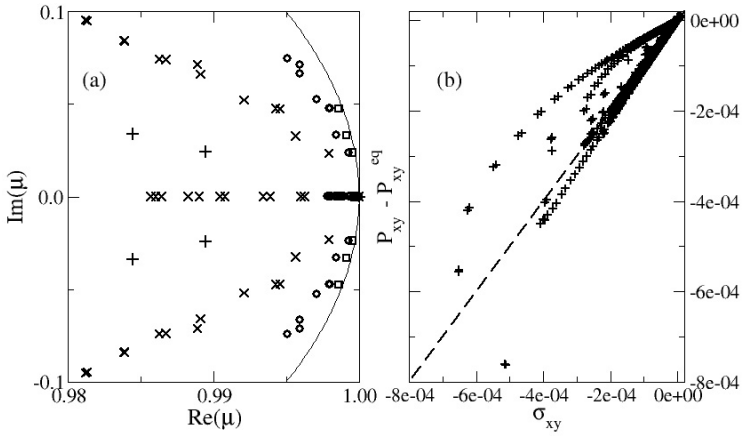



Fig. 9: (a) Leading eigenvalues of the minimal kinetic model at steady state (square: Kn=$10^{-4}$, circle: Kn=$10^{-3}$, X: Kn=$10^{-2}$, +: Kn=$10^{-1}$); (b) Scatter plot as in Fig. 4(a) for a state perturbed away from the steady state along the leading eigenvector (Kn=$10^{-3}$).

*Acknowledgement.* Discussions with A.N. Gorban are gratefully acknowledged. The work of I.V.K. was partially supported by the Swiss Federal Office of Energy (BFE) under the project No. 100862. The work of S.S.C. was supported by the ETH grant 0-20280-05. The work of I.G.K. was partially supported by an NSF-ITR grant and by DOE.

## References


1. S. Ansumali: *Minimal Kinetic Modeling of Hydrodynamics*. PhD thesis, ETH Zurich, (2004)
2. S. Ansumali, S.S. Chikatamarla, C.M. Frouzakis, K. Boulouchos: Entropic Lattice Boltzmann Simulation of the Flow Past Square Cylinder. *Int. J. Mod. Phys. C* **15** (3), 435–445 (2004)
3. S Ansumali, I.V. Karlin: Entropy Function Approach to the Lattice Boltzmann Method. *J. Stat. Phys.* **107** (1-2), 291–308 (2002)
4. S. Ansumali, I.V. Karlin: Kinetic Boundary Condition for the Lattice Boltzmann Method. *Phys. Rev. E* **66** (2), 026311 (2002)
5. S. Ansumali, I.V. Karlin: Consistent Lattice Boltzmann Method. *Phys. Rev. Lett.* **95**, 260605 (2005)
6. S. Ansumali, I.V. Karlin, C.E. Frouzakis, K.B. Boulouchos: Entropic Lattice Boltzmann Method for Microflows. *Physica A* **359**, 289–305 (2006)
7. S. Ansumali, I.V. Karlin, H.C. Öttinger: Minimal Entropic Kinetic Models for Simulating Hydrodynamics. *Europhys. Lett.* **63** (6), 798–804 (2003)
8. S. Ansumali, I.V. Karlin, H.C. Öttinger: Thermodynamic Theory of Incompressible Hydrodynamics. *Phys. Rev. Lett.* **94**, 080602 (2005)
9. B.F. Armaly, F. Durst, J.C.F. Pereira, B. Schonung: Experimental and Theoretical Investigation of Backwardfacing Flow. *J. Fluid Mech.* **127**, 473–496 (1983)
10. A. Beskok, G.E. Karniadakis: *Microflows: Fundamentals and Simulation* (Springer, Berlin 2001)
11. B.M. Boghosian, J. Yepez, P.V. Coveney, A.J. Wagner: Entropic Lattice Boltzmann Methods. *Proc. Roy. Soc. Lond.* **457**, 717–766 (2001)
12. S. Chapman, T.G. Cowling: *The Mathematical Theory of Non-Uniform Gases* (Cambridge University Press, Cambridge 1970)
13. H. Chen, S. Chen, W. Matthaeus: Recovery of the Navier-Stokes Equation Using a Lattice-Gas Boltzmann method. *Phys. Rev. A* **45**, R5339–R5342 (1992)
14. H. Chen, S. Kandasamy, S. Orszag, R. Shock, S. Succi, V. Yakhot: Extended-Boltzmann Kinetic Equation for Turbulent Flows. *Science* **301**, 633–636 (2003)
15. S. Chen, G.D. Doolen: Lattice Boltzmann Method for Fluid Flows. *Annu. Rev. Fluid Mech.* **30**, 329 (1998)
16. S.S. Chikatamarla, S.Ansumali, I.V. Karlin: Grad's Approximation for Missing Data in Lattice Boltzmann Simulations. *Europhys. Lett.*, in press, 2006
17. A.N. Gorban, I.V. Karlin: Method of Invariant Manifolds and Regularization of Acoustic Spectra. *Transport Theory Stat. Phys.* **23**, 559–632 (1994)
18. A.N. Gorban, I.V. Karlin: Short-wave Limit of Hydrodynamics: A Soluble Example. *Phys. Rev. Lett.* **77**, 282–285 (1996)
19. A.N. Gorban, I.V. Karlin: *Invariant Manifolds for Physical and Chemical Kinetics*, vol. 660 in Lect. Notes Phys. (Springer, Berlin Heidelberg 2005)
20. H. Grad: On the Kinetic Theory of Rarefied Gases. *Comm. Pure Appl. Math.* **2**, 331–407 (1949)
21. X. He, S. Chen, G.D. Doolen: A Novel Thermal Model for the Lattice Boltzmann Method in Incompressible Limit. *J. Comput. Phys.* **146** (1), 282–300 (1998)
22. F. Higuera, S. Succi, R. Benzi: Lattice Gas-Dynamics with Enhanced Collisions. *Europhys. Lett.* **9**, 345–349 (1989)
23. J.-Z. Jiang, J. Fan, C. Shen: Statistical Simulation of Micro-cavity Flows. *23rd Int. Symposium on Rarefied Gas Dynamics*, pages 784–790, (2003)

24. L.K. Kaiktsis, G.E. Karniadakis, S.A. Orszag: Onset of Three-dimensionality, Equilibria, and Early Transition in Flow Over a Backward-facing Step. *J. Fluid Mech.* **231**, 501 (1991)
25. I.V. Karlin, A. Ferrante, H.C. Öttinger: Perfect Entropy Functions of the Lattice Boltzmann Method. *Europhys. Lett.* **47**, 182–188 (1999)
26. I.V. Karlin, A. Gorban, S. Succi, V. Boffi: Maximum Entropy Principle for Lattice Kinetic Equations. *Phys. Rev. Lett.* **81**, 6–9 (1998)
27. I.G. Kevrekidis, C.W. Gear, J.M. Hyman, P.G. Kevrekidis, O. Runborg, C. Theodoropoulos: Equation-free: Coarse-grained Multiscale Computation Enabling Microscopic Simulators to Perform System-level Analysis. *Comm. Math. Sci.* **1**, 715–762 (2003)
28. J. Kim, P. Moin: Application of a Fractional-step Method to Incompressible Navier-Stokes Equations. *J. Comput. Phys.* **59**, 308 (1985)
29. T. Lee, C.-L. Lin: Rarefaction and Compressibility Effects of the Lattice-boltzmann-equation Method in a Gas Microchannel. *Phys. Rev. E* **71**, 046706 (2005)
30. R.B Lehoucq, D.C Sorensen, C. Yang: *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods.* (SIAM 1998)
31. X. D. Niu, C. Shu, Y.T. Chew: Lattice Boltzmann BGK Model for Simulation of Micro Flows. *Europhys. Lett.* **67**, 600–606 (2003)
32. Y.H. Qian, D. d'Humieres, P. Lallemand: Lattice BGK Models for Navier-Stokes Equation. *Europhys. Lett.* **17**, 479–484 (1992)
33. X. Shan, X. He: Discretization of the Velocity Space in the Solution of the Boltzmann Equation. *Phys. Rev. Lett.* **80**, 65 (1998)
34. M. Slemrod: Renormalization of the Chapman-Enskog Expansion: Isothermal Fluid Flow and Rosenau Saturation. *J. Stat. Phys.* **91**, 285–305 (1998)
35. S. Succi, R Benzi, M. Vergassola: The Lattice Boltzmann-Equation –Theory and Applications. *Phys. Rep.* **222**, 145–197 (1992)
36. S Succi, I. V. Karlin, H. Chen: Role of the $H$ theorem in Lattice Boltzmann Hydrodynamics. *Rev. Mod. Phys.* **74**, 1203 (2002)
37. S. Succi, M. Sbragaglia: Analytical Calculation of Slip Flow in Lattice Boltzmann Models with Kinetic Boundary Conditions. *Phys. Fluids* **17** (9), 093602 (2005)
38. K. Theodoropoulos, Y.-H. Qian, I.G. Kevrekidis: "Coarse" Stability and Bifurcation Analysis Using Timesteppers: a Reaction-diffusion Example. *Proc. Natl. Acad. Sci.* **97** (18), 9840–9843 (2000)
39. F. Toschi, S. Succi: Lattice Boltzmann Method at Finite Knudsen Numbers. *Europhys. Lett.* **69**, 549–555 (2005)
40. S. Ubertini, S. Succi: Recent Advances of Lattice Boltzmann Techniques on Unstructured Grids. *Progress in Computational Fluid Dynamics* **5**, 85–96 (2005)
41. Y. Zhang, R. Qin, D.R. Emerson: Lattice Boltzmann Simulation of Rarefied Gas Flows in Microchannels. *Phys. Rev. E* **71**, 047702 (2005)

# Numerical and Analytical Spatial Coupling of a Lattice Boltzmann Model and a Partial Differential Equation

P. Van Leemput, W. Vanroose, and D. Roose

Department of Computer Science, Katholieke Universiteit Leuven,
Celestijnenlaan 200A, B-3001 Heverlee, Belgium,
{pieter.vanleemput, wim.vanroose, dirk.roose}@cs.kuleuven.be

**Summary.** This article is concerned with the spatial coupling of a lattice Boltzmann model (LBM) and the finite difference discretization of the corresponding partial differential equation (PDE). At the interface, we have a one-to-many problem since the macroscopic PDE variables have to be mapped to more LBM variables. We show how this mapping can be done either analytically, using results from the Chapman-Enskog expansion or numerically, using a fixed point iterative scheme. The results are illustrated for different diffusive systems on a one-dimensional domain.

## 1 Introduction

A dynamical system can be described by various models, each operating on a different level of abstraction. On the macroscopic level, there are partial differential equations (PDEs) that describe the system's evolution in terms of a few macroscopic variables, like density, velocity, etc. On a finer level, there are mesoscopic or pseudo particle models, like lattice Boltzmann models (LBMs) that use idealized particle distribution functions on a regular grid to describe the system. On the truly microscopic level, one has molecular dynamics and kinetic Monte Carlo methods that model the interactions between particles individually.

The choice for a particular model depends on several criteria. Macroscopic-level models, like PDEs, typically have a small dimensional state space and in general allow large time steps during simulation. However, they often fail to describe the dynamics of complex systems. Mesoscopic models like LBMs on the other hand allow the incorporation of complex physics in a more bottom-up way than macroscopic models but typically require more variables and smaller time steps. Furthermore, they can treat irregular domain boundaries in a natural way. Similar advantages apply to microscopic models, but simulation with these models can be very expensive and often becomes prohibitive. Finally, when modeling a system with a higher level model fails because it can

not be written in terms of variables at that particular level of abstraction only, i.e. when the higher level model does not close, a lower level model describing the same physics in more detail should be used.

Sometimes, the level of detail required to model a physical system changes from region to region and different models have to be used on different parts of the domain. At the interface between the models, there will be a mismatch in the kind (and number) of variables used by the different models. There, the variables have to mapped to one another. Many such *hybrid* models, which couple a microscopic particle method to a macroscopic continuum method, have already been well developed, see e.g. [2, 6, 10, 11, 12, 15] and references therein.

In this article, we will spatially couple a LBM and a PDE model describing the same diffusive system in different regions of space on a one-dimensional domain. The PDE is discretized using finite differences and has the particle density as the sole macroscopic variable. For this setup, the corresponding LBM has three times as much variables (the particle distribution functions). Since there are more LBM than PDE variables, we have a one-to-many problem at the interface where we have to map densities to distribution functions. The inverse mapping of distribution functions to densities is straightforward because density is defined as the sum of the distribution functions.

Albuquerque *et al.* [1] used the Chapman-Enskog expansion to write the missing distribution functions at the interface as a functional of the density variable only. We will use the same concept but a different implementation. For cases where these functionals are not available or difficult to obtain analytically, the constrained runs scheme developed by Gear and Kevrekidis [8, 7] can be used to obtain these functionals numerically. This scheme performs a series of short microscopic (here LBM) simulations and resets the lowest order velocity moment (density) to its initial value while leaving the higher order moments unchanged. Van Leemput *et al.* showed in [17] that the application of the scheme to the LBM discussed here, produces a numerical approximation of the Chapman-Enskog relations that is correct up to first order.

The work described in this article is a step in the development of efficient methods for the coupling of LBM and PDE models. In the discussion, we have made some simplifying assumptions, e.g. we used the same time step and grid spacing for both the PDE and LBM. Taking different time steps can further optimize the methods presented here.

This article is organized as follows. In Sect. 2, we discuss different LBMs and the corresponding PDEs. Section 3 describes the constrained runs scheme applied to diffusive LBM. The issues concerning the coupling of the LBM and PDE are discussed in Sect. 4. In Sect. 5 we present numerical results on a) the FitzHugh-Nagumo reaction-diffusion system, b) a pure diffusion example and c) a growth-diffusion example where the LBM reaction term depends on both the density and the velocities of the particles. Section 6 summarizes the main conclusions.

## 2 Models for One-Dimensional Diffusive Systems

In Sect. 2.1, we describe the finite difference discretization of the partial differential equation for one-dimensional reaction-diffusion systems. In Sect. 2.2, we describe the corresponding lattice Boltzmann BGK model with a reaction term depending on density only. Section 2.3 shows that both models are equivalent when the macroscopic solution is smooth. In Sect. 2.4, we briefly describe a lattice Boltzmann BGK model with a velocity dependent reaction term to simulate growth-diffusion systems. The corresponding PDE is also given.

### 2.1 Partial Differential Equation (PDE)

In a one-dimensional reaction-diffusion system, the partial differential equation (PDE) describing the evolution of the particle density (concentration) $\rho(x,t)$ as a function of space $x$ and time $t$ is given by

$$\frac{\partial \rho(x,t)}{\partial t} = D\frac{\partial^2 \rho(x,t)}{\partial x^2} + F(\rho(x,t)) \tag{1}$$

where $D$ is the diffusion coefficient and $F(\rho(x,t))$ a macroscopic reaction force term which depends on $\rho(x,t)$ only.

To find a solution of (1), the equation is discretized using finite differences (forward difference in time and central difference in space) to obtain

$$\begin{aligned}\rho(x,t+\Delta t) = \rho(x,t) + \frac{\Delta t\, D}{\Delta x^2}(\rho(x+\Delta x,t) - 2\rho(x,t) + \rho(x-\Delta x,t)) \\ + \Delta t\, F(\rho(x,t))\end{aligned} \tag{2}$$

with $\Delta x$ and $\Delta t$ the corresponding space and time steps.

### 2.2 Lattice Boltzmann Model (LBM)

Boltzmann models describe the evolution of a distribution function $f(x,v,t)$ that represents the number of particles that move with certain velocity $v$ at position $x$ and time $t$. Lattice Boltzmann models (LBM) [3, 13] use discretized distribution functions $f_i(x,t)$ with velocity $v_i$ that are defined on a space-time lattice with grid spacing $\Delta x$ in space and $\Delta t$ in time. On a one-dimensional domain, only three values are considered for the velocity (D1Q3 model):

$$v_i = c_i\frac{\Delta x}{\Delta t}, \quad c_i = i \in \{-1,0,1\} \tag{3}$$

with $c_i$ the dimensionless lattice velocity.

The lattice Boltzmann evolution law for the distribution functions is

$$f_i(x + c_i\Delta x, t+\Delta t) = (1-\omega)f_i(x,t) + \omega f_i^{eq}(x,t) + R_i(x,t) \tag{4}$$

for $i \in \{-1, 0, 1\}$. The right hand side of (4) updates the values $f_i(x,t)$ to *post-collision* values $f_i^\star(x,t^\star)$ (with $t < t^\star < t + \Delta t$). Afterwards, these values propagate to a neighboring lattice site according to their velocity direction (left hand side of (4)). Diffusive collisions are modeled by the Bhatnagar-Gross-Krook (BGK) collision term $-\omega(f_i(x,t) - f_i^{eq}(x,t))$ in (4) as a relaxation to a *local diffusive* equilibrium

$$f_i^{eq}(x,t) = \frac{1}{3}\,\rho(x,t). \tag{5}$$

The BGK relaxation coefficient $\omega$ in (4) will be defined in Sect. 2.3. Reactions are modelled by the term $R_i(x,t)$ in (4) as [14, 5]

$$R_i(x,t) = \frac{\Delta t}{3}\,F(\rho(x,t)) \tag{6}$$

with $F(\rho(x,t))$ defined in (1). Here, it is assumed that reactions occur at the local diffusive equilibrium [4].

The particle density $\rho(x,t)$, i.e. the macroscopic variable (cf. (1)), is defined as the zeroth order velocity moment of the distribution functions

$$\rho(x,t) = \sum_{i=-1}^{1} f_i(x,t) = \sum_{i=-1}^{1} f_i^{eq}(x,t), \tag{7}$$

where the second equality expresses that the BGK diffusive collisions locally conserve density (compare (5)).

In a similar way, we define the dimensionless first and second order velocity moments (up to the factor 1/2 for the second order moment) as

$$\phi(x,t) = \sum_{i=-1}^{1} c_i\, f_i(x,t) \qquad \xi(x,t) = \frac{1}{2}\sum_{i=-1}^{1} c_i^2\, f_i(x,t) \tag{8}$$

We will refer to these moments as the "momentum" $\phi$ and (kinetic) "energy" $\xi$ (although these are non-conserved quantities in a diffusive system). The state of the LBM at time $t$ and position $x$ is then fully determined by either the distribution functions $\mathbf{f} = [f_{-1}\ f_0\ f_1]'$ or the moments $\boldsymbol{\varrho} = [\rho\ \phi\ \xi]'$. By definition,

$$\begin{bmatrix} \rho \\ \phi \\ \xi \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} f_{-1} \\ f_0 \\ f_1 \end{bmatrix} \quad \Leftrightarrow \quad \boldsymbol{\varrho} = M\,\mathbf{f} \tag{9}$$

and vice versa $\mathbf{f} = M^{-1}\boldsymbol{\varrho}$ (one-to-one relationship).

### 2.3 Relations between LBM and PDE

When the solution of the LBM varies slowly on a macroscopic length and time scale, we can show that the LBM from Sect. 2.2 reduces to the PDE

introduced in Sect. 2.1 using a multiscale Chapman-Enskog expansion [3]. Under this condition, both models describe the same macroscopic behavior. To this end, we define a small tracer parameter $\epsilon$ and the scaling $x_\epsilon = \epsilon x$, $t_\epsilon = \epsilon^2 t$ such that

$$\frac{\partial}{\partial x} = \epsilon \frac{\partial}{\partial x_\epsilon} \quad \text{and} \quad \frac{\partial}{\partial t} = \epsilon^2 \frac{\partial}{\partial t_\epsilon}. \tag{10}$$

For reaction-diffusion problems, we further assume that the reaction term $R_i(x,t)$ in (4) is of second order, i.e. $R_i = \epsilon^2 R_{i,\epsilon}$ [3, 5], which explains the choice in (6).

A second order Taylor expansion of the term $f_i(x + c_i \Delta x, t + \Delta t)$ in (4) around $f_i(x,t)$ leads to

$$\begin{aligned} c_i \Delta x \frac{\partial f_i}{\partial x} + \Delta t \frac{\partial f_i}{\partial t} + \frac{c_i^2 \Delta x^2}{2} \frac{\partial^2 f_i}{\partial x^2} + c_i \Delta x \Delta t \frac{\partial^2 f_i}{\partial x \partial t} + \frac{\Delta t^2}{2} \frac{\partial^2 f_i}{\partial t^2} \\ = -\omega (f_i - f_i^{eq}) + R_i \end{aligned} \tag{11}$$

Introducing the tracer scaling (10) and dropping the subscript $\epsilon$ notation, we obtain

$$\begin{aligned} \epsilon c_i \Delta x \frac{\partial f_i}{\partial x} + \epsilon^2 \Delta t \frac{\partial f_i}{\partial t} + \epsilon^2 \frac{c_i^2 \Delta x^2}{2} \frac{\partial^2 f_i}{\partial x^2} + \epsilon^3 c_i \Delta x \Delta t \frac{\partial^2 f_i}{\partial x \partial t} + \epsilon^4 \frac{\Delta t^2}{2} \frac{\partial^2 f_i}{\partial t^2} \\ = -\omega (f_i - f_i^{eq}) + \epsilon^2 R_i \end{aligned} \tag{12}$$

The distribution functions $f_i(x,t)$ are expanded in terms of increasingly higher order contributions $f_i^{[0]}$, $f_i^{[1]}$, ... as follows

$$f_i = f_i^{[0]} + \epsilon f_i^{[1]} + \epsilon^2 f_i^{[2]} + \dots \tag{13}$$

We substitute (13) in (12) and keep only terms up to second order to obtain

$$\begin{aligned} \epsilon c_i \Delta x \frac{\partial f_i^{[0]}}{\partial x} + \epsilon^2 c_i \Delta x \frac{\partial f_i^{[1]}}{\partial x} + \epsilon^2 \Delta t \frac{\partial f_i^{[0]}}{\partial t} + \epsilon^2 \frac{c_i^2 \Delta x^2}{2} \frac{\partial^2 f_i^{[0]}}{\partial x^2} \\ = -\omega (f_i^{[0]} + \epsilon f_i^{[1]} + \epsilon^2 f_i^{[2]} - f_i^{eq}) + \epsilon^2 R_i \end{aligned} \tag{14}$$

This equation should hold for each order separately. Equating the zeroth order terms leads to

$$f_i^{[0]} = f_i^{eq} = \frac{1}{3}\rho \tag{15}$$

The part of order $\epsilon$ leads to the following expression for the first order correction $f_i^{[1]}(x,t)$

$$f_i^{[1]} = -\frac{c_i \Delta x}{\omega} \frac{\partial f_i^{[0]}}{\partial x} = -\frac{c_i \Delta x}{3\omega} \frac{\partial \rho}{\partial x}, \tag{16}$$

Gathering the terms of order $\epsilon^2$ in (14), we have

$$c_i \Delta x \frac{\partial f_i^{[1]}}{\partial x} + \frac{c_i^2 \Delta x^2}{2} \frac{\partial^2 f_i^{[0]}}{\partial x^2} + \Delta t \frac{\partial f_i^{[0]}}{\partial t} = -\omega f_i^{[2]} + R_i \tag{17}$$

Substitution of (15), (16) and (6) results in the following expression for the second order contribution $f_i^{[2]}(x,t)$

$$f_i^{[2]} = -\frac{c_i^2 \Delta x^2}{6\omega^2}(\omega - 2)\frac{\partial^2 \rho}{\partial x^2} + \frac{\Delta t}{3\omega}\left(F(\rho) - \frac{\partial \rho}{\partial t}\right) \tag{18}$$

When we sum (18) over all velocities, we obtain

$$\frac{\partial \rho}{\partial t} = -\frac{\Delta x^2}{3\omega \Delta t}(\omega - 2)\frac{\partial^2 \rho}{\partial x^2} + F(\rho) \tag{19}$$

where we used the fact that $\sum_i f_i^{[2]} = 0$ (sum up (13) and use (7)). Comparing (19) to (1), we obtain the relation between $D$ and $\omega$ (cf. [14])

$$\omega = \frac{2}{1 + 3\,D\frac{\Delta t}{\Delta x^2}}\ . \tag{20}$$

Expansion (13) together with (15), (16) and (18) can be used to represent the state of the LBM system. Note that, since we dropped the index $\epsilon$ notation in the derivation, the actual macroscopic derivatives are obtained by combining (10) with (13). Using (9), the equivalent higher order moments $\phi(x,t)$ and $\xi(x,t)$ can be computed as functionals of the density $\rho(x,t)$ only

$$\begin{aligned} \phi &= -\frac{2\Delta x}{3\omega}\frac{\partial \rho}{\partial x} + O(\epsilon^3), \\ \xi &= \frac{\rho}{3} - \frac{\Delta t}{6\omega}\left(F(\rho) - \frac{\partial \rho}{\partial t}\right) + O(\epsilon^4)\ . \end{aligned} \tag{21}$$

These functionals, and by extension (15), (16) and (18), are called *slaving relations*.

Note that, using the PDE (19) itself, (18) can be rewritten as

$$f_i^{[2]} = -\frac{\Delta t}{6\omega}(3c_i^2 - 2)\left(F(\rho) - \frac{\partial \rho}{\partial t}\right) = -\frac{\Delta x^2}{18\omega^2}(\omega - 2)(3c_i^2 - 2)\frac{\partial^2 \rho}{\partial x^2}\ . \tag{22}$$

### 2.4 Velocity Dependent Mesoscopic Reactions: Growth-Diffusion

Many microscopic systems have reaction rates that depend on the velocities of the colliding particles. One example is the ionization reaction that appears in electron transport through a molecular gas. During transport, the electrons collide with the molecules and transfer part of their energy. Fast electrons slow down by kicking additional electrons out of the molecule during reactive collisions. Slow electrons collide elastically and only change their direction. This can only be modeled by a velocity dependent reaction term.

The LBM (4) in Sect. 2.2 is not suitable to describe such system because the reaction term (6) depends on the density only. However, the classical

lattice Boltzmann BGK equation can be extended to include a more general velocity dependent reaction term

$$f_i(x+c_i\Delta x, t+\Delta t)-f_i(x,t)=-\omega(f_i(x,t)-f_i^{eq}(x,t))+\Delta t\sum_j A_{ij}f_j(x,t). \quad (23)$$

At each time step, the rate $A_{ij}$ denotes the chance that either a particle with speed $c_j$ ends up with speed $c_i$ or that a particle has been created or destroyed during the reaction. The microscopic velocities are again crudely discretized. As in Sect. 2.2, we implemented the D1Q3 model (3) on a one-dimensional domain.

In this article, we discuss a limited class of models with velocity dependent reaction rates. We specifically look at a problem that gives rise to a growth-diffusion PDE on a macroscopic scale when the solution of (23) is slowly varying. As described in [18], this reduced model can be derived through a Chapman-Enskog expansion similar to the one outlined in Sect. 2.2, and is given by

$$\frac{\partial\rho(x,t)}{\partial t} = D\frac{\partial^2\rho(x,t)}{\partial x^2} + \alpha\rho(x,t)\ , \quad (24)$$

where both $D$ and the growth rate $\alpha$ depend on the microscopic reaction rates $A_{ij}$ and relaxation parameter $\omega$. As in Sect. 2.1, (24) can be discretized using finite differences.

## 3 Constrained Runs Scheme

Given only the density values on the domain, the full state of the LBM can be represented by the slaving relations (15), (16) and (18) as described in Sect. 2.3. Assuming that such relations are unavailable or difficult to obtain analytically, the *constrained runs scheme* [8, 7] can be used to approximate these relations numerically.

The application of the constrained runs scheme to the LBM from Sect. 2.2 or Sect. 2.4 is outlined in Algorithm 1. Given the initial density profile $\rho^{(0)}$, an initial guess for $f_i(x,t)$ is computed using e.g. (5). The LBM is then repeatedly used to evolve the state for a short time $\tau$. After each such simulation the transformation (9) is used to reset the lowest moment of the distribution functions to the initial density profile.

The constrained runs scheme can be defined as a map

$$\boldsymbol{\varrho}^{(k+1)} = \mathcal{C}_\tau(\boldsymbol{\varrho}^{(k)})\ ; \quad k=0,1,2,\ldots,K \quad (25)$$

on the state vector $\boldsymbol{\varrho}^{(k)} = [\rho^{(0)}\ \phi^{(k)}\ \xi^{(k)}]'$; with $k$ the iteration number and $\tau$ the simulation time of the inner microscopic model, here the LBM. Since $\rho^{(k+1)}$ is reset to $\rho^{(0)}$ after each step, the map effectively iterates on the higher order moments $\phi$ and $\xi$ to obtain a fixed point.

**Algorithm 1** Constrained runs scheme for a one-dimensional diffusive LBM

**Required:** $\rho^{(0)} = \rho(x,0)$
  $f_i^{(0)} = w_i\,\rho^{(0)}$ ; $\sum_{i=-1}^{1} w_i = 1$, e.g. $w_i = 1/3$ — Choose $f_i^{(0)}$ s.t. (7) holds
  **repeat**
    $\mathbf{f}^{(k+1)} = \mathrm{LBM}(\mathbf{f}^{(k)})$ — LBM simulation (4) over time $\tau$
    $\boldsymbol{\varrho}^{(k+1)} = M\mathbf{f}^{(k+1)}$ — Corresponding $\phi^{(k+1)}$ and $\xi^{(k+1)}$ (9)
    $\rho^{(k+1)} = \rho^{(0)}$ — Reset macroscopic variables
    $\mathbf{f}^{(k+1)} = M^{-1}\boldsymbol{\varrho}^{(k+1)}$ — Map back (9)
  **until** convergence heuristic $< \theta$ , with $\theta \ll 1$

A straightforward choice for the convergence heuristic in Algorithm 1 is

$$||\phi^{(k+1)}(x) - \phi^{(k)}(x)||_2 < \theta \quad \text{and} \quad ||\xi^{(k+1)}(x) - \xi^{(k)}(x)||_2 < \theta \tag{26}$$

with $\theta$ a user-defined tolerance ($\theta \ll 1$). Figure 1 sketches the evolution of the procedure.

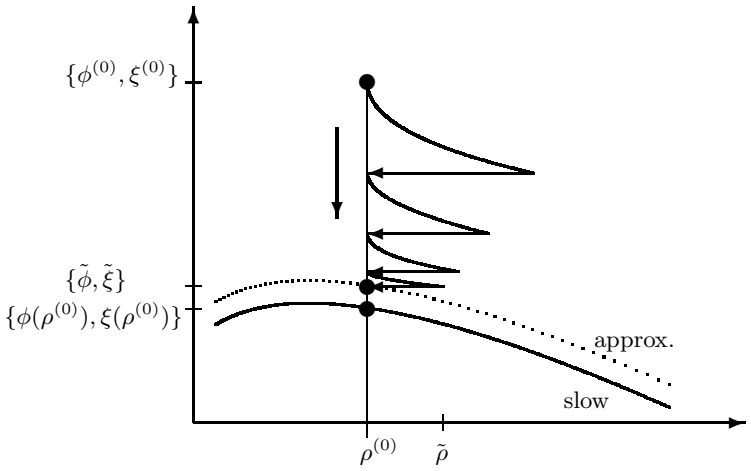



Fig. 1: Sketch of the evolution of the constrained runs scheme. The higher order moments $\phi$ and $\xi$ are plotted with respect to the macroscopic variable $\rho$. The density $\rho$ is reset to the given $\rho^{(0)}$ after each LBM simulation. The constrained runs scheme iterates towards a fixed point $\{\tilde{\phi}, \tilde{\xi}\}$ that is an approximation to the slaved state $\{\phi(\rho^{(0)}), \xi(\rho^{(0)})\}$ (21). The fixed point lies on an "approximate" manifold while the exact solution lies on the slow manifold described by $\rho$. The value $\tilde{\rho}$ is the density corresponding to $\{\tilde{\phi}, \tilde{\xi}\}$ (before the final reset to $\rho^{(0)}$) and will be useful as an estimate for the error.

In [17] we analyzed the application of the constrained runs scheme for the initialization of the LBM for one-dimensional reaction-diffusion problems (see Sect. 2.2). We have proven that the scheme is unconditionally stable and convergent. Below, we restate the main theorems.

**Theorem 1 (Stability theorem).** *The constrained runs scheme for the lattice Boltzmann BGK model that describes a one-dimensional reaction-diffusion system with either periodic, no-flux or Dirichlet boundary conditions (and with the reaction term depending on $\rho$ only* (6)*), is unconditionally stable.*

*Proof (Outline).* The eigenvalues $\mu$ of the linearization (the Jacobian matrix) of one step of the fixed point iterator (25) determine the stability of Algorithm 1. If all eigenvalues in the spectrum $\sigma(\mathcal{C}_\tau)$ lie within the unit circle, i.e. $\forall\, \mu \in \sigma(\mathcal{C}_\tau) : |\mu| < 1$, the iteration is stable. For the three types of boundary conditions considered, we prove in [17] that these eigenvalues lie on a circle centered at the origin with radius $|1 - \omega|$. The constrained runs iteration is unconditionally stable because $0 < \omega < 2$ [13] (compare (20)) and thus always $|\mu| = |1 - \omega| < 1$.

**Theorem 2 (Asymptotic convergence factor).** *As a corollary of the above proof, the asymptotic convergence factor $\eta := \max\{|\mu| : \forall \mu \in \sigma(\mathcal{C}_\tau)\}$ is equal to $|1 - \omega|$.*

**Theorem 3 (Convergence theorem).** *The constrained runs algorithm for the LBM described in Theorem 1 converges to a first order correct approximation $\{\tilde{\phi}, \tilde{\xi}\}$ of the slaved state* (21)*. The approximation error depends on $\tilde{\rho} - \rho^{(0)}$, where $\rho^{(0)}$ is the initial density and $\tilde{\rho}$ is the internal simulated-upon density corresponding to $\{\tilde{\phi}, \tilde{\xi}\}$ (before the final reset to $\rho^{(0)}$).*

The proof is quite technical and given in [17]. Using the one-to-one relationship between $\boldsymbol{\varrho}$ and $\mathbf{f}$ (9), the expressions for the constrained runs fixed point $\{(\rho^{(0)}), \tilde{\phi}, \tilde{\xi}\}$ from [17] can be written as follows

$$\tilde{f}_i = \frac{1}{3}\rho^{(0)} - \frac{c_i \Delta x}{3\omega}\frac{\partial \rho^{(0)}}{\partial x} - \frac{\Delta t}{6\omega}(3c_i^2 - 2)\left(F(\rho^{(0)}) - 3\frac{(\tilde{\rho} - \rho^{(0)})}{\Delta t}\right) . \qquad (27)$$

When we compare this expression with the expansion of $f_i$ (13) in its Chapman-Enskog components $f_i^{[0]}$, $f_i^{[1]}$ and $f_i^{[2]}$ (15)–(18), we see that (27) is indeed correct up to first order in the Chapman-Enskog expansion. Due to the approximation error (the third term) in (27), the fixed point $\tilde{f}_i$ (or equivalently $\{\tilde{\phi}, \tilde{\xi}\}$) lies on an *"approximate" manifold* instead of the true slow manifold described by the macroscopic variable $\rho$ (cf. Fig. 1). To make this error as small as possible, we choose $\tau = \Delta t$ (one LBM time step).

## 4 Spatial Coupling

### 4.1 Problem Specification

In this section, we describe how to couple the PDE and LBM from Sect. 2. In our setup, shown in Fig. 2, the one-dimensional domain is split into two

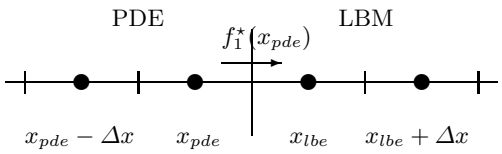



Fig. 2: Spatial coupling of a PDE (left) and LBM (right) on a one-dimensional domain. The interface lies in between two lattice sites. The evolution of the solution in the LBM region requires the propagating (post-collision) distribution value $f_1^\star(x_{pde}, t^\star)$ coming from the PDE domain. Since the PDE only evolves density values $\rho(x,t)$, this value is unavailable.

non-overlapping sublattices. Another option, using one overlapping lattice site is discussed in [1]. The PDE is applied to the left sublattice and the LBM to the right sublattice. We use the same lattice spacing $\Delta x$ and time step $\Delta t$ for both the PDE and LBM, i.e. the simplest coupled space-time lattice. Since the LBM is a mesoscopic model (as opposed to a truly microscopic model), using the same $\Delta x$ is very reasonable. On the other hand, using the same $\Delta t$ is an important simplification. More efficient coupling schemes, especially those that allow for different $\Delta t$, will be the subject of future research.

Since the PDE and LBM use a different set of variables, namely $\rho$ versus $\mathbf{f} = [f_{-1}\, f_0\, f_1]'$, we have to be careful how to exchange information at the interface during time simulation. To evolve the PDE in $x_{pde}$, the value $\rho(x_{lbe}, t) = \rho(x_{pde} + \Delta x, t)$ (see Fig. 2) is needed in (2). This value is computed from the LBM variables $f_i(x_{lbe}, t)$ using (7).

The inverse problem, where we have to transfer information from the PDE to the LBM region is more difficult. To evolve the LBM (4) in $x_{lbe}$ from $t$ to $t + \Delta t$, we need to map a single density value onto three corresponding distributions. This is formally stated as

$$\rho(x_{pde}, t) \mapsto f_i(x_{pde}, t)\ ; \quad i \in \{-1, 0, 1\}\ . \tag{28}$$

Since (7) should hold, this leaves two degrees of freedom.

The initialization of a LBM from a given density profile as described in [17] faces the same one-to-many problem, but there the problem concerns the whole domain instead of a single lattice site.

We will use the analytical slaving relations (15) and (16) or the corresponding numerical approximation (27) by the constrained runs scheme to derive the distributions at the interface given only the value of $\rho(x_{pde}, t)$. In the next sections we discuss these two strategies.

### 4.2 Implementation using First Order Perturbations

In order to evolve the solution in the leftmost site $x_{lbe}$ of the LBM sublattice, we need the value of the distribution function $f_1(x_{pde}, t)$ coming from the

PDE lattice that propagates into the LBM sublattice (see Fig. 2). This value is missing, but can be computed as follows.

The key observation is that the PDE (2) simulates "directly" from $t$ to $t+\Delta t$, while the LBM (4) executes in two phases: first, collisions and reactions to go from $t$ to $t^\star$ and secondly, propagation of the post-collision distributions $f_i^\star$ to get from $t^\star$ to $t+\Delta t$. Thus we actually need the post-collision value $f_1^\star(x_{pde}, t^\star)$ instead of the value $f_1(x_{pde}, t)$.

First, we compute the distribution $f_1(x_{pde}, t)$ corresponding to the PDE density $\rho(x_{pde}, t)$ at time $t$, using the first order perturbations (15) and (16)

$$\begin{aligned} f_1(x_{pde}, t) &= f_1^{[0]}(x_{pde}, t) + f_1^{[1]}(x_{pde}, t) \\ &= \frac{1}{3}\rho(x_{pde}, t) - \frac{\Delta x}{3\,\omega}\,\frac{\rho(x_{lbe}, t) - \rho(x_{pde} - \Delta x, t)}{2\Delta x} \end{aligned} \tag{29}$$

In (29), the derivative $\partial\rho(x_{pde}, t)/\partial x$ is approximated with central differences. The value $\rho(x_{lbe}, t)$ is obtained from the LBM domain using (7).

Afterwards, the corresponding post-collision value $f_1^\star(x_{pde}, t^\star)$ is computed from $f_1(x_{pde}, t)$ (29) using the LBM:

$$f_1^\star(x_{pde}, t^\star) = (1-\omega) f_1(x_{pde}, t) + \frac{\omega}{3}\rho(x_{pde}, t) + \frac{\Delta t}{3} F(\rho(x_{pde}, t)) \tag{30}$$

Finally, it is this value that is propagated to $x_{lbe}$, i.e.

$$f_1(x_{lbe}, t + \Delta t) = f_1^\star(x_{pde}, t^\star) \tag{31}$$

Note that the outgoing post-collision value $f_{-1}^\star(x_{lbe}, t^\star)$ that enters the PDE domain is never used.

### 4.3 Implementation using Constrained Runs

As explained in Sect. 3, the numerical computation of $f_i(x, t)$ from a given $\rho(x, t)$ by the constrained runs scheme is accurate up to first order. As an alternative to the procedure from Sect. 4.2, we can thus replace (29) with Algorithm 1 and apply (30) and (31) to the result.

As already mentioned in Sect. 4.1, Algorithm 1 solves the one-to-many problem for the LBM on the full domain, whereas the mapping problem for spatial coupling (28) is an issue in a single lattice site $x_{pde}$ only.

Since information in the (explicit) LBM (4) propagates over only one lattice site in each iteration, Algorithm 1 requires initial density values on a sublattice with at least $2K+1$ lattice sites, symmetrically distributed around $x_{pde}$, with $K$ the number of iterations needed for convergence of the algorithm. We can impose arbitrary boundary conditions on this sublattice because the boundary information will not have reached $x_{pde}$ within $K$ iterations.

Alternatively, one can drop the outer lattice sites (and distribution functions) during propagation in each iteration to obtain a *funneled* scheme. Here

we keep only the information streaming towards $x_{pde}$. Again it is important that there are at least $2K+1$ initial sites, symmetrically positioned around $x_{pde}$. Note that this funneled scheme decreases the amount of work with a factor two. On the other hand, this implementation requires changes to the propagation step of the LBM in Algorithm 1, which may not be desirable.

Of course, for the above implementations to work, the number of constrained runs $K$ has to be obtained first. To this end, one could do a preliminary run on (part of) the domain with Algorithm 1 and observe its convergence (see also [17]).

Depending on the implementation, the amount of work needed by the scheme is either $K^2$ or $K^2/2$. If $K$ is large and the full domain is small, this is an expensive overhead since the scheme has to be used in between each time step $\Delta t$. Of course, for situations where the analytical slaving relations (29) are unknown or difficult to obtain analytically, it is the only alternative.

## 5 Numerical Results

### 5.1 FitzHugh-Nagumo Reaction-Diffusion System

We will apply the proposed coupled PDE/LBM method to the FitzHugh-Nagumo (FHN) reaction-diffusion system on a one-dimensional domain. For this problem, both the PDE and LBM are known and valid on the full domain. The system consists of two species: an activator and an inhibitor.

The PDE system describes the evolution of the activator $\rho^{ac}(x,t)$ and inhibitor $\rho^{in}(x,t)$ concentration (density) and is given by

$$\begin{cases} \dfrac{\partial \rho^{ac}}{\partial t} = D^{ac}\dfrac{\partial^2 \rho^{ac}}{\partial x^2} + \rho^{ac} - (\rho^{ac})^3 - \rho^{in} \ , \\ \dfrac{\partial \rho^{in}}{\partial t} = D^{in}\dfrac{\partial^2 \rho^{in}}{\partial x^2} + \varepsilon(\rho^{ac} - a_1\rho^{in} - a_0) \ . \end{cases} \tag{32}$$

For each species, the LBM is described in Sect. 2.2. Given the PDE reaction terms in (32), the LBM reaction terms are defined as (6)

$$\begin{aligned} R_i^{ac}(x,t) &= \frac{\Delta t}{3}\left(\rho^{ac}(x,t) - (\rho^{ac})^3(x,t) - \rho^{in}(x,t)\right), \\ R_i^{in}(x,t) &= \frac{\Delta t}{3}\,\varepsilon\left(\rho^{ac}(x,t) - a_1\rho^{in}(x,t) - a_0\right), \quad i \in \{-1,0,1\}. \end{aligned} \tag{33}$$

For our numerical tests, we choose the parameter values as $D^{ac} = 1$, $D^{in} = 4$, $a_0 = -0.03$, $a_1 = 2$ and $\varepsilon = 0.05$. The domain has length $L = 20$. At the boundaries of the domain, we impose homogeneous Neumann (or no-flux) boundary conditions which are implemented in the LBM using the halfway bounce-back scheme [9]. For both models, the lattice points $x$ lie at the midpoints of 200 lattice intervals, such that $\Delta x = 0.1$. We choose the time step $\Delta t = 0.001$ [16].

### 5.2 Spatial Coupling of the FHN PDE and LBM

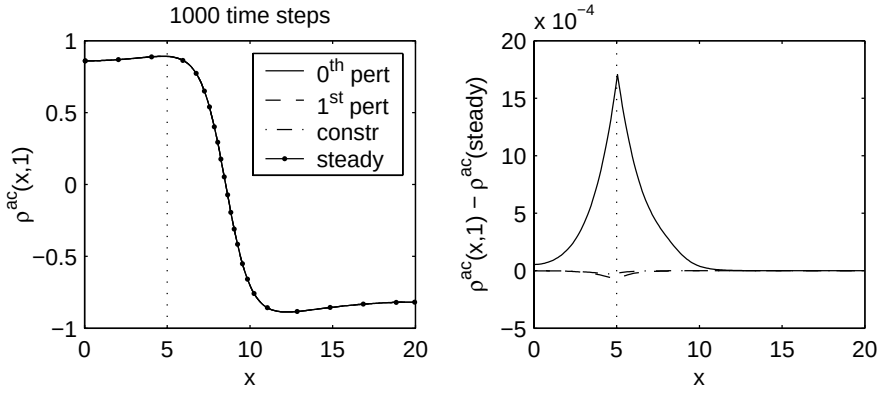



Fig. 3: Solution of the coupled FHN PDE/LBM model after 1000 time steps. We used either zeroth order perturbations (34), first order perturbations (Sect. 4.2) or constrained runs (Sect. 4.3) at the interface. The activator density $\rho^{ac}(x,t)$ at $t = 1000\Delta t$ is shown left and the difference with respect to the reference solution, the LBM steady state on the full domain, is shown right. The dotted line shows the position of the interface.

In this section, we spatially couple the FHN PDE (left) and FHN LBM (right) models from Sect. 5.1. The domain is divided as in Fig. 2 with the interface positioned at $x = 5$ (in between two lattice sites). Correspondingly, we solve predominately with the LBM. We initialize the coupled model with the LBM steady state computed on the full domain (see [16]). As described in Sect. 4, we use either first order perturbations or constrained runs at the interface. For our problem, the number of constrained iterations needed for convergence is $K = 25$ (see [17]).

As an illustration, we also compare our results with a modification of the scheme described in Sect. 4.2. We drop the spatial derivatives from (29) to obtain

$$f_1(x_{pde}, t) = f_1^{[0]}(x_{pde}, t) = \frac{1}{3}\rho(x_{pde}, t) \tag{34}$$

and replace (29) by (34). Afterwards, we proceed as in Sect. 4.2. We call this modification the zeroth order coupling scheme.

We first look at the short time behavior of the coupled problem. Figure 3 shows the results after 1000 time steps. On the time scale considered, the difference between the densities cannot be resolved visually. The corresponding error is shown in the right panel. Here, we see that the zeroth order coupling scheme results in a significant interfacial error, while the error with first order

or constrained runs coupling is much smaller. Clearly, neglecting first order contributions in the coupling scheme is not a good option.

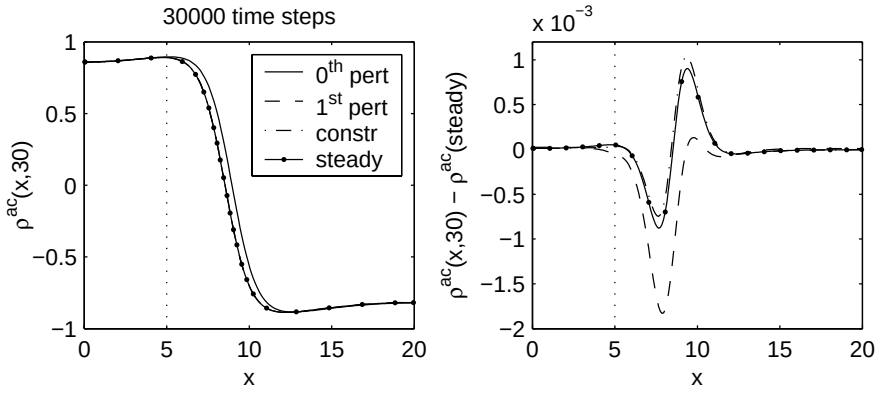



Fig. 4: The solution of the coupled FHN PDE/LBM model for different coupling mechanisms after 30000 time steps. Shown on the left is the activator density $\rho^{ac}(x, 30)$. In the right panel, we compare the corresponding errors to the difference between the PDE and LBM density computed on the full domain. The latter is marked by the label "steady".

Next, we look at the long term effects of this interfacial error. In Fig. 4, we show the solution after 30000 time steps. The left panel shows that the error of the zeroth order coupling has now propagated over the domain and shifted the solution globally to the right. In fact, all coupled models converge towards a steady state that is different from the steady state obtained with one model on the full domain.

The right panel shows that the modeling errors as a result of the coupling with first order and constrained runs are comparable to the modeling error between the LBM and PDE solution computed on the full domain. These errors are most pronounced in the region where the solution varies strongly (cf. Sect. 2.3). Note that we compared to the PDE steady state reference solution on the full domain here, as we chose to solve predominately with the LBM in the coupled model.

From the numerical experiments, we can also learn something about the relation between the error and the spatial derivatives of the solution at the interface. The (small) interfacial error in Fig. 3 for the coupling with both first order perturbations or the constrained runs scheme comes from the second order term $f_1^{[2]}(x_{pde}, t)$ (18) that is neglected in the computation of the distribution function at the interface. This term is related to the second spatial derivative $\partial^2\rho(x_{pde}, t)/\partial x^2$ of the solution at the interface (see (22)). Since this

second derivate is nonzero at the interface for our FHN example, we observe a local interfacial error.

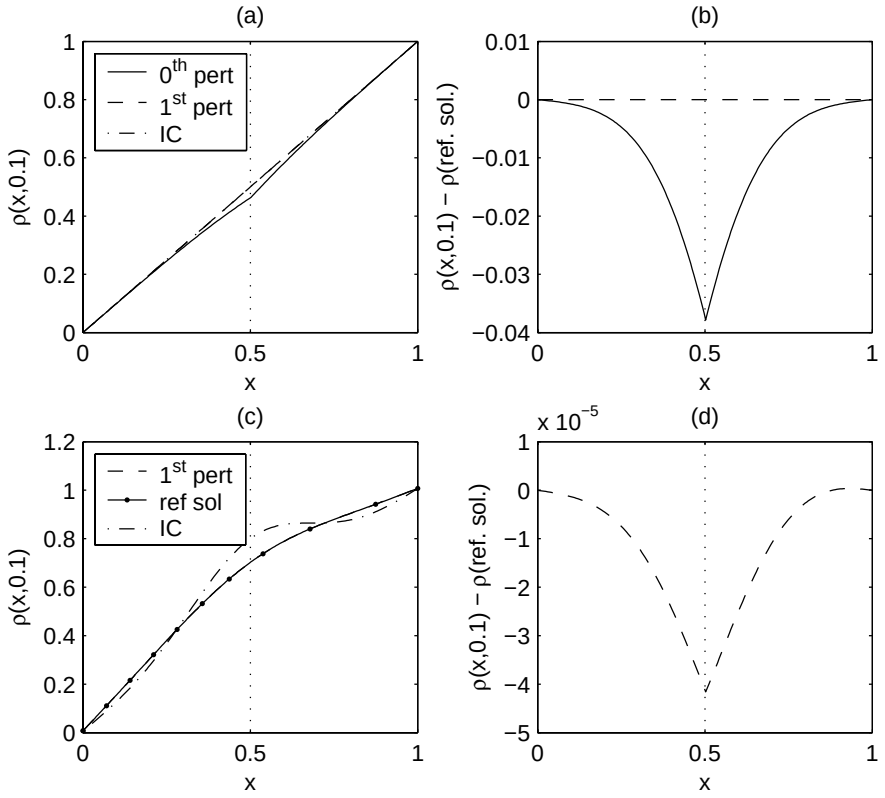



Fig. 5: Illustration on how the error depends on the spatial derivatives of the solution. The example is a pure diffusion system with Dirichlet boundary conditions on the domain $[0, 1]$ for two types of initial conditions. The coupled PDE/LBM system is simulated for 10000 time steps. The position of the interface is marked by a dotted line. Figures (a) and (b) show the solution and the error when the initial condition is a straight line connecting the two boundary conditions. Figures (c) and (d) show respectively the reference solution and solution with first order coupling and the corresponding error for an initial profile with nonzero second derivatives. First order coupling is correct for a solution with only first derivatives, but has a small error when the solution has a nonzero second derivative.

In Fig. 5 we perform an experiment to illustrate this relation. Here, we consider a pure diffusion model problem with Dirichlet boundary conditions $\rho(0, t) = 0$ and $\rho(1, t) = 1$. The domain $[0, 1]$ is discretized with 200 lattice points and the parameters are chosen as $D = 0.2$ and $\Delta t = 0.00001$. The

interface is located at $x = 0.5$. The steady state solution is a straight line connecting the density values at the boundary. We simulate the coupled system for 10000 time steps. We impose two different initial conditions: the steady state solution (Fig. 5 (a)-(b)) or an initial condition with a nonzero second derivative (Fig. 5 (c)-(d)). In the first case, Fig. 5 (b) shows that the error is zero when first order coupling is used, since $\partial^2\rho/\partial x^2 = 0$ here. In the second case, the solution at $t = 10000\Delta t$ has a nonzero second derivative and the first order coupling error behaves similar to the error of the zeroth order coupling in Fig. 5 (b). Note the different scale of Fig. 5 (b) and (d).

Of course, when simulating the system in Fig. 5 (c)-(d) for long enough time, the interfacial error for the first order coupling will become zero since the solution converges to the steady state as in Fig. 5 (a)-(b). For the FHN example however, the steady state has a nonzero second derivative at the interface and both the local interfacial error and the resulting global modeling error evolve to a constant nonzero value.

Figures 3 and 4 suggest that the coupling with constrained runs is more accurate than coupling with first order perturbations. This can be explained by comparing (27) to (22). We see that the part corresponding to the reaction term in (22) is approximated correctly by (27) while it does not appear in the first order coupling scheme. At least for our example, this results in a higher accuracy for the constrained runs coupling.

### 5.3 Spatial Coupling of a Growth-Diffusion PDE and LBM

In this section, we will spatially couple the LBM (23) and the finite difference discretization of the corresponding PDE (24) for the growth-diffusion system from Sect. 2.4. The setup on the one-dimensional domain is described in Fig. 2. We used the D1Q3 scheme and a reaction matrix

$$A = [A_{ij}] = \begin{bmatrix} -R & 0 & 0 \\ 1.1R & 0 & 1.1R \\ 0 & 0 & -R \end{bmatrix} \tag{35}$$

with $R = 0.02$. The relaxation parameter is $\omega = 1.6160$. This LBM leads to a macroscopic growth-diffusion problem (24) with $D = 0.1856$ and $\alpha = 1.315 \cdot 10^{-3}$. Grid parameters are $\Delta x = 1/399$ and $\Delta t = 2.5 \cdot 10^{-6}$. The interface is positioned at $x = 0.25$.

Instead of using the analytical slaving relations derived through a tedious Chapman-Enskog expansion [18], we will use the constrained runs scheme to derive the distributions at the interface numerically, given the density value. Although we only proved stability and convergence of the constrained runs scheme for the BGK LBM with density dependent reaction term (6) in [17] (see Sect. 3), we expect that the scheme can be applied more generally also.

Figure 6 shows the density profile $\rho(x, t)$ and corresponding error of the coupled PDE/LBM model for the growth-diffusion system from Sect. 2.4. As

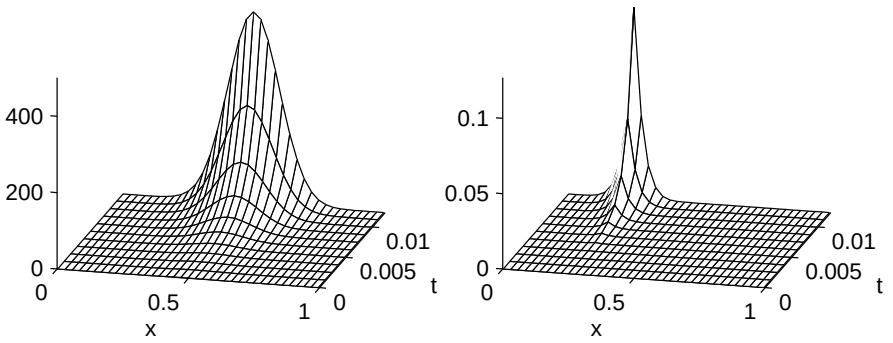



Fig. 6: The solution of the coupled PDE/LBM model for the growth-diffusion system from Sect. 2.4 as a function of space and time. The left panel shows the density $\rho(x,t)$ while the right panel shows the corresponding absolute error. We used constrained runs at the interface, which is positioned at $x = 0.25$.

explained in Sect. 5.2, an interfacial error exists because the solution has a nonzero second derivative at the interface. Since the solution grows over time, the error grows as well.

## 6 Conclusions and Future Work

In this article, we have studied the coupling of a lattice Boltzmann model (LBM) and a partial differential equation (PDE) describing the same diffusive system, each on a part of a one-dimensional domain. We discretized the PDE using finite differences.

At the interface between the two models, we have a one-to-many problem: the PDE variables (here density) have to be mapped to more LBM variables (the distribution functions). We showed that this can be done both analytically, using results from the Chapman-Enskog expansion, or numerically, using the constrained runs scheme [8, 7]. We use the results from [17], where it is shown that this scheme applied to the LBM under discussion approximates the results from the Chapman-Enskog expansion correctly up to first order. We illustrate these concepts for several diffusive systems.

We show that the solutions of the coupled model are comparable in accuracy to the solution of either the PDE or LBM model on the full domain. We also show that the error made at the interface is proportional to the second derivative of the solution at the interface. The latter is related to the error term for the constrained runs scheme and to the second order term in the Chapman-Enskog expansion.

For certain classes of problems, the current approach is not sufficient and coupling that is correct to higher order is required. In this case, higher order

terms in the Chapman-Enskog expansion could be used if these analytic expressions are available. If these are not available, constrained run schemes as proposed in [7] in the context of ODEs, could be useful.

In the current discussion, we used the same space and time step $\Delta x$ and $\Delta t$ for both the PDE and LBM sublattice. The focus was on the details of the coupling and how information is exchanged between the two sublattices. The use of different $\Delta x$ and $\Delta t$ values in both regions, optimized to local stability properties, would be a further development. When e.g. the time step of the PDE is a multiple of the LBM time step, interpolation of densities between two PDE time steps is then required to provide the necessary information to the LBM region.

We foresee several applications of a coupled PDE/LBM simulation approach. One example is plasma physics where detailed reaction rates in localized regions determine the macroscopic behavior. Here, the system has a solution that varies rapidly in a localized region of the domain but behaves smoothly in the remainder. In the former region, the approximations made to derive a PDE model will break down and the LBM with the complete reaction details has to be used. In the smooth region, in contrast, the PDE approximations are valid. It is important that the interface is situated in a region where the solution is still slowly varying such that the slaving relations from the Chapman-Enskog expansion hold.

*Acknowledgement.* The work of P. Van Leemput is supported by project G.0130.03 funded by the Fund for Scientific Research - Flanders and by the Belgian Programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office. The work of W. Vanroose is supported by a DWTC return grant from the Belgian Federal Science Policy Office.

## References

1. P. Albuquerque, D. Alemani, B. Chopard, P. Leone: Coupling a lattice Boltzmann and a finite difference scheme. In: *International Conference on Computational Science – ICCS 2004*, vol. 3039 of *Lecture Notes in Computer Science*, ed. by M. Bubak, G. D. van Albada, P. M. Sloot, J. Dongarra, pages 540–547 (Springer, Berlin Heidelberg New York 2004)
2. F.J. Alexander, A.L. Garcia, D.M. Tartakovsky: Algorithm refinement for stochastic partial differential equations: I. Linear diffusion. *J. Comput. Phys.* **182**, 47–66 (2002)
3. B. Chopard, A. Dupuis, A. Masselot, P. Luthi: Cellular automata and lattice Boltzmann techniques: An approach to model and simulate complex systems. *Adv. Complex Systems* **5** (2/3), 103–246 (2002)
4. D. Dab, J.-P. Boon, Y.-X. Li: Lattice-gas automata for coupled reaction-diffusion equations. *Phys. Rev. Lett.* **66** (19), 2535–2538 (1991)
5. S.P. Dawson, S. Chen, G.D. Doolen: Lattice Boltzmann computations for reaction-diffusion equations. *J. Chem. Phys.* **98** (2), 1514–1523 (1993)

6. A.L. Garcia, J.B. Bell, W.Y. Crutchfield, B.J. Alder: Adaptive mesh and algorithm refinement using direct simulation Monte Carlo. *J. Comput. Phys.* **154**, 134–155 (1999)
7. C. W. Gear, T. J. Kaper, I. G. Kevrekidis, A. Zagaris: Projecting to a slow manifold: Singularly perturbed systems and legacy codes. *SIAM J. Appl. Dynamical Systems* **4** (3), 711–732 (2005)
8. C.W. Gear, I.G. Kevrekidis: Constraint-defined manifolds: A legacy code approach to low-dimensional computation. *J. Scientific Computing* **25** (1), 17–28 (2005) Preprint available at `physics/0312094`
9. I. Ginzbourg, P.M. Adler: Boundary flow condition analysis for the three-dimensional lattice Boltzmann model. *J. Phys. II France* **4**, 191–214 (1994)
10. N.G. Hadjiconstantinou. Hybrid atomistic-continuum formulations and the moving contact-line problem. *J. Comput. Phys.* **154**, 245–265 (1999)
11. P. Le Tallec, F. Mallinger: Coupling Boltzmann and Navier-Stokes equations by half fluxes. *J. Comput. Phys.* **136**, 51–67 (1997)
12. P. Le Tallec, M.D. Tidriri: Convergence analysis of domain decomposition algorithms with full overlapping for the advection-diffusion problems. *Math. of Computation* **68** (226), 585–606 (1999)
13. Y.H. Qian, D. D'Humières, P. Lallemand: Lattice BGK models for Navier-Stokes equation. *Europhys. Lett* **17** (6), 479–484 (1992)
14. Y.H. Qian, S.A. Orszag: Scalings in diffusion-driven reaction $A + B \rightarrow C$: Numerical simulations by lattice BGK models. *J. Stat. Phys.* **81** (1/2), 237–253 (1995)
15. S. Tiwari, A. Klar: An adaptive domain decomposition procedure for Boltzmann and Euler equations. *J. Comput. & Appl. Math.* **90**, 223–237 (1998)
16. P. Van Leemput, K. Lust, I.G. Kevrekidis: Coarse-grained numerical bifurcation analysis of lattice Boltzmann models. *Physica D: Nonlinear Phenomena* **210** (1–2), 58–76 (2005)
17. P. Van Leemput, W. Vanroose, D. Roose: Initialization of a lattice Boltzmann model with constrained runs. Technical Report TW 444, (Katholieke Universiteit Leuven, Department of Computer Science, December 2005) Submitted to J. Comput. Phys.
18. W. Vanroose. Analysis of a lattice Boltzmann model for planar streamer fronts. in preparation, (2006)

# Modelling and Control Considerations for Particle Populations in Particulate Processes Within a Multi-Scale Framework

N. Bianco and C. D. Immanuel*

Department of Chemical Engineering, Centre for Process Systems Engineering, Imperial College London, South Kensington, London SW7 2AZ, UK

**Summary.** This article deals with a class of distributed parameter systems, the so-called particulate processes that are modelled by population balances. Population balances have been employed in modelling chemical, physical and biological processes for over 40 years. The population balance equation is a hyperbolic partial differential equation that presents challenges in numerical solution. Recent advances in the understanding of the underlying mechanisms of the particulate processes enables formulation of more comprehensive population balance models for these complex processes by the incorporation of multi-scale representations for the kernels of the constituent rate processes. These multi-scale modelling ventures lead to additional numerical challenges for model solution. Further, the purposes of these comprehensive models is towards use for control of distributions in these processes. This control becomes challenging in view of the different scales represented by the manipulated and controlled variables and in view of the underlying process complexity. This article first presents an efficient numerical solution technique to handle multi-scale population balance models, and then discusses a potential model-based strategy for control of distributions in particulate processes.

## 1 Introduction

Particulate processes are characterised by a distribution of particles and constitute a considerable fraction of chemical engineering unit operations, falling under the general class of distributed parameter systems. The most common examples of distributed particulate processes are polymerisation processes that are characterised by a distribution of chain lengths of the polymeric entities; crystallisation and granulation processes that are characterised by a size distribution of the crystals; and dispersed phased systems including emulsions that are characterised by a particle/droplet size distribution (Refer to [4] for insight into a wide range of distributed parameter systems including novel nano-structured materials and microfluidics).

* Corresponding author, c.immanuel@imperial.ac.uk

The incorporation of particle-level details in models of particulate processes is both important and feasible. The feasibility of such ventures is attributed to the availability of intricate measurements on the one hand, and the enhanced computation capabilities on the other hand. The importance of such ventures is attributed to the direct need for the control of population distributions on the one hand and the need for high fidelity models commensurate with the state-of-the-art knowledge and measurement capabilities on the other hand. These ventures translate into the development of particle population distribution models.

To elaborate on the need for such population distribution models further:

- There is a strong incentive for the control of particle populations in particulate processes [13, 4, 2]. For instance, in processes that involve particle size distribution, certain applications necessitate broad or even multimodal distributions to tailor the packing densities and other end-use properties. In certain other processes such as those that involve distributions of polymeric chains, there is a necessity to ensure a narrow distribution of chain lengths for increased strength and cohesion. These clearly highlight the need for control of particle and molecular entities. The distribution control problem is clearly presented by Semino & Ray in the form of comprehensive controllability studies to determine the ability to influence distributions and their shapes [41, 42]. Subsequently, several researchers have shown the importance and the feasibility of this problem [13, 4, 2].
- Another class of processes are cellular biological processes, wherein, the complexity of the process and the need for high-fidelity models warrant the introduction of information at the cellular levels and even at the molecular level for application in various systems engineering tasks [8, 15, 43]

With the incorporation of particle-level information, such process models assume a multi-scale character. The particle population is determined by a complex array of particle-level phenomena of nucleation (birth), continuous growth, discrete growth (aggregation and breakage), and death. Each of these particle-level phenomenon exhibits strong dependence on one or more of the population traits, the so-called internal coordinates, the most common of these being the particle size, cell age and molecular weight. Each of these internal coordinates covers a wide range of values in their representative systems, thereby leading to computation-intensive models. Figure 1 depicts the multi-scale model and the control problem that underlie particulate processes. It has to be mentioned that these process models are usually less multi-scale in character compared to the models that will be necessary for material and product design applications. Nevertheless, the distributed character of the process-level variables still renders these models with a high computational requirements. Thus, there is still a need for efficient computation to enable online applications such as process control.

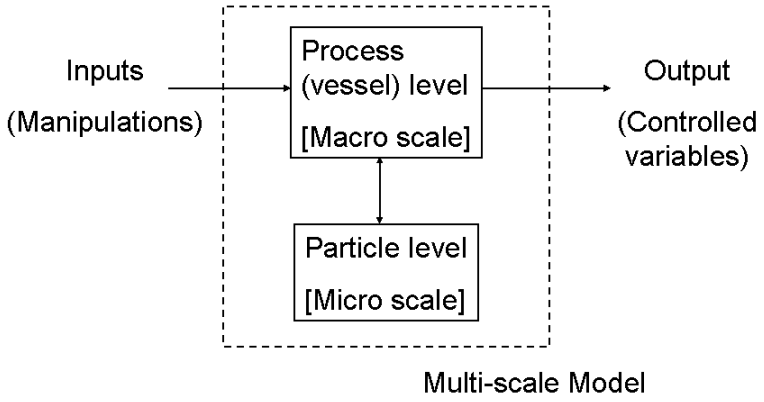



Fig. 1: Depiction of the multi-scale modelling and control problems that underlie particulate processes

Different computational methods have been adopted to tackle this problem. These can be broadly classified as follows, employing the spirit of previous classifications [25]:

- Particle-ensemble approach: This approach is focussed only on the particle level, either in the form of statistical Monte Carlo simulations or in the form of discrete element methods. Such models infer behaviour at the process (vessel) level only through the particle-level solutions.
- Segregated or sequential approach: A segregated modelling approach employs particle-level models to determine phenomenological equations which are then employed in process-level models. Such models will usually be restricted in application, as in most systems the particle-level phenomena are dynamic and vary during the process. Thus, a segregated determination of the phenomenological laws will be less suitable for those cases. However, if applicable, this approach is the best recourse.
- Multi-scale or simultaneous approach: The simultaneous multi-scale models seek to solve both the particle-level and process (vessel)-level models together. Although there are several ways of performing this combined modelling, a simple approach is to appropriately simplify the model at one of the two levels, depending upon the intended objectives and the needed accuracy levels. See [25] for a more detailed and elaborate classification of such approaches.

The approach adopted in this study is based on using the guiding theories of the underlying particle-level physics (such as kinetic theory, colloids theory *etc.*) to *transform* the particle-level model, to a form which can then be *embedded* into the process-level model, borrowing the terminologies of [25].

The process-level behaviour in particulate processes is exemplified by the particle size distribution, and is described employing the so-called population balance equations [39]. There are several research challenges associated with population balances. The first of these is the difficulty in identifying closed-form kernels or constitutive relationships for the underlying particle-level rate processes (for example, particle nucleation, growth, aggregation etc.). The second challenge lies in the numerical solution of the population balance models, which are characterised by a computation-intensive and multi-scale character. The third challenge lies in the control of such processes. Although control is mainly sought at the macroscopic level of the particle size distribution, it will be shown later that this is best achieved through control of the underlying particle-level rate processes.

With regard to the first challenge mentioned above, different particulate processes are at different levels of advancement. For example, processes such as the crystallisation, precipitation and emulsion polymerisation processes have a strong fundamental understanding for the development of particle-level models, while processes such as granulation and several biological processes have reduced advances in this front. This issue is elaborated in section 2. The numerical solution techniques are presented in section 3, and the control issues are addressed in section 4.

## 2 Population Balance Models

The general population balance equation (PBE) is naturally suited to model processes characterised by birth, aging and death phenomena such as particulate processes. The PBE is given by:

$$\frac{\partial}{\partial t}\zeta(\eta,t)+\frac{\partial}{\partial \eta}\left(\zeta(\eta,t)\Re_{growth}\right)=\Re_{birth}(\eta,t)-\Re_{death}(\eta,t)+\Re_{realignment}(\eta,t) \tag{1}$$

$$\Re_{realignment}(\eta,t)=\Re_{formation}(\eta,t)-\Re_{depletion}(\eta,t) \tag{2}$$

$$\Re_{formation}(\eta,t)=\frac{1}{2}\int_{\eta'=\eta_{nuc}}^{\eta-\eta_{nuc}}\beta(\eta',\eta-\eta')\zeta(\eta',t)\zeta(\eta-\eta',t)\ \mathrm{d}\eta' \tag{3}$$

$$\Re_{depletion}(\eta,t)=\int_{\eta'=\eta_{nuc}}^{\eta_{max}}\beta(\eta',\eta-\eta')\zeta(\eta',t)\zeta(\eta-\eta',t)\ \mathrm{d}\eta' \tag{4}$$

where $\zeta(\eta,t)$ is called the population density function. This represents the distribution of particles with respect to the internal coordinates $\eta$. The internal coordinates could represent one or more distributed variables such as particle size, cell age, cellular metabolites, molecular weight, intraparticle composition, *etc.* The terms $\Re_{nuc}$, $\Re_{growth}=\frac{\mathrm{d}\eta}{\mathrm{d}t}$, $\Re_{realignment}$ and $\Re_{death}$ account

for particle-level phenomena of nucleation (birth), continuous growth, discrete growth, and death, respectively, and constitute the source of the model's multi-scale character. As mentioned in the introduction, the particle-level phenomena are converted, employing theories such as kinetic and colloids theory, into phenomenological laws (constitutive equations). This is typically exemplified by the aggregation (coagulation) phenomenon, which is a discrete growth phenomenon, and is illustrated next for the emulsion polymerisation problem.

Emulsion polymerisation is a heterogeneous multi-phase polymerisation method in which the polymer is produced in the form of sub-micron particles that are dispersed in the continuous phase (water). The internal coordinate $\eta$ is the particle size (volume) $V$, and the particle density $\zeta$ is defined as $\zeta(\eta, t) = F_V(V, t)$ such that $F_V(V, t)\mathrm{d}V$ is the moles of particles of volume between $V$ and $V + \mathrm{d}V$ per unit volume of the continuous phase ($V_{aq}$). The particle population distribution is determined by three major particle-level phenomena of nucleation, growth due to polymerisation and swelling, and inter-particle coagulation. The particle size distribution (PSD) is a major determinant of the polymer properties.

In emulsion polymerisation, the coagulation phenomenon is driven by colloidal forces of attraction and repulsion that influence the particles. The forces of attraction are usually the van der Waals' forces, while the forces of repulsion are attributed to the presence of surfactant chains adsorbed onto the surface of the particles. In the case of ionic surfactants, surfactant adsorption results in charged particles (with 'like' charges), thereby causing inter-particle repulsion. In the case of non-ionic surfactants, the bulkiness of the surfactant chains adsorbed onto the particles causes steric repulsion (space constraints). The strength of the repulsion forces depends on the level of adsorption (particle coverage), which in turn is determined by the PSD, as depicted in Figure 2. Thus, the inter-particle force balances on the particle pairs will need to be performed dynamically while solving the process-level population balance model. In addition, the forces themselves are dependent on the particle size.

As mentioned previously, in this work, the dynamic force/potential balances is performed in an implicit semi-analytical framework to obtain the phenomenological law representing the coagulation kernel $\beta$, also called the intrinsic coagulation rate. In order to undergo coagulation, the net attractive force acting on the particle pairs must overcome an activation barrier as governed by the kinetic transition state theory. This activation barrier is represented in terms of a stability ratio W for the particle pair $(r, r')$ given as follows:

$$W(r, r') = (r + r') \int_{D=(r+r')}^{\infty} \frac{e^{\frac{\Psi(D)}{kT}}}{D^2} \, \mathrm{d}(D) \tag{5}$$

where $k$ is the Boltzmann constant, $T$ is the temperature of the emulsion, and $\Psi(D) = \Psi_R(D) - \Psi_A(D)$ is the net potential acting on the particles, $\Psi_R$ and $\Psi_A$ representing the repulsive and attractive potentials respectively. The terms $r$ and $r'$ represent the radii of the two particles involved. The coagulation kernel

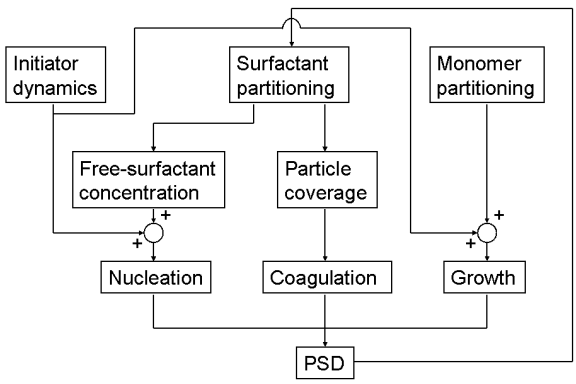



Fig. 2: Schematic representation of the need for both process-level and particle-level details in models of emulsion polymerisation

can then be calculated as follows:

$$\beta(r, r') = c\frac{4\pi D_0(r + r')}{W} \tag{6}$$

where $c$ is an adjustable constant and $D_0$ is the diffusion co-efficient. See detailed modelling of coagulation kernel in [19] and references therein on the underlying colloids theory.

Thus, the dynamic inter-particle force balance is reduced to the above phenomenological law, facilitating the solution of the multi-scale model. However, the model is still computation-intensive, both in the calculation of $\beta$ (which needs to be computed for each combination of the particle pairs), and the terms appearing in the PBE. Thus, efficient strategies are necessary for the solution of the resulting multi-scale models.

Also, as depicted in Figure 2, in addition to the amount of surfactant utilised for stabilisation, the PSD also determines the amount of free surfactant which is then available for the nucleation of new particles. Further, PSD also affects the monomer partitioning among the particles, thereby affecting the growth phenomenon. These clearly elucidate the need for a simultaneous incorporation of details at the particle-level and the process level in models of the emulsion polymerisation process.

## 3 Solution Techniques for Population Balance Models

### 3.1 Literature Review

The method of moments has been successfully applied and developed to handle population balances problems [37, 9, 10, 11]. The developments focus on

making the method applicable to systems with size-dependent particle-level behaviour and to overcome problems with unclosed system of equations. However, the application is most suitable to those problems where distribution control is not of direct interest. A notable and representative application is presented by Chiu & Christofides [3] for a model of a continuous crystallizer, and the study describes the application of an approximate inertial manifold method to extract a low-dimensional model from the discretised Galerkin's version of the PBE. However, when the distribution itself is the controlled variable of interest, the method of moments is less suitable. In this regard, a relatively new method that relies on the moments called the quadrature method of moments is presented by Marchisio & Fox [34] for application to problems with size-dependent kernels and for complex systems which are not easily amenable to the more common method of moments.

The general solution techniques for PBE models can be classified into two major types, one of which approaches the PBE as any general partial differential equation (PDE) [26, 7, 33], while the other considers the underlying physics of the evolution of PSD in formulating a solution strategy [16]. With regard to the first class, a general PDE is usually solved by either

- a finite difference scheme that approximates the differentials in terms of differences or
- a method of weighted residuals that solves the PDE by casting the solution in terms of pre-determined (chosen) basis functions and seeking to minimise the approximation error at fixed points along the independent variables.

Both these methods have been employed for PBE models. Crowley *et al.* [6] present a detailed evaluation of the relative merits and drawbacks in the application of the finite difference methods for the solution of PBE. Melis *et al.* [36] employ the finite difference method for the solution of a complex and realistic example system in emulsion polymerisation. Ma *et al.* [32] present a high-resolution finite difference method that exploits the relative advantages of the upwind and Lax-Wendroff methods with respect to numerical diffusion and oscillation. Hu *et al.* [17] apply the finite different scheme for a seeded batch crystallizers adopting a method that could be classified as a Eulerian-Lagrangian framework. Bennett and Rohani [1] adopt a finite difference method for a crystallisation problems, based on a combination of the Lax-Wendroff and Crank-Nicholson techniques to minimise spurious oscillations and maintain stability due to the implicit forms of the finite differences. As regards the method of weighted residuals, applications range from the Galerkin's method to orthogonal collocation on finite elements to wave-let based basis functions [7, 33, 31, 20, 35, 27]

The PBE is unique in relation to partial differential equations, attributed to the strong size-dependence (or more generally, dependence on the independent variables) of the particle rate processes of nucleation, growth, death, and discrete realignment phenomena. This uniqueness, coupled with

the computation-intensive character of the phenomenological laws as well as the terms in the PBE, motivates the development of custom-built solution techniques for the PBE. A seminal work in this regard was presented by Hounslow *et al.* [16], which was adopted and/or extended in a wide range of studies [29, 28, 26].

A so-called hierarchical two-tier solution technique which falls under the second category above was presented recently by Immanuel & Doyle [23]. Figure 3 shows a schematic representation of the proposed algorithm. At each time step, the first tier of the algorithm involves the calculation of the rates of nucleation, growth, and coagulation individually, while holding the PSD constant. Then the PSD is updated in the second tier employing the rates calculated in the first tier. The calculation then proceeded to the next time step. Employing such a hierarchical algorithm provides the following advantages:

- an ability to simplify both the phenomenological laws and the rate terms of the PBE, thereby reducing the on-line computational requirements
- a possible time scale separation, and hence a reduction in the stiffness of the model equations

In that work, a fine discretisation along the particle size was employed to ensure accuracy. The study presented in section 3.2 proposes a method to cope with the varying size-dependence of the underlying rate processes by employing a multi-level discretisation of the distribution domain.

Ramkrishna and Mahoney [40] highlight that despite years of research and advancement, population balances still remain at the forefront of research. This is in view of the plethora of novel applications necessitating multi-dimensional population balances and applications to stochastic systems. Ma *et al.* [32] successfully applied the high-resolution algorithms mentioned previously for two-dimensional problems; one of the first reports on the solution of multi-dimensional PBE. The hierarchical two-tier technique has also been applied to solve three-dimensional problems in granulation processes [18], with some early results reported for applications in 6-D problems characterising bioprocesses [24].

### 3.2 Two-Level Discretisation-Based Two-Tier Hierarchical Algorithm

The hierarchical two-tier algorithm mentioned previously is based on the discretisation of the particle population into finite elements or bins. The two-tier framework enables a consideration of the different rate processes of nucleation, growth, coagulation, *etc.* separately from each other. Figures 4 and 5 show that while the coagulation model is less sensitive to the bin width, the nucleation and growth models necessitate a fine discretisation. Fortuitously, the coagulation model is the more computation-intensive one and hence will benefit from a coarser discretisation. Thus, a two-level discretisation is proposed,

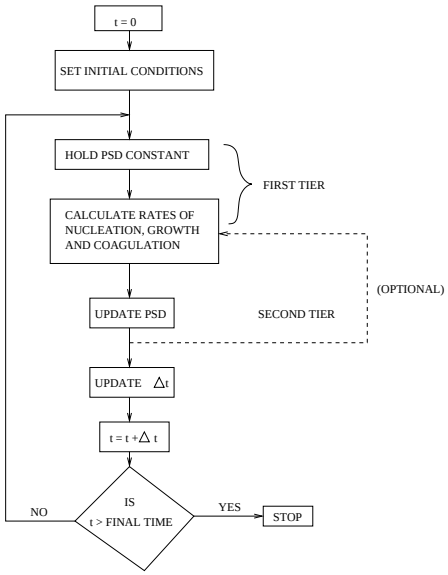



Fig. 3: The schematic of the hierarchical two-tier algorithm [23]. The first tier involves the calculation of the rates of nucleation, growth and coagulation individually while holding the PSD constant. The PSD is then updated in the second tier. Iteration over these two tiers can be employed to exchange information between the two tiers.

with the growth rates being solved employing a finer discretisation and the coagulation rates being solved employing a coarser discretisation. Nucleation is restricted to the smallest finer grid. Figure 6 shows the resultant two-level discretisation of the particle size domain.

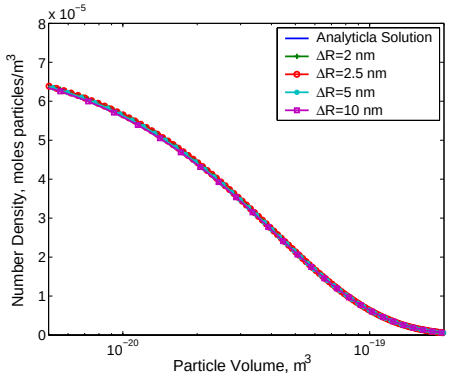



Fig. 4: The variation of the endpoint number density under coagulation-only conditions, employing different grid widths. The analytical solution is also shown.

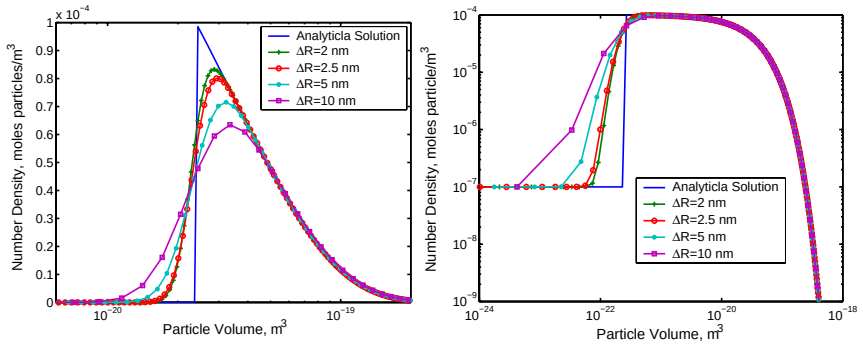



Fig. 5: Validation of growth and nucleation model. (a) The variation the endpoint number density under growth-only conditions, employing different grid widths. The analytical solution is also shown. (b) The variation the endpoint number density under coagulation-free conditions, employing different grid widths. The analytical solution is also shown.

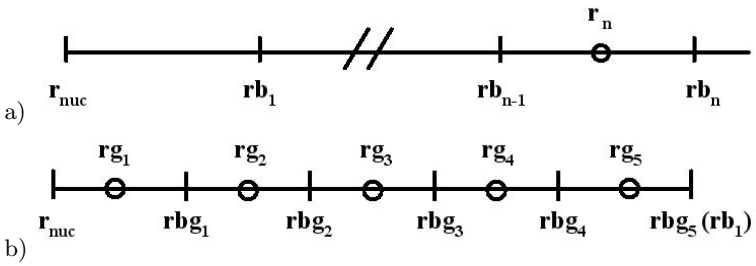



Fig. 6: Two-level finite element discretisation of the particle size domain. (a) Finite element discretisation on the coarser level. $r_{nuc}$ is the size of the nucleus, $r_i$ is the representative size for finite element $i$ on the coarser level, $rb_i$ is the upper boundary of finite element $i$ on the coarser level. (b) Finite element discretisation of a given coarser grid at the finer level. $rg_i$ is the representative size for finite element $i$ on the finer level, $rbg_i$ is the upper boundary of finite element $i$ on the finer level.

In the proposed two-level discretisation algorithm, two particle densities are defined, one at the coarser level $F^c(r,t)$ and the other at a finer level $F^f(r,t)$. The coagulation events are updated at the coarser level as follows:

$$F^c_{new,i} = F^c_i - \Delta t\,(H(i_{upper} - i)\Re_{formation_i} - H(i_{cut-off} - i)\Re_{depletion_i}) \quad (7)$$

where $F^c_i$ is the total particles in the $i^{th}$ coarse bin. $H$ is the heaviside function defined as $H(i) = 1$ for $i > 0$ and $H(i) = 0$ for $i <= 0$; and $i_{upper}$ and $i_{cut-off}$ account for the underlying physics and represent the number of the largest coarse bins in which particles can form due to coagulation and particles can

participate in coagulation, respectively. In the above implementation, a first order Euler method is depicted for the integration for ease of illustration. A more elaborate higher order integration technique might be employed, such as the sixth order Simpson's rule given below. An Euler integration method will be depicted for the rest of the equations, purely for ease of illustration.

$$F^c_{new,i} = F^c_i - H(i_{upper} - i)(\frac{\Delta t}{4})(14\Re_{formation_{i,1}} + 64\Re_{formation_{i,2}} + 24\Re_{formation_{i,3}} + 64\Re_{formation_{i,4}} + 14\Re_{formation_{i,5}}) \left(\frac{1}{45}\right) - H(i_{cut-off} - i)(\frac{\Delta t}{4})(14\Re_{depletion_{i,1}} + 64\Re_{depletion_{i,2}} + 24\Re_{depletion_{i,3}} + 64\Re_{depletion_{i,4}} + 14\Re_{depletion_{i,5}}) \left(\frac{1}{45}\right) \quad (8)$$

In Equation (8), $\Re_{formation_{i,j}}$ and $\Re_{depletion_{i,j}}$ are the total rates of formation and depletion, respectively, of particles in $i^{th}$ coarse grid at the $j$th time sub-step. Note that for the Euler integration method (Equation (7)), $j = 1$ as there is only one time sub-step within any given major time interval. The nucleation and growth events are updated at the finer level to obtain the PSD as follows:

$$F^f_{new,1} = F^f_1 - F^f_1 \frac{(rb_1 - ri_1)}{\Delta R_1} + (\Delta t)\,(\Re_{nuc}) \quad (9)$$

and

$$F^f_{new,i} = F^f_i - F^f_i \frac{(rb_i - ri_i)}{\Delta R_i} + F^f_{i-1} \frac{(rb_{i-1} - ri_{i-1})}{\Delta R_{i-1}} \quad (10)$$

where $F^f_i$ represents the total number of particles in the $i^{th}$ fine grid, $ri_i$ is the cut-off size for growth event in the $i^{th}$ fine grid computed as follows:

$$ri^3_i = rb^3_i - (\Delta t)\,(\Re_{Growth_i}) \quad (11)$$

where $\Re_{Growth_i}$ is the growth rate during this time instance in bin $i$.

Having accounted for the coagulation phenomenon at the coarser level and the nucleation and growth phenomena at the finer level, the next step is to combine the effects of all the three phenomena. This is done *via* two information exchanges as follows. The information from the finer level is transferred to the coarser level to update the PSD on the coarser level as follows:

$$F^c_{new,j} = F^c_j + F^f_{(j-1)N_{FE}} \frac{Growth_{(j-1)N_{FE}}}{\Delta V_{(j-1)N_{FE}}} - F^f_{jN_{FE}} \frac{Growth_{jN_{FE}}}{\Delta V_{jN_{FE}}} + H(j_{upper} - j)(\Delta t)(\Re_{formation_j}) - H(j_{cut-off} - j)(\Delta t)(\Re_{depletion_j}) \quad (12)$$

In the above equation, $N_{FE}$ is the number of finer grids per coarser grid.

It is imperative to also exchange information from the coarser level to the finer level. This second exchange of information between the levels is achieved in a straightforward manner. Upon updating the particle densities in the coarser grids employing Equation (12), the particle density in each fine grid '$i$' within any given coarse grid $j$ is updated as follows:

$$F^f_{new,i} = F^f_i \frac{F^c_{new,j}}{\Sigma^{N_{FE}}_{k=1} F^f_k} \tag{13}$$

The proposed two-level discretisation algorithm is implemented in the emulsion polymerisation model. In emulsion polymerisation, the coagulation kernel $\beta(r, r')$ is a strongly size-dependent quantity. Thus, instead of calculating the coagulation rate constant at a representative point within each coarse grid, an averaging procedure is employed by involving each of the finer bin within any given coarse bin, as shown in Equation (14). Thus, the coagulation rate constant within any given coarse grid is averaged based on finer-level discretisation, although the actual coagulation computations (the integrals shown in Equations (3) and (4)) are performed at the coarser level.

$$\beta^c_{i,j} = \frac{\int^{V_i}_{V'=V_{i-1}} \int^{V_j}_{V''=V_{j-1}} \beta^f_{V',V''}\, \mathrm{d}V'\, \mathrm{d}V''}{\int^{V_i}_{V'=V_{i-1}} \int^{V_j}_{V''=V_{j-1}} \mathrm{d}V'\, \mathrm{d}V''}$$

$$\beta^c_{i,j} = \frac{\sum^{N_{FE}}_{k=1} \sum^{N_{FE}}_{l=1} \beta^f_{(i-1)N_{FE}+k,(j-1)N_{FE}+l} \times \Delta V g_{(i-1)N_{FE}+k} \Delta V g_{(j-1)N_{FE}+l}}{\Delta V_i \Delta V_j} \tag{14}$$

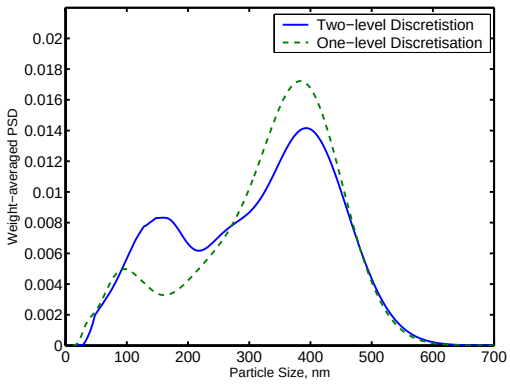



Fig. 7: The comparison of the endpoint weight-averaged PSD based on the one-level discretisation and the two-level discretisation

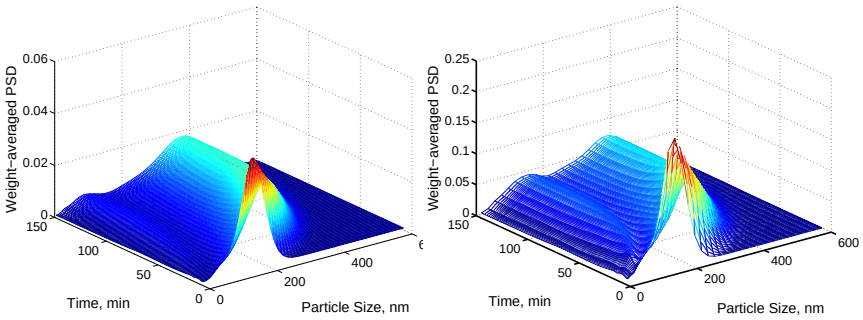



Fig. 8: The comparison of the evolution of PSD based on the one-level discretisation and the two-level discretisation. (a) The evolution of PSD based on hierarchical two-tier algorithm with one-level discretisation. (b) The evolution of PSD based on two-level discretisation algorithm.

Figure 7 shows the comparison of the endpoint weight-averaged PSD based on the hierarchical two-tier algorithm with one-level discretisation and the two-level discretisation algorithm. Note that the plots are averaged distributions, with the errors in absolute distributions being below the measurement accuracy. In the two-level algorithm, the finer grids are based on $\Delta r_f = 2nm$, while the coarser grids are based on $\Delta r_c = 10nm$. The one-level discretisation employs uniform fine grids of $\Delta r = 2nm$ each for the solution of each of nucleation, growth and coagulation models.

Figure 8 shows the evolution of the PSD throughout the course of batch based on one-level discretisation and the two-level discretisation algorithms, which again indicates a very good qualitative agreement. The technique has been validated for several other operating conditions (not presented here). The single-level discretisation results based on the finer discretisation ($\Delta r = 2\ nm$) have been validated previously against experimental results (see [19]). This in turn implies the validity of the present two-level discretisation method.

## 4 Distribution Control Considerations

### 4.1 Problem Definition and Literature Review

The control objective in particulate processes is the tailoring of the PSD to specified targets. A motivational study in this regard is the controllability analysis of Semino and Ray [41, 42], which was followed by several seminal contributions to this field. Some of the major challenges in the development of distribution control for particulate processes are

- limited degrees of freedom, thereby necessitating optimal strategies

- complexity of the underlying optimisation problem such as nonconvexity and possible discontinuity in employing a multi-scale process model for control
- sparsity of measurements
- complexity in the underlying processes such as an integrated and interacting nature of the particle-level processes and irreversibility traits.

To cope with the last issue, a multi-objective strategy lends itself naturally [22]. It seeks to attain the target PSD while providing a decoupling of the underlying particle processes, in an optimal manner with the limited actuation available. This perspective is elucidated in Figure 9, wherein the manipulations at the vessel (process) level are used to tailor particle-level behaviour so as to obtain the desired population distribution. These studies are presented in section 4.2. A representative literature account on distribution control is presented for the rest of this section.

The studies on distribution control can be grouped into the following major classes. Studies that deal with recipe optimisation for batch/semi-batch processes to attain target distributions; studies that address feedback control either in a batch-to-batch sense or in an in-batch sense. Crowley *et al.* [7] present one of the first studies in both open-loop and in-batch feedback control of PSD in emulsion polymerisation processes. Subsequent studies [5, 12] described batch-to-batch feedback control studies also for emulsion polymerisation as a means to account for inevitable model mismatch. Kalani & Christofides [26] present a nonlinear output feedback control in combination with a Luenberger-type observer for aerosol processes involving simultaneous chemical reaction, nucleation, condensation and coagulation. The method of moments is used to derive a low-order model and a presupposition of the form of distribution (log-normal) is made so as to avoid closure problems. Flores-Cerrillo & MacGregor [14] present a mid-course correction strategy as a means of in-batch feedback control. The correction strategy is implemented if significant deviation is expected from the target PSD mid-course of the process. Lee *et al.* [30] present a batch-to-batch iterative control strategy for precipitation processes. The strategy is defined in the typical model predictive control framework and have demosntrated the use of either a nonlinear model or a linearized model to perform the optimization. Park *et al.* [38] describe a model-predictive control applied for on-line in-batch feedback control of PSD, employing low-order partial least squares model for the process. The MPC acts over open-loop optimal control trajectories to account for model-mismatch and process disturbances. Similar open-loop control trajectory generation is presented in other studies [21].

### 4.2 Open-Loop Recipe Optimisation Studies

In this section, open-loop optimisation studies aimed at developing the operating policies for semi-batch emulsion polymerisation to attain target PSD is discussed. Two different strategies are presented as follows

- employing a detailed multi-scale model to directly develop the optimal recipes that match the target PSD
- an exploitation of the multi-phase process physics in combination with a suitable (multi-scale) model to develop the optimal recipe, as depicted in Figure 9.

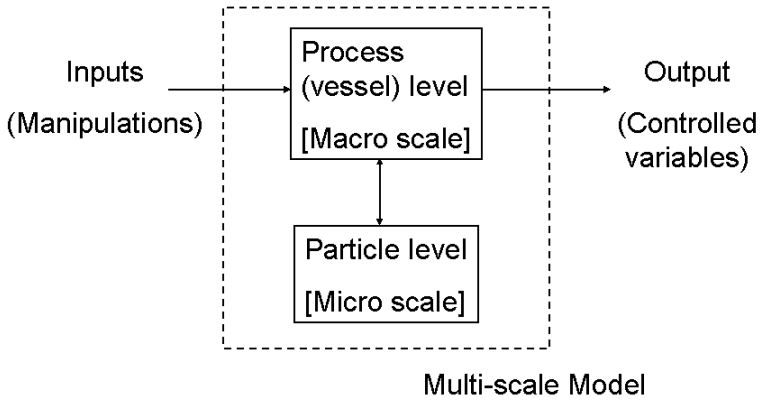



Fig. 9: A hierarchical control srategy proposed for particulate processes

The feed rates of the reagents (surfactant, monomers and initiators) suitably discretised along time into piece-wise constants are used as the decision variables (manipulations). Three objective functions are considered in the present study, to meet the two strategies indicated above: $\theta_1 = \int_{r_{nuc}}^{r_{max}} W(r,t_f) - W_{ref}(r))^2 \ \mathrm{d}r$, $\theta_2 = \int_{t=0}^{t_f} (sc(t) - sc_{ref}(t))^2 \ \mathrm{d}t$ and $\theta_3 = \int_{t=0}^{t_f} w_i(np(t) - np_{ref}(t))^2 \ \mathrm{d}t$. These were used to build four different formulations of the optimisation problem as listed below. Among these, formulations 1-3 fall under the first strategy, while formulation 4 falls under the second strategy.

1. *Single objective* $\quad Minimise\ \theta_1$
2. $Weighted\ sum \quad Minimise\ \theta = w_1\theta_1 + w_2\theta_2 + w_3\theta_3$
3. $Weighted\ min - max \quad Minimise\ \theta = max(w_1\theta_1, w_2\theta_2, w_3\theta_3)$
4. *Pareto optimisation* $\quad Minimise\ \theta_1$ under the conditions:
   a) $\theta_2 \leq \epsilon_2$
   b) $\theta_3 \leq \epsilon_3$
   c) $\theta_2 \leq \epsilon_2 \ \ and \ \ \theta_3 \leq \epsilon_3$

The first formulation is a single objective formulation accounting for the error between the end-point and the target weight-averaged PSDs ($\theta_1$). The next two are multi-objective strategies in which terms accounting for solids

content ($\theta_2$) and number of particles ($\theta_3$) were included. A weighted sum of $\theta_1$, $\theta_2$ and $\theta_3$ and a weighted maximum among $\theta_1$, $\theta_2$ and $\theta_3$ were the objective functions for these two cases, respectively. Thus, each of these formulations rely on the multi-scale model to indirectly tailor the particle-level behaviour so as to meet the process-level behaviour (target PSD).

In the fourth case, a complete set of pareto solutions were also obtained through *epsilon*-constraint formulations applied to one or two of the above objective functions when implemented in the two and three-objective formulation respectively. In view of the highly integrated nature of the process, coupled with the process irreversibility traits with regard to the underlying particle rate processes, it is important to implement a decoupling of the particle rate processes of nucleation and growth for effective feedback control. This is the rationale behind the multi-objective optimization formulation 4. The objectives are the attainment of the target PSD while also meeting target nucleation rates and growth rates. The design nucleation and growth rates are represented as profiles of total particles and solids content over the course of the batch.

The copolymer system vinyl acetate (VAc)- butyl acrylate (BuA) was considered, which is a commercial adhesive polymer. The VAc and surfactant feed rates were used as manipulated variables. The duration of the batch is divided into different time intervals, and these manipulated variables are discretised as piece-wise constants within these time intervals. A Sequential Quadratic Programming (SQP) algorithm is used to solve the underlying optimisation problem. FORTRAN subroutines from the numerical algorithm group (NAG) are employed for this purpose.

As seen in Figure 10, the optimisation formulations 1-3 are effective in identifying operating policies that reaches the target PSD. Figure 11 shows the corresponding sub-optimal feed rates of the reagents.

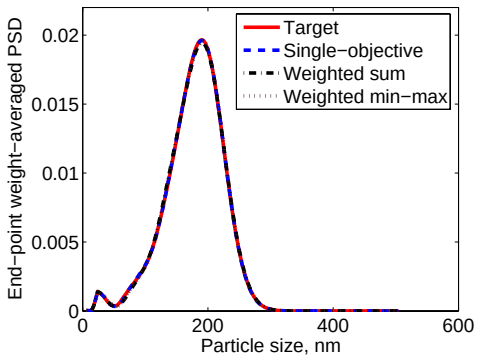



Fig. 10: Comparison of the optimal weight-averaged PSD with the target PSD based on optimisation problem formulations 1-3.

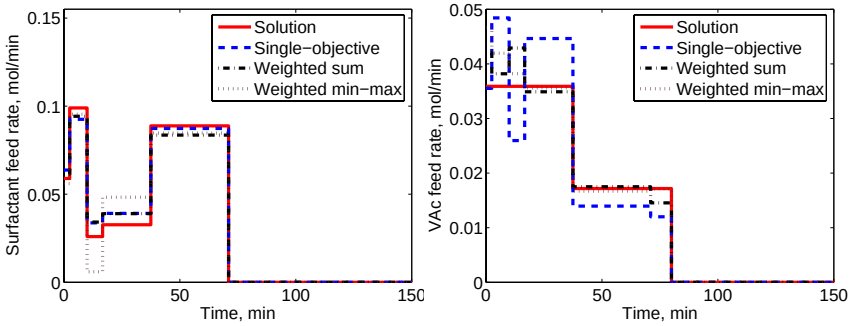



Fig. 11: Comparison of the sub-optimal decision variables with the actual solutions based on formulations 1-3 (a) Surfactant solution (b) Vinyl acetate monomer.

Pareto optimisation studies were performed through formulation 4. The three formulations 4(a), 4(b) and 4(c) indicated above were implemented. Pareto solutions are a set of non-dominated solutions, *i. e.*, each solution is better than every other solution in the set at least with respect to one of the objectives. Of all the pareto solutions computed following the above strategy we picked out only those that were closest to the utopia point. The utopia point is where all the objective functions taken in account assume the corresponding smallest value in the set. One such pareto set is shown in Figure 12 for the case of the two-objective epsilon-constraint problem formulation 4(b).

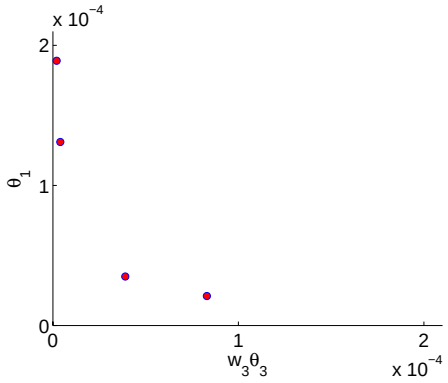



Fig. 12: Set of pareto solutions obtained applying case 4(b).

Figures 13 and 14 show that it is possible to attain the target PSD as well as bring about the required decoupling (independent control of nucleation and growth phenomena) with the limited resources. Further investigation reveals

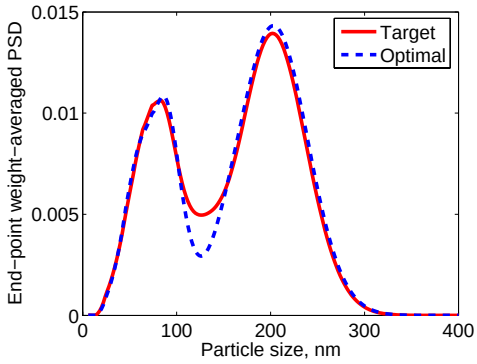



Fig. 13: Comparison of the optimal weight-averaged PSD with the target PSD based on the three-objective optimisation problem formulation 4(c).

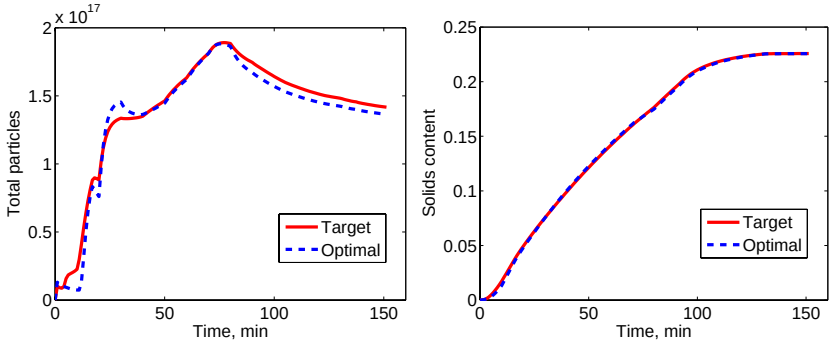



Fig. 14: Comparison of the profiles of total number of particles and solids content obtained from the three-objective optimisation problem formulation 4(c) (a) Total particles (b) Solids content.

that any further reduction in the actuation (the usage of only surfactant feed as the manipulated variable) results in loss of independent control of the nucleation and growth phenomena. Figure 15 shows the input profiles corresponding to the solution presented in Figures 13 and 14.

## 5 Summary and conclusions

In this article, modelling and control studies for population distributions in particulate processes were presented. The models assume a multi-scale character, with the macro-scale being represented by the population dynamics at the vessel (process) level, and the micro/meso scale being represented by the particle behaviour. A population balance is employed at the macro-level, with the particle-level behaviour being *embedded* within, using the terminology of

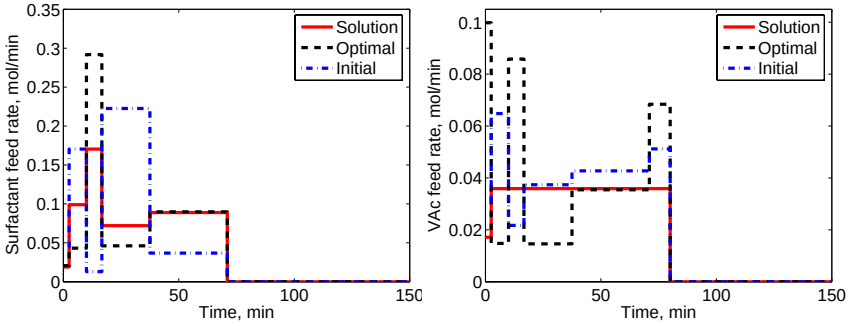



Fig. 15: Comparison of the sub-optimal decision variables with the actual solutions based on the three-objective optimisation problem formulation 4(c) (a) Surfactant solution (b) Vinyl acetate monomer.

Ingram *et al.* [25]. The embedding is achieved by employing kinetic and colloids theory to *transform* particle-level behaviour models into phenomenological laws (constitutive equations). This approach renders the process model feasible for application in process decision making (open-loop optimisation, in-batch and batch-to-batch feedback control).

The embedded multi-scale model is still computation-intense, both due to the terms that appear in the PBE (complex integrals) and the calculations involved in the phenomenological laws. Thus, an efficient solution strategy is presented [23, 44]. This is based on a discretisation of the population distribution into finite elements (bins) and the employment of a two-tier integration framework to incorporate the particle-level behaviour into the population balance. Two hypothetical discretisation are defined, one at a coarser level and the other at a finer level. The computation-intensive but less sensitive coagulation phenomenon is solved at the coarser level, while the more sensitive nucleation and growth phenomena are solved at the finer level. A two-way information exchange is employed to extract the particle size distribution from the two hypothetical distributions.

Finally, on the control side, open-loop recipe optimisation studies are presented for the attainment of target PSD. Two different approaches are presented to solve this problem. The first is a straightforward control of the macro-level behaviour employing the multi-scale model. The second is the manipulation of the particle-level behaviour to control te population distribution. An inferential control strategy is employed in the implementation of the latter approach. This approach also enables decoupling of the particle-level processes and hence facilitate feedback control.

Future work in this topic will be focussed along the following areas:

- Generalisation of the strategies presented here for manipulation and transformation of microscale models to embed them into macroscale process models, to enable efficient process control [24]
- Further extension of the proposed novel implementation of the hierarchical two-tier numerical solution algorithm to multi-dimensional population balance problems
- Development and implementation of feedback control for PSD, both by direct manipulation of the particle-level behaviour and indirectly through the multi-scale models in a model predictive control framework.

## References

1. M.K. Bennett, S. Rohani: Solution of population balance equations with a new combined Lax-Wendroff/Crank-Nicholson method. *Chem. Eng. Sci.* **56**, 6623–6633 (2001)
2. R.D. Braatz, S. Hasebe: Particle Size and Shape Control in Crystallization Processes. In: *AIChE Symposium Series: Chemical Process Control-VI*, vol. 98 of *326*, ed. by J.B. Rawlings, B.A. Ogunnaike, J.W. Eaton, pages 307–327 (2001)
3. T. Chiu and P.D. Christofides: Nonlinear Control of Particulate Processes. *AIChE J.* **45** (6), 1279–1297 (1999)
4. P.D. Christofides: Control of nonlinear distributed process systems: Recent developments and challenges. *AIChE J* **47** (3), 514–518 (2001)
5. T.J. Crowley, C.A. Harrison, F.J. Doyle III: Batch-to-batch Optimization of PSD in Emulsion Polymerization using a Hybrid Model. In: *Proc. 2001 American Control Conf.*, pages 981–986 (Arlington, VA 2001)
6. T.J. Crowley, E.S. Meadows, F.J. Doyle III: Numerical issues in solving population balance equations for particle size distribution control in emulsion polymerization. In: *Proc. 1999 American Control Conf.* (1999)
7. T.J. Crowley, E.S. Meadows, E.Kostoulas, F.J. Doyle III: Control of Particle Size Distribution described by a Population Balance Model of Semibatch Emulsion Polymerization. *J. Proc. Cont.* **10**, 419–432 (2000)
8. P. Daoutidis, M.A. Henson: Dynamics and Control of Cell Populations in Continuous Biorroeact. In: *AIChE Symposium Series: Chemical Process Control-VI*, volume 98 of *326*, ed. by J.B. Rawlings, B.A. Ogunnaike, J.W. Eaton, pages 274–289 (2001)
9. R.B. Diemer, J.H. Olson: A moment methodology for coagulation and breakage problems: Part 1 - analytical solution of the steady state population balance. *Chem. Eng. Sci.* **57**, 2193–2209 (2002)
10. R.B. Diemer, J.H. Olson: A moment methodology for coagulation and breakage problems: Part 2 - moment models and distribution reconstruction. *Chem. Eng. Sci.* **57**, 2211–2228 (2002)
11. R.B. Diemer, J.H. Olson: A moment methodology for coagulation and breakage problems: Part 3 - generalized daughter distribution function. *Chem. Eng. Sci.* **57**, 4187–4198 (2002)
12. F.J. Doyle III, C.A. Harrison, T.J. Crowley: Hybrid model-based approach to batch-to-batch control of particle size distribution in emulsion polymerization. *Comp. Chem. Eng.* **27**, 1153–1163 (2003)

13. F.J. Doyle III, M. Soroush, C. Cordeiro: Control of Product Quality in Polymerization Processes. In: *AIChE Symposium Series: Chemical Process Control-VI*, volume 98 of *326*, ed. by J.B. Rawlings, B.A. Ogunnaike, J.W. Eaton, pages 290–306 (2001)
14. J. Flores-Cerrillo, J.F. MacGregor: Control of Particle Size Distribution in Emulsion Semi-batch Polymerization using Mid-course Correction Policies. *Ind. Eng. Chem. Res.* **41**, 1805–1814 (2002)
15. K. Gadkar, J. Varner, F.J. Doyle III: Model Identification of Signal Transduction Networks from Data Using a State Regulator Problem. *IEE Systems Biol.* **2**, 17–30 (2005)
16. M.J. Hounslow, R.L. Ryall, V.R. Marshall: A discretized population balance for nucleation, growth and aggregation. *AIChE J.* **34**, 1821–1832 (1988)
17. Q. Hu, S. Rohani, A. Jutan: Modelling and optimization of seeded batch crystallizers. *Comp. Chem. Eng.* **29**, 911–918 (2005)
18. C.D. Immanuel: Population Balance Model for Cellular Processes in Biological Systems: Biochemical and Biomedical Applications (Cambridge, MA 2004)
19. C.D. Immanuel, C.F. Cordeiro, S.S. Sundaram, F.J. Doyle III: Population Balance PSD Model for Emulsion Polymerization with Steric Stabilizers. *AIChE J.* **49** (6), 1392–1404 (2003)
20. C.D. Immanuel, C.F. Cordeiro, S.S. Sundaram, E.S. Meadows, T.J. Crowley, F.J. Doyle III: Modeling of Particle Size Distribution in Emulsion Co-Polymerization: Comparison with Experimental Data and Parametric Sensitivity Studies. *Comp. Chem. Engng.* **26** (7-8), 1133–1152 (2002)
21. C.D. Immanuel, F.J. Doyle III: Open-loop Control of Particle Size Distribution in Semi-batch Emulsion Co-Polymerization using a Genetic Algorithm. *Chem. Eng. Sci.* **57** (20), 4415–4427 (2002)
22. C.D. Immanuel, F.J. Doyle III: A Hierarchical Strategy for Control of Particle Size Distribution Using Multi-Objective Optimization. *AIChE J.* **49** (9), 2383–2399 (2003)
23. C.D. Immanuel, F.J. Doyle III: Computationally-Efficient Solution of Population Balance Models Incorporating Nucleation, Growth and Coagulation. *Chem. Eng. Sci.* **58** (16), 3681–3698 (2003)
24. C.D. Immanuel, F.J. Doyle III: Solution Technique for a Multi-dimensional Population Balance Model Describing Granulation Processes. *Powder Tech.* **156**, 213–225 (2005)
25. G.D. Ingram, I.T. Cameron, K.M. Hangos: Classification and analysis of integrating frameworks in multiscale modelling. *Chem. Eng. Sci.*, **59**, 2171–2187 (2004)
26. A.Kalani, P.D. Christofides: Nonlinear control of spatially inhomogeneous aerosol processes. *Chem Eng. Sci* **54**, 2669–2678 (1999)
27. C.Kiparissides, A. Alexopoulos, A. Roussos, G. Dompazis, C. Kotoulas. Population balance modeling of particulate polymerization processes . *Ind. Eng. Chem. Res.* **43**, 7290–7302 (2004)
28. S. Kumar, D. Ramkrishna. On the Solution of Population Balance Equations by Discretization - II. A Moving Pivot Technique. *Chem. Eng. Sci.* **51** (8), 1333–1342 (1996)
29. S. Kumar, D. Ramkrishna: On the Solution of Population Balance Equations by Discretization-I. A Fixed Pivot Technique. *Chem. Eng. Sci.* **51** (8), 1311–1332 (1996)

30. K. Lee, J. H. Lee, D.R. Yang, A.W. Mahoney: Integrated run-to-run and on-line model-based control of particle size distribution in semi-batch precipitation reactor. *Comp. Chem. Engng.* **26** (7-8), 1117–1131 (2002)
31. Y. Liu, I.T. Cameron: A new wavelet-based method for the solution of the population balance equation. *Chem. Eng. Sci.* **56** 5283–5294 (2001)
32. D.L. Ma, D.K. Tafti, R.D. Braatz. High resolution simulation of multi-dimensional crystallization. *Ind. Eng. Chem. Res.* **41**, 6217–6223 (2002)
33. A.W. Mahoney, D. Ramkrishna: Efficient solution of population balance equations with discontinuities by finite elements. *Chem. Eng. Sci.* **57**, 1107–1119 (2002)
34. D.L. Marchisio, J. Pikturna, L. Wang, R.D. Vigil, R.O. Fox: Quadratic Method of Moments for Population Balances in CFD Applications: Comparison with Experimental Data. *Chem. Engg. Trans.* **1**, 305–310 (2002)
35. E.S. Meadows, T.J. Crowley, C.D. Immanuel, F.J. Doyle III: Non-isothermal Modeling and Sensitivity Studies for Batch and Semi-batch Emulsion Polymerization of Styrene. *Ind. Eng. Chem. Res.* **42** (3), 555–567 (2003)
36. S. Melis, M. Kemmere, J. Meuldijk, G. Storti, M. Morbidelli: A model for the coagulation of polyvinyl acetate particles in emulsion. *Chem. Eng. Sci.* **55**, 3101–3111 (2000)
37. K. Mitra, K. Deb, S.K. Gupta: Multiobjective dynamic optimization of an industrial nylon 6 semibatch reactor using genetic algorithm. *J. Applied Polym. Sci.* **69**, 69–87 (1998)
38. M-Y Park, M.T. Dokucu, F.J. Doyle III: Regulation of the Emulsion Particle Size Distribution Using Partial Least Squares Model-Based Predictive Control. *Ind. Eng. Chem. Res.* **43**, 7227–7237 (2004)
39. D. Ramkrishna: *Population Balances* (Academic Press, San Diego 2000)
40. D. Ramkrishna, A.W. Mahoney: Population balance modeling. Promise for the future. *Chem. Eng. Sci.* **57**, 595–606 (2002)
41. D. Semino, W.H. Ray: Control of Systems Described by Population Balance Equations - I. Controllabity Analysis. *Chem. Eng. Sci.* **50** (11), 1805–1824 (1995)
42. D. Semino, W.H. Ray: Control of Systems Described by Population Balance Equations - II. Emulsion Polymerization with Constrained Control Action. *Chem. Eng. Sci.* **50** (11), 1825–1839 (1995)
43. J. Stelling, U. Sauer, Z. Szallasi, F.J. Doyle III, J. Doyle: Robustness of Cellular Functions. *Cell* **118**, 675–685 (2004)
44. N. Sun, C.D. Immanuel: Efficient solution of population balance models employing a hierarchical solution strategy based on a multi-level discretization. *Trans. Inst. Meas. Contr.* **5**, 347–366 (2005)

# Diagnostic Goal-Driven Reduction of Multiscale Process Models

E. Németh[1], R. Lakner[2], and K. M. Hangos[1]

[1] Process Control Research Group, Systems and Control Research Laboratory, Computer and Automation Institute HAS H-1518 Budapest, P.O. Box 63, Hungary, {nemethe,hangos}@scl.sztaki.hu
[2] Department of Computer Science, University of Veszprém H-8201 Veszprém, P.O. Box 158, Hungary, lakner@almos.vein.hu

**Summary.** Fault detection and diagnosis in large-scale process systems is of great practical importance and present various challenging research problems at the same time. One of them is the computational complexity of the algorithms that causes an exponential growth of the computational resources (time and memory) with increasing system sizes [21]. One remedy of this problem is to decompose the system model and effectively focus on its relevant sub-model when doing the fault detection, isolation and loss prevention.

Multi-scale modelling is an emerging interdisciplinary field that offers a systematic way of constructing, analyzing and solving dynamic models of large-scale complex systems [22]. The aim of this paper is to propose a model reduction approach based on multi-scale modelling of process systems for diagnostic purposes. Because lumped or concentrated parameter process models are the most important and widespread class of process models for control and diagnostic applications, therefore we also restrict ourselves to this case.

## 1 Introduction

### 1.1 Multiscale Modelling in Process Systems Engineering

The field of multiscale modelling is quite broad that spans many disciplines, including physics, chemistry, bio-chemistry, mathematics, statistics, image processing, chemical and mechanical engineering, as well as materials science. Some of them, like physics, chemical kinetics (see e.g. [5], [6]) and image processing use advanced bottom-up "coarse graining" techniques to decompose the system into different in magnitude time scales and simplify the solution by using the fast scale and slow scale dynamics separately. Here the starting point of the analysis and reduction is a detailed model on a small length scale that is decomposed into two models, one for the fast and another for the slow time scale, respectively.

In engineering one uses a different, top-down approach to construct a multi-scale model that provides an effective way both of decomposing and handling the available information in large-scale complex systems. Here one first identifies the relevant scales, most often the relevant length or detail scales for the problem, constructs appropriate sub-models for each scale and then somehow organizes the information exchange between these sub-models to obtain a multiscale model.

From the viewpoint of dynamic simulation, the problem of multiscale modelling is seen as how to design a simulator architecture that employs dynamically coupled codes in such a way that no numerical instabilities occur [20], [3]. The methods and tools of linear and nonlinear systems theory has been applied there to develop a sufficient condition for the numerical stability of the coupled codes and to design suitable filters for the coupling.

The systematic multiscale modelling approach in process systems engineering (see e.g. [9], [4], [10]) provides us with a natural mechanism-driven hierarchical decomposition of the underlying process model with any related information and an integration framework to organize the information exchange between the partial models. This is the multiscale modelling methodology that forms the basis of our approach in this paper.

### 1.2 Reduction of Process Models

The most effective way of focusing to a part of a dynamic system relevant to our purposes is to apply model reduction or model simplification techniques. Many of such is reported in the systems and control engineering, as well as in the process systems engineering literature both for linear and nonlinear models. The engineering approach to model reduction starts with a detailed model that is intended to be used for a given purpose. Then one can try to leave out the unnecessary parts in the model, that is, to reduce the model such that the properties relevant for the given purpose still remain unchanged.

In linear system theory, there is a well-known model reduction technique for linear time-invariant state-space models that is based on controllability and observability indices, called gramians and on linear state transformations to construct a balanced realization. This method is purely black-box in its nature because the physical meaning of the state variables in the reduced model is completely lost. [7] generalize this method for stable nonlinear systems for nonlinear model-based predictive control purposes.

The reduction of nonlinear lumped dynamic process models that exhibit multiple time scale behaviour is usually done by using singular perturbation technique [17]. The singular and regular perturbation analysis gives an insight into the system structure and thereby one can eliminate superfluous model equations or reduce them to other forms [19].

The same singular perturbation analysis was used to investigate the nonlinear dynamics of process systems with recycle [14] and that of integrated process networks with multi-rate reactions [1]. The engineering conditions of

time scale separation in these process systems have been established and reduced models have been developed from the overall model for the fast and slow time scales separately.

Model structure simplification methods [16] offer a grey-box alternative to model reduction where the number of state variables is reduced using steady-state and/or variable lumping transformations and the physical meaning of the remaining state variables remains unchanged. The model reduction approach proposed in this paper can be regarded as an extension to this method.

Almost all of the reported model reduction and simplification methods apply analytical or combinatorial techniques that are difficult to apply in a multiscale context where naturally both analytical and qualitative methods are used. Therefore, our aim was to propose a mixed method for model reduction of multiscale process models for diagnosis that effectively utilizes the model structure inherent in the multiscale nature of the model.

## 2 The Model Reduction Problem

The basic notions for formulating and solving the model reduction problem are briefly described in this section.

### 2.1 Process System Models and their Modelling Goals

*Process systems* are a sub-class of systems that *obey the law of thermodynamics*. This implies that dynamic process models based on first engineering principles are constructed by using the first and second law of thermodynamics.

The construction of process models starts by formulating a *modelling problem statement* in the following general form.

**Given** : a process system together with a ***modelling goal***
that can be process design, process control or diagnosis to mention just a few,
**Construct** : a model ***for the modelling goal***.

**Construction principle.** Process models for control and diagnostic purposes are constructed from the dynamic conservation balances of mass, component masses and energy (see e.g. [8]). The terms in the conservation balances correspond to the various *mechanisms* that are taken into account:

- in-convection (inlet term),
- out-convection (outlet term),
- (interphase) transfer,
- sources (and sinks including chemical reactions)

**Modelling goal and its effect on the model.** It is important to emphasize that the modelling goal has a determining effect on the model to be constructed: it determines the model performance variables, the mechanisms to be taken into account and the desired accuracy of the model in terms of its performance variables, as well as the type (static – dynamic, deterministic – stochastic, lumped parameter – distributed parameter) of the model. There is no matured systematic way of how to take the effect of the modelling goal on the model into account, just a few early steps about the "goal-directed modelling" [15] is available.

### 2.2 The Structure of Process Models

Lumped process models that are constructed on first engineering principles possess a well defined structure. The *model equations* form a set of DAEs, where the

- *conservation balances are the differential equations* that are equipped with
- constitutive equations being algebraic equations.

The structure of the model equations gives rise to the classification of the *variables of a state-space model* that has been derived from a process model based on first engineering principles.

- The *state variables* are the intensive pairs of conserved extensive quantities (mass, concentrations, temperature) [8] for which conservation balances are constructed, while
- the *inputs, outputs and disturbance* variables are problem formulation dependent.

### 2.3 Functionally Equivalent Process Models

Two process models of the same process system are called functionally equivalent, if they fulfill the same modelling goal [15]. If the modelling goal is formulated in terms of performance indices that are predicates defined on the performance variables, than the functional equivalence can be determined algorithmically.

Suitably defined size indices that give the size of a model in a generalized sense can be used to relate functionally equivalent models and to determine which one is of less size, i.e. "reduced" compared to the other. Fig. 1 shows how the performance and size indices relate to each other.

### 2.4 The Problem Statement of Model Reduction

The aim of model reduction is to find a "simple" model that is functionally equivalent with a given detailed "original" one but it is easier to handle computationally, i.e. it is of smaller size than the original one.

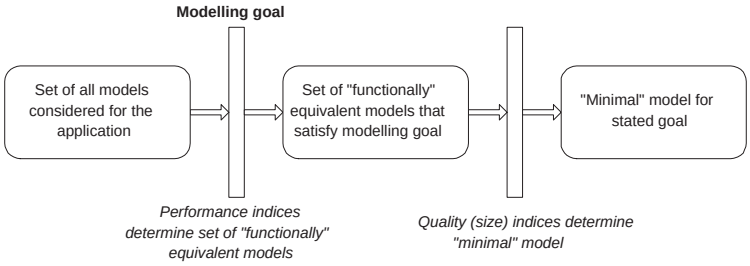



Fig. 1: Functionally equivalent models and their reduction

If one wants to obtain reduced models for diagnostic purposes, then the following assumptions can be made about the process system and its model:

$\mathcal{A}1$ Lumped dynamic process models are considered for both the original and the reduced model.

$\mathcal{A}2$ The process system has smooth nonlinearities and we want to describe its behaviour in the neighbourhood of a steady-state operating point.

Under the above assumptions it is possible to formulate the problem statement of model reduction to this special case as follows.

**Given**:

- a *detailed process model* that originates from first engineering principles
- a set of *prescribed input-output scenarios* that define the required performance of the models,
- a set of *simplicity indices* as generalized size indices being the number of state variables (dynamic conservation balances) and the linearity of the model

**Compute**: the *"simplest" reduced model* that is

- functionally equivalent with the detailed process model in terms of the prescribed input-output scenarios, and it is
- of the smallest size with respect to the simplicity indices.

The reduced model with the above two properties can be called *minimal model* [15] with respect to the specified simplicity indices. Note that minimal models are not necessarily unique.

### 2.5 Elementary Model Reduction Steps

Under the assumptions $\mathcal{A}1$ and $\mathcal{A}2$ one can consider two basic elementary model reduction steps that are applicable to nonlinear state-space models: the

reduction of the number of state equations by separating the different time-modes (fast, medium and slow modes, for example) and the linearization of the state equations around a steady-state point.

**Reduction of the number of state equations.** State equations originate from dynamic conservation balances in a process model derived from first engineering principles with the constitutive equations substituted. If balances for different conserved quantities, such as component masses and energy, or balances over balance volumes with largely different total holdup are present, then one observes the *separation of the time-constants*. There are also formal methods based on singular perturbation analysis (see e.g. [14], [1] and the references therein) to find the separated characteristic times and to obtain dynamic models from the original detailed model separately for each time level by model reduction.

If one considers the elementary steps of model reduction as model transformations, specifically projections acting on the set of lumped dynamic models of a process system, then the transformation applicable in the case of time-scale separation can be characterized by a ***steady-state assumption for a "fast" or "slowly changing" variable*** $\chi$. Formally one applies the transformation

$$\frac{\mathrm{d}\chi}{\mathrm{d}t} \approx 0 \tag{1}$$

to a set of model equations. As a result of the transformation, the differential equation that originates from the dynamic conservation balance for $\chi$ becomes algebraic, and thus it should be substituted to the remaining differential equations. Thus the number of state equations (and the number of state variables) decreases by one, and $\chi$ formally disappears from the equations (see [16] for the details).

**Linearization.** Linearization around a steady-state point is a basic and well-known model simplification (or reduction) operation [8]. The applicability condition is that the original model should have smooth nonlinearities and the linearized model will be valid only within a more-or-less narrow region of a nominal operating point.

One may apply linearization for the total model or for a part of the model only depending on the nonlinearities along the variable-coordinate directions.

**Properties of the elementary model reduction steps.** It can be shown that both the reduction of the number state equations and the linearization as model transformations possess the following basic properties:

- They *preserve the basic dynamic properties* (e.g. controllability, observability and stability) of the model [16].
- Their actual result is strongly *steady-state point dependent*.
- They are both applicable in the *"initial phase" of a fault* when the system is in the neighbourhood of a steady-state nominal operating point.

## 3 Prediction-Based Diagnosis and Loss Prevention

Prediction of a system's behaviour is used for deriving the consequences of a state of the system in time that is usually performed in process engineering by dynamic simulation. With the help of prediction, however, the faulty mode of the system can also be detected based on the comparison between the real plant data and the predicted values generated by a suitable dynamic model. This type of fault detection and diagnosis is called prediction-based diagnosis [21]. Because of the high complexity of multiscale process models, however, the computational load of performing the prediction can be substantial. Therefore, the need of model reduction arises.

### 3.1 Modelling Goal for Diagnosis

If one intends to construct a process model for diagnostic purposes this modelling aim has important implications on the model and its variables as follows.

- The model should be dynamic and should be able to produce *dynamic input-output behaviour* with the measurable quantities as output variables.
- One usually defines *symptoms* from the measurable output signals, which are qualitative performance output signals of the model.
- The *actuator input* variables correspond to manipulable input variables, that can be used for preventing dangerous consequences of the considered faults.
- Root causes are usually considered as *disturbances* that may determine the "failure modes" of the system with possibly different process models (i.e. a hybrid model is often needed).

### 3.2 Elements of Prediction-Based Diagnosis

The most important notions of prediction-based diagnosis are briefly summarized here.

**Symptoms.** Similarly to medical diagnosis, the diagnosis of process systems is usually based on symptoms. Loosely speaking, symptoms are deviations from a well-defined "normal behaviour". Symptoms are formally described by using Boolean-valued *predicates* that contains a *measurable variable*, e.g. $T$, such as

$$p_{T_{low}} = (T < T_{ss})$$

Because of the dependence of the symptom on a measurable signal, its value is *time-dependent*, and can be regarded as a qualitative-valued *performance output* of the process model.

A *family of symptoms* is a set of symptoms that are defined over the same measurable variable.

**Diagnostic scenario.** Similarly to an input-output scenario that is a finite record of related input and output signals, a *diagnostic scenario is a timed sequence of symptoms from the same family* (i.e. over the same measurable output signal).

If one associates to the underlying measurable variable the symptoms defined thereon as qualitative variables, then a *diagnostic scenario can be regarded as a qualitative-valued output signal* of the system.

**Root cause.** In model-based fault detection and diagnosis one usually assigns a so called root cause to every faulty mode of the system, the variation of which acts as a cause of the fault. Root causes are most often not measurable and have discrete value (indicator variables) thus a root cause is described as an unmeasurable *disturbance* in a process model for diagnosis.

**Preventive action.** If a fault occurs it is usually possible to take actions in the initial phase of the transient to avoid serious consequences or to try to drive the system back to its original normal state. Dedicated *input signal(s)* serve for this purpose separately for each fault (identified by its root cause) where the *preventive action is a prescribed scenario for the manipulated input signal*.

### 3.3 Prediction for Diagnosis

In model-based diagnosis *the model of the process system is assumed to describe the behaviour of the system in each of the considered faulty mode.* This is a quite severe requirement, thus one usually narrows the domain of the model by ***assuming that the "normal" operating mode of the system is steady-state and only the initial deviations are considered***.

The model is used for predicting *diagnostic scenarios* of the measured output variables for at least two purposes related to diagnosis.

- **Fault isolation**
  When the occurrence of a fault is detected the first task is to find out which is the root cause of the fault, i.e. to isolate the fault. For this task one uses the observed *diagnostic scenario* and tries to match it with the generated diagnostic scenarios by using *every possible root cause*. The generation of the possible diagnostic scenarios can be done by model-based prediction using the dynamic model of the process system.
- **Testing preventive actions**
  Having isolated the fault, i.e. assigned a root cause to the observed diagnostic scenario, one has to determine the course of actions to remedy the situation. This can be done by performing *"what-if"* type prediction by using the dynamic model of the process system and applying *every possible action*. The selection of the suitable preventive action can then be performed by comparing the final state of the system with the "normal" operation.

Because process systems are highly nonlinear and their model can be drastically changed depending on the actual fault mode, simple **reduced** models are needed separately for every *(root cause, diagnostic scenario, preventive action)* triplets.

### 3.4 External Diagnostic Knowledge: HAZOP and FMEA

**HAZOP.** Hazard and operability study (HAZOP), formalized by Imperial of Chemical Industries (ICI) at the end of the 1960s, is the most widely used methodology for hazard identification. HAZOP [13] is a systematic procedure for determining the causes of process deviations from normal behaviour and their consequences. The main idea behind HAZOP is that hazards in process plants arise as a result of deviations from normal operating conditions. A group of experts systematically identifies every conceivable deviations in a plant, finds all the possible abnormal causes, and the adverse consequences of that deviation. During the HAZOP these deviations are systematically analyzed by applying guide expressions (for example NONE, MORE OF, LESS OF, PART OF, MORE THAN, OTHER, ...) in conjunction with process variables and parameters. Driven by these guide words, failure causes and their effects are listed in a systematical way.

The results of the HAZOP analysis are collected in a HAZOP result table. A HAZOP analysis table (the structure of which is shown in Table 1) defines logical (static) cause-consequence relationships between symptoms and potential causes that can be traced to root causes of the deviation. These can be used for fault detection and isolation.

| **Guide word** | **Deviation** | **Possible causes** | **Consequences** | **Action required** |
|---|---|---|---|---|
| Fresh Feed Flow | NONE | (1) Feed hopper empty | ◇ loss of production<br>◇ shift in GSD<br>◇ decrease in recycle and output | a) feed the hopper<br>b) check the hopper |
| | | (2) Feed chute blockage | ◇ Covered by (1) | c) check the hopper |
| | ⋮ | ⋮ | ⋮ | ⋮ |

**Symptoms** : Guide word $\bigoplus$ Deviation
**Root Cause** : Possible causes
**Action** : Preventive actions
**Scenario** : Consequences

Table 1: An example of a HAZOP result table

**FMEA.** Fault mode and effect analysis (FMEA) [12] is a qualitative analysis of hazard identification, universally applicable in a wide variety of industries.

Its use in the process industries has been more limited with HAZOP as one of the main contenders for the preferred hazard identification tool. FMEA is a tabulation of each piece of equipment, noting the various modes by which the equipment can fail, and the corresponding consequences (effects) of the failures. FMEA focuses on individual components and their failure modes. Thus, each failure mode is only considered once, and all of its effects and controls are listed together. This allows a more accurate assessment of the risk associated with each component failure

## 4 Multiscale Process Models and Model Reduction

In the case of large and/or complex systems, the use of a multiscale modelling [9, 10] approach is recommended. The basis of multiscale modelling is to divide a complex problem into a family of sub-problems that exist at different scales. Multiscale models of a system can be organized along various scales depending on the system and on the intended use of the model. Generally, we distinguish between the length, time and detail scales, but diagnosis requires having a multiscale model organized along the time scale [18].

A multiscale model is then an ordered collection of partial models or sub-models that are connected by a so-called *integration framework*. The serial, subroutine-like organization [9], i.e. the simplest way the integration framework integrates the partial models, is used in this paper.

### 4.1 Multiscale Modelling: the Length and Time Scales

Traditionally, multiscale models are built along the length scale because the mechanisms that determine a model based on first engineering principles drive the model building. The levels of the multiscale model are found if one looks on the separation of the characteristic scales, if such separation exists.

In the case of process plant we generally distinguish at least the *molecular level* for chemical kinetics, *particle level* when applicable, *operating unit level* and *plant level* along the length scale. We try, if possible, build a multiscale model where we can solve the sub-models or partial models separately, i.e. when the sub-models are integrated using a *serial integration framework* [9]. This means that the solution of a sub-model in a lower level is used for developing so called *correlations*, i.e. static algebraic relationships between the variables on the higher level. In a recent PhD thesis [11] the kernel functions of the granule bed level model are found by solving the equations of motion for a set of granules in the bed.

For model-based diagnosis, however, we focus on the time-dependent behaviour of our system, thus we have to arrange our sub-models along the *time scale*. Fortunately, the characteristic times that define the levels along the time scale in a multiscale model usually follow the separation of characteristic lengths, i.e. the characteristic time belonging to a particular length scale level

is an order of magnitude larger than that belonging to a higher length level. For example, the characteristic time on a molecular level is in the order of seconds, while the characteristic time constants on an operating unit level are in the order of minutes (when particle level does not exist).

For a particular process system one can construct a so-called ***scale-map*** that relates the identified time and length scales and connects them to the mechanisms considered in the model. Fig. 3 shows an example of a scale-map constructed for a granulator circuit.

It will be important for model-based diagnosis of multiscale process systems, that ***each variable is associated to a level in a model hierarchy*** that is determined by the mechanism or governing conservation balance/constitutive equation the particular variable is assigned to (i.e. is determined by). One can refine the scale-map of a multiscale process system by denoting regions associated to particular variables that are important from the viewpoint of diagnosis. Such a refined scale-map is called the ***model structure map***. An example of such a refined scale-map is seen in Fig. 6.

### 4.2 Diagnosis of Multiscale Models: the Model Reduction Problem

As we have already seen in subsection 3.2, there are at least three characteristic variables that determine a *diagnostic scenario*: a root cause, a symptom variable that generates the symptoms and an input signal that is used for implementing preventive actions. Associated to these variables we have the following levels of interest along a scale, usually along the length scale:

- root *cause* level,
- *target* or symptom level,
- *control* or preventive action level.

Of course, the ideal case is when all the above three levels coincide, that is we select symptom variables and preventive actions from the same level as the root cause level is on.

Now we are ready to formulate the variant of the problem statement of model reduction applicable for multiscale models as follows.

**Given:**

1. a *multiscale process model* that is able to describe the system's behaviour under the considered faulty model (generated by the set of root causes),
2. a nominal *steady-state operating point*,
3. a *(root cause, symptom variable, preventive action)* triplet,
4. *reference input-output behaviour*: in the form of diagnostic scenarios with and without preventive actions,
5. *simplicity index*: the number of state variables and linearity.

**Determine:**

- a *reduced process model* on a single time scale level that is linear and generates the given reference behaviour for fault effect prediction.

**Conceptual steps of the solution:**

(1) Determine the *target level on the time scale* from the variables present in the symptoms of interest. If there is more than one time-level corresponding to the target length level, choose the one that corresponds to the given diagnostic scenarios.
(2) Select the dynamic conservation balances from the target, cause and control levels from the length scale that belong to the target level on the time scale. These together with their constitutive equations form the *scale-reduced nonlinear model*.
(3) Linearize the above nonlinear model (possibly in a DAE form) around the given nominal steady-state operating point. Form a standard linear state-space model by substituting the algebraic equations into the differential ones. The obtained model determines the structure of the *scale-reduced linear model*.
(4) Determine the model parameters of the above obtained scale-reduced linear model by standard parameter estimation methods either by using real measured data or by using data generated by simulation.

## 5 Case Study: Model Reduction of a Granule Bed for Diagnosis

The proposed model reduction method is demonstrated on a commercial fertilizer granulation drum example, where the granulator system is used for the production of mono-ammonium and di-ammonium phosphate (MAP, DAP) [4].

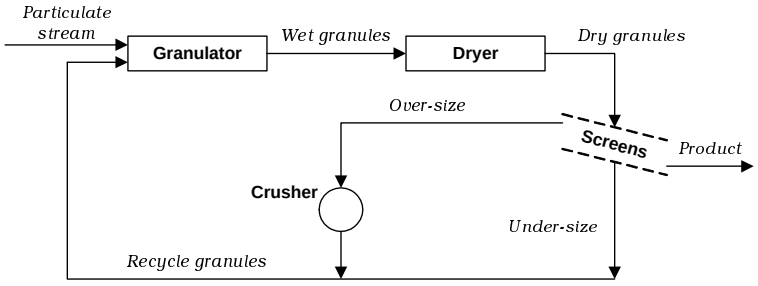



Fig. 2: A flowsheet of a granulator circuit

A typical granulation plant consists of a granulator drum, a dryer, screens and a crusher arranged in a granulator circuit as seen in Fig. 2. Details of the technology and the equipments involved are found elsewhere (see e.g. [2, 4]).

### 5.1 Scale-Map of a Granulator Circuit

If one considers the characteristic lengths or times of the various phenomena taking part in a granulator circuit then 5 scale levels can be distinguished along both scales as shown in Fig. 3 (adopted from [10]). The lower 4 levels belong to the granulator drum itself, where we concentrate our attention.

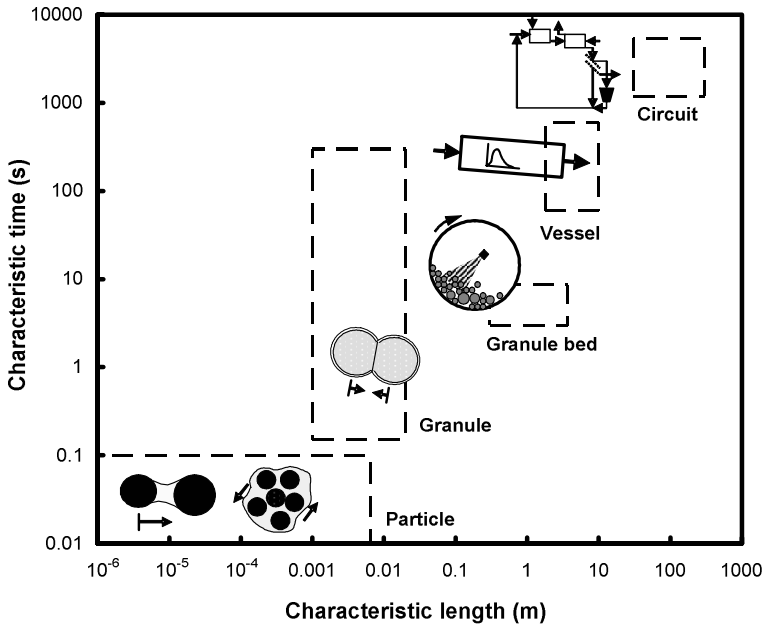



Fig. 3: Scale-map of a granulator circuit

It is important to observe that there is a close relationship between the characteristic length and time scales of a granulator circuit. Generally, *the higher the level of length is the higher level of characteristic times corresponds to it.* An exception to this rule is the time-length scale relationship of the levels **Granule bed**, **Vessel** and **Granule**. Here the mechanisms characterizing the granules have a direct substantial influence on the dynamic behaviour of the variables on higher levels.

There is a clear separation of the characteristic lengths with a little overlap between the **Particle** and **Granule** levels. This enables to separate the

overall model into sub-models along the time scale, this is how the multiscale modelling of granulation processes takes place (see e.g. [11])

### 5.2 Granulator Drum: Levels, Variables and Mechanisms

As seen in Fig. 3 there are four scales or levels that can be identified in a granulator drum as follows.

- **Vessel** level: the whole vessel
  *Variables*: concentrations in the granules (pseudo-solid) and in the binder (liquid) phases
  *Mechanisms*: convection and phase transfer
- **Granule bed** level: slice of the vessel
  *Variables*: particle size distribution (PSD) of granules, *component*: a size range
  *Mechanism*: phase transfer, particle (solids) convection
- **Granule** level: a single granule
  *Variables*: size and composition of the granules together with their position in time
  *Mechanisms*: reaction, agglomeration, breakage, growth, coupled with collisions with each other
  and with the mechanical parts of the equipment
- **Particle** level: particle and binder
  *Variables*: size, shape and porosity of the particles
  *Mechanisms*: inter-granule processes, adsorption-desorption, reaction on the surface etc.

**Variables and symptoms of the granulator drum model**
Table 2 shows the identified variables and symptoms of the drum based on the results of the HAZOP studies.

Based on the scale-map of the granulator drum one can associate the symptoms and their variables listed in Table 2 to the levels of the multiscale model seen in Fig. 4.

### 5.3 Granulator Drum: the Model Structure Map

In order to develop the refined scale-map of the granulator drum to be used for identifying the target, cause and control levels on the time scale we need to have the detailed model of the system. The following model equations are considered adopted from [2].
*Component mass balances in the liquid phase*:
the change of *MAP, DAP* and $H_2O$ mass over time

$$\frac{\mathrm{d}m_{MAP}}{\mathrm{d}t} = F_{L,in}^{MAP} + F_{SL}^{MAP} - F_{L,out}^{MAP} - c_1 \cdot r_{MAP/DAP} \tag{2}$$

| Variable | Symptom |
|---|---|
| Binder Flow | NONE |
| Binder Flow | MORE |
| Binder Flow | LESS |
| Binder Viscosity | MORE |
| Binder Viscosity | LESS |
| Solids Feed PSD | NARROW |
| Solids Feed PSD | WIDE |
| Solids Feed Flow | NONE |
| Solids Feed Flow | MORE |
| Solids Feed Flow | LESS |
| Solids Feed Size | MORE |
| Solids Feed Size | LESS |
| Granulator Drum Speed | NONE |
| Granulator Drum Speed | MORE |
| Granulator Drum Speed | LESS |
| Granulator Exit Distribution | NARROW |
| Granulator Exit Distribution | WIDE |
| Granulator Exit Flow | NONE |
| Granulator Exit Flow | MORE |
| Granulator Exit Flow | LESS |
| Granulator Exit Size | MORE |
| Granulator Exit Size | LESS |

Table 2: The list of variables and symptoms connected to the drum

$$\frac{\mathrm{d}m_{DAP}}{\mathrm{d}t} = F_{L,in}^{DAP} + F_{SL}^{DAP} - F_{L,out}^{DAP} - \dot{m}_{crystals} + c_2 \cdot r_{MAP/DAP} \tag{3}$$

$$\frac{\mathrm{d}m_{H_2O}}{\mathrm{d}t} = F_{L,in}^{H_2O} + (1-\varphi)F_{SL}^{H_2O} - F_{L,out}^{H_2O} - F_{H_2O}^{evap} \tag{4}$$

*Overall mass balances*

- liquid phase:

$$\frac{\mathrm{d}M_L}{\mathrm{d}t} = F_{L,in} + F_{SL} + F_{NH_3} - F_{H_2O}^{evap} - F_{L,out} - \dot{m}_{crystals} \tag{5}$$

- solid phase:

$$\begin{aligned}\frac{\mathrm{d}M_S(i)}{\mathrm{d}t} &= F_{S,in}(i) + F_{SL}^{MAP,sol}(i) + F_{SL}^{DAP,sol}(i) - F_{S,out}(i) + \\ &+ \dot{m}_{crystals}(i) + Agg(i) + Lay(i) - Break(i)\end{aligned} \tag{6}$$

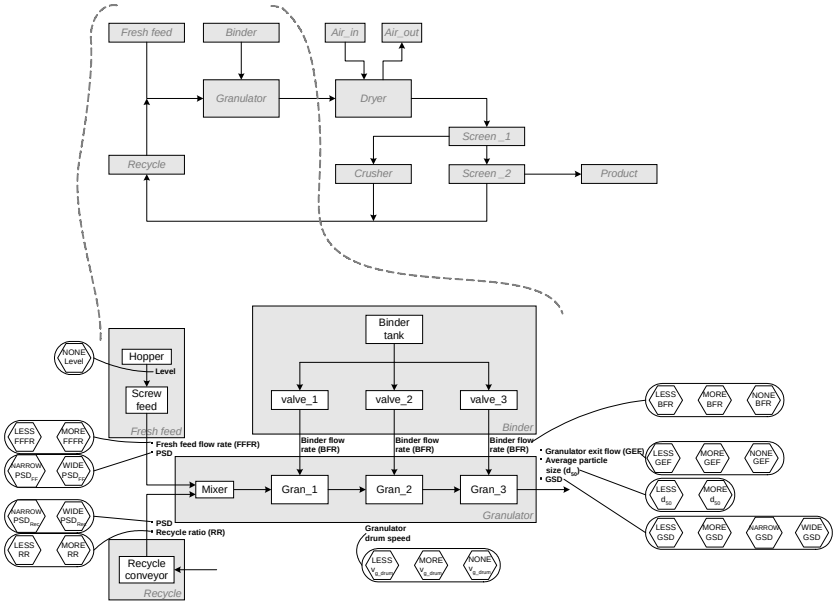



Fig. 4: The hierarchy of symptoms

*Energy balances*

- liquid phase:

$$\frac{\mathrm{d}E_L}{\mathrm{d}t} = E_{L,in} + E_{SL} + E_{NH_3} + c_2 \cdot \Delta H \cdot r_{MAP/DAP} - E_{H_2O}^{evap} - E_{L,out} - \\ - E_{LS} - \dot{m}_{crystals} \cdot \Delta H_{crys} \tag{7}$$

- solid phase:

$$\frac{\mathrm{d}E_S}{\mathrm{d}t} = E_{S,in} + E_{SL}^{MAP,sol} + E_{SL}^{DAP,sol} - E_{S,out}(i) + E_{LS} + E_{crystals} \tag{8}$$

In the above equations the notation in Table 3 is used:

**State-space model** It is seen from the above model equations that Eqs. (2) - (8) form the state equations of the state-space model with the state vector

$$x = \left[ m_{MAP}\ m_{DAP}\ m_{H_2O}\ M_L\ M_S(i)\ E_L\ E_S \right]^T \tag{9}$$

The input variables (to be manipulated or to have as disturbances) are

$$u = [\ \begin{array}{cccccc} F_{SL}^{MAP} & F_{SL}^{DAP} & F_{SL}^{NH_3} & F_{SL}^{H_2O} & F_{SL}^{MAP,sol} & F_{SL}^{DAP,sol} \\ T_{SL} & F_{SL}^{MAP} & F_{SL}^{DAP} & F_{SL}^{NH_3} & F_{SL}^{H_2O} & T_{L,in} \\ F_{NH_3} & T_{NH_3} & F_{S,in}(i) & T_{S,in} & T_{S,in}^{H_2O} & \end{array}\ ]^T$$

| Variable | Meaning |
|---|---|
| $m_{MAP}$ | mass of MAP in the liquid phase |
| $F_{L,in}^{MAP}$ | flow of MAP into the liquid phase |
| $F_{SL}^{MAP}$ | flow of MAP solution with the slurry stream |
| $F_{L,out}^{MAP}$ | flow of MAP with the liquid phase out of the drum section |
| $c_1$ | coefficient of reaction rate |
| $r_{MAP/DAP}$ | reaction rate between ammonia and MAP |
| $m_{DAP}$ | mass of DAP in the liquid phase |
| $F_{L,in}^{DAP}$ | flow of DAP into the liquid phase |
| $F_{SL}^{DAP}$ | flow of DAP solution with the slurry stream |
| $F_{L,out}^{DAP}$ | flow of DAP with the liquid phase out of the drum section |
| $\dot{m}_{crystals}$ | mass rate of crystallization onto existing solid phase |
| $c_2$ | coefficient of reaction rate |
| $m_{H_2O}$ | mass of $H_2O$ in the liquid phase |
| $F_{L,in}^{H_2O}$ | flow of $H_2O$ into the liquid phase |
| $\varphi$ | flash fraction of water from slurry flow as it exits the spray nozzle |
| $F_{SL}^{H_2O}$ | flow of $H_2O$ solution with the slurry stream |
| $F_{L,out}^{H_2O}$ | flow of $H_2O$ with the liquid phase out of the drum section |
| $F_{H_2O}^{evap}$ | flow of water evaporated |
| $M_L$ | mass holdup of liquid phase |
| $F_{L,in}$ | flow of liquid phase in the drum section |
| $F_{SL}$ | flow of slurry into the liquid phase section |
| $F_{NH_3}$ | flow of ammonia uptaken into the liquid phase |
| $F_{L,out}$ | flow of liquid phase out of the drum section |
| $M_S(i)$ | mass holdup of the solids in each particle size interval $i$ |
| $F_{S,in}(i)$ | flow of solids in each particle size interval $i$ flowing in each section |
| $F_{SL}^{MAP,sol}(i)$ | flow of MAP crystals from the slurry stream deposited onto each size range $i$ |
| $F_{SL}^{DAP,sol}(i)$ | flow of DAP crystals from the slurry stream deposited onto each size range $i$ |
| $\dot{m}_{crystals}(i)$ | mass rate of crystallization into each size range $i$ onto existing solid phase |
| $F_{S,out}(i)$ | flow of solids in each particle size interval $i$ flowing out of each section |
| $Agg(i)$ | birth and death agglomeration in each particle size interval $i$ |
| $Lay(i)$ | layering in each particle size interval $i$ |
| $Break(i)$ | breakage into each particle size interval $i$ |
| $E_L$ | energy content in the liquid phase |
| $E_{L,in}$ | energy content in flow of liquid phase in the drum section |
| $E_{SL}$ | energy content in flow of slurry into the liquid phase section |
| $E_{NH_3}$ | energy content in flow of ammonia uptaken into the liquid phase |
| $\Delta H$ | heat of reaction |
| $E_{H_2O}^{evap}$ | energy content in flow of water evaporated |
| $E_{L,out}$ | energy content in flow of liquid phase in the drum section |
| $E_{LS}$ | energy transferred between the liquid phase and the solid phase |
| $\Delta H_{crys}$ | heat of crystallization |
| $E_S$ | energy content in the solids |
| $E_{S,in}$ | energy content in total flow of solids in each drum section |
| $E_{SL}^{MAP,sol}$ | energy content in flow of MAP crystals in the slurry stream |
| $E_{SL}^{DAP,sol}$ | energy content in flow of DAP crystals in the slurry stream |
| $E_{S,out}(i)$ | energy content in total flow of solids out of each drum section |
| $E_{crystals}$ | energy content of crystallization |

Table 3: The list of variables in the granulator drum model

The structure of the state equations can be described by the following linear qualitative differential equation:

$$\dot{x} = Ax + Bu$$

where the matrices $A$ and $B$ are structure matrices with either fixed 0 or nonzero $\star$ elements as follows:

$$A = \begin{bmatrix} 0 & 0 & \star & \star & \star & 0 & 0 \\ \star & \star & \star & \star & \star & \star & 0 \\ 0 & 0 & \star & \star & \star & \star & 0 \\ \star & \star & \star & \star & \star & \star & 0 \\ \star & \star & 0 & \star & \star & \star & 0 \\ \star & \star & \star & \star & \star & \star & \star \\ \star & \star & 0 & \star & \star & \star & \star \end{bmatrix} \tag{10}$$

$$B = \begin{bmatrix} \star & 0 & 0 & 0 & 0 & 0 & 0 & \star & 0 & 0 & 0 & 0 & \star & 0 & 0 & 0 & 0 \\ 0 & \star & 0 & 0 & 0 & 0 & 0 & \star & 0 & 0 & 0 & \star & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \star & 0 & 0 & \star & 0 & 0 & 0 & \star & 0 & 0 & 0 & 0 & 0 & 0 \\ \star & \star & \star & \star & 0 & 0 & \star & \star & \star & \star & \star & 0 & \star & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \star & \star & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \star & 0 & 0 & \\ \star & \star & \star & \star & 0 & 0 & \star & \star & \star & \star & \star & \star & \star & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & \star & \star & \star & 0 & 0 & 0 & 0 & 0 & 0 & \star & \star & \star & \end{bmatrix} \tag{11}$$

The above structure shows that the overall model is highly coupled in the general case, with little chance to apply black-box model reduction techniques (such as `modred` in MATLAB).

### 5.4 Fault Scenarios and Time-Scale Separation

In order to investigate the possibilities of time-scale separation and model reduction, and their dependence on the root cause of the faults, two fault scenarios have been considered:

(A) the increase in the binder flow ("Binder_flow=MORE") that acts primarily on the overall mass and energy variables,
(B) the increase in the width of the particle size density ("PSD=WIDE") that acts on every mechanisms and balance considered,

In both cases we have waited till a steady-state operating condition developed and then issued a step-like disturbance in the relevant variable to the system and observed the transient responses.

**(A) Binder Flow – MORE** the total amount of ammonia feed ($F_{NH_3}$) to the granulator section is increased at $t = 700s$. The simulation result can be seen in Fig. 5.

A time-scale separation between the masses of the particle size classes (slow variables) and the rest of the state variables can be observed in Fig. 5 with at least an order of magnitude difference in the dominant time constant. This is in a good agreement with our engineering expectations that the increase in the $NH_3$ feed acts primarily on the overall mass and energy variables and only a slower, secondary effect is expected in the PSD variables on the granule level.

*Time-scale separation* Based on the step-response scenarios above, we can construct a refined scale-map of the drum model (see in Fig. 6) by indicating

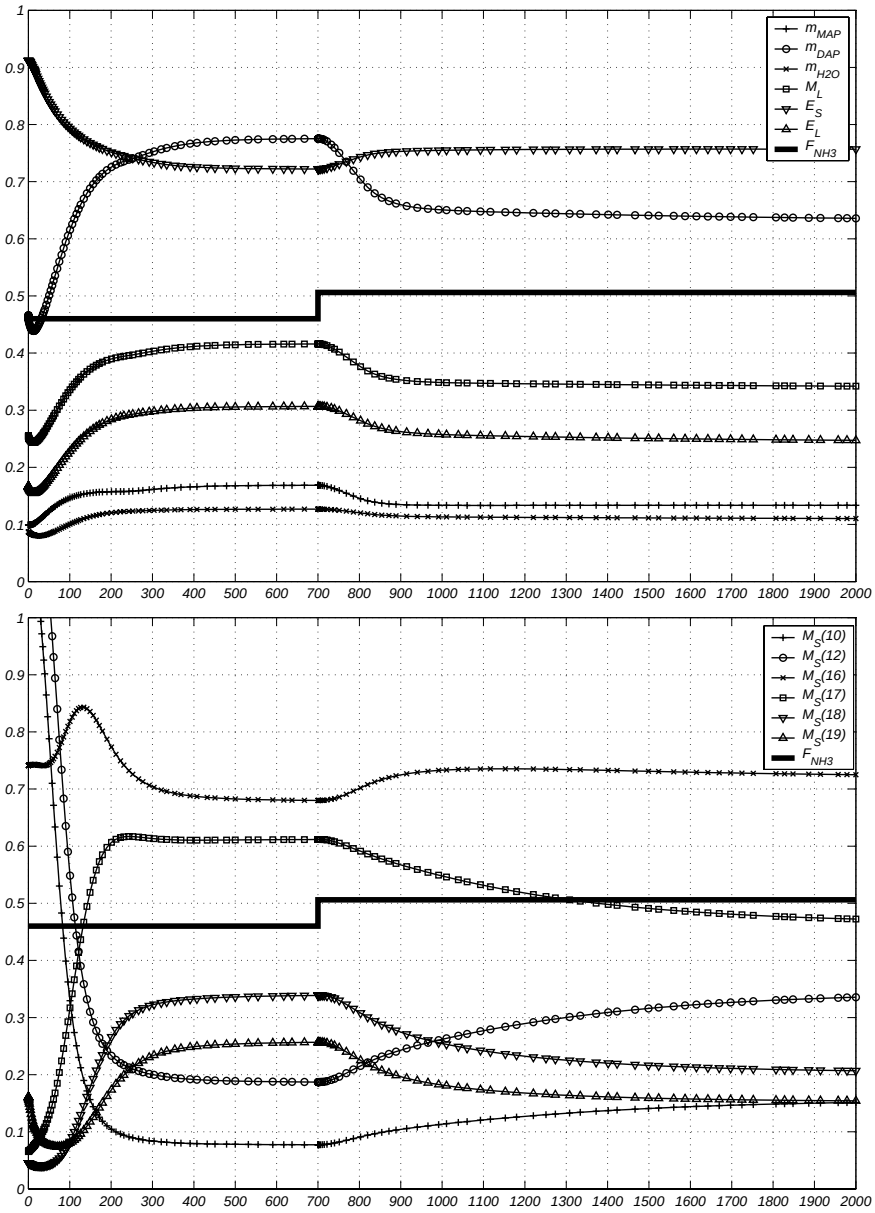



Fig. 5: Simulation result of the increased binder flow rate

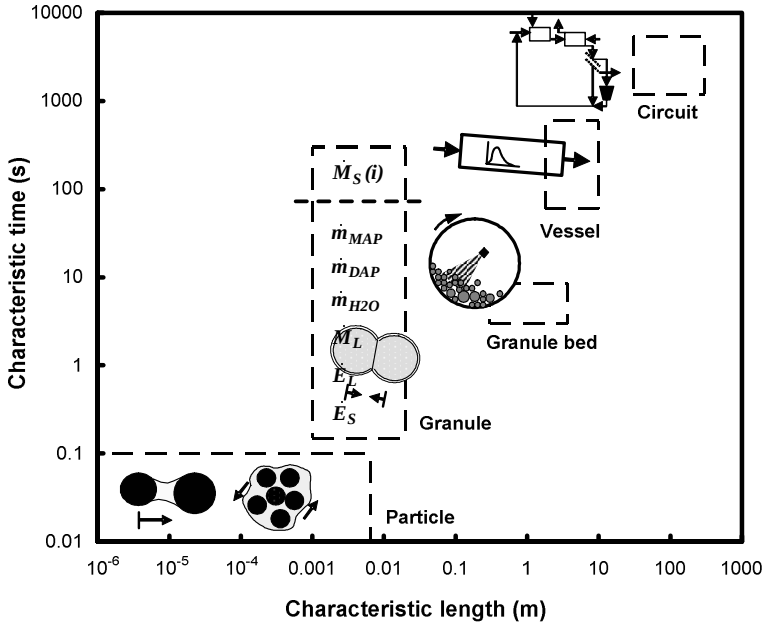



Fig. 6: The model structure map (refined scale-map) of a granulator circuit

the time-scale of the change in the particular variables effected by the root cause of increasing the binder flowrate.

**(B) Solids Feed PSD – WIDE** the particle size distribution of solids feed to the granulator is made wider (smaller size particles) at $t = 700s$. The Figure 7 shows the transient of the state variables.

There is no time-scale separation in this case.

## 6 Conclusions

A systematic approach is proposed in this paper in order to reduce the number of state variables and parameters of multiscale process models for prediction-based fault detection and diagnosis. The method requires a well-documented multiscale process model that is able to describe all considered faulty modes of the system. In addition, the list of faulty modes together with their characteristic time scales and dominant mechanisms are needed that drive the reduction procedure applied for each faulty mode.

The proposed model structure-driven reduction method applies a refined scale-map, the so called model structure map to form the scale-reduced nonlinear model of the system that only contains the dynamic balance equations on the target time scale(s) with all the other dynamic equations reduced by

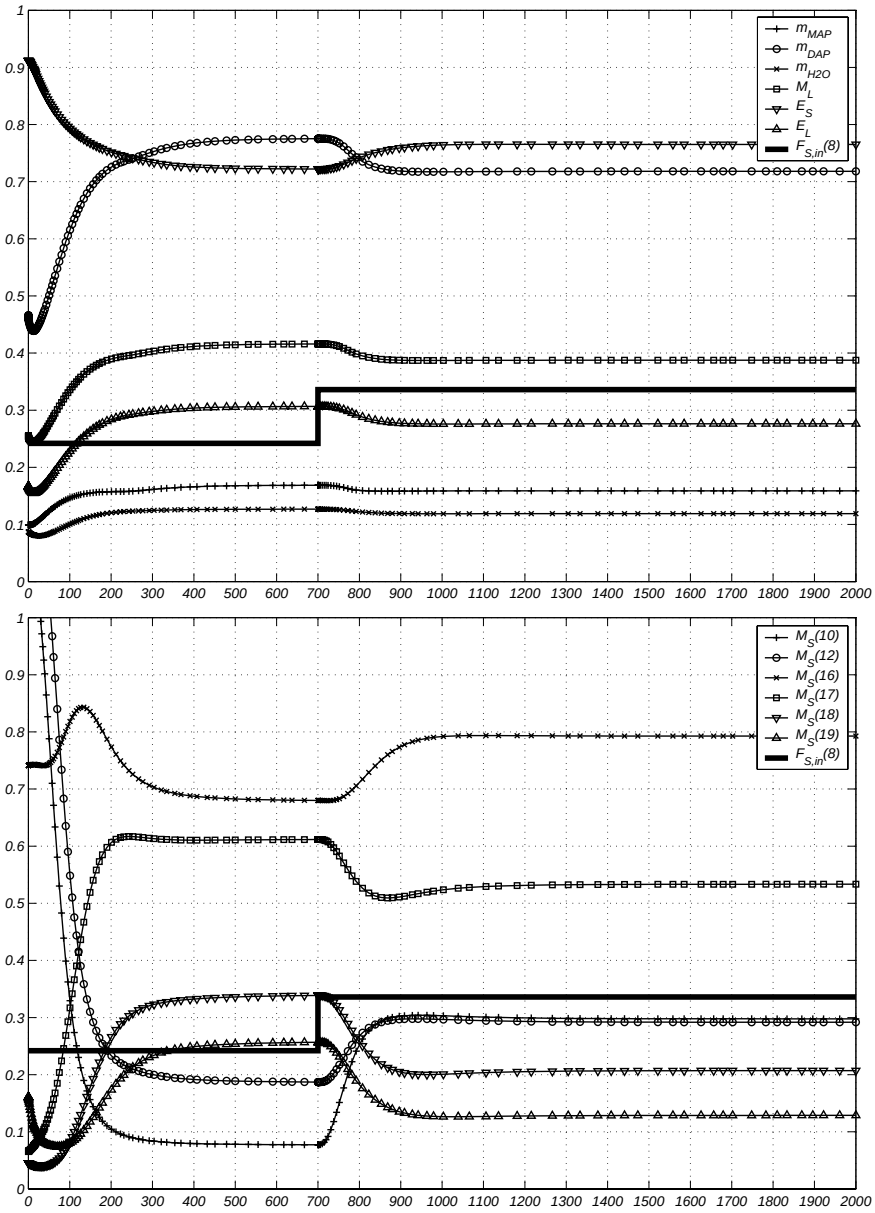



Fig. 7: Simulation result of the PSD changing of the granulator solids feed

steady-state assumptions. Thereafter local linearization is applied to obtain the final reduced model.

The resulting set of reduced models preserves the engineering meaning of the remaining state variables, while the number of state variables and parameters is decreased significantly.

The proposed model can only be applied if the following conditions hold.

- The system is in the initial phase of a fault.
- The normal operating point is steady-state.
- There is a time-scale separation in the set of dynamic conservation balance equations to be reduced.

The proposed method is illustrated on the example of a granule bed in a granulator circuit producing mono-ammonium and di-ammonium phosphate.

*Acknowledgement.* This work has been supported by the Hungarian National Research Fund through grant T042710 and T047198 which is gratefully acknowledged. Also Australian Research Council International Linkage Award LX0348222.

## References

1. M. Baldea, M. Daoutidis: Dynamics and control of integrated process network with multi-rate reactions. In: *IFAC World Congress*, (Prague, Czech Republic 2005)
2. N. Balliu: An object-oriented approach to the modelling and dynamics of granulation circuits. PhD Thesis, School of Engineering, The University of Queensland, Brisbane (2004)
3. R.D. Braatz, et al: Perspectives on the design and control of multiscale systems. *J. Process Control* **in Press** (2005)
4. I.T. Cameron, F.Y. Wang, C.D. Immanuel, F. Stepanek: Process systems modelling and applications in granulation: A review. *Chem. Eng. Sci.* **60**, 3723–3750 (2005)
5. A.N. Gorban, I.V. Karlin: Method of Invariant Manifold for Chemical Kinetics. *Chem. Eng. Sci.* **58**, 4751-4768 (2003)
6. A.N. Gorban, I.V. Karlin: Quasi-equilibrium closure hierarchies for the Boltzmann equation. *Physica A* **360**, 325-364 (2006)
7. J. Hahn, T.F. Edgar, W. Marquardt: Controllability and observability covariance matrices for the analysis and order reduction of stable nonlinear systems. *J. Process Control* **13**, 115–127 (2003)
8. K.M. Hangos, I.T. Cameron: *Process modelling and model analysis* (Academic Press, London 2001)
9. G.D. Ingram, I.T. Cameron, K.M. Hangos: Classification and analysis of integrating frameworks in multiscale modelling. *Chem. Eng. Sci.* **59**, 2171–2187 (2004)
10. G.D. Ingram, I.T. Cameron: Challenges in Multiscale Modelling and its Application to Granulation Systems. *Development in Chemical Engineering and Mineral Processing* **12** (3/4), 293–308 (2004)

11. G.D. Ingram: Challenges in Multiscale Modelling (to be made more precise). PhD Thesis, The University of Queensland, Brisbane (2005)
12. W. Jordan: Failure modes, effects and criticality analyses. In: *Proceedings of the Annual Reliability and Maintainability Symposium*, (IEEE Press 1972) pp 30–37
13. R.E. Knowlton: *Hazard and operability studies: the guide word approach*, (Vancouver: Chematics International Company 1989)
14. A. Kumar, P. Daoutidis: Nonlinear dynamics and control of process systems with recycle. *J. Process Control* **12**, 475–484 (2002)
15. R. Lakner, K.M. Hangos, I.T. Cameron: On minimal models of process systems. *Chem. Eng. Sci.* **60**, 1127–1142 (2005)
16. A. Leitold, K.M. Hangos, Z. Tuza: Structure simplification of dynamic process models. *J. Process Control* **12**, 69–83 (2002)
17. E.C. Martinez, B. Drozdowicz: Multiscale approach to real-time simulation of stiff dynamic systems. *Comp. & Chem. Eng.* **13**, 767–778 (1989)
18. E. Németh, I.T. Cameron, K.M. Hangos: Diagnostic goal driven modelling and simulation of multiscale process systems. *Comp. & Chem. Eng.* **29**, 783–796 (2005)
19. G.A. Robertson, I.T. Cameron: Analysis of dynamic process models for structural insight and model reduction – Part 1. Structural identification measures. *Comp. & Chem. Eng.* **13**, 767–778 (1989)
20. E. Rusli, T.O. Drews, R.D. Braatz: Systems analysis and design of dynamically coupled multiscale reactor simulation codes. *Chem. Eng. Sci.* **59**, 5607–5613 (2004)
21. V. Venkatasubramanian, R. Rengaswamy, S.N. Kavuri: A review of process fault detection and diagnosis Part II: Qualitative models and search strategies. *Comp. & Chem. Eng.* **27**, 313–326 (2003)
22. D.G. Vlachos: A review of multiscale analysis: Examples from systems biology, materials engineering, and other fluid-surface interacting systems. *Advances in Chemical Engineering* **30**, 1–61 (2005)

# Understanding Macroscopic Heat/Mass Transfer Using Meso- and Macro-Scale Simulations

D. V. Papavassiliou

School of Chemical, Biological and Materials Engineering
The University of Oklahoma, USA,
dvpapava@ou.edu

**Summary.** Scalar (heat or mass) transfer in the macroscale is the result of microscale diffusion and convection effects. Our fundamental hypothesis is that heat or mass transfer behavior can be synthesized from the behavior of a single, instantaneous, point source of heat or mass, and that understanding this behavior leads to an improved understanding of transport. Based on this concept, a simulation technique has been developed that involves the tracking of trajectories of heat or mass markers in a flow field, and then applying simple statistical methods to extract information about the macroscopic temperature or concentration field. The motion of these scalar markers is decomposed into a convection part, which is calculated using macroscopic flow simulations, and a diffusion part, which is simulated using a mesoscopic Monte-Carlo approach. Three different cases where this simulation methodology can been applied are presented, each one with different physics and with distinct applications. The first case is about heat transfer without convection (applied to the determination of the effective thermal conductivity of a nanocomposite material), the second case is the case of heat transfer in laminar flow (with applications in microfluidics), and the third case is the case of heat transfer with strong convective effects (applied to turbulent heat transport). The combination of a macroscopic and a mesoscopic simulation applied here allows the simulation of heat or mass transfer in cases that other conventional approaches are not feasible, and it allows the investigation of the physics of heat or mas transport in a more natural way.

## 1 Introduction

The transport of a passive scalar (either heat or mass) involves multiple length and time scales. For example, heat transfer at the macroscale is a manifestation of molecular level phenomena, such as transfer of heat at the phonon level, and of macroscale phenomena, such as convection due to flow structures that characterize the flow. A fully multiscale approach to scalar transport would, thus, have to incorporate the full spectrum of scales starting from the

molecular level and ending at the time and length scale of the process. Even though such an approach is truly fundamental, it is unlikely to be necessary (or feasible) for most cases of scalar transfer.

Methodologies for the multiscale simulation of flow have been recently reviewed [1], with a focus on particle methods, such as vortex and probability density function methods. The present paper focuses on the development of a simulation methodology specifically for heat or mass transport in different situations, where different physics may dominate.

The methodology is in essence a Lagrangian one, in which the trajectories of scalar (heat or mass) markers are followed in a flow field that is simulated with a macroscopic simulation. For example, the behavior of a heated plane can be synthesized by the behavior of an infinite number of continuous sources of heat that cover the plane [2, 3]. A practical advantage of this type of simulations is that one can reveal interesting physics about the transport process, in addition to the fact that simulations can be conducted in cases where it is difficult to apply other conventional methods. The fundamental concept is that heat or mass transport in the macroscale is the result of the behavior of single, instantaneous sources of heat or mass. If this behavior is known, then macroscopic transport properties can be obtained. The simulation technique that has been developed based on this hypothesis (our group refers to it as the *Lagrangian scalar tracking method*, LST) involves the tracking of trajectories of heat or mass markers in a flow field, and the application of statistical methods to extract information about the macroscopic temperature or concentration field. The motion of these scalar markers that carry with them either heat or a substance is decomposed into a convection part, which is calculated using macroscopic flow simulations, and a diffusion part, which is simulated using a mesoscopic Monte-Carlo approach. The molecular diffusion part depends on the properties of the fluid medium (i.e., the Prandtl or Schmidt number). The macroscopic flow simulation is preferably a direct method, such as a direct numerical simulation for turbulent flows (see among others [4] - [13]) or a lattice Boltzmann method (see among others [14] - [18]), and it provides the convection part of the transport.

In this paper we will examine the LST methodology in three different cases (with three distinct physical applications) and will emphasize the problems that can arise due to the differences in scale. First is the case of heat transfer without convection. The specific physical problem is that of heat transfer in Carbon nanotube composites, where it has been found experimentally that the effective thermal conductivity is much lower than what is theoretically expected. Second is the case of heat transfer in laminar, steady state flow. The physical problem involves heat convection in microchannels that can have walls lined up with Carbon nanotubes, or other high thermal conductivity material. The third case is that of heat or mass transfer from the wall in turbulent flow – a case in which the flow field is changing rapidly.

## 2 Numerical Methodology

The basic assumption is that the macroscopic heat or mass transfer is the result of the behavior of an infinite number of continuous sources of heat or mass. Since it is not numerically feasible to simulate an infinite number of sources, a grid of sources that are located at the hot surface within the computational domain is employed. Scalar markers are released into the simulated flow field from these grid points at one instant of the simulation. The convective part of the marker motion is calculated using the fluid velocity $V$ at the particle position and integrating in time. The time integration can be a simple Euler scheme [19] or an Adams-Bashforth scheme [20], depending on the complexity of the flow field. The molecular part of the motion is simulated by imposing a random jump at the end of each convection step. This random jump takes values from a normal probability density function that has a zero mean and a standard deviation that depends on the fluid properties. For example, in the case of laminar flow, the equation of motion for each marker in each space direction $x$ is given by

$$x_{t+1}^p = x_t^p + V_t^p \Delta t + Z\sigma \tag{1}$$

where $x_{t+1}^p$is the displacement of the marker relative to its source at time $t+1$, $V_t^p$ is the velocity of the fluid in the $x$ direction at position $x_t^p$, $\Delta t$ is the time step, $Z$ is a random number following a standard normal distribution and $\sigma$ is the standard deviation of the normal distribution that describes the Brownian motion of the markers. The molecular motion is calculated based on Einstein's [21] theory for the dispersion of particles with Brownian motion, which relates the rate of dispersion to the molecular diffusivity $D$

$$\frac{\mathrm{d}\overline{(x^p)^2}}{\mathrm{d}t} = 2D \tag{2}$$

for the case of dispersion in the direction $x$. Therefore, the standard deviation of the distribution that describes the molecular motion is given by $\sigma = \sqrt{2D\ \Delta t}$. Effects of the fluid properties in the transport process can be incorporated into the calculations by modifying $\sigma$.

There are different types of numerical error associated with the stochastic tracking of markers, when the tracking is done as described above. First, there is error associated with the calculation of the velocity field. This error depends on the numerical method utilized in the macroscopic scale for the simulation of the flow. Using direct numerical simulations (DNS) for cases of rather complicated flows (e.g., turbulence) can minimize this type of error. However, for steady laminar flows, or slowly developing flows, other conventional numerical methodologies can work well. Second is the error associated with the calculation of the velocity of the fluid at a marker location. Since the tracking of particles is an off-lattice numerical technique, it would be a rare occasion to find a marker situated on a lattice point of the numerical grid

used for the velocity field simulations. In our experience, simple interpolation schemes (linear or second order interpolation) work well for laminar flows. However, for turbulent flows, this error can be significant, and higher order interpolation schemes are necessary. Yeung and Pope [22] and Balachandar and Maxey [23] have examined this issue for homogeneous isotropic turbulence, and Kontomaris et al. [20] have investigated the accuracy of various interpolation schemes for anisotropic turbulent flows simulated with pseudospectral methods. The third source of error is the time integration scheme for the equation of marker motion. It is related to the error of the Euler method or the Adams-Bashforth method, and it can be controlled by reducing the time step of the simulations. Finally, there is error associated with the number of markers used in the calculations. Our early work with turbulent flows, which utilized databases that tracked 16,129 markers per run, addressed this issue by examining the statistics of the marker trajectories (i.e., the mean position and the mean dispersion in different directions) by repeating the calculations with half the markers ([3, 24]). More recently, Mitrovic and Papavassiliou [25] obtained results with one order of magnitude more markers for each simulation (145,161 markers). That work showed that results of acceptable accuracy can be obtained with the sample size of 16,129 markers, but it also showed that when more markers are tracked, the statistics that characterize the marker trajectories become smoother.

The behavior of each one of the point sources of heat or mass is finally used to synthesize the macroscopic scalar field, and to extract the Eulerian macroscopic parameters that characterize the transport process. The building block for this synthesis is the joint and conditional probability density function, $P_1$, for a marker to be at a location $\overrightarrow{x}$ at time $t$ given that it was released at location $\overrightarrow{x}_o$ at time $t_o$. This probability density function can be interpreted physically as temperature or concentration, if the physical properties of the fluid (density and heat capacity) are assumed to be independent of the location in the domain and of the number of markers [26, 27]. For the case of an instantaneous scalar source, this is the temperature or the concentration of a *puff* of markers. Probability $P_1$ is calculated by simply counting the number of markers in bin-cells that cover the computational domain. By integrating (or, in the discrete case, summing up) $P_1$ from time $t_o$ to a final time $t_f$, the behavior of a continuous source, represented by the probability function $P_2$, can be obtained, where

$$P_2\left(\overrightarrow{x^p} = \overrightarrow{x} - \overrightarrow{x}_o,\ t_f \mid \overrightarrow{x}_o\right) = \sum_{t=t_o}^{t_f} P_1\left(\overrightarrow{x^p},\ t \mid \overrightarrow{x}_o,\ t_o\right) \tag{3}$$

The marker cloud emitted from this continuous source, called a *plume*, is a series of instantaneous clouds, each of which is released at every time unit. Simulations done for turbulent channel flow [28, 24, 29, 30], have shown very good agreement between results obtained with LST and both experiments and simulations.

For the specific case of heat transfer in a channel, where heat is added through the channel walls, the temperature profile can be synthesized using a series of continuous line sources covering one (the bottom), or two walls of the channel (both the top and the bottom). Heat flux added to the bottom wall can be simulated by integrating $P_2$ over the streamwise direction

$$T(x_f, y) \equiv \sum_{x=x_o}^{x_f} P_2\left(x - x_o, y - y_o,\ t_f \mid \overrightarrow{x}_o\right) = \sum_{x=x_o}^{x_f} \sum_{t=t_o}^{t_f} P_1\left(x - x_o, y - y_o,\ t \mid \overrightarrow{x}_o,\ t_o\right) \quad (4)$$

When $t_f \to \infty$ and $x_f \to \infty$, Equation (4) provides the fully developed temperature profile, $T(y)$, in the channel.

Given the above discussion, the following questions may arise: What is the smallest or highest diffusivity (or in other words the Prandtl, *Pr*, or Schmidt, *Sc*, numbers) that can be simulated with such a method? How large should the time step for the mesoscopic simulation (i.e., for the particle time-stepping) be relative to the time step of the macroscopic simulation? How long should the tracking last to safely assume that $t_f \to \infty$? What should the boundary conditions be for the marker movement? How can one validate the LST simulation results? Responses to these questions will be given as specific cases of LST are discussed in subsequent sections. However, some general comments can be offered at this point.

In principal, the methodology is a Monte-Carlo approach for the determination of the probability density function that characterizes the trajectory of a single scalar marker emitted instantaneously from a point source (Equation 3). A Monte-Carlo approach should also be employed at a higher level, one that would include repetitions of each numerical experiment by generating more than one realizations of the flow field. Of course, generating many flow realizations for a computationally intensive type of flow might not always be reasonable.

Regarding the range of *Pr* that can be simulated, there should not be a limit. This is quite important, considering especially the case of turbulent flow, where steep mean temperature gradients close to the wall do not allow the simulation of high *Pr* fluids and large production of temperature fluctuations in the center of channels does not allow the simulation of low *Pr* fluids. In fact, Eulerian direct numerical simulations for anisotrpic turbulent flow have to-date been accomplished only for $0.025 \leq \mathrm{Pr} \leq 10$ [31] - [34] and one case of $\mathrm{Pr} = 54$ [35]. For homogeneous isotropic turbulent flows, recent Eulerian DNS have examined cases of higher *Sc* [36]. Brethouwer et al., [37] studied the range of $0.04 \leq Sc \leq 144$, and Yeung and coworkers [38] examined the case of $0.125 \leq Sc \leq 64$, and later on [39, 40] the case of $Sc \leq 1024$. For Small *Pr* (i.e., large *D,* and, thus, large molecular diffusion terms in Equation 1) one would need to use appropriately small time steps. An empirical guideline for the determination of the time step of the macroscopic and the mesoscopic

simulation is to use comparable time steps. Regarding the duration of the simulations, if the flow is constrained, the simulation can be run until the puff is uniformly distributed across the computational domain. The boundary conditions for the tracking are important, and will be discussed when we examine the case of microfluidics. Chandrasekhar [41] has also explored the statistics of mass tracers and Brownian motion, and his manuscript can be used as a resource for developing LST simulations. Finally, regarding the validation of the results, comparisons with experiments are the best tests. However, comparisons with cases where theoretical results are available are necessary, and can provide insights regarding the LST implementation for each physical problem that needs to be simulated.

## 3 Conductive Transport – the Case of Composite Materials

The exceptional physical properties of Carbon nanotubes (CNs) (e.g., the tubes are known to exhibit very high thermal and electrical conductivity, to possess very high tensile strength, and to have a large Young's modulus [42]) promise the development of new applications and the synthesis of new materials. For multi-walled CNs, the thermal conductivity is about 3000 W/(K.m) [43] and for single-walled CNs it is about 6000 W/(K.m) [44], values that make them comparable to diamonds in room temperature conditions. Based on the predictions of Maxwell's formula [45] or of the simple "mixture law", one would expect CN composites with low conductivity polymers (i.e., conductivity less than 0.5 W/mK) to demonstrate an effective conductivity tens of times higher than the conductivity of the polymer matrix, even for 1.0wt% CN. However, experimental evidence has shown that the addition of up to 2.0wt% of CNs into industrial epoxy composites increases their thermal conductivity by up to 125% [46]. The problem for the development of composite materials that can have high thermal conductivities (a problem that was actually known since 1941 but was not readily recognized to be relevant) is the Kapitza heat resistance [47] that exists at the interfaces between different solids [48] and between solid and liquid in contact [49]. Since heat is transferred by phonons in insulator solids, mismatch of phonon frequencies (acoustic mismatch) causes this additional resistance that could be large even at room temperatures [50].

Molecular dynamics simulations have been applied to obtain a fundamental understanding of the Kapitza effect [51] - [54]. Non-equilibrium molecular dynamics (NEMD) [55] for heat transfer between a liquid and a solid showed that, at non-wetting conditions, the Kapitza resistance is about 20-30 times higher than at wetting conditions. The value of Kapitza length can reach 50 molecular diameters (about $\sim$17 nm). Studies for solid-solid interfaces [56] also showed that thermal resistance of the composite is dominated by the Kapitza resistance. The Kapitza resistance becomes a first order effect as the size of

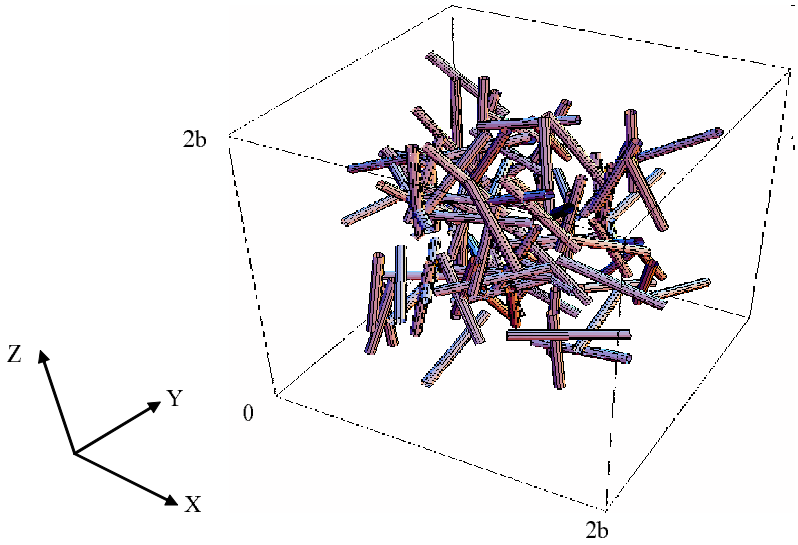



Fig. 1: A model of nanotube composite materials. Shown here is a realization of the case of disordered nanotubes at a filling fraction 4% and $L/D = 7.50$. Hot (cold) walkers enter the system on the surface $x = 0$ ($x = 2b$) and performed a random walk in the matrix.

the high thermal conductivity enclosure in the composite becomes smaller [57], which is specifically the case of CN composites.

Molecular dynamics are computationally expensive and typically can provide detailed information for a single CN in a composite. They can also provide an estimation of the value of the Katitza resistance. However, the LST methodology has been applied to investigate the macroscopic effects of several parameters (such as the CN dispersion pattern, the CN aspect ratio, the CN volume fraction in the composite) on the effective thermal conductivity of the composite, given the Kapitza resistance at the matrix-CN interface [58]. The algorithm is computationally efficient for calculating the effective thermal conductivity in a Multiwall CN composite. Thermal markers were released in a composite matrix material with finite thermal conductivity, which included cylindrical enclosures of very high thermal conductivity (see Figure 1). The heat markers moved only due to molecular diffusion (i.e., the convective term in Equation 1 was zero). The probability that allowed a thermal marker to enter a CN enclosure was proportional to the thermal resistance at the CN-matrix interface. Once inside the CN, a marker was allowed to randomly move to any location occupied by the CN, taking, thus, into account the fact that the CN thermal conductivity is orders of magnitude larger than that of the matrix material. Tomadakis and Sotirchos [59, 60] have used a similar algorithm to investigate cylindrical enclosures with different properties than the

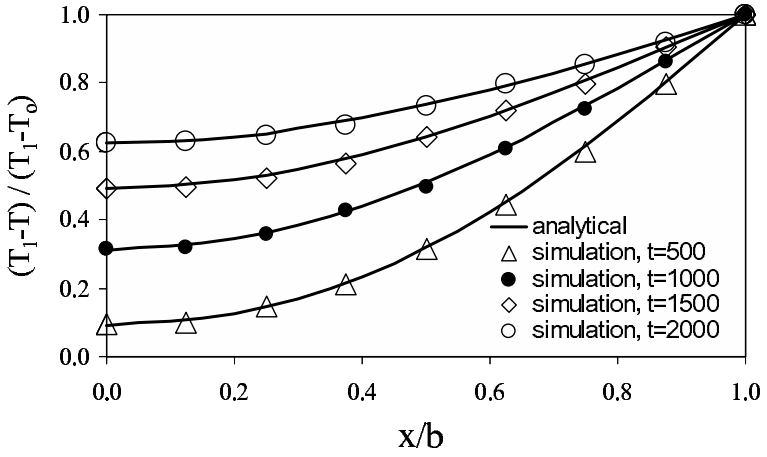



Fig. 2: Simulation results compared to the analytical solution (Equation 5) for heat transfer through a heat slab of width $2b$. The computational domain is a cube with $128^3$ bins and 90,000 walkers released in the domain.

matrix material, but did not take thermal resistance into account for their simulations.

In order to test the numerical approach for heat transfer without convection, it was first applied to the heating of a computational domain of side $2b$ with isothermal boundaries and with constant heat flux; cases for which analytical solutions are available [45, 61]. There were no CNs in the solid phase for these test runs. The simulations were tested with three different numbers of walkers (10,000, 40,000 and 90,000), with three different numbers of bins ($64^3$, $128^3$ and $256^3$) and with three different time increments (0.1, 0.25 and 0.5 in dimensionless units). The computational results agreed very well to the analytical solutions (within 0.5%). Figure 2 shows the simulation results and the analytical solution for the case of heat transfer from both sides of a slab that is initially at temperature $T_o$, and whose sides change to temperature $T_1$ at time $t = 0$. The analytical solution for this case is given as [61]

$$\frac{T_1 - T}{T_1 - T_o} = 2\sum_{n=0}^{\infty} \frac{(-1)^n}{\left(n + \frac{1}{2}\right)\pi} \exp\left[-\left(n + \frac{1}{2}\right)^2 \pi^2 \frac{Dt}{b^2}\right] \cos\left[\left(n + \frac{1}{2}\right)\frac{\pi y}{b}\right] \quad (5)$$

The same procedure was then used for composites. Considering the computational time needed for each simulation and the computational errors involved, it was found that using 90,000 walkers, a time increment of 0.25 and $128^3$ bins was sufficient.

In order to calculate a value of the effective thermal conductivity for the composites, it is convenient to simulate the case of heat transfer with constant

heat flux through the domain with hot and cold planes at the two sides of the domain (at $x = 0$ and at $x = 2b$, respectively). This can be done by having "hot" walkers enter the domain at $x = 0$ and "cold" walkers (carrying negative energy) enter at $x = 2b$. The theoretical solution of this problem at steady state is a linear temperature profile whose slope is inversely proportional to the medium conductivity. This time-independent result is trivial to fit, in contrast with the changing exponential profiles of a time-dependent problem.

The assumption that a marker is distributed uniformly throughout the space occupied by a CN, once the marker crosses into the CN, is the mesoscopic result of the Brownian random movement of a thermal marker inside a CN. In other words, the random Brownian phonon transfer inside the CN appears as ballistic heat transport on the time scale for conduction in the matrix material. If one wanted to simulate with Brownian movement the marker motion once inside the CNs, a separate time step that would be several orders of magnitude smaller than that used for the movement through the matrix material would be necessary (recall that $\sigma = \sqrt{2D\ \Delta t}$). The determination of the correct value of the probability that a marker has to bounce back into the matrix (or back into the CN) when the random jump of the particle makes it cross the matrix-CN (or the CN-matrix) interface is another issue where scale considerations become important. The LST algorithm was developed so that once a walker in the matrix reached the interface between the matrix and a CN, the walker moved into the CN phase with a probability $f_{m-CN}$, which represented the thermal resistance of the interface (and it stayed at its previous position in the matrix with a probability $(1 - f_{m-CN})$. Similarly, once a walker was inside a CN, the walker either re-distributed randomly within the CN at the end of a time step (with a probability $(1 - f_{CN-m})$ ), or would cross into the matrix phase with a probability $f_{CN-m}$. Even though it was assumed that the thermal resistance is the same for a heat walker traveling from the matrix to a Carbon nanotube and from a Carbon nanotube to the matrix phase, it was found that $f_{m-CN} \neq f_{CN-m}$. The reason is that the previous assumption (that of a uniform distribution of a marker once inside a CN) removed a length scale from the problem (that of the length of the thermal walker movement inside the CN). Therefore, the exit probability $f_{CN-m}$ has to be weighted so that the flux of walkers into the CNs is equal to the flux of the walkers exiting the CNs at thermal equilibrium. In order to maintain equilibrium, the two probabilities must be related as

$$f_{CN-m} = c\frac{\sigma A_C}{V_C} f_{m-CN} \tag{6}$$

where $A_c$ and $V_c$ are the surface area and the volume of a nanotube, respectively, $\sigma$ is the standard deviation of the random jump in the matrix, and $c$ is a constant that depends on the shape of the high conductivity enclosures. We now have recovered a length scale, $V_c/A_c$, in the problem. The value of $c$ can be found theoretically, or can be found computationally by conducting numerical experiments at thermal equilibrium. According to the acoustic mis-

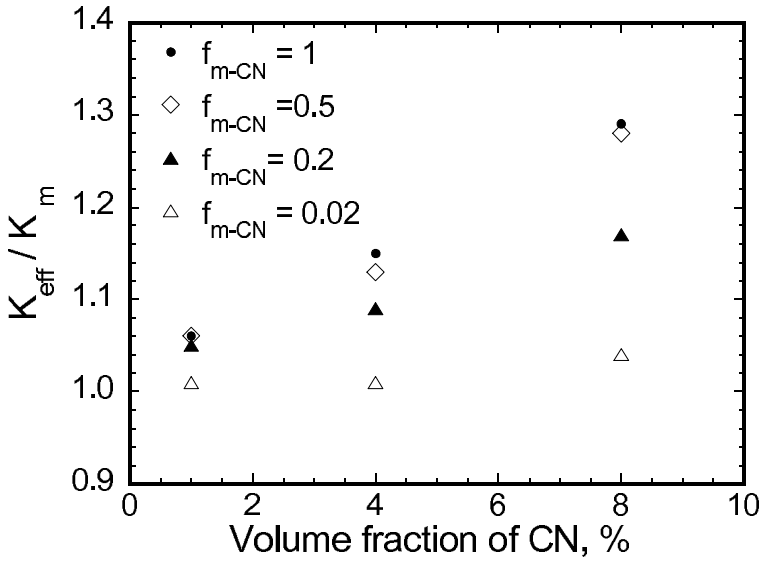



Fig. 3: Effective thermal conductivities of randomly-dispersed nanotube composites as a function of the volume fraction of nanotubes at $L/D$=3.75. For each value of thermal resistance and volume fraction of nanotubes, the thermal conductivity is the average of three simulation runs with different initial nanotube random distributions (Figure taken from Duong et al., 2005).

match theory [50], the average probability for transmission of phonons across the interface into the dispersion, $f_{m-CN}$, is given by

$$f_{m-CN} = \frac{4}{\rho\ C_p\ C_m\ R_{bd}} \tag{7}$$

where $\varrho$ is the matrix density, $C_p$ is the matrix specific heat, $C_m$ is the velocity of sound in the matrix, and $R_{bd}$ is the thermal boundary resistance.

Figure 3 shows the effective thermal conductivities of randomly-dispersed nanotube composites where the ratio of the nanotube length, $L$, over the nanotube diameter, $D$, is 3.75. When either the volume fraction of the nanotubes increases or the thermal resistance decreases, walkers move from the matrix into the CN easier. In addition, the walkers can move much faster inside a CN than in the matrix due to its high thermal conduction. Hence the effective thermal conductivity increases.

## 4 Convective Transport – the Case of Laminar Flow

The second case where the Lagrangian method has been applied is that of heat transfer in laminar, steady state flow. The physical problem involves heat

convection in microchannels that can have walls lined up with Carbon nanotubes, or other high thermal conductivity material. Understanding of heat or mass transfer phenomena at the microscale (defined in this case as geometries smaller than 1 mm) is important for the development of applications that can take advantage of recent micro- and nano-technology advances. The main issue for the study of flow and transport in this scale is that correlations that apply in the macroscale do not necessarily apply in the microscale [62]. There is a need to develop new appropriate correlations, as well as a need to develop techniques for the investigation of phenomena in microfluidics.

The lattice Boltzmann method, LBM, efficiently links microscopic flow phenomena to the macroscopic behavior of a fluid. It is a practical, easilly parallelizable method that can be used for microfluidics as well as macroscopic scale flows, and it has the additional advantage that it is effective for the simulation of flows in complicated geometries (e.g., for flow in porous media [16]). However, the simulation of scalar transfer with LBM is not as common or as easy as the simulation for flow. Thermal LBM models have been developed, initially for a narrow range of temperatures [63, 64]. Later on, models that introduced a separate internal energy distribution function to calculate the temperature field appeared in the literature [65] - [67].

The methodology suggested here is to combine an LBM (macroscopic level simulation) with the tracking of scalar markers in the flow field (mesoscopic level). A single simulation of the flow field can be used in conjunction with several mesoscopic simulations, each one of which corresponds to different types of fluids and/or to different thermal boundary conditions [19]. This case can illustrate some of the errors involved in the numerical methodology, as well as the development of a thermal LBM method for heat transfer in microfluidics.

The specific LBM algorithm used in this work applies a multi-speed model consisting of a 3-dimensional 15-component velocity (this configuration of the model is known as D3Q15, see [68]). The simulation grid consists of $nx$, $ny$ and $nz$ nodes in the $x$, $y$ and $z$ directions, respectively. Each fluid node consists of discrete packets of fluid with density, $\varrho$, represented by a particle distribution function, $f_i$, $(i = 0, .., 14)$. These fifteen components of the distribution function belong to one of three types, the rest position $(i = 0)$, class I $(i = 1, \ldots, 6)$, and class II $(i = 7, \ldots, 14)$ type components. Initially, the density at a fluid node is distributed according to the ratio $16 : 8 : 1$ among the three types of components. Periodic boundary conditions are applied at the non-wall faces. The no-slip velocity condition at the walls is simulated by a bounce-back scheme [69, 70] with Ziegler's suggestion [71] of shifting the wall boundary into the fluid by one-half mesh unit in order to achieve second-order accurate results.

The particle distribution function, $f_i$, is calculated as a function of space and time from the discretized Boltzmann equation [72]

$$f_i\left(\overrightarrow{x} + \overrightarrow{e}_i \Delta t, t + \Delta t\right) = f_i\left(\overrightarrow{x}, t\right) + \Omega_i\left(\overrightarrow{x}, t\right) + f f_i \tag{8}$$

where $\overrightarrow{e}_i$ is the velocity in the 15 directions of the numerical grid, $t$ and $\Delta t$ are the time and the time step, $\Omega_i$ is the collision operator and $ff_i$ is the component of the forcing factor. The terms on the right hand side of Equation (8) constitute the three fractional steps of the lattice Boltzmann algorithm, namely the streaming, collision and forcing steps. At time t, during the streaming step the rest component ($f_0$) remains at the center of the face-centered cubic lattice, while the 6 class I components ($f_i, i = 1, \ldots, 6$) move to the nearest neighboring nodes and the 8 class II components ($f_i, i = 7, \ldots, 14$) move along the diagonal towards the 8 corners of the cubic lattice. Boundary conditions are then applied to restore the fluid moving towards the wall nodes by bouncing it back in the opposite direction (no-slip). In the second step, collision rules are applied to relax the fluid distribution function on a node after a collision with fluid from neighboring nodes back to equilibrium. This step is calculated according to the Bhatnagar, Gross and Krook approximation [73, 74]

$$\Omega_i\left(\overrightarrow{x}, t\right) = -\frac{1}{\tau}\left(f_i - f_i^{eq}\right) \tag{9}$$

In the above equation, $\tau$ is the relaxation time calculated as $\tau = 3\nu + 0.5$, where $\nu$ is the kinematic viscosity of the fluid. The equilibrium distribution functions are derived from the conservation equations of mass and momentum, and are given below for rest ($i = 0$), class I ($i = 1, \ldots, 6$) and class II ($i = 7, \ldots, 14$) components:

$$\begin{aligned} f_0^{(eq)} &= \varrho\left[\frac{1}{8} - \frac{\overrightarrow{U}^2}{3}\right] \\ f_i^{(I,eq)} &= \varrho\left[\frac{1}{8} + \frac{1}{3}\left(\overrightarrow{e}_i \cdot \overrightarrow{U}\right) + \frac{1}{2}\left(\overrightarrow{e}_i \cdot \overrightarrow{U}\right)^2 - \frac{1}{6}\overrightarrow{U}^2\right] \\ f_i^{(I,eq)} &= \varrho\left[\frac{1}{64} + \frac{1}{24}\left(\overrightarrow{e}_i \cdot \overrightarrow{U}\right) + \frac{1}{16}\left(\overrightarrow{e}_i \cdot \overrightarrow{U}\right)^2 - \frac{1}{48}\overrightarrow{U}^2\right] \end{aligned} \tag{10}$$

In the third fractional step, a pressure drop or forcing factor is added to the fluid components moving in the positive streamwise direction and subtracted from those moving in the negative x direction . We used the method described by Noble [75] to calculate the fraction of the forcing factor applied on each of the fifteen components in our 3D grid.

After the completion of the three fractional steps shown in Equation (8) the macroscopic properties of the fluid, such as density and velocity, are calculated from the conservation equations of mass and momentum, respectively, given by

$$\varrho = \sum_{i=0}^{14} f_i \tag{11}$$

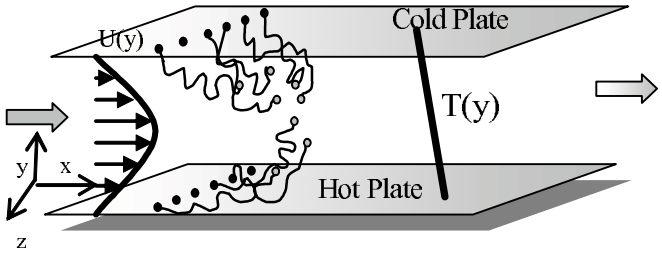



Fig. 4: Schematic of a channel with flow in the x direction and heat transfer between a cold and a hot wall. The height of the channel is 2H.

$$\varrho\overrightarrow{U} = \sum_{i=0}^{14} f_i \overrightarrow{e}_i \tag{12}$$

The velocity of the fluid at each fluid node is calculated by dividing Equation (12) with Equation (11). The simulations were carried out till steady state velocity profiles were obtained throughout the domain.

In order to simulate the case of heat trasnfer in the flow field, the position of all the markers that are released from the bottom wall of the channel is tracked in time (see Figure 4 for a schematic of the physical situation). Their convective and diffusive motion is tracked till their mean position in the $y$ direction reaches the center of the channel ($H$). This would ensure that the markers are uniformly distributed throughout the height of the channel. Also, the variance of this uniform distribution in the $y$ direction should reach the value $4H^2/12$.

An error analysis of the algorithm has been conducted to determine the effects of the number of markers, the time step and the bin size on the simulation results [19]. The physical problems used for validation included the heating of a semi-infinite solid whose surface is suddenly raised to a specified temperature $T_w$ and the problem of isoflux heat transfer between a hot and a cold channel wall. The first problem has a known analytical solution that is based on the error function. The temperature profile at small times is given by $T/|T_w| = erfc\left(y/\sqrt{4Dt}\right)$ [45]. However, this problem does not include flow, and in that respect is similar to the problem discussed in the previous section. The only difference is that one needs to introduce a weight function, $w(t)$, which weights the contributions of each puff to the final temperature of a plume in a way such that the temperature at the surface of the slab is constant through time (i.e., the number of thermal markers at the surcase remains constant). Since the particle movement in the $x$ direction was not important, the puffs and plumes propagated only in the $y$ direction. The weight functions were then calculated for all $t$ at the slab surface, and the temperature was given by

$$T(y) = \int_{t_o}^{t_f} P_1\left(y, t - t_o | t_o\right) w(t)\, \mathrm{d}t \tag{13}$$

In the case of isoflux heat transfer between a hot and a cold channel wall (Figure 4), the bottom wall of the channel was continuously heated and the top wall was continuously cooled at the same rate. Theoretically, the temperature profile is expected to be a straight line, with maximum and minimum temperatures at the hot and cold plates, respectively. At the center, the temperature is expected to be the mean of the maximum and the minimum temperatures. Assuming that the channel is symmetric, the temperature profile $T_{isoflux}$ can be calculated numerically from the temperature calculated in Equation (4) using the relation

$$T_{isoflux}\left(x_f, y\right) = T\left(x_f, y\right) - T\left(x_f, 2H - y\right) \tag{14}$$

A scaling issue that requires special attention in these cases is to make sure that the ratio of $\sigma$ (the standard deviation of the probability distribution that describes the molecular movement of the markers) to the width of the channel is small (smaller than $10^{-2}$). When $\sigma$ is larger, peculiar results might be obtained, such as an increase of the relative error with increasing the number of heat markers.

Temperature profiles calculated for a microchannel are presented in Figure 5. The simulated microchannel is heated only from the bottom wall with a step change of the wall heat flux. It has dimensions $(1.5 \times 0.5 \times 0.5)10^{-3}$ m$^3$, pressure drop 1000 Pa/m and Reynolds number $Re = 18$ . The temperature is non-dimensionalized as follows

$$T^+(x, y) = \frac{T(x, y)}{T^*} = -\,\mathrm{Pr}\,\frac{T(x, y)}{\left(\mathrm{d}T/\mathrm{d}y^+\right)_w} \tag{15}$$

where $\left(\mathrm{d}T/\mathrm{d}y^+\right)_w$ is the slope of the temperature at the wall of the channel and $y^+ = u^* y/\nu$ ($u^*$ is the bulk velocity of the fluid). Different $Pr$ fluids were simulated covering the range of liquid metals ($Pr = 0.1$), gases ($Pr = 1$) and water ($Pr = 6$). Cases A, C and E had the same constant heat flux from the bottom wall, while cases B, D, and F had higher heat flux at a specific location $x_s$ in order to simulate the case where the bottom wall of the channel has an area of high thermal conductivity. The strength of the plume in this location was claculated to be 10 times higher than the strength of the sources in the rest of the channel wall, corresponding to a material that has 0.9wt % Carbon nanotubes. The temperature profile increased with Prandtl number at all $x$ locations in the channel. The temperature exhibits a maximum value at the wall, where heat markers are released, and decreases with the distance from that wall and is significant till half channel height, $H$. The temperature profiles shown in Figure 5 are *upstream* of the location that has higher thermal conductivity. When nanotubes are dispersed on the

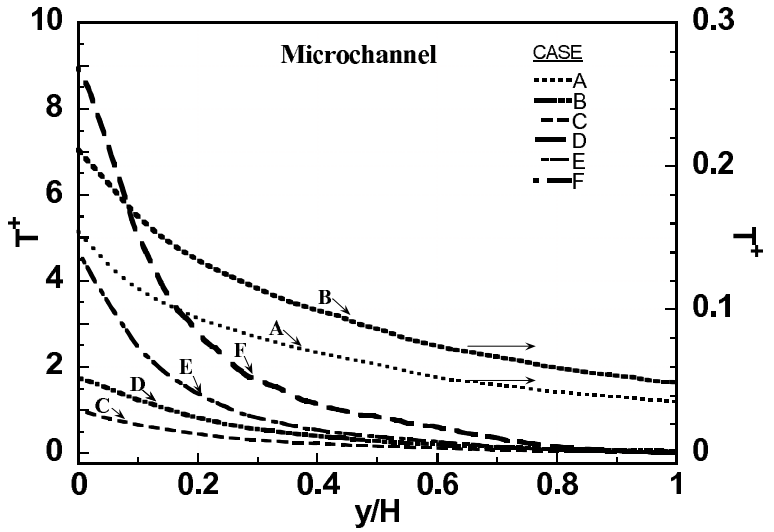



Fig. 5: Effect of backscattering of heat markers on the dimensionless temperature profile in a microchannel. The temperature is shown at x/H=0.139 upstream from the location of the high thermal conductivity zone on the microchannel wall. Cases A,B have Pr=0.1, Cases C,D have Pr=1, and Cases E,F have Pr=6.

wall downstream from the location of step change in wall heat flux ($x_s > x_o$), the temperature at and around the location of nanotubes increases. Both the temperature profile and slope at the wall increase with the presence of nanotubes. The dimensionless temperature, $T^+$, which is proportional to the ratio of these quantities (see Equation 15), increases. Figure 5 shows, therefore, that back-scattering of temperature can be important in microfluidics, even though an Eulerian computation might not accurately simulate an effect like this. The temperature profiles can be further used to calculate other heat transfer parameters, like the heat transfer coefficient as a function of the $Pr$ [76].

Finally, in the case of microfluidics applications, special attention needs to be paid to the Knudsen number of the flow. If it falls in the first regime of fluid rarefaction, where $Kn < 10^{-3}$ [77], then the Navier-Stokes equation and no-slip boundary conditions are applicable for that flow. In our numerical simulations, the flow field in the microscale geometries has a Knudsen number value on the order of $10^{-6}$, and thus the Navier-Stokes equation and no-slip boundary conditions are inherently applicable.

## 5 Convective Transport – the Case of Turbulent Transport

The third case where we employ LST is that of heat or mass transfer from the wall in turbulent flow – a case in which the flow field is changing rapidly. The fluid velocity is calculated using a direct numerical simulation (DNS), which determines the convective part of the scalar marker motion. The molecular contribution of the motion is simulated as in the previous two cases, with the important difference that the simulations of flow and of molecular diffusion are run simultaneously. During each time step of the turbulence DNS, a molecular diffusion step has to be completed. However, it is advisable (for passive scalars) to incorporate diffusion steps corresponding to different diffusivities during the same velocity step, simulating, thus, several types of fluids with the same velocity field. These simulations can provide results for cases where the macroscopic solution of the heat or mass balance equation is not feasible (for example for very high values of the Schmidt number). In terms of physics, questions that have been a matter of debate between turbulence theorists over a long time can be answered, such as the dependence of turbulent transport on the Schmidt number close to a solid surface.

Currently, Eulerian simulations are limited in the range of $Pr$ or $Sc$ that they can cover because, in order to resolve all the scales of motion and temperature [78], the number of grid points has to be analogous to $Pr^{3/2}Re^{9/4}$, where $Re$ is the Reynolds number of the flow. An increase of $Pr$ by one order of magnitude means an increase of the number of grid points by about thirty times. Large eddy simulation (LES) has also been used to study turbulent mass transfer for Schmidt numbers in the range $1 \leq Sc \leq 200$ [79]. An important issue with LES for anisotropic turbulence is that different scales of motion contribute to heat transfer at different distances from the wall, and that these scales are $Pr$ dependent [80] making the determination of the spectrum cutoff point difficult.

The particular DNS algorithm used in our group is based on the pseudospectral method of Lyons et al. [5]. It can simulate fully developed turbulent flow in an infinite channel for Poiseuille and for plane Couette flow. The Navier-Stokes equations are integrated in time using the fractional step method introduced by Orszag and Kells [81] with the pressure field correction suggested by Marcus [82]. The fluid is considered to be incompressible and Newtonian. In the case of a channel, the flow is driven by a constant mean pressure gradient, and for the case of plane Couette flow it is driven by the shear motion of the two moving walls of the channel [8, 30]. The Reynolds number, defined with the centerline mean velocity and the half-height of the channel for the Poiseuille flow channel, and defined with half the velocity difference between the two walls and the half channel height for the Couette flow channel, was 2660 for both. For the Poiseuille channel, the simulation was conducted on a $128 \times 65 \times 128$ grid in $x, y, z$, and the dimensions of the computational box were $4\pi H \times 2H \times 2\pi H$, where $H = 150$ in wall units. (In wall

turbulence, quantities are commonly made dimensionless with the so called *viscous* scales, i.e., the friction wall velocity $u^* = \sqrt{\tau_w/\varrho}$ where $\tau_w$ is the shear stress at the wall, the friciton length scale $l^* = \frac{\nu}{u^*}$, and the friction time scale $t^* = \frac{l^*}{u^*}$. Dimensionless quantities obtained with such scaling are called *viscous* or *wall* quantities. All quantities in this section are in wall units.) For the Couette flow channel, the simulation was conducted on a $256 \times 65 \times 128$ grid, and the dimensions of the computational box were $8\pi H \times 2H \times 2\pi H$, where $H = 150$. The flow is regarded as periodic in the $x$ and $z$ directions, with the periodicity lengths equal to the dimensions of the computational box in these directions. The time step for the calculations of the hydrodynamic field and the Lagrangian tracking was $\Delta t = 0.25$ and $\Delta t = 0.25$ for the Poiseuille and Couette channels, respectively. Both simulations were first allowed to reach a stationary state before the heat markers were released.

We focus here only on results that relate to the calculation of the heat transfer coefficient. Other heat transfer parameters can be found in references from our group cited herein. The heat transfer coefficient, $h$, is found in numerous research papers and technical textbooks to be given in the form of a correlation for the Nusselt number

$$Nu = ARe^b Pr^c \tag{16}$$

where $A, b, c$ are constants that depend on the type of flow (e.g., flow in a pipe or a channel, flow around an immersed object, etc.). This type of correlation originates from applications of dimensional analysis in transport phenomena. However, experimental data have demonstrated scatter around this correlation, implying that there is another functional relationship between the dimensionless numbers, $Nu = f(Re, Pr)$, (see Churchill's insightful discussion about this issue [83]). Regarding $Pr$ dependence, there is a controversy in the literature among investigators who argue for a heat transfer coefficient that goes as $h \sim Pr^{-3/4}$ and those who argue for $h \sim Pr^{-2/3}$. This argument has its origin in the fundamental issue of the asymptotic behavior of the eddy diffusivity, $E_c$, very close to the wall. If $E_c \sim y^3$ as $y \to 0$ (see Monin and Yaglom's monograph [84]), then $h \sim Pr^{-2/3}$ but if $E_c \sim y^4$ as $y \to 0$ (see [85]), then $h \sim Pr^{-3/4}$. To further complicate the issue, there is experimental evidence that the exponent is neither of the above; instead it has a value between –3/4 and –2/3 [86] and there is theoretical analysis that accounts for the turbulence space-time correlation close to the solid-fluid interface and for the diffusive dumping of the temperature fluctuations, and which suggests that the exponent should be –7/10 [87].

Our LST results are obtained with a consistent methodology over a wide range of $Pr$, and have produced predictive correlations for the heat transfer coefficient as a function of $Pr$ for both Poiseuille [29] and plane Couette flow [88]. Figures 6 and 7 show the LST obtained data for the channel flow and the plane Couette flow, respectively. The results of Figure 6 can be summarized for channel flow and for $0.01 < Pr < 50000$ as follows [29]:

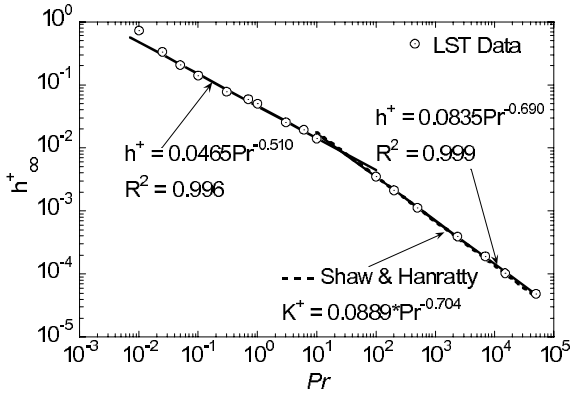



Fig. 6: Heat transfer coefficient as a function of Pr for turbulent channel flow. The values have been obtained with a consistent methodology for all Pr. (Figure taken from Mitrovic et al., 2004).

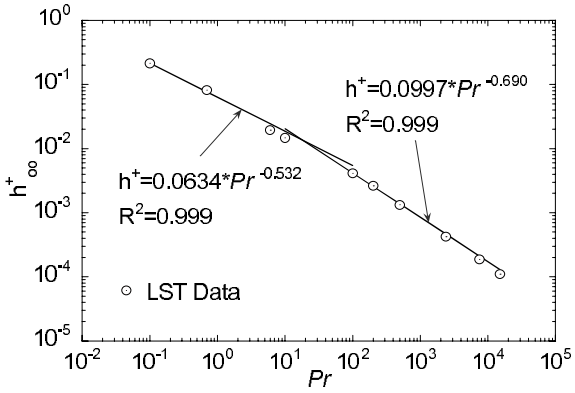



Fig. 7: Heat transfer coefficient as a function of Pr for turbulent Couette flow. The values have been obtained using a DNS in conjunction with the Lagrangian tracking of scalar markers in the flow field for all Pr. (Figure taken from Le and Papavassiliou, 2006).

$$h^+ = 0.0465\ Pr^{-0.510} \qquad for \qquad Pr \leq 10$$

$$h^+ = 0.0835\ Pr^{-0.690} \qquad for \qquad Pr \geq 100 \tag{17}$$

and

$$h^+ = \frac{0.0835\,\mathrm{Pr}^{-0.690}}{\left[1+\left(\frac{25.85}{\mathrm{Pr}}\right)^2\right]^{0.090}} \qquad for \quad all \quad Pr$$

For Couette flow, the corresponding correlations were found to be (based on simulations for $0.1 < Pr < 15000$ [88])

$$h^+ = 0.0634\ Pr^{-0.532} \qquad for \quad Pr \leq 10$$

$$h^+ = 0.0997\ Pr^{-0.690} \qquad for \quad Pr \geq 100 \tag{18}$$

and

$$h^+ = \frac{0.0997\,\mathrm{Pr}^{-0.690}}{\left[1+\left(\frac{17.56}{\mathrm{Pr}}\right)\right]^{0.158}} \qquad for \quad all \quad Pr$$

An important finding is that the heat transfer coefficients for plane Couette flow show the same trend as for Poiseuille channel flow. The exponential values (constant $c$ in Eqn. 16) are the same or close to those in Poiseuille channel flow, but the pre-exponential factors are higher. The interpretation of this observation is that the mechanism of turbulent transport from the wall is the same in both cases, i.e., only a part of the spectrum (the smaller wavenumbers part) of the turbulent velocity field contributes to turbulent transport from the wall, and this part depends on the fluid $Pr$ (as $Pr$ increases, a smaller part of the spectrum contributes, see [80, 25]). However, the turbulent velocity field is different in Couette and Poiseuille flow, with turbulence intensities being higher in Couette flow, and this fact manifests itself as a larger pre-exponential factor.

## 6 Summary and Conclusions

The three case studies discussed here illustrate that the combination of macroscopic and mesoscopic simulations, as described above, can allow the simulation of heat or mass transport in cases that other conventional techniques are not feasible, and that it can provide a natural way of investigating the physics of scalar transport. For the case of transport without convection, the effective thermal conductivity of composites with nanotubes dispersed in a continuous matrix can be computed over a wide range of thermal resistance, CN volume fraction and nanotube aspect ratio. The proposed algorithm is efficient in that it removes the need to perform random walks within the CNs. Scale effects need to be taken into account carefully for the numerical treatment of the CN-matrix interface. For the case of laminar flow and microfluidics, a methodology for conducting thermal lattice Boltzmann simulations applicable to passive scalar transport was implemented. Scale effects are important

when the bounce back boundary condition for the thermal markers is implemented, and the molecular diffusion random jumps should be at least two orders of magnitude smaller than the dominant length scale of the microchannel. Important physics, such as the $Pr$ dependence of the scalar field and the back-scattering of heat can be investigated. For the case of turbulent flow, the proposed numerical methodology can allow the conduction of simulations that are not currently trivial to pursue with other methods, in addition to providing significant insights to the mechanism of turbulent heat transfer.

This manuscript would be incomplete without refering to other work that has utilized the same or very similar methodology for the investigation of problems that are different than those described here. These problems include the simulation of heat transfer across the interface between a turbulent gas and a turbulent liquid [89, 90], the simulation of mass transfer in low $Re$ fluids and non-Cartesian geometries for bubble dissolution in the presence of surfactants [91], and the simulation of the effects of the flow on chemical reactions [92]. Finally, Mito and Hanratty [27] have studied the behavior of markers released at different elevations in a turbulent flow channel in order to calculate the time scales associated with the marker movement. Their goal was to use these time scales to solve a modified Langevin equation for the prediction of the velocity field along the trajectories of the markers, and subsequently use that velocity field to predict the marker trajectories (in other words, they substituted the DNS velocity in an LST procedure with the velocity field resulting from the solution of the Langevin equation).

*Acknowledgement.* The support of NSF under CTS-0209758 and under EPS-0132534 is gratefully acknowledged. The Donors of The Petroleum Research Fund, administered by the American Chemical Society, should also be acknowledged for support of this research through grant PRF# 39455-AC9. This work was also supported by the National Computational Science Alliance under CTS-040023 and utilized the NCSA IBMp690 and the NCSA SGI/CRAY Origin2000. Computational support was also offered by the University of Oklahoma Center for Supercomputing Education and Research (OSCER). Finally, discussions with Drs. Lloyd Lee, Kieran Mullen and Hai Duong, as well as with Bojan Mitrovic, Phuong Do, and Nishitha Thummala should be acknowledged.

## References

1. P. Koumoutsakos: Multiscale flow simulations using particles. *Annu. Rev. Fluid Mech.* **37**, 457–487 (2005)
2. D.V. Papavassiliou, T.J. Hanratty: The use of Lagrangian methods to describe turbulent transport of heat from the wall *Ind. Eng. Chem. Res* **34**, 3359–3367 (1995)
3. D.V. Papavassiliou, T.J. Hanratty: Transport of a passive scalar in a turbulent channel flow. *Int. J. Heat Mass Transfer* **40** (6), 1303–1311 (1997)

4. J. Kim, P. Moin, R. Moser: Turbulence statistics in fully developed channel flow at low Reynolds numbers. *J. Fluid Mech.* **177**, 133–166 (1987)
5. S.L. Lyons, T.J. Hanratty, J.B. McLaughlin: Large-scale computer simulation of fully developed turbulent channel flow with heat transfer. *Int. J. Numer. Methods Fluids* **13**, 999–1028 (1991)
6. N. Kasagi, N. Shikazono: Contribution of direct numerical simulation to understanding and modeling turbulent transport *Proc. R. Soc. Lond. A* **451**, 257–292 (1995)
7. C. Xu, Z. Zhang, J.M.J. den Toonder, F.T.M. Nieuwstadt: Origin of high kurtosis levels in the viscous sublayer. Direct numerical simulation and experiment, *Phys. Fluids* **8** (7), 1938–1944 (1996)
8. D.V. Papavassiliou, T.J. Hanratty, Interpretation of large scale structures in a turbulent plane Couette flow. *Int. J. Heat and Fluid Flow* **18**, 55–69 (1997)
9. P. Moin, K. Mahesh: Direct Numerical Simulation: A tool in turbulence research, *Annu. Rev. Fluid Mech.* **30**, 539–578 (1998)
10. A. Gunther, D.V. Papavassiliou, M.D. Warholic, T.J. Hanratty: Turbulent flow in a channel at low Reynolds number. *Exp. in Fluids* **25**, 503–511 (1998)
11. R.D. Moser, J. Kim, N.N. Mansour: Direct numerical simulation of turbulent channel flow up to Re=590. *Phys. Fluids* **11** (4), 943–945 (1999)
12. P. Vedula, P.K. Yeung: Similarity scaling of acceleration and pressure statistics in numerical simulations of isotropic turbulence. *Phys. Fluids* **11** (5), 1208–1220 (1999)
13. H. Abe, H. Kawamura, H. Choi: Very large-scale structures and their effects on the wall shear-stress fluctuations in a turbulent channel flow up to Re-tau=640. *J. Fluids Eng. – Trans. ASME* **126** (5), 835–843, (2004)
14. D. Grunau, S. Chen, K. Eggert, A lattice Boltzmann model for multiphase fluid flows. *Phys. Fluids A* **5** (10), 2557–2562 (1993)
15. D.R. Noble, S.Y. Chen, J.G. Georgiadis, R.O. Buckius: A consistent hydrodynamic boundary condition for the lattice Boltzmann method. *Phys. Fluids* **7** (1), 203–209 (1995)
16. S. Chen, G.D. Doolen: Lattice Boltzmann method for fluid flows. *Annu. Rev. Fluid Mech.* **30**, 329–364 (1998)
17. D.A. Wolf-Gladrow: *Lattice-gas cellular automata and lattice Boltzmann models*, Lecture Notes in Mathematics 1725 (Springer, Berlin 2000)
18. S. Succi: *The Lattice Boltzmann Equation for fluid dynamics and beyond* (Oxford University Press, Oxford 2001)
19. N. Thummala, D.V. Papavassiliou: Simulation of heat transfer with LBM and Lagrangian methods for microfluidic applications, paper HT2005-72313, *CD-ROM Proceedings 2005 (July 17-22) ASME Summer Heat Transfer Conference* (San Francisco, CA 2005)
20. K. Kontomaris, T.J. Hanratty, J.B. McLaughlin: An algorithm for tracking fluid particles in a spectral simulation of turbulent channel flow. *J. Comput. Phys.* **103**, 231–242 (1993)
21. A. Einstein: Uber die von der molekular-kinetischen Theorie der Warme geforderte Bewegung von in ruhenden Flussigkeiten suspendierten Teilchen. *Ann. d. Phys.* **17**, 549 (1905)
22. P.K. Yeung, S.B. Pope: An algorithm for tracking fluid particles in numerical simulations of homogeneous turbulence. *J. Comp. Phys.* **79**(2), 373–416 (1988)
23. S. Balachandar, M.R. Maxey: Methods for evaluating fluid velocities in spectral simulations of turbulence. *J. Comp. Phys.* **83** (1), 96–125 (1989)

24. D.V. Papavassiliou: Scalar dispersion from an instantaneous line source at the wall of a turbulent channel for medium and high Prandtl number fluids. *Int. J. Heat and Fluid Flow* **23** (2), 161–172 (2002)
25. B.M. Mitrovic, D.V. Papavassiliou: Transport properties for turbulent dispersion from wall sources, *AIChE J.,* **49** (5), 1095–1108 (2003)
26. P.G. Saffman: On the effect of the molecular diffusivity in turbulent diffusion. *J. Fluid Mech.* **8**, 273–283 (1960)
27. Y. Mito, T.J. Hanratty: Lagrangian stochastic simulation of turbulent dispersion of heat markers in a channel flow, *Int. J. Heat Mass Transfer* **46**(6), 1063–1073 (2003)
28. D.V. Papavassiliou: Turbulent transport from continuous sources at the wall of a channel, *Int. J. Heat Mass Transfer* **45**, 3571–3583 (2002)
29. B.M. Mitrovic, P.M. Le, D.V. Papavassiliou: On the Prandtl or Schmidt number dependence of the turbulence heat or mass transfer coefficient. *Chem. Eng. Sci.* **59**(3), 543–555 (2004)
30. P.M. Le, D.V. Papavassiliou: Turbulent dispersion from elevated sources in channel and Couette flow. *AIChE J.* **51** (9), 2402–2414 (2005)
31. S.L. Lyons, T.J. Hanratty, J.B. McLaughlin: Direct numerical simulation of passive heat transfer in a turbulent channel flow. *Int. J. Heat Mass Transfer* **34** (4/5), 1149–1161 (1991)
32. Y. Na, D.V. Papavassiliou, T.J. Hanratty: Use of Direct Numerical Simulation to study the effect of Prandtl number on temperature fields. *Int. J. Heat and Fluid Flow* **20** (3), 187–195 (1999)
33. H. Kawamura, K. Ohsaka: DNS of turbulent heat transfer in channel flow with low to medium-high Prandtl number fluid. *Int. J. Heat and Fluid Flow* **19**, 482–491 (1998)
34. I. Tiselj, E. Pogrebnyak, L. Changfeng, A. Mosyak, G. Hetsroni: Effects of wall boundary condition on scalar transfer in a fully developed turbulent flume,*Phys. Fluids* **13** (4), 1028–1039 (2001)
35. G. Hetsroni, I. Tiselj, R. Bergant, A. Mosyak, E. Pogrebnyak: Convection velocity of temperature fluctuations in a turbulent flume. *J. Heat Transfer-Trans. ASME* **126** (5), 843–848 (2004)
36. R.A. Antonia, P. Orlandi: Effect of Schmidt number on small-scale passive scalar turbulence. *Appl. Mech. Rev.* **56** (6), 615–632 (2003)
37. G. Brethouwer, J.C.R. Hunet, F.T.M. Nieuwstadt: Micro-structure and Lagrangian statistics of the scalar field with a mean gradient in isotropic turbulence. *J. Fluid Mech.* **474**, 193–225 (2003)
38. P.K. Yeung, S. Xu, K.R. Sreenivasan: Schmidt number effects on turbulent transport with uniform mean scalar gradient. *Phys. Fluids* **14** (2), 4178–4191 (2002)
39. P.K. Yeung, S. Xu, D.A. Donzis, K.R. Sreenivasan: Simulations of three-dimensional turbulent mixing for Schmidt numbers of the order 1000. *Flow Turbulence and Combustion* **72** (2-4), 333–347 (2004)
40. M.S. Borgas, B.L. Sawford, S. Xu, D.A. Donzis, P.K. Yeung: High Schmidt number scalars in turbulence: Structure functions and Lagrangian theory. *Phys. Fluids* **16** (11), 3888–3899 (2004)
41. S. Chandrasekhar: Stochastic problems in Physics and Astronomy. *Rev. of Modern Physics* **15** (1), 1–89 (1943)
42. M. Meyyappan: *Carbon Nanotubes Science and Applications* (CRC Press, Boca Raton 2005)

43. P. Kim, L. Shi, A. Majumdar, P.L. McEuen: Thermal transport measurements of individual multiwalled nanotubes. *Phys. Rev. Lett.* **87**, 215502-1 (2001)
44. S. Berber, Y.K. Kwon, D. Tomanek: Unusually high thermal conductivity of Carbon nanotubes. *Phys. Rev. Lett.* **84**, 4613 (2000)
45. R.B. Bird, W.S. Stewart, E.N. Lightfoot: *Transport Phenomena*, 2nd Edition, pp. 282, 376 and 397 (John Wiley & Sons Inc., New York, 2002)
46. M.J. Biercuk, M.C. Llaguno, M. Radosavljevic, J.K. Hyun, A.T. Johnson: Carbon nanotube composites for thermal management. *Appl. Phys. Lett.* **80**, 2767–2779 (2002)
47. P.L. Kapitza: The Study of Heat Transfer in Helium II. *J. Phys. USSR* **4**, 181–210 (1941)
48. D.G. Cahill, W.K. Ford, K.E. Goodson, A. Majumdar, H.J. Maris, S.R. Phillpot, Nanoscale thermal transport. *J. Appl. Phys.* **93** (2), 793–818 (2003)
49. J.A. Eastman, S.R. Phillpot, S.U.S. Choi, P. Klebinski: Thermal transport in nanofluids. *Annu. Rev. Mater. Res.* **34**, 219–246 (2004)
50. E.T. Swartz, R.O. Pohl: Thermal-Boundary Resistance. *Rev. of Modern Physics* **61** (3), 605–668 (1989)
51. J-L. Barrat, F. Chiaruttini: Kapitza resistance at the liquid-solid interface. *Mol. Phys.* **101**, 1605–1610 (2003)
52. C-J. Twu, J-R. Ho: Molecular dynamics study of energy flow and Kapitza conductance across an interface with imperfection formed by two dielectric thin films. *Phys. Rev. B* **67**, 205400 (2003)
53. P. Chantrenne, J-L Barrat: Finite size effects in determination of thermal conductivities: Comparing molecular dynamics results with simple models. *J. Heat Transf., ASME Transactions* **126**, 577–585 (2004)
54. Q. Tang: A molecular dynamics simulation: the effect of finite size on the thermal conductivity in a single crystal silicon. *Mol. Phys.* **102** (18), 1959–1964 (2004)
55. A. Maiti, G.D. Mahan, S.T. Pantelides: Dynamical simulations of nonequilibrium processes – Heat flow and the Kapitza resistance across grain boundaries. *Solid St. Commun.* **102**, 517–521 (1997)
56. S. Shenogin, L. Xue, R. Ozisik, P. Keblinski, D.G. Cahill: Role of thermal boundary resistance on heat flow in Carbon-nanotube composites. *J. Appl. Phys.* **95**, 8136–44 (2004)
57. M.S. Toprak, C. Stiewe, D. Platzek, S. Williams, L. Bertini, E. Muller, C. Gatti, Y. Zhang, M. Rowe, M. Muhammed: The impact of nanostructuring on the thermal conductivity of thermoelectric CoSb3. *Adv. Funct. Mater.* **14** (12), 1189–1196 (2004)
58. H.M. Duong, D.V. Papavassiliou, L.L. Lee, K.J. Mullen: Random walks in nanotube composites: Improved algorithms and the role of thermal boundary resistance. *Appl. Phys. Lett.* **87**, 013101 (2005)
59. M.M. Tomadakis, S.V. Sotirchos: Transport properties of random arrays of freely overlapping cylinders with various orientation distributions. *J. Chem. Phys.* **98**, 616–626 (1993)
60. M.M. Tomadakis, S.V. Sotirchos: Transport through random arrays of conductive cylinders dispersed in a conductive matrix. *J. Chem. Phys.* **104**, 6893–6900 (1996)
61. H.S. Carslaw, J.C. Jaeger: *Conduction of Heat in Solids* 2nd edition, p.97 (Oxford University Press 1959)

62. J. Judy, D. Maynes, B.W. Webb: Characterization of frictional pressure drop for liquid flows through microchannels, *Int. J. Heat Mass Transf.* **45**, 3477–3489 (2002)
63. F.J. Alexander, S. Chen, J.D. Sterling: Lattice Boltzmann thermohydrodynamics. *Physical Review E* **47**, 2249–2252 (1993)
64. X. Shan: Solution of Rayleigh-Bénard convection using a Lattice Boltzmann method. *Physical Review E* **55**, 2780–2788 (1997)
65. X. He, S. Chen, G.D. Doolen: A novel thermal model for the Lattice Boltzmann method in incompressible limit. *J. Comp. Phys.* **146**, 282–300 (1998)
66. B.J. Palmer, D.R. Rector: Lattice Boltzmann algorithm for simulating thermal flows in compressible fluids. *J. Comp. Phys.* **161**, 1–20 (2000)
67. G.H. Tang, W.Q. Tao, Y.L. He: Simulation of fluid and heat transfer in a plane channel using the Lattice Boltzmann method. *Int. J. Modern Physics B* **17** (1&2), 183–187 (2003)
68. Y.H. Qian, D. d'Humieres, P. Lallemand: Lattice BGK models for Navier–Stokes equation *Europhysics Letters* **17** (6), 479–484 (1992)
69. S. Wolfram: Cellular automation fluids, 1: Basic Theory: *J. Stat. Phys.* **45**, 471–526 (1986)
70. P. Lavallee, J.P. Boon, A. Noullez: Boundaries in Lattice gas flows: *Physica D* **47**, 233–240 (1991)
71. D.P. Ziegler: Boundary conditions for Lattice Boltzmann simulations. *J. Stat. Phys.* **71**, 1171–1177 (1993)
72. G.R. McNamara, G. Zanetti: Use of the Boltzmann equation to simulate lattice-gas automata. *Phys. Rev. Lett.* **61** (20), 2332–2335 (1988)
73. P.L. Bhatnagar, E.P. Gross, M. Krook: A model for collision processes in gases. I. small amplitude processes in charged and neutral one-component systems. *Phys. Rev.* **94** (3), 511–525 (1954)
74. J.M.V.A. Koelman: A simple lattice-Boltzmann scheme for Navier–Stokes fluid flow. *Europhysics Letters* **15** (6), 603–607 (1991)
75. D.R. Noble: Lattice Boltzmann study of the interstitial hydrodynamics and dispersion in steady inertial flows in large randomly packed beds, PhD Thesis, University of Illinois, Urbana–Champaign, Illinois, 1997.
76. N. Thummala: Convective heat transfer in microfluidics using Lagrangian methods and lattice Boltzmann simulations, MS Dissertation, University of Oklahoma, Norman (2004).
77. N-T. Nguyen, S.T. Wereley, *Fundamentals and applications of microfluidics* (Artech House, Boston 2002)
78. H. Tennekes, J.L. Lumley: *A First Course In Turbulence* p. 96 (MIT Press, Boston 1972)
79. I. Calmet, J. Magnaudet: Large-Eddy simulation of high-Schmidt number mass transfer in a turbulent channel flow. *Phys. Fluids* **9** (2), 438–454 (1997)
80. Y. Na, T.J. Hanratty: Limiting behavior of turbulent scalar transport close to a wall. *Int. J. Heat Mass Trans.* **43** (10), 1749–1758 (2000)
81. S.A. Orszag, L.C. Kells: Transition to turbulence in plane Poisseuille and plane Couette flow. *J. Fluid Mech.* **96**, 159–205 (1980)
82. P.S. Marcus: Simulation of Taylor–Couette flow. *J. Fluid Mech.* **146**, 45–64 (1984)
83. S.W. Churchill: Progress in the thermal sciences: AICHE Institute Lecture. *AIChE J.* **46** (9), 1704–1722 (2000)

84. A.S. Monin, A.M. Yaglom: *Statistical Fluid Mechanics: Volume 1, Mechanics of Turbulence*, pp. 279–282, (MIT Press, Cambridge, MA 1965)
85. V.G. Levich: *Physicochemical Hydrodynamics* (Prentice-Hall, Englewood Cliffs, NJ 1962)
86. D.A. Shaw, T.J. Hanratty: Turbulent Mass Transfer Rates to a wall for large Schmidt numbers. *AIChE J.* **23** (1), 28–37 (1977)
87. C.A. Petty: A statistical theory for mass transfer near interfaces. *Chem. Eng. Sci.* **30**, 413–418 (1975)
88. P.M. Le, D.V. Papavassiliou: Turbulent heat transfer in plane Couette flow. *J. of Heat Transf., Trans. ASME* **128**, 53–62 (2006)
89. Y. Hasegawa, N. Kasagi: The effect of Schmidt number on air-water interface mass transfer. In: *Proceedings, 4th Int Conference on Multiphase Flow, May 2001*, (CD-ROM) (New Orleans, Louisiana 2001)
90. Y. Hasegawa, N. Kasagi, H. Hanazaki: Direct numerical simulation of passive scalar transfer across a turbulent gas-liquid interface. In: *Proceedings, First International Symposium on Advanced fluid Information, October 2001*, 696–701 (Sendai, Japan 2001)
91. S.S. Ponoth, J.B. McLaughlin: Numerical simulation of mass transfer for bubbles in water. *Chem. Eng. Sci.*. **55**, 1237–1255 (2000)
92. B.M. Mitrovic, D.V. Papavassiliou: Effects of a first-order chemical reaction on turbulent mass transfer. *Int. J. Heat Mass Transfer* **47** (1), 43–61 (2004)

# An Efficient Optimization Approach for Computationally Expensive Timesteppers Using Tabulation

A. Varshney and A. Armaou

Department of Chemical Engineering, Pennsylvania State University, University Park, PA-16802, USA, armaou@psu.edu

**Summary.** A methodology is outlined for the efficient solution of dynamic optimization problems when the system evolution is described by computationally expensive timestepper-based models. The computational requirements issue is circumvented by extending the notion of in situ adaptive tabulation to stochastic systems. Conditions are outlined that allow unbiased estimation of the mapping gradient matrix and, subsequently, expressions to compute the ellipsoid of attraction are derived. The proposed approach is applied towards the solution of two representative dynamic optimization problems, (a) a bistable reacting system describing catalytic oxidation of $CO$ and, (b) a homogeneous chemically reacting system describing dimerization of a monomer. In both cases, tabulation resulted in significant reduction in the solution time of the optimization problem.

## 1 Introduction

In recent years, there has been an increased focus towards atomistic/particle simulation techniques for process modeling in place of traditional continuum or mean-field approaches. The advantage of using atomistic models is their capability to describe phenomena whose characteristic length and time scales are much smaller than those for which the continuum approximation holds. Simulation methods such as Molecular Dynamics (MD), kinetic Monte-Carlo (kMC), Lattice-Boltzmann (LB) etc., have been utilized to model homogeneous reacting systems [7, 8], biological systems [23, 26, 29], microstructure evolution during thin-film growth [22], crack propagation [31] and fluid flow [21] to name a few. However, issues related to noise and high computational requirements have limited their applicability for process optimization and control.

Dynamic optimization or open-loop optimal control has been the focus of extensive research in recent years. The objective is to search for optimal input trajectories for dynamic plants which optimize certain performance measures. With the availability of detailed process models, considerable research effort is

being directed towards efficient solution of the corresponding optimal control problems [34, 35, 5, 15]. We are interested in the efficient solution of optimal control problems for processes whose evolution is described by microscopic simulations. Apart from the issue of noise inherent to microscopic simulations, such optimization problems are also subject to evolution constraints which are unavailable in closed form. The standard approach for the solution of such problems is to compute the objective functional as a "black-box", and employ direct search algorithms such as Hooke-Jeeves, Nelder-Mead, pattern search etc., to compute the optimal control trajectory [28, 2]. An alternative methodology for global optimization for nonlinear programs constrained by "nonfactorable" constraints (constraints defined by a computational model for which no explicit analytical representation is available) was proposed in [24]. However, the above approaches are inefficient if the computation of the cost functional (or the black-box simulation) is expensive, which is usually the case for atomistic simulations. There are a number of approaches specially designed for problems where computation of the objective function is expensive [3, 14], but most of them have been employed for noise-free systems which limits their applicability for the current problem.

Motivated by the above, we present an approach that addresses the issue of computational requirements during process optimization when the available process model is in the form of black-box timestepper. The approach employs tabulation of process data within a database, which is constructed during the solution of the dynamic optimization problem using situ adaptive tabulation (ISAT) method. Since ISAT tabulates the process data and process sensitivities during the simulations, it is computationally less demanding than direct tabulation since only the realizable region, which is the region of the parameter space traversed during the computations, is tabulated and which is usually a small subset of the whole state space. The critical issue of the computation of process sensitivities is addressed through the use of finite differences with common random numbers and conditions are outlined that allow their unbiased estimation. Subsequently, standard search algorithms can be employed for the solution of the optimization problem. We present two applications of the proposed approach in the context of a bistable reacting system and a reversible dimerization process. For the former case, we compute optimal time-varying profiles of the manipulated input which transforms the state of the system from one stationary state to another. For the latter case, we compute optimal time-varying profiles of the manipulated input such that output concentrations of the desired products are close to the set-point. We observe that incorporation of ISAT resulted in computational savings by two orders of magnitude.

The manuscript is organized as follows. In the next section, we formulate the dynamic optimization problem for processes whose evolution is described by black-box timesteppers. Subsequently, we describe the ISAT algorithm and derivative estimation using finite difference with common random numbers

which allows the extension of ISAT for stochastic systems. We conclude with the application of ISAT algorithm in the context of two illustrative examples.

## 2 Problem Formulation

We investigate the issue of dynamic optimization problem formulation for a class of systems described by the following discrete-time description [6]:

$$\begin{gathered}\mathbf{X}(t_{i+1}) = \mathit{\Pi}(\mathbf{X}(t_i), \delta t_i, \omega; \boldsymbol{\theta}_i) \\ \delta t_i = t_{i+1} - t_i\end{gathered} \tag{1}$$

where $\mathbf{X}(t_{i+1})$ and $\mathbf{X}(t_i) \in \Omega_1 \subset \mathbb{R}^n$ are the vector of states of the system at time instants $t_{i+1}$ and $t_i$, respectively, $t_{i+1}, t_i \in [0\ T]$, $\boldsymbol{\theta}_i \in \Omega_1 \subset \mathbb{R}^p$ is the control input vector which is constant for $t \in (t_i\ t_{i+1}]$ and $\omega$ is a random walk defined over some measurable space. Most of dynamic systems, continuous or discrete, can be expressed in the form given by Eq. 1 when only an input-output relationship is required. For example, spatially distributed parabolic partial differential equations (PDEs) arising frequently while modeling transport-reaction processes assume the above form where the right-hand side (RHS) is obtained from the appropriate spatial and temporal discretization of the PDE. For systems which are modeled using atomistic simulations, such as kMC, the RHS represents the corresponding evolution rule. The function $\mathit{\Pi}(\cdot)$ in this case is fundamentally different from the one obtained by discretizing PDEs as it is unavailable in closed-form. The latter systems, for which the function $\mathit{\Pi}(\cdot)$ is a "black-box", are the primary motivation of the current work. We assume the following smoothness assumption with respect to parameters for the process $\mathbf{X}(t_i)$:

**Assumption 2.1** *The stochastic process* $\mathbf{X}(t_i, \boldsymbol{\theta}, \mathbf{x})$ *with* $\mathbf{X}(0, \boldsymbol{\theta}, \mathbf{x}) = \mathbf{x}$ *defined over the probability space* $[\Omega, \Sigma, P_{\boldsymbol{\theta}}]$ *is twice continuously differentiable with respect to* $\boldsymbol{\theta} \in \Theta$ *and* $\mathbf{x} \in \mathbb{R}^n$ *for all* $\omega \in \Omega$ *with probability one.*

We are interested in computing an optimal time-varying profile of the control input, $\theta^*(t)$, such that a particular goal for the *averaged* process dynamics is realized. Such profile can be obtained from the solution of the following dynamic optimization problem:

$$\begin{gathered}\min_{\theta(t)} \int_0^{t_f} \mathcal{Q}(E(\mathbf{X}), \theta)\, \mathrm{d}t + \mathcal{W}(|E(\mathbf{X}(t_f)) - \bar{\mathbf{X}}(t_f)|) \\ s.t. \\ \mathbf{X}(t_{i+1}) = \mathit{\Pi}(\mathbf{X}(t_i), \delta t_i, \omega; \boldsymbol{\theta}_i),\ \delta t_i = t_{i+1} - t_i \\ g_d(\mathbf{X}, \theta) \leq 0\end{gathered} \tag{2}$$

where $\mathcal{Q}$ is a scalar cost function, $\mathcal{W}$ is an appropriate final-time penalty function, and $g_d$ denotes the set of inequality constraints on state and manipulated

variables. Discretizing the time interval $[0, t_f]$ into $N$ sub-intervals and assuming $\theta(t)$ to be piecewise constant during each sub-interval, we can obtain a finite-dimensional approximation to the above dynamic optimization problem. However, the equality constraints cannot be handled explicitly during optimization due to their unavailability in closed-form. The standard approach for the solution of the above optimization problem is to compute the objective functional as a black-box during optimization and employ derivative-free optimization algorithms such as Nelder-Mead, Hooke-Jeeves, pattern search [16], etc. to compute $\theta^*(t)$. However, if the computation of the objective functional is expensive, which requires the simulation of the timestepper for the period $[0, t_f]$, the solution time required may become prohibitive. To address this issue, we extend the applicability of ISAT to accomplish efficient simulation of the stochastic timestepper, resulting in efficient solution of the optimization problem.

## 3 In Situ Adaptive Tabulation

The in-situ adaptive tabulation (ISAT) scheme was originally developed for deterministic systems in the context of efficient implementation of combustion chemistry [25]. In this section, a brief overview of the original algorithm is provided (for details refer to [25, 33]). Consider a dynamically evolving process with the following state-space description:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}) = \mathbf{f}(\boldsymbol{\phi}) \tag{3}$$

where $\mathbf{x} \in \Omega_1 \subset \mathbb{R}^n$ is the vector of state variables and $\mathbf{u} \in \Omega_2 \subset \mathbb{R}^p$ is the vector of control variables; The vector $\boldsymbol{\phi}$ is defined as $\boldsymbol{\phi} = [\mathbf{x}\ \mathbf{u}]^T$, $\boldsymbol{\phi} \in \Omega = \Omega_1 \times \Omega_2$. We define $\boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_0)$ to be a nonlinear integral operator representing the evolution of the system from initial state $\boldsymbol{\phi}_0$ at time $t_0$ to state $\boldsymbol{\phi}(t_0+\tau) = \boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_0)$ at time $t_0+\tau$ (the time-step, $\tau$, will be henceforth referred as ISAT-reporting horizon). To reduce computational costs it is desired to approximate $\boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_q)$ due to a "nearby" (in a sense that will become clear later) state $\boldsymbol{\phi}_q$, based on the knowledge of $\{\boldsymbol{\phi}_0, \boldsymbol{\mathcal{R}}\}$. One way to address this issue is to tabulate a large number of doublets $\{\boldsymbol{\phi}, \boldsymbol{\mathcal{R}}(\boldsymbol{\phi})\}$ regularly spanning the whole realizable region $\Omega$ into a database (an $(n+p)$-dimensional mesh), and subsequently interpolate within this database to estimate $\boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_q)$. The interpolation error that is incurred can be controlled through refining the mesh. However, the generation of the database, which is usually done in a *pre-processing* phase, can become cumbersome if the dimensionality of state-space (i.e. $\Omega$) is large.

We, define the *accessed*-region, $\Omega_a$ ($\Omega_a \subset \Omega$), as the set of all states $\boldsymbol{\phi}$ that occur in the calculation. A crucial observation is that the accessed region is much smaller than the realizable region. Exploiting this fact, ISAT constructs the database *online* and hence tabulates only the accessed region, $\Omega_a$. Moreover, to control the interpolation errors, the *mapping gradient matrix* is also

computed (and tabulated) which is defined as:

$$A_{ij}(\phi) \equiv \frac{\partial \mathcal{R}_i(\phi)}{\partial \phi_j} \tag{4}$$

Consider a tabulated triplet $\{\boldsymbol{\phi}_p, \boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_p), \mathbf{A}(\boldsymbol{\phi}_p)\}$. A linear interpolation for $\boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_q)$ can be obtained as:

$$\begin{gathered} \boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_q) \approx \boldsymbol{\mathcal{R}}^l(\boldsymbol{\phi}_q) \equiv \boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_0) + \delta\boldsymbol{\mathcal{R}}^l \\ \delta\boldsymbol{\mathcal{R}}^l \equiv \mathbf{A}(\boldsymbol{\phi}_p)\delta\boldsymbol{\phi} + \mathcal{O}(\mid \delta\boldsymbol{\phi} \mid^2), \quad \delta\boldsymbol{\phi} = \boldsymbol{\phi}_q - \boldsymbol{\phi}_p \end{gathered} \tag{5}$$

The error induced due to the interpolation can be analyzed as follows. Assume $\boldsymbol{\phi}_p$ and $\boldsymbol{\phi}_q$ are such that $\mid \boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_q) - \boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_p) \mid \leq \epsilon_{tol}$. It follows from above that:

$$\delta\boldsymbol{\phi}^T \mathbf{A}^T(\boldsymbol{\phi}_p)\mathbf{A}(\boldsymbol{\phi}_p)\delta\boldsymbol{\phi} \leq \epsilon_{tol}^2 \tag{6}$$

Eq.6 defines a hyper-ellipsoid (referred as Ellipsoid of Attraction, EOA) centered at $\boldsymbol{\phi}_p$ with principle axes given by elements of the diagonal matrix $\boldsymbol{\Sigma}$ such that $\boldsymbol{Q}^T\boldsymbol{\Sigma}Q$ is the singular value decomposition of $\mathbf{A}$. Now given any query $\boldsymbol{\phi}_q$, if there exist a tabulated $\boldsymbol{\phi}_p$ such that Eq.5 is valid, the error due to interpolation will be less than $\epsilon_{tol}$. If such a $\boldsymbol{\phi}_p$ is not found in the database, direct integration of Eq.3 is performed and stored in the database.

The matrix $\mathbf{A}(\boldsymbol{\phi})$ can be related to sensitivity coefficients. The first-order sensitivity coefficients with respect to the initial conditions are defined as:

$$B_{ij}(\boldsymbol{\phi}_0, t) \equiv \frac{\partial \phi_i(t)}{\partial \phi_0^j} \tag{7}$$

From the above, it can be seen that

$$\mathbf{A}(\boldsymbol{\phi}_0) = \mathbf{B}(\boldsymbol{\phi}_0, \tau). \tag{8}$$

ISAT is implemented in practice using a binary tree. Ideally, once a query point, $\boldsymbol{\phi}_q$, is generated, one would like to determine $\boldsymbol{\phi}_0$ that is *closest* to $\boldsymbol{\phi}_q$ for interpolation purposes. However, a complete database search for $\boldsymbol{\phi}_0$ could be expensive, especially if the database size is large. To circumvent this problem, the database is organized as a binary tree comprising of *leafs* and *nodes*. Each node contains the information regarding a convex region that is *likely* to be spanned by the corresponding leafs, which, in turn, contain the *record* comprising of $\boldsymbol{\phi}_0$, $\boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_0)$, $\mathbf{A}(\boldsymbol{\phi}_0)$, $\boldsymbol{Q}$ and $\sigma$. The convex region contained within each node is characterized by a vector $\mathbf{v}$ and scalar $a$ such that leafs pertaining to *sub-region* $\mathbf{v}^T\boldsymbol{\phi} < a$ are on the left and leafs corresponding to sub-region $\mathbf{v}^T\boldsymbol{\phi} > a$ are on the right. The division into a number of convex regions allows efficient search of the point within the database which most likely to be nearest to the query point.

When the database is probed with a query ($\boldsymbol{\phi}_q$), three distinct possibilities may arise:

1. $\boldsymbol{\phi}_q$ lies within EOA of $\boldsymbol{\phi}_0$. In this case, presented in Fig. 1a, the corresponding integral map based on interpolation around $\boldsymbol{\phi}_0$ (Eq. 5) is returned.
2. $\boldsymbol{\phi}_q$ lies outside EOA of $\boldsymbol{\phi}_0$; $\boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_q)$ is computed through simulation and post simulation it is observed that $|\boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_q) - \boldsymbol{\mathcal{R}}^l(\boldsymbol{\phi}_q)| < \epsilon_{tol}$. In this case, presented in Fig. 1b, the EOA around $\boldsymbol{\phi}_0$ is *grown* to include $\boldsymbol{\phi}_q$; the calculated $\boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_q)$ is returned.
3. $\boldsymbol{\phi}_q$ lies outside EOA of $\boldsymbol{\phi}_0$; $\boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_q)$ is computed through simulation and post simulation it is observed that $|\boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_q) - \boldsymbol{\mathcal{R}}^l(\boldsymbol{\phi}_q)| > \epsilon_{tol}$. In this case, presented in Fig. 1c, the database is augmented by a record for $\boldsymbol{\phi}_q$ and the original leaf, $\boldsymbol{\phi}_0$ is replaced by a node. The records for $\boldsymbol{\phi}_0$ and $\boldsymbol{\phi}_q$ are stored as left and right leafs, respectively, of the new node; the calculated $\boldsymbol{\mathcal{R}}(\boldsymbol{\phi}_q)$ is returned.

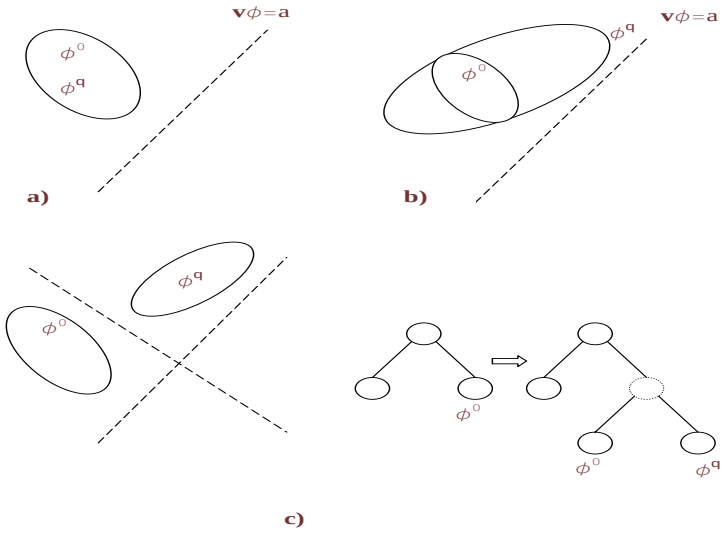



Fig. 1: Three possibilities that will arise once the ISAT database is probed with a query. a) $\phi^q$ lies within EOA of $\phi^0$. b) $\phi^q$ lies outside EOA of $\phi^0$, but $|\mathcal{R}(\phi^q) - \mathcal{R}^l(\phi^q)| < \epsilon_{tol}$. c) $\phi^q$ lies outside EOA of $\phi^0$, and $|\mathcal{R}(\phi^q) - \mathcal{R}^l(\phi^q)| > \epsilon_{tol}$.

In contrast to deterministic black-box systems, the problem of derivative estimation for stochastic black-box systems is complex due to the issues of bias and variance. For example, finite difference approximations cannot be directly employed in Eq.7 to obtain first-order sensitivity matrix. An extensive amount of literature is available addressing this issue; important techniques include Finite Difference/Finite Difference with Common Random Numbers (FD/FDC)

[18, 36, 9, 4], Infinitesimal Perturbation Analysis (IPA) [17, 12, 13, 36, 30] and Likelihood Ratio estimation (LR) [27, 10, 11, 17]. In our implementation of stochastic-ISAT, FDC was employed for derivative estimation. In the following subsection we discuss sufficient conditions for unbiasedness and finite-variance in derivative estimation using FDC.

### 3.1 FDC Derivative Estimation and EOA

Consider a stochastic process $\mathbf{X}(t, \boldsymbol{\theta}),\ \mathbf{X} \in \mathbb{R}^n\ \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ defined over a probability space $(\Omega, \Sigma, P_{\boldsymbol{\theta}})$ and let $\mathbf{X}(t, \boldsymbol{\theta}, \omega) \mid t \geq 0,\ \omega \in \Omega$ denote a sample path. For the ease of notation we assume $n = m = 1$ for the rest of the discussion. Let $X'(t, \theta_0)$ be the derivative $\dfrac{\partial X(t,\theta)_{\theta=\theta_0}}{\partial \theta}$ for some $\theta_0 \in \Theta$, assuming it exists. FD and FDC estimates of the derivative $X'(t, \theta_0)$ are defined as follows:

$$\begin{aligned} \bar{X}'^{,FD}(t, \theta_0) &= \frac{1}{N} \sum_{i=1}^{N} X_i'^{,FD}(t, \theta_0) \\ X_i'^{,FD}(t, \theta_0) &= \frac{X(t, \theta_0 + \delta\theta, \omega') - X(t, \theta_0, \omega)}{\delta\theta} \end{aligned} \tag{9}$$

$$\begin{aligned} \bar{X}'^{,FDC}(t, \theta_0) &= \frac{1}{N} \sum_{i=1}^{N} X_i'^{,FDC}(t, \theta_0) \\ X_i'^{,FDC}(t, \theta_0) &= \frac{X(t, \theta_0 + \delta\theta, \omega) - X(t, \theta_0, \omega)}{\delta\theta} \end{aligned} \tag{10}$$

An immediate issue arising due to the above definitions is the appropriate choice of $N$ and $\delta\theta$ that would guarantee satisfactory unbiasedness and accuracy of the derivative estimates. To make these concepts more precise, we define the following *loss function* [18]:

**Definition 1.** *The loss function associated with a derivative estimator $\bar{X}'(t, \theta_0)$ based on $N$ samples is defined as:*

$$\begin{aligned} R_N = E[\bar{X}'(t, \theta_0) - X'(t, \theta_0)]^2 = VAR(\bar{X}'(t, \theta_0)) + B_N^2 \\ B_N = E[\bar{X}'(t, \theta_0)] - X'(t, \theta_0) \end{aligned} \tag{11}$$

*where the first term denotes the variance of the derivative estimators and the second term denotes the bias. Also the convergence rate is said to be $\mathcal{O}(f(N))$ if $R_N \in \mathcal{O}(f(N))$.*

For the variance of FD and FDC estimators, we state the following result from [9]:

**Theorem 1.** *Suppose that $X(t, \theta, \omega)$ is described by Eq.1 and assumption 2.1 holds, then for $\theta_0 \in \Theta$, $VAR[X(t, \theta_0 + \delta\theta, \omega') - X(t, \theta_0 + \delta\theta, \omega)]$ is*

*(i)* $\mathcal{O}(1)$ *if* $\omega$ *and* $\omega'$ *are independent.*
*(ii)* $\mathcal{O}(\delta\theta^2)$*, if* $\omega = \omega'$*.*

Theorem 1 states that variance of FD derivative estimators tends to infinity as $\delta\theta \to 0$. However, using FDC the variance can be made vanishingly small. Next, we state the following theorem from [18] to establish the convergence of FDC estimates:

**Theorem 2.** *Suppose that assumption 2.1 holds and* $\Psi(\omega)$ *defined as*

$$\Psi(\omega) = \begin{cases} \sup_{\theta\in\Theta} \mid \bar{X}'(\theta,\omega) \mid, \text{ if } \omega \in \Omega; \\ 0 \qquad\qquad\qquad\quad \text{otherwise}, \end{cases}$$

*is such that* $\Psi(\omega) \leq \Gamma(\omega)$ *for some function* $\Gamma : \Omega \to \mathbb{R}$*. If* $\Gamma(\omega)$ *satisfies* $\int_\Omega [\Gamma(\omega)]^2 \, \mathrm{d}P(\omega) < \infty$*, then*

$$R_N^{FDC} = \sigma_{FDC}^2/N + [X''(\xi^+)\delta\theta/4]^2 \tag{12}$$

*where* $\theta_0 \leq \xi^+ \leq \theta_0 + \delta\theta$*. As a consequence, the convergence rate of FDC estimates is* $\mathcal{O}(N^{-1/2})$ *provided that* $\delta\theta \in \mathcal{O}(N^{-1/2})$*. In the limit* $\delta\theta \to 0$*, the bias also vanishes.*

Combination of theorems 1 and 2 forms the theoretical rationale behind the computation of derivatives based on finite differences with common random numbers. From the simulation point of view, FDC can be implemented by resetting the random seed of the random number generator while evaluating $X(t, \theta_0 + \delta\theta, \omega)$ and $X(t, \theta_0, \omega)$.

We now formally define the EOA for systems governed by equations of the form Eq.1:

**Definition 2.** *Let* $\mathbf{X}(t, \boldsymbol{\alpha}, \mathbf{x})$*,* $\mathbf{X}(\cdot) \in \mathbb{R}^n$ *be a stochastic process governed by Eq.1 such that* $\mathbf{X}(0, \boldsymbol{\alpha}, \mathbf{x}) = \mathbf{x}$ *and let* $\mathcal{G} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^{n\times(n+p)}$ *be the first order sensitivity matrix defined as:*

$$\mathcal{G} = \left[\frac{\partial \mathbf{X}}{\partial \mathbf{x}} \;\; \frac{\partial \mathbf{X}}{\partial \boldsymbol{\alpha}}\right]. \tag{13}$$

*Then the state* $\mathbf{z}' = [\mathbf{x}' \;\; \boldsymbol{\alpha}']^T, \mathbf{z} \in \mathbb{R}^{n+p}$ *lies within the Ellipsoid of Attraction of* $\mathbf{z} = [\mathbf{x} \;\; \boldsymbol{\alpha}]^T$ *if:*

$$(\mathbf{z}' - \mathbf{z})^T \mathcal{G}^T \mathcal{G} (\mathbf{z}' - \mathbf{z}) \leq \epsilon_{tol}^2 \tag{14}$$

In the next section, we present two applications of the above scheme when the underlying dynamical system is modeled by a timestepper based description.

## 4 Applications

We initially consider a kinetic model describing $CO$ oxidation by $O_2$ on a catalytic surface [1]. The model involves Langmuir adsorption for $CO$, dissociative adsorption of $O_2$ and second-order surface reaction to produce $CO_2$,

which desorbs instantaneously. The overall reaction can be summarized as $A + 1/2B_2 \rightarrow AB$, where $A$, $B$ and $C$ represent $CO$, $O$ and $CO_2$, respectively. The mean-field relationship between the surface coverage of $A$ and $B$, denoted as $\theta_A$ and $\theta_B$ respectively, and adsorption, desorption and surface reaction rates are obtained as [1]:

$$\begin{aligned} \mathrm{d}\theta_A/\mathrm{d}t &= \alpha(1-\theta_A-\theta_B) - \gamma\theta_A - 4k_r\theta_A\theta_B \\ \mathrm{d}\theta_B/\mathrm{d}t &= 2\beta(1-\theta_A-\theta_B)^2 - 4k_r\theta_A\theta_B, \end{aligned} \tag{15}$$

where $\alpha$, $\beta$, $\gamma$ and $k_r$ denote the adsorption rate of $A$ and $B$, desorption rate of $A$ and surface reaction rate, respectively. The system has been reported to exhibit multiple steady-states for a range of values of $\beta$ [1]. In the absence of mean-field equations, the stochastic simulation algorithm (SSA) [7] can be employed to obtain the profiles of $\theta_A(t)$ and $\theta_B(t)$. SSA samples the underlying chemical master equation and provides unbiased realizations of the system which converge to the mean-field solution when the number of particles employed during simulation tends to infinity. SSA proceeds, given $\mathbf{X}(t)$ to be the state of the system at any time instant $t$, by choosing *exactly one* of the events, $j$, to occur between $t$ and $t + \mathrm{d}t$ with a probability proportional to their current propensity functions, $a_j$, and then jumping forward in time by an interval, $\delta t$, during which *exactly no* event occurs, given by:

$$\delta t = \frac{-\ln(r)}{a_0(x)},\ a_0(x) = \sum_{j=1}^{M} a_j(x) \tag{16}$$

The above probabilistic description of the system is equivalent to a timestepper, and we employ the tabulation scheme presented in the previous section for the solution of the following constrained dynamic optimization problem [1]:

$$\begin{gathered} \min_{\beta(t)} = F(\beta(t)) \\ F = \int_0^{NT} (\beta(t)-\beta_{ss})^2(1-0.3e^{-t})T\sum_{i=1}^{N}\delta(t-iT)\,\mathrm{d}t + \\ 50[1-e^{-R(|\theta_A(t_f)-\theta_{A,ss,f}|-\epsilon)}e^{-R(|\theta_B(t_f)-\theta_{B,ss,f}|-\epsilon)}] \\ s.t., \\ \theta_A(t=0)=\theta_{A,ss,i},\ \theta_B(t=0)=\theta_{B,ss,i} \end{gathered} \tag{17}$$

The optimization objective is to compute an optimal adsorption rate profile, $\beta(t)$, such that state of the system switches from an initial stable stationary state, $\theta_{ss,i}$, to another stable stationary state, $\theta_{ss,f}$, in time $t_f = NT$, traversing an unstable stationary state, $\theta_{ss,u}$ (see Table 1 for the values of relevant parameters). The constraints for the optimization problem arise from the (unavailable in closed form) dynamic evolution rule (SSA in the present case). To solve the optimization problem, the interval $[0, t_f]$ was parameterized into $N$ sub-intervals and with a piecewise-constant variation of the manipulated

Table 1: CO oxidation process parameters

| Parameter | Value | Steady states | |
|---|---|---|---|
| $k_r$ | 1.0 | $\theta_{A,ss,i}$ | .13944 |
| $\alpha$ | 1.6 | $\theta_{A,ss,u}$ | .67526 |
| $\gamma$ | 0.04 | $\theta_{A,ss,f}$ | .97101 |
| $\beta_{ss}$ | 3.5 | $\theta_{B,ss,i}$ | .63553 |
| $t_f$ | 5s | $\theta_{B,ss,u}$ | .11452 |
| | | $\theta_{B,ss,f}$ | .00137 |

input. For a given $\beta(t)$, the underlying timestepper was integrated in time to compute the objective functional $F$, so that the dynamic equality constraints are implicitly satisfied during the optimization. A pattern-search algorithm [32, 19, 20]† was employed for the solution of the above minimization problem. In order to reduce the computation of the timestepper, the ISAT algorithm was used. Initially, the problem was solved for $N = 10$ and $T = 0.5$s with an empty initial database. The database was concurrently built during the solution of the optimization problem and necessary interpolations, whenever possible, were performed during the dynamic simulation of the system. The ISAT reporting horizon was $\tau = 0.01$s, so that the database was queried 500 times per objective functional evaluation. The optimal profile for the manipulated input is shown in Fig. 2 and the resulting optimal trajectories for $\theta_A(t)$ and $\theta_B(t)$ are shown in Fig. 3. In Fig. 3 the profiles of $\theta_A(t)$ and $\theta_B(t)$ are also compared with the ones resulting from exact SSA simulations, which shows the accuracy of ISAT interpolations. In Fig. 4, the number of database interpolations and the number of timestepper evaluations performed are plotted as a function of objective functional evaluation during optimization. It is observed that, initially the number of timestepper evaluations required for the computation of $F$ is high, which, however, continuously decreases as the database size increases. A similar trend is observed for the wall-clock time spent per $F$ evaluation which is plotted in Fig. 5. The average number of timestepper evaluations and CPU time-spent per $F$ evaluation were 24 and 1.61s respectively, which is significantly lower than the corresponding values without interpolation, namely, 500 and 24s, respectively (see also Table 2). The CPU time reported are for a Pentium IV 3.01 GHz processor.

Subsequently, the optimization problem was solved with $N = 50$ and $T = 0.1$s to obtain an improved temporal resolution in $\beta(t)$. The database created previously was employed during subsequent computations. The resulting optimal profile for adsorption flux, $\beta(t)$, is shown in Fig. 6 and the number of database interpolations and number of timestepper evaluations performed as a function of $F$ evaluation during optimization are plotted in Fig. 7. The

† Software for pattern-search algorithm is available in Direct Search toolbox of MATLAB.

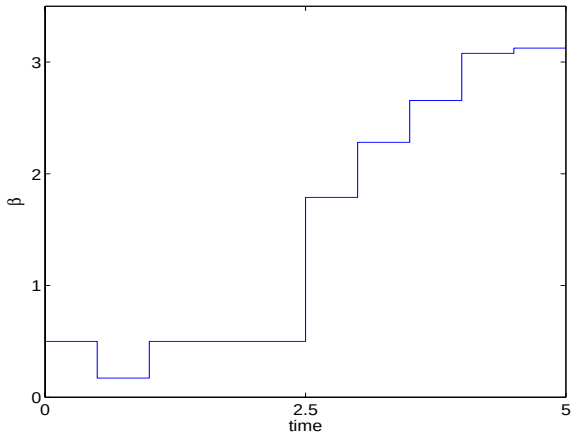



Fig. 2: Optimal control profile for $N = 10$.

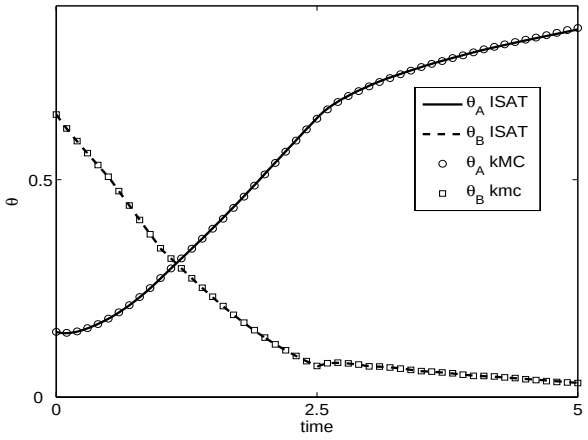



Fig. 3: Trajectories of $\theta_A(t)$ and $\theta_B(t)$ under the optimal profile of $\beta(t)$; $N = 10$.

optimal control trajectory shows a noisy behavior which is a manifestation of noisy behavior of the dynamic system resulting to performance deterioration of pattern search algorithms. The advantage of preexisting database is clearly evident in this case as the average number of timestepper evaluations per $F$ calculation reduced to 7 compared to 24 in the previous case. In Fig. 8, the CPU time-spent per $F$ evaluation is plotted, which is dominated by the time spent during database retrieval. Efficient data mining schemes can further decrease the required CPU time per iteration.

For the second application, we consider the following reaction sequence which describes the formation of a stable "dimer" from a monomer [8]:

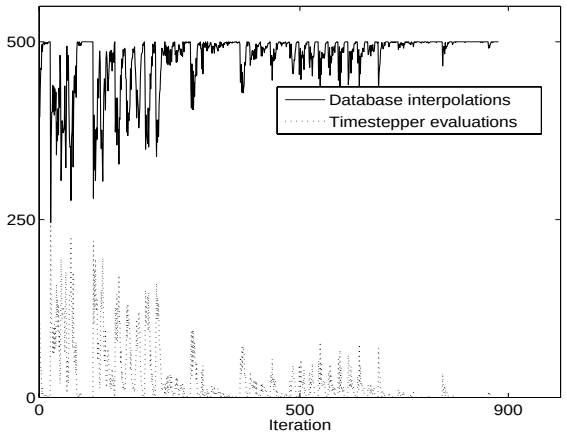



Fig. 4: Number of timestepper evaluations and database interpolations per objective function computation for $N = 10$.

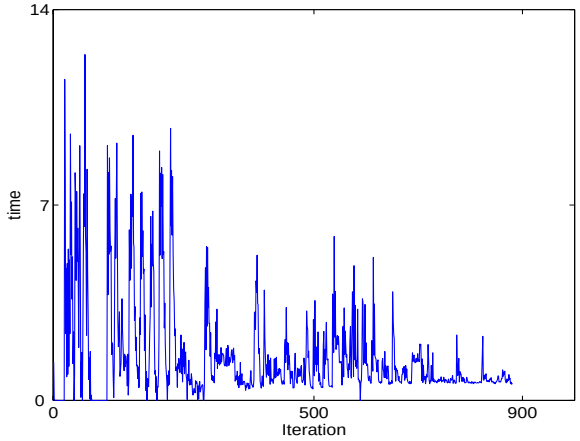



Fig. 5: CPU time spent per objective function computation for $N = 10$.

$$\begin{aligned} S_1 &\xrightarrow{c_1} 0 \\ S_1 + S_1 &\xrightarrow{c_2} S_2 \\ S_2 &\xrightarrow{c_3} S_1 + S_1 \\ S_2 + A &\xrightarrow{c_4} S_3 \end{aligned} \tag{18}$$

In the above scheme, the reversible dimerization of monomer $S_1$ into an unstable dimer $S_2$ is superimposed on the irreversible isomerization of $S_1$. $S_2$, in turn, isomerizes into a stable form $S_3$ in presence of an isomerizing agent $A$. It is assumed that the $A$ is present in excess. The assumed rate constants and the initial conditions are:

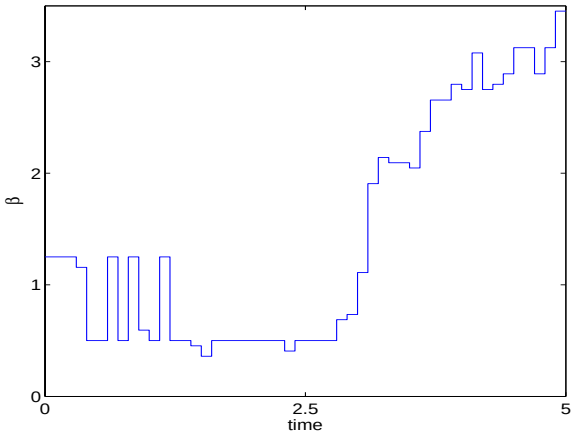



Fig. 6: Optimal control profile for $N = 50$.

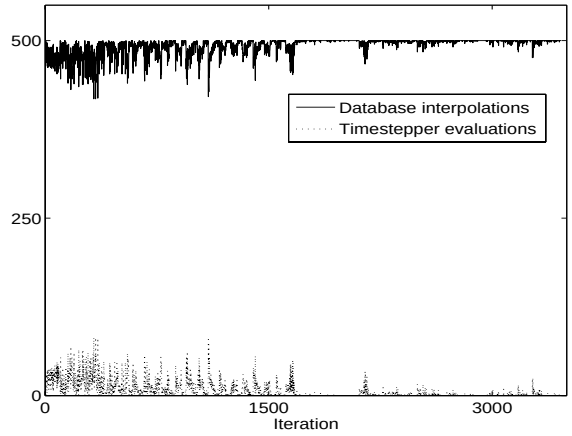



Fig. 7: Number of timestepper evaluations and database interpolations per objective function computation for $N = 50$.

$$\begin{aligned} &c_1 = 1, \;\; c_2 = 0.002, \;\; c_3 = 0.5, \;\; c_4 = 0.04a \\ &X_1(0) = 10^6, \;\; X_2(0) = X_3(0) = 0. \end{aligned} \tag{19}$$

where $a$ is the concentration of $A$.

Fig. 9 presents the system evolution for the parameter values of Eq.19, computed using SSA algorithm. It can be seen that the system dynamics exhibits a two-timescale behavior with the concentration of $S_1$ falling steeply in the beginning, followed by a slow evolution. One of the major drawbacks while simulating the above system using SSA algorithm is that the computational requirements are considerably high. One can employ $\tau$-leaping technique [8] to accelerate computations, however we propose that the stochastic-ISAT al-

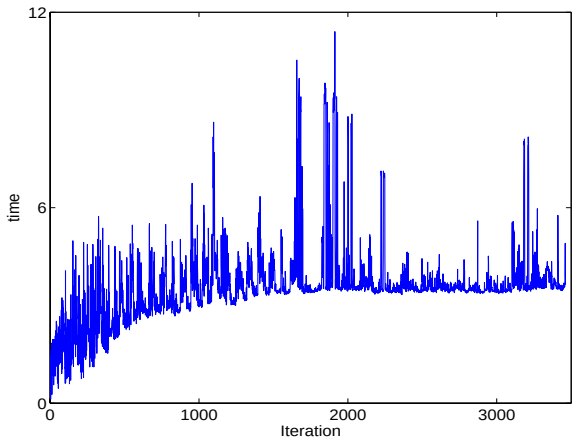



Fig. 8: CPU time (in seconds) per objective function computation for $N = 50$.

gorithm may prove advantageous. We demonstrate this proposal through solution of a representative optimization problem formulated as keeping the concentration of stable dimer $S_3$ close to a set-point at the end of process operation $t_f$, by minimally varying the concentration of the isomerizing agent $A$. Mathematically the problem can be formulated as:

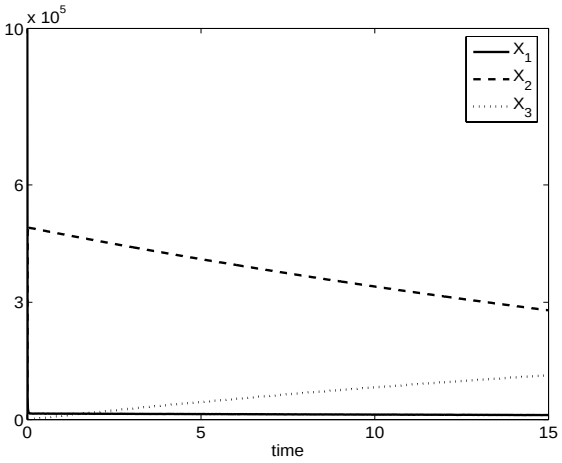



Fig. 9: Concentration profiles of $S_1$, $S_2$ and $S_3$ for nominal parameter values.

$$\begin{aligned} &\min_{A(t)} w_1(X_3(t_f) - \bar{X}_3)^2 + w_2 \int_0^{t_f} A(t)^2 \, \mathrm{d}t \\ &s.t. \\ &\mathbf{X}(t_{i+1}) = \Pi(\mathbf{X}(t_i), \delta t_i, \omega; \boldsymbol{\theta}_i) \\ &\delta t_i = t_{i+1} - t_i \end{aligned} \tag{20}$$

where $\bar{X}_3$ is the set-point.

Fig. 10 presents the optimal trajectory of the concentration of the isomerizing agent $A$ as obtained from the solution of the optimization problem of Eq.20. The corresponding concentration profiles of the reacting species are shown in Fig. 11. The set-point for concentration of $S_3$ is also shown. It is observed that by optimally varying the concentration of $A$, the concentration of $S_3$ can be stabilized to the desired set point. The figure also compares the optimal trajectories of $X_1$, $X_2$ and $X_3$ obtained from ISAT with those obtained from SSA, which demonstrates the accuracy of ISAT interpolations. In Figs. 12 and 13, the performance parameters of the ISAT algorithm, namely, the number of time-steps evaluated using SSA, number of time-steps interpolated from the database and the solution time required per $F$ evaluation, are plotted for each iteration during optimization. Similarly to the previous example, a marked reduction in timestepper evaluations and wall-clock time is observed as the optimization progresses and wall-clock time is largely the database retrieval time near the termination of the optimization. Fig. 13 also presents the average CPU time required per iteration which is two orders of magnitude less than the time required using SSA alone. Performance statistics are summarized in Table 2.

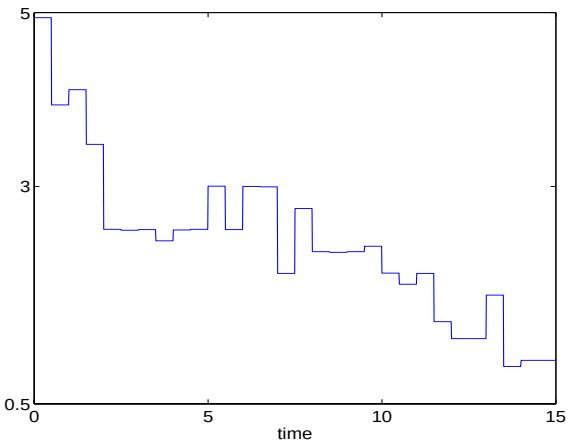



Fig. 10: Optimal profile of concentration of $A$ obtained through the solution of optimization problem of Eq.20.

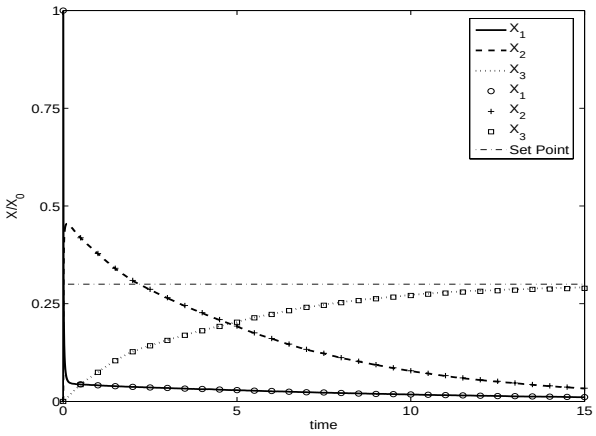



Fig. 11: Concentration profiles of $S_1$, $S_2$ and $S_3$ under optimal variation of concentration of $A$. Accuracy of ISAT interpolation (solid lines) is compared with SSA simulations (solid circles).

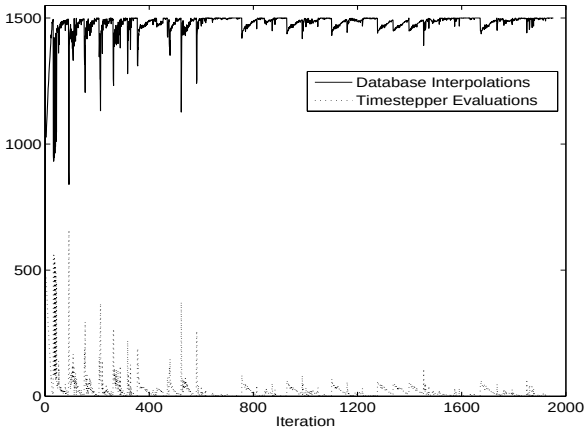



Fig. 12: Number of function evaluations and number of database interpolations as a function of progress of optimization.

## 5 Conclusion

The current work outlines a methodology for the efficient solution of optimal control problems arising in the context of systems described by computationally expensive timestepper based models. The issue of computational requirements for the system evolution is circumvented by extending the notion of in situ adaptive tabulation to stochastic systems. Conditions are outlined that allow unbiased estimation of the mapping gradient matrix and, subsequently,

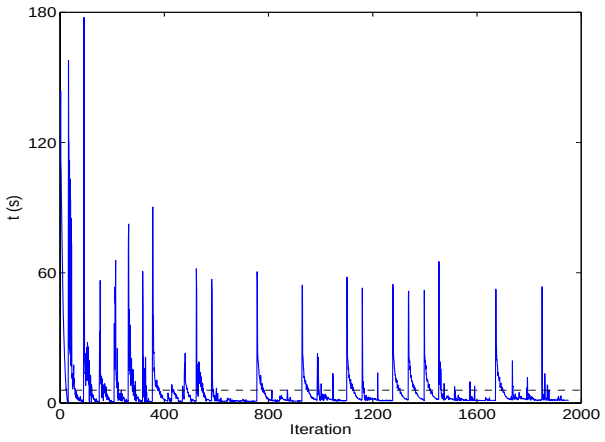



Fig. 13: Time spent per iteration as a function of progress of optimization. The dashed line represents average time spent per iteration.

Table 2: Optimization summary

| | | |
|---|---|---|
| Example 1 | Timestepper Evaluation/Iteration | 24 |
| $N = 10$ | Database interpolation/Iteration | 476 |
| | CPU Time/Iteration | 1.61 s |
| | CPU Time required/Iteration without interpolation | 24 s |
| Example 1 | Timestepper Evaluation/Iteration | 7 |
| $N = 50$ | Database interpolation/Iteration | 493 |
| | CPU Time/Iteration | 3.5 s |
| | CPU Time required/Iteration without interpolation | 24 s |
| Example 2 | Timestepper Evaluation/Iteration | 25 |
| | Database interpolation/Iteration | 1475 |
| | CPU Time/Iteration | 5.93 s |
| | CPU Time required/Iteration without interpolation | 304 s |

CPU times are for a Pentium IV 3.02 GHz processor

expressions to compute the ellipsoid of attraction are derived. The proposed approach was applied towards the solution of representative dynamic optimization problems for a bistable reacting system describing catalytic oxidation of $CO$ and a homogeneous chemically reacting system describing dimerization of a monomer. In both cases, tabulation resulted in significant reduction in the solution time of the optimization problem.

*Acknowledgement.* Financial support from the American Chemical Society, initiation award, Pennsylvania State University Dean's fund, and Pennsylvania department of education are gratefully acknowledged.

## References

1. A. Armaou, I.G. Kevrekidis: Equation-free optimal switching policies using coarse time-steppers. *International Journal of Robust and Nonlinear Control*, in press (2005)
2. A. Armaou, C.I. Siettos, I.G. Kevrekidis: Time-steppers and 'coarse' control of distributed microscopic processes. *Int. J. Robust & Nonlin. Contr.* **14**, 89–111 (2004)
3. A.J. Booker, J.E. Dennis, P.D. Frank, D.B. Serafini, V. Torczon, M.W. Trosset: A rigorous framework for optimization of expensive functions by surrogates. *NASA/CR-1998-208735*, pages 1–19 (1998)
4. L. Dai: Rate of convergence for derivative estimation of discrete-time Markov chains via finite-difference approximation with common randon numbers. *SIAM J. Appl. Math.* **57**, 731–751 (1997)
5. W.F. Feehery, P.I. Barton: Dynamic optimization with equality path constraints. *Ind. Eng. Chem. Res.* **38**, 2350–2363 (1999)
6. C.W. Gear, J.M. Hyman, P.G. Kevrekidis, O. Runborg, K. Theodoropoulos, I.G. Kevrekidis: Equation-free multiscale computation: enabling microscopic simulators to perform system-level tasks. *Comm. Math. Sciences* **1**, 715–762 (2003)
7. D.T. Gillespie: A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.* **22**, 403–34 (1976)
8. D.T. Gillespie: Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.* **115**, 1716–1733 (2001)
9. P. Glasserman, D.D. Yao: Some guidelines and guarantees for common random numbers. *Management Science* **38**, 884–908 (1992)
10. P.W. Glynn: Likelihood ration gradient estimation for stochastic systems. *Communications of the ACM* **33**, 76–84 (1990)
11. P.W. Glynn, P. L'ecuyer: Likehood ratio gradient estimation for stochastic recursions. *Adv. Appl. Prob.* **27**, 1019–1053 (1995)
12. P. Heidelberger, X.R. Cao, M.A. Zazanis, R. Suri: Convergence properties of infinitesimal perturbation analysis estimates. *Management Science* **34**, 1281–1302 (1988)
13. Y.C. Ho: Performance evaluation and perturbation analysis of discrete event dynamic systems. *IEEE Transactions on Automatic Control* **AC-32**, 563–572 (1987)
14. D.R. Jones, M. Schonlau, W.J. Welch: Efficient global optimization of expensive black-box functions. *J. Global Opt.* **13**, 455–492 (1998)
15. C.T. Kelley, E.W. Sachs: Solution of optimal control problems by a pointwise projected newton method. *SIAM Journal on Control and Optimization* **33** (6), 1731 (1995)
16. T.G. Kolda, R.M. Lewis, V. Torczon: Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review* **45**, 385–482 (2003)

17. P. L'Ecuyer: A unified view of the IPA, SF and LR gradient estimation techniques. *Management Science* **36**, 1364–1383 (1990)
18. P. L'Ecuyer, G. Perron: On the convergence rates of IPA and FDC derivative estimators. *Operations Research* **42**, 643–656 (1994)
19. R.M. Lewis, V. Torczon: Pattern search algorithms for bound constrained minimization. *SIAM J. Optimization* **9**, 1082–1099 (1999)
20. R.M. Lewis, V. Torczon: Pattern search algorithms for linearly constrained minimization. *SIAM J. Optimization* **10**, 917–947 (2000)
21. J. Li, D. Liao, S. Yip: Nearly exact solution for coupled continuum/MD fluid simulation. *J. Comp. Mat. Des.* **6**, 95–102 (1999)
22. Y. Lou, P.D. Christofides: Estimation and control of surface roughness in thin film growth using kinetic Monte-Carlo methods. *Chem Eng. Sci.* **58**, 3115–3129 (2003)
23. H.H. McAdams, A. Arkin: Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci.* **94**, 814–819 (1997)
24. C.A. Meyer, C.A. Floudas, A. Neumaier: Global optimization with nonfactorable constraints. *Ind. Eng. Chem. Res.* **41**, 6413–6424 (2002)
25. S.B. Pope: Computationally efficient implementation of combustion chemistry using *in situ* adaptive tabulation. *Combust. Theory & Modelling* **1**, 41–63 (1997)
26. C.V. Rao, A.P. Arkin: Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *J. Chem. Phys.* **118**, 4999–5010 (2003)
27. M.I. Reiman, A. Weiss: Sensitivity analysis of simulations via likelihood ratios. *Operations Research* **37**, 830–844 (1989)
28. C.I. Siettos, A. Armaou, A.G. Makeev, I.G. Kevrekidis: Microscopic/stochastic timesteppers and coarse control: a kinetic Monte Carlo example. *AIChE J.* **49**, 1922–1926 (2003)
29. R. Srivastava, L. You, J. Yin: Stochastic vs. deterministic modeling of intracellular viral kinetics. *J. Theor. Biol.* **218**, 309–321 (2002)
30. R. Suri: Perturbation analysis: The state of the art and research issues explained via the GI/G/l queue. *Proc. IEEE* **77**, 114–137 (1989)
31. E.B. Tadmor, G.S. Smith, N. Bernstein, E. Kaxiras: Mixed finite element and atomistic formulation for complex crystals. *Phys. Rev. B* **59**, 235–245 (1999)
32. V. Torczon: On the convergence of pattern search algorithms. *SIAM J. Optimization* **7**, 1–25 (1997)
33. A. Varshney, A. Armaou: Multiscale optimization of thin-film growth using hybrid PDE/kMC process systems. *Chem. Eng. Sci.* **60** 6780–6794 (2005)
34. S. Vasantharajan, J. Viswanathan, L.T. Biegler: Reduced successive quadratic programming implementation for large-scale optimization problems with smaller degrees of freedom. *Comp. Chem. Eng.* **14**, 907–915 (1990)
35. V.S. Vassiliadis, R.W.H. Sargent, C.C. Pantelides: Solution of a class of multistage dynamic optimization problems, parts I & II. *Ind. & Eng. Chem. Res.* **33**, 2111–2133 (1994)
36. M.A. Zazanis, R. Suri: Convergence rates of finite-difference sensitivity estimates for stochastic systems. *Operations Research* **41**, 694–703 (1993)

# A Reduced Input/Output Dynamic Optimisation Method for Macroscopic and Microscopic Systems

C. Theodoropoulos and E. Luna-Ortiz

School of Chemical Engineering and Analytical Science, University of Manchester, Manchester M60 1QD UK, k.theodoropoulos@manchester.ac.uk

**Summary.** Efficient optimisation algorithms based on model reduction methods are essential for the effective design of large-scale macroscopic, microscopic and multiscale systems. A model reduction based optimization scheme for input/output dynamic systems is presented. It is based on a multiple shooting discretization of the dynamic constraints. The reduced optimization framework is developed by combining an Newton-Picard Method [51], which identifies the (typically) low-dimensional slow dynamics of the (dissipative) model in each time subinterval of the multiple shooting discretization, with reduced Hessian techniques for a second reduction to the low-dimensional subspace of the control parameters. Optimal solutions are then computed in an efficient way using only low-dimensional numerical approximations of gradients and Hessians. We demonstrate the capabilities of this framework by performing dynamic optimization using an explicit tubular reactor transient model and by estimating kinetic parameters of a biochemical system whose dynamics are given by a microscopic Monte Carlo simulator.

## 1 Introduction

Multi-scale models that effectively couple a range of time and length scales are currently paid increasing attention, since they can significantly enhance the understanding of complex processes, through increased insight on the intricate inter-relationships between different system components e.g. [1, 2] and can ultimately lead to products of high quality under stringent specifications. Although the multi-scale concept is not new (e.g. boundary layer theory in transport phenomena [3]), the increasing computational power available nowadays (faster processors and larger memory capacity) along with the use of parallel techniques [4, 5, 6] has made possible the construction of complex multiscale algorithms for a variety of applications (e.g. [7, 8, 9, 10, 11, 12, 13, 14]). Multi-scale models can link the the molecular/mesoscale with the macroscale for example kinetic Monte Carlo (kMC) with computational fluid dynamics (CFD) codes e.g. [15, 16], or different molecular scales (DFT/quantum mechanics/molecular dynamics with kMC e.g [17]) or even macroscopic fine and

coarse scales (e.g. reservoir models [4], multiscale partial differential equation (PDE) models [18] etc).

Models of macroscopic systems (dynamic or steady state) are typically based on systems of non-linear partial differential equations (PDEs) which require numerical solution and after spatial discretization to computational nodes lead to large scale systems. Microscopic models on the other hand are based on the evolution of molecular states in time. Discrete microscopic models (molecular dynamics, Monte Carlo) involve a large number of molecules in order to cover only very small volumes or surfaces ( often in the order of nm) at very small timesteps. Such models are therefore very computationally expensive. Furthermore, if the dynamic transition to each molecular state is only given by a number of evolution rules (e.g. transition probabilities, intermolecular forces etc. which can be obtained from ab initio calculations and quantum mechanics) the governing system equations will not be available in closed form.

Efficient process design and product quality control require system optimisation/ open-loop optimal control. While modern powerful computers make the simulations of complex multi-scale models a feasible yet tedious computational task, optimisation requires a large number of function evaluations, making it extremely computationally costly. Hence, coarse-graining and model reduction techniques seem to be the only way to make optimisation tasks efficient for multi-scale models. A number of approaches have been proposed in the literature that can lead to efficient multiscale models amenable to optimisation at different levels/scales. A brief overview of recent optimisation efforts for multiscale systems is given below. Adjoint analysis has been used as an efficient optimisation method for large-scale computational fluid dynamic systems [19, 20] (and references within) and also to compute reduced sensitivity equations from large-scale fixed-point procedures [21]. In [18] adjoint analysis is expanded to the optimisation of multi-scale dynamic PDE-based systems using a number of regularisation techniques [22, 23] in order to tune the optimisation algorithm and to target the objective function towards the time and length scales of interest. The gradient is also appropriately preconditioned by *filtering* the adjoint field, from which sensitivity information is extracted, through the use of different brackets according to the scales that need to be emphasized (or filtered-out). The resulting method results in a significant speed-up of the optimisation procedure.

A multiscale analytical sensitivity approach (MASA) was proposed in [24] based on homogenization theory. The sensitivity problem is constructed through direct differentiation and asymptotic analysis is performed for each scale. The sensitivities thus obtained for inelastic periodic composites models are in general more accurate than the ones obtained by brute-force central finite differences since they are significantly influenced by the size of the numerical perturbation used. Large computational resources are still required for the optimisation of large-scale systems.

A multiscale optimisation approach was developed in [25] for the design of bioremediation processes. The method is based on the application of Sequential Quadratic Programming (SQP) on multiscale PDE-based simulators operating at 4 different mesh levels. Fine-level derivatives are efficiently calculated at the coarser level and interpolated back to the fine level and this procedure was termed v-cycling. The problem is first solved at the coarser levels and the (interpolated) solution is used as an initial guess for the finer levels, while derivatives are calculated either by direct numerical differentiation or by v-cycling. The method achieves significant computational speed-up for the optimisation of bioremediation systems. The development of dynamic optimisation algorithms for multiscale PDE-based models is addressed in [26], through Galerkin-based multiscale discreatization in wavelet coordinates. Iterative solvers with simple Jacobi preconditioning are used for the arising large-scale linear problems and a nested iteration scheme is constructed to handle the different levels. For each level, a tailored iterative scheme is derived based on the structure of the modelling equations.

The developments described above were focused in essence on multi-level PDE-based systems. The following optimisation procedures were developed for systems coupling microscopic/molecular with macroscopic scales.

A combination of funnelling algorithms for large-scale geometries and terrain methods [28] for rough/noisey small-scale geometries was proposed in [27] in order to efficiently obtain global optima of multi-scale systems. Funnelling algorithms take advantage of the topological similarities between different large-scale systems and represent the topology of the system (e.g. an energy surface) by simple exponential functions that have the same global minima with the original system. The terrain methods are based on the idea that local stationary points are connected along valleys and ridges. After locating one stationary point one can move to another following the eigendirections of the connecting topology using predictor-corrector methods. Since the terrain methods can deal with rugged (and potentially noisey) systems the combined approach is promising for the optimisation of coupled macroscopic/microscopic multiscale systems.

An atomic scale finite element method equivalent is developed in [29] in order to model atomic-scale systems, which uses atoms instead of computational nodes. The interactions between neighbouring atoms are represented by appropriate stiffness matrices. The structure of the AFEM elements depends on the atomic configuration of the system and is different for different systems. It has been found to be faster than the conjugate gradient method and can be directly combined with conventional finite elements, modelling macroscopic continuum systems to create efficient multi-scale models. AFEM is using both first and second order derivatives of the system's energy and is therefore appropriate for handling multi-scale optimisation problems.

In [30] reduced multi-scale models are constructed from coupled PDE-based macroscopic system reduced through the Karhunen-Loeve expansion [31] and a kinetic Monte Carlo-based microscopic system where in situ adap-

tive tabulation has been used where stationary states of the microscopic simulation are tabulated on the fly and system information is extracted through interpolations. The method has been used effectively for the optimisation of a multi-scale chemical vapour deposition reactor.

In [32] the problem of estimating kinetic parameters of molecular kMC-based models through optimization has been addressed. The important kinetic parameters have been estimated through sensitivity analysis, scaled to discriminate response changes from inherent noise. A response surface technique was used for the optimisation procedure. The reaction rates have been parametrized as polynomials of the sensitive kinetic parameters and response surface are computed through a number of factorial numerical experiments. The computational expense in this method lies in the construction of the response surfaces. Kinetic parameters are then computed by minimising the error between the systems response and experimental observation using stochastic optimisation, namely simulated annealing [34]. Also, a sensitivity algorithm including a stochastic term for the computation of the system's sensitivities has been coupled to the above algorithm in order to increase the accuracy of kinetic parameter estimation [9].

All the methods described above either assume that the system's model is intimately known and all equations are available in closed form or that simplified reduced models based on a few simple equations can be extracted from all the levels of the multi-scale ones. "Equation-free" methods (see e.g. [35] and references within) act upon black-box timesteppers and enable them to perform system-level tasks, such as computation of unstable steady states, stability/bifurcation analysis and control e.g. [36, 37]. Microscopic simulators can be effectively handled through *restriction* (obtaining coarse states from the microscopic/molecular ones, through averaging, filtering, smoothing) and *lifting* procedures (obtaining molecular states from the coarse ones with the aid of appropriate distribution functions. These methods have been used to efficiently address a large number of microscopic/molecular and multi-scale systems.

In this work we present an optimisation methodology based on model reduction that can be employed for the efficient dynamic optimisation of input/output large-scale macroscopic and microscopic simulators. This optimisation approach belongs to the family of equation-free methods and is an extension of our recent work on steady state optimisation for input/output large-scale simulators [38].

## 2 Reduced Dynamic Optimisation for Input/Output Simulators

Consider the following dynamic optimization problem:

$$\min_{u(t),z} \int_{t_0}^{t_f} \Phi\left(t, u\left(t\right), z\right) \mathrm{d}t \tag{1a}$$

$$\text{s.t. } \frac{\partial u}{\partial t} = F\left(t, u\left(t\right), z\right) \tag{1b}$$

$$u\left(t_0\right) = u_0 \tag{1c}$$

where $u \in \mathbb{R}^n$ are the state variables, $t$ is time, $F \in \mathbb{R}^{n+dof} \rightarrow \mathbb{R}^n$ are the dynamic model equations, and $z \in \mathbb{R}^{dof}$ are the control parameters which are coefficients of a certain parametrization of a continuous control profile function $c\left(t\right)$ to be adjusted by the optimization procedure. Given a set of initial conditions $u_0$ and values of the coefficients $z$, the control profile $c\left(t\right)$ can be evaluated at any point of the time horizon and the state vector $u = u\left(t\right)$ is uniquely determined by the solution of the differential system.

A variety of solution methods have been developed for this problem. A compilation of these methods can be found in e.g. [43, 44, 45]. In this work, we focus on NLP methods, based on some (total or partial) discretization of the infinitesimal dynamic constraints. Orthogonal collocation is a common method to handle this (possible) highly nonlinear transient system. It discretizes the differential equations over the time reporting horizon $[t_0, t_f]$, computing solutions of the dynamic system at each discrete time and the control parameters being estimated by the optimization algorithm [46, 47]. Collocation is very convenient for the implementation of multi-level approaches discussed in the previous section. Also, the estimation of the first and second order information can be obtained at little cost [42]. However, the direct use of input/output dynamic simulators is not possible. Input/output optimisation methods are useful when process models are only available in "black-box" form. Consequently, all sensitivity information not explicitly available, and is often, too expensive to approximate by numerical perturbations of the solver. This is the case when commercial or scientific software is used to simulate dynamic processes. and also, as discussed in the previous section, the case of multiscale/microscopic models not being available in closed form Stochastic optimisation methods can be readily used with such black-box codes (e.g. [33]), but for distributed processes with many state variables, function evaluations become prohibitively expensive. If a gradient-based approach is preferred, a "black-box" a shooting scheme should be adopted to solve the optimal control problem. There are two shooting methods. Single shooting which suffers from lack of robustness and is prone to numerical instabilities [40] and multiple shooting [55] which is a stable method. In multiple shooting, the time horizon $[t_0, t_f]$ is divided into (possible many) subintervals and the solution of the differential system is computed over each subinterval. Continuity of the dynamic profiles is forced by linking the initial conditions of each subinterval with the final values of the previous subinterval, generating continuity constraints (multiple shooting equations).

Depending on the number of the time subintervals chosen, the optimization formulation leads to a very large-scale problem. Extra function evalu-

ations are needed to compute the sensitivities required by the optimization method. The most expensive part of this scheme is, precisely, the computation of the derivatives, which can take up to 95% of the computational time [42]. Specialized software [53] and automatic differentiation [49] can be used to compute these sensitivities more efficiently. To reduce the complexity of the large-scale optimization problem, reduced Hessian methods can be employed in the case of relatively few degrees of freedom. In [50], a multiple shooting procedure combined with orthogonal collocation is developed using a partially reduced SQP strategy, in which the full space is projected onto the reduced subspace of differential and control variables; it looks to be efficient specially when handling inequality constraints but the method appears to require important modifications to the existing optimization solvers. Here, we present a computationally-efficient method for optimizing dynamic models in a "black-box" fashion extending ideas presented in [38], where we dealt with steady state optimization. We perform a reduced multiple shooting procedure [51], in which we compute reduced block Jacobians corresponding to the (few) dominant eigenmodes of each subinterval and solve the corresponding continuity constraints. Then, using reduced Hessian methods [54, 52] we perform a second projection onto the small subspace of the (also few) control parameters, leading to a number of small unconstrained quadratic subproblems that are solved at each major iteration without the construction of unnecessary high-dimensional Jacobians and Hessians.

The remaining sections are organised as follows: First, we show how the dynamic problem (1) can be converted to a NLP problem using the multiple shooting discretization in the dynamic constraints. This, leads to a large-scale optimization problem. Then, we develop our reduced multiple shooting algorithm which computes only low order derivative matrices. Finally, two numerical examples are provided to demonstrate the capabilities of this framework to handle large distributed and microscopic dynamic systems.

## 3 Multiple Shooting Approach for Dynamic Optimization

Let us consider the problem given by (1). In multiple shooting, the time reporting horizon $[t_0, t_f]$ is partitioned in $N$ subintervals

$$[t_i, t_{i+1}] \text{ for } i = 0, \ldots, N-1$$

where $t_N = t_f$. The differential equations (1b) are solved over each subinterval. The $i$-th initial condition at $t_i$ of each subinterval is given by intermediate variables $u_i$, and solution of each subinterval at $t_{i+1}$ is given by

$$u_{i+1} = G\left(t_i, u_i, t_{i+1}, z\right) \tag{2}$$

where $G\left(t_i, u_i, t_{i+1}, z\right)$ is a non-expansive map for $u_{i+1}$. In order to achieve a continuous dynamic trajectory along the time horizon, continuity constraints

are imposed:

$$r_{i+1}\left(t_i, u_i, t_{i+1}, u_{t+1}, z\right) = u_{i+1} - G\left(t_i, u_i, t_{i+1,} z\right) = 0. \tag{3}$$

Then, in multiple shooting, the following nonlinear system has to be solved:

$$\begin{gathered} r_1\left(t_0, u_0, t_1, u_1, z\right) = u_1 - G\left(t_0, u_0, t_1, z\right) \\ r_2\left(t_1, u_1, t_2, u_{2,} z\right) = u_2 - G\left(t_1, u_1, t_2, z\right) \\ \vdots \\ r_{N-1}\left(t_{N-2}, u_{N-2}, t_{N-1}, u_{N-1,} z\right) = u_{N-1} - G\left(t_{N-2}, u_{N-2}, t_{N-1}, z\right) \\ r_N\left(t_{N-1}, u_{N-1}, t_N, u_{N,} z\right) = u_N - G\left(t_{N-1}, u_{N-1}, t_N, z\right) \end{gathered} \tag{4}$$

for the unknown (state) variables $u_i, i = 1, \ldots, N$ since $u_0$, the initial conditions are given. To solve (4), Newton's method is applied, leading to linearized system (with $\Delta u_0 = 0$):

$$\begin{bmatrix} J_u & J_z \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta z \end{bmatrix} = -r \tag{5}$$

where $J_u$, is the Jacobian of the continuity equations (3) with respect to the state variables $u_i$, $J_z$ is the Jacobian with respect to the control variables, $\Delta u = (\Delta u_1, \ldots, \Delta u_N)$, and $\Delta z = (\Delta z_1, \ldots, \Delta z_{dof})$. The matrices $J_u$and $J_u$ have the following structure

$$J_u = \begin{bmatrix} I & 0 & 0 & \cdots & 0 \\ -\frac{\partial G(t_1, u_1, t_2, z)}{\partial u_1} & I & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & 0 & -\frac{\partial G(t_{N-2}, u_{N-2}, t_{N-1}, z)}{\partial u_{N-2}} & I & 0 \\ 0 & \cdots & 0 & -\frac{\partial G(t_{N-1}, u_{N-1}, t_N, z)}{\partial u_{N-1}} & I \end{bmatrix} \tag{6}$$

and

$$J_z = \begin{bmatrix} \frac{\partial G(t_0, u_0, t_1, z)}{\partial z} \\ \vdots \\ \frac{\partial G(t_{N-1}, u_{N-1}, t_N, z)}{\partial z} \end{bmatrix}. \tag{7}$$

Each Jacobian block $\frac{\partial G(t_i, u_i, t_{i+1}, z)}{\partial u_i}$ is a $n \times n$ matrix, and its computation requires the solution of $n$ initial value problems per subinterval. In the same way, the block $\frac{\partial G(t_i, u_i, t_{i+1}, z)}{\partial z}$ is a $n \times dof$ matrix and requires the solution of $dof$ initial value problems per subinterval. Clearly, the numerical computation of $J_u$ is far too expensive if the dynamic system is large. The estimation of these sensitivities is the most consuming part of the algorithm. Specialized software for sensitivity estimation of dynamic systems [53], and automatic differentiation (AD) of numerical integrators [49] can be employed to compute efficiently and fast these derivatives. In [57], other strategies are discussed to compute the sensitivities in the context of shooting techniques. In a forthcoming publication ([58]) we discuss how the optimization approach presented

here can be combined with automatic differentiation methods. However the use of AD is not always possible for black-box integrators.

In the same fashion, the objective function can be discretized in $N$ time subintervals such that

$$\int_{t_0}^{t_f} \Phi\left(t, u\left(t\right), z\right) \mathrm{d}t = \sum_{i=0}^{N-1} \Phi\left(t_i, u_i, t_{i+1}, u_{i+1}, z\right) = \Phi\left(t_0, u_0, t_1, u_1, z\right) + \cdots + \Phi\left(t_{N-1}, u_{N-1}, t_N, u_N, z\right).$$

The infinitesimal optimization problem (1) can be now formulated as a finite dimensional large-scale constrained optimization problem of dimension $N(n) + dof$:

$$\min_{u_1,\ldots,u_N,z} \sum_{i=0}^{N-1} \Phi\left(t_i, u_i, t_{i+1}, u_{i+1}, z\right) \tag{8a}$$

$$\text{s.t. } r_{i+1}\left(t_i, u_i, u_{i+1}, t_{i+1}, z\right) = 0 \ \forall i = 0, ..., N-1 \tag{8b}$$

$$u\left(t_0\right) = u_0 \tag{8c}$$

The above large-scale nonlinear optimization problem can be solved using any standard optimization solver such as SQP [56]. SQP requires the computation of the large-scale sparse Jacobians stated previously. For systems of moderate size this can work very well. However, as the dimension $N$ or $n$ is increased the computational cost becomes prohibitively. In [62], this problem is tackled by reducing the number of sensitivity solutions needed to compute the derivatives capitalizing on the sparsity of the Jacobian of the continuity equations. Furthermore, the computation of the derivatives is carried out using specialized sensitivity software [53]. However, this reduction may compromise the stability properties of multiple shooting.

Since the most expensive part of the algorithm is the numerical computation of $\frac{\partial G(t_i, u_i, t_{i+1}, z)}{\partial u_i}$, there is a strong motivation to develop an algorithm in which the explicit construction and storage of each block is avoided, without losing stability from the multiple shooting discretization and consequently, reducing the computational cost of solving large optimization problems.

## 4 The Newton-Picard-Based Dynamic Optimisation Scheme

In this section, we discuss the computational framework which avoids the explicit computation of the block matrix $\frac{\partial G(t_i, u_i, t_{i+1}, z)}{\partial u_i}$ combined with reduced Hessian techniques to solve the large-scale optimization problem (8).

We capitalize on the fact that the dominant (slow) dynamics of many systems is low-dimensional. In other words there is a separation of timescales in the eigenspectrum of the system which allows partitioning of the system into slow (possibly unstable) and fast (stable) subspaces. Then, the explicit calculation of the block matrices can be avoided since only the action on the low-dimensional dominant subspace is required, A Newton-Picard procedure developed in [51] is employed in order to dynamically compute the low-order dominant subspaces in this multiple shooting framework. It was originally constructed to compute of periodic solutions of PDEs by performing Newton iterations on the computed subspace of the dominant eigendmodes and Picard iterations on its orthogonal complement. Here we adapt this Newton-Picard scheme in order to compute solutions of the continuity constraints then combine with reduced Hessian techniques to solve the dynamic optimization problem (1).

Let us recall that in multiple shooting we have the non-expansive map (2):

$$u_{i+1} = G\left(t_i, u_i, t_{i+1}, z\right)$$

for $i = 0, \ldots, N-1$. For each time subinterval, we can define subspaces $\mathbf{P}$ and $\mathbf{Q}$ of $\mathbb{R}^n$, where $\mathbf{P}$ is an invariant subspace of $G_{u_i}\left(t_i, u_i, t_{i+1}, z\right)$, and $\mathbf{Q} = \mathbb{R}^n - \mathbf{P}$ is the orthogonal complement of $\mathbf{P}$. Then $u_i$ can be decomposed in

$$u_i = p_i + q_i \tag{9}$$

with $p_i = P_i u_i \in \mathbf{P}$ and $q_i = Q u_i \in \mathbf{Q}$. Then we can decompose (2) to get:

$$p_{i+1} = w\left(p_i, q_i, z\right) = P_i G\left(t_i, p_i + q_i, t_{i+1}, z\right) \tag{10a}$$

$$q_{i+1} = g\left(p_i, q_i, z\right) = Q_i G\left(t_i, p_i + q_i, t_{i+1}, z\right). \tag{10b}$$

The system (10a) can be solved using Newton's method and Picard iterations to solve (10b). This leads to the coupled stabilized iteration (with $\Delta p_0 = 0$):

$$W_p \Delta p = -s \tag{11a}$$

$$q_{i+1} = g\left(p_i, q_i, z\right) \tag{11b}$$

where

$$W_p = \begin{bmatrix} I & 0 & \cdots & 0 & 0 \\ -w_{p_1} & I & 0 & \vdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & -w_{p_2} & I & 0 \\ 0 & \cdots & 0 & -w_{p_{N-1}} & I \end{bmatrix} \tag{12}$$

$$\Delta p = \begin{bmatrix} \Delta p_1 \\ \Delta p_2 \\ \vdots \\ \Delta p_{N-1} \\ \Delta p_N \end{bmatrix} \tag{13}$$

$$s = \begin{bmatrix} p_1 - w\left(p_0, q_0, z\right) \\ p_2 - w\left(p_1, q_1, z\right) \\ \vdots \\ p_{N-1} - w\left(p_{N-2}, q_{N-2}, z\right) \\ p_N - w\left(p_{N-1}, q_{N-1}, z\right) \end{bmatrix} \tag{14}$$

with $w_{p_i} = P_i \frac{\partial w(p_i, q_i, z)}{\partial p_i} P_i$.

To solve the system (11), the computation of $P_i$ and $Q_i$ is required. Let $V_i \in \mathbb{R}^{n \times m}$ an orthonormal basis whose columns span the low-dimension invariant subspace $\mathbf{P}$. Here, without loss of generality, we assume that the number of dominant eigenvalues is the same in every time subinterval, hence the dimension $m$ is the same in each subinterval. Then, the projectors are given by

$$P_i = V_i V_i^T \tag{15a}$$

$$Q_i = I - V_i V_i^T \tag{15b}$$

with $V_i^T V_i = I \in \mathbb{R}^{m \times m}$. Now, a set of reduced variables, $v_i$, can be introduced for the representation of $p_i \in \mathbf{P}$ in the basis $V_i$

$$v_i = V_i^T p_i = V_i^T u_i \tag{16}$$

where $v_i \in \mathbb{R}^m$, $p_i = V_i v_i$ and $u_i = V_i v_i + q_i$. Using (16), the iteration (11a) can be written as

$$\overline{W}_v \Delta v = -\overline{s} \tag{17}$$

where

$$\overline{W}_v = \begin{bmatrix} I & 0 & 0 & \cdots & 0 \\ -H_1 & I & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & 0 & -H_{N-2} & I & 0 \\ 0 & \cdots & 0 & -H_{N-1} & I \end{bmatrix} \tag{18}$$

$$\Delta v = \begin{bmatrix} \Delta v_1 \\ \Delta v_2 \\ \vdots \\ \Delta v_{N-1} \\ \Delta v_N \end{bmatrix} \tag{19}$$

$$s = \begin{bmatrix} v_1 - V_1^T G\left(t_0, u_0, t_1, z\right) \\ v_2 - V_2^T G\left(t_1, u_1, t_2, z\right) \\ \vdots \\ v_{N-1} - V_{N-1}^T G\left(t_{N-2}, u_{N-2}, t_{N-1}, z\right) \\ v_N - V_N^T G\left(t_{N-1}, u_{N-1}, t_N, z\right) \end{bmatrix} \tag{20}$$

with $H_i = V_{i+1}^T \frac{\partial G(t_i, u_i, t_{i+1}, z)}{\partial u_i} V_{i+1} = V_{i+1}^T \frac{\partial w(p_i, q_i, z)}{\partial p_i} V_{i+1}$ for $i = 0, \ldots, N-1$. Notice that the system has been reduced from $N\left(n\right)$ to $N(m)$ where $m \ll n$. Then, iteration (11b) can be expressed as

$$q_{i+1} = G\left(t_i, u_i, t_{i+1}, z\right) - V_{i+1}\left[V_{i+1}^T G\left(t_i, u_i, t_{i+1}, z\right)\right], \tag{21}$$

here, we have used $g\left(p_i, q_i, z\right) = \left(I - V_{i+1} V_{i+1}^T\right) G\left(t_i, p_i + q_i, t_{i+1}, z\right)$. To avoid the explicit computation of the blocks $\frac{\partial G(t_i, u_i, t_{i+1}, z)}{\partial u_i}$, we compute directly the product $\left[\frac{\partial G(t_i, u_i, t_{i+1}, z)}{\partial u_i} V_{i+1}\right]$ by numerical central differentiation

$$\begin{aligned} \left[\frac{\partial G\left(t_i, u_i, t_{i+1}, z\right)}{\partial u_i} V_{i+1}\right] \approx \frac{1}{2\varepsilon} \left[G\left(t_i, u_i + \varepsilon V_{i+1}, t_{i+1}, z\right)\right. \\ \left. - G\left(t_i, u_i - \varepsilon V_{i+1}, t_{i+1}, z\right)\right] \end{aligned} \tag{22}$$

for each column of $V_{i+1}$.

The most expensive part of this Newton-Picard algorithm is the computation of the basis $V_i$. Subspace iterations have been proposed as a reliable computation method [48]. In particular, we have chosen to use the algorithm implemented in [41] which includes deflation and locking procedure, and allows the direct computation of the directional derivatives $\left[\frac{\partial G(t_i, u_i, t_{i+1}, z)}{\partial u_i} V_{i+1}\right]$. When the dimension of the control parameters is smaller than the dimension of the *reduced* state variables, i.e. $dof \ll N(m)$, we use reduced Hessian methods [54, 52] to reduce the computational cost required to solve the optimization problem and we obtain (by analogy from [38]) the *reduced* coordinate basis

$$\overline{Z} = \begin{bmatrix} -\overline{W}_v^{-1} \overline{W}_z \\ I \end{bmatrix} \tag{23}$$

where $\overline{Z} \in \mathbb{R}^{N(m) \times dof}$, $\overline{W}_v \in \mathbb{R}^{N(m) \times N(m)}$ and with $\overline{W}_z \in \mathbb{R}^{N(m) \times dof}$. The matrix $\overline{W}_z$ can be easily obtained by

$$\overline{W}_z = \begin{bmatrix} V_1^T & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_N^T \end{bmatrix} \begin{bmatrix} \frac{\partial G(t_0, u_0, t_1, z)}{\partial z} \\ \vdots \\ \frac{\partial G(t_{N-1}, u_{N-1}, t_N, z)}{\partial z} \end{bmatrix}. \tag{24}$$

Then, the reduced QP subproblem [38] becomes:

$$\min_{d_{\overline{Z}}} \left(\overline{Z}^T \nabla f\right)^T d_{\overline{Z}} + \frac{1}{2} d_{\overline{Z}}^T \left(\overline{Z}^T B \overline{Z}\right) d_{\overline{Z}} \tag{25a}$$

$$\text{s.t.} \left[\begin{pmatrix} u^L - u^k \\ z^L - z^k \end{pmatrix}\right] \leq \overline{Z} d_{\overline{Z}} \leq \left[\begin{pmatrix} u^U - u^k \\ z^U - z^k \end{pmatrix}\right]. \tag{25b}$$

Here, we notice that only $m$ Lagrange multipliers are needed per subinterval, (i.e. $\phi_{i+1} = V_{i+1}^T \lambda_{i+1}$) to compute the projected Hessian. This reduced Lagrange multipliers can be computed from

$$\overline{W}_v \phi = - \begin{bmatrix} V_1^T & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_N^T \end{bmatrix} Y^T \nabla f \tag{26}$$

where $\phi \in \mathbb{R}^{N(m)}$, $Y^T \in \mathbb{R}^{n \times (N(n)+dof)}$ and $\nabla f \in \mathbb{R}^{N(n)+dof}$.

A flowchart of the reduced dynamic optimisation procedure is provided in Fig. (1). The major benefits of this algorithm are the computation of the coordinate basis by inverting only low-order Jacobians and the computation of reduced Hessians with only a small number of Lagrange multipliers.

## 5 Numerical Examples

### 5.1 Dynamic Optimization of a Tubular Reactor

In order to illustrate the features of the proposed framework, first a PDE-based system is chosen. It is a tubular reactor with pseudohomogeneous axial dispersion, where a simple exothermic irreversible first order reaction, $A \rightarrow B$, occurs [67]. The mathematical model consists of two nonlinear parabolic partial differential equation given, in dimensionless form for the reactant concentration $x_1$ and temperature $x_2$, by

$$\frac{\partial x_1}{\partial t} = \frac{1}{Pe_1} \frac{\partial^2 x_1}{\partial y^2} - \frac{\partial x_1}{\partial y} + Da\,(1 - x_1) \exp\left(\frac{x_2}{1 + \frac{x_2}{\gamma}}\right) \tag{27a}$$

$$\frac{\partial x_2}{\partial t} = \frac{1}{LePe_2} \frac{\partial^2 x_2}{\partial y^2} - \frac{1}{Le} \frac{\partial x_2}{\partial y} - \frac{\beta}{Le} x_2 + CDa\,(1 - x_1) \exp\left(\frac{x_2}{1 + \frac{x_2}{\gamma}}\right) + \frac{\beta x_{2_w}(t)}{Le} \tag{27b}$$

with boundary conditions

$$\left.\frac{\partial x_1}{\partial y}\right|_{y=0} = Pe_1 x_1 \tag{28a}$$

$$\left.\frac{\partial x_2}{\partial y}\right|_{y=0} = Pe_2 x_2 \tag{28b}$$

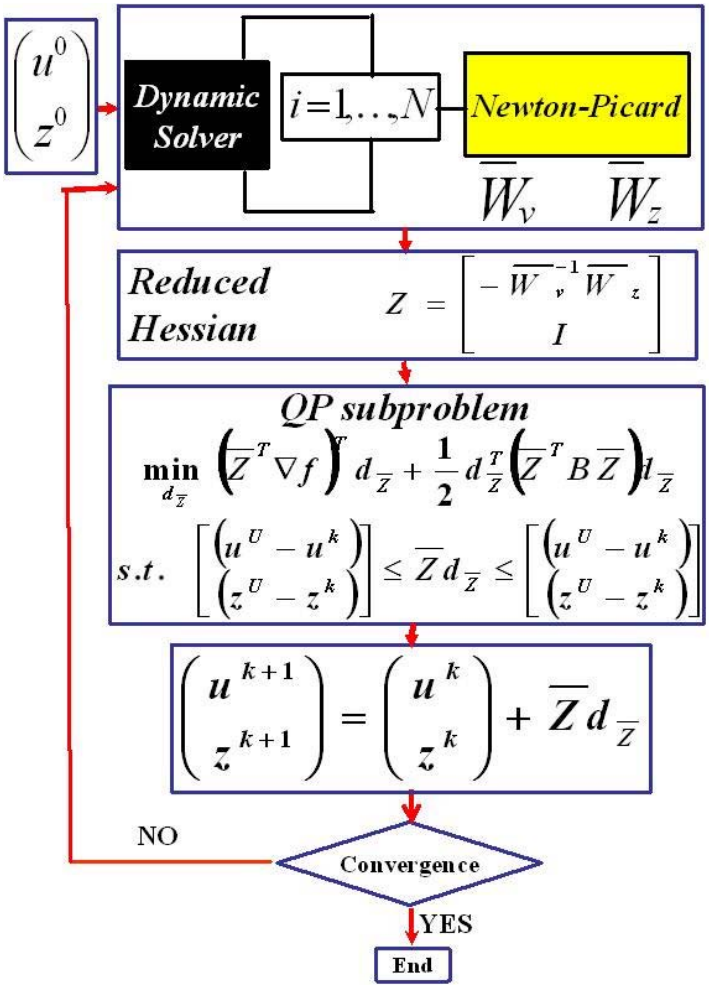



Fig. 1: The reduced optimisation algorithm. The continuity constraints (multiple shooting equations) are solved in the Newton-Picard framework. Low-dimensional Jacobians are then used to compute the null-space basis for optimization.

$$\left.\frac{\partial x_1}{\partial y}\right|_{y=1} = 0 \tag{29a}$$

$$\left.\frac{\partial x_2}{\partial y}\right|_{y=1} = 0 \tag{29b}$$

and initial conditions

$$x_1(0) = x_2(0) = 0 \tag{30}$$

where $Da$ is the Damköhler number, $Le$ the Lewis number, $Pe_1$ and $Pe_2$ are the Peclet numbers for mass and heat transport, $\beta$ a dimensionless heat transfer coefficient, $C$ is the dimensionless adiabatic temperature rise, and $y$ the dimensionless longitudinal coordinate. The dimensionless wall temperature $x_{2_w}(t)$, is adjusted indirectly by controlling the flowrate of three cooling jackets. The spatial profile of the control variable is defined as:

$$x_{2_w}(y) = \sum_{j=1}^{3} \left[H(y - y_{j-1}) - H(y) - y_j\right] x_{2_{w_j}}(t) \tag{31}$$

where $H(\cdot)$ is the Heaviside function, $y_0 = 0$, $y_1 = 1/3$, $y_2 = 2/3$, $y_3 = 1$ and $x_{2_{w_j}}(t)$, $j = 1, \ldots, 3$ is the dimensionless temperature at each cooling zone. The control function for each cooling zone has been approximated using the parametrization presented in [59]. This parametrisation can generate various types of continuous dynamic control profiles of two types of curves given by:

*Type I*

$$x_{2_{w_j}}(t) = x_{2_{w_j}}(t_f) - \left[x_{2_{w_j}}(t_f) - x_{2_{w_j}}(t_0)\right] \left[1 - \frac{t}{t_f}\right]^{A_1} \tag{32}$$

*Type II*

$$x_{2_{w_j}}(t) = x_{2_{w_j}}(t_0) - \left[x_{2_{w_j}}(t_0) - x_{2_{w_j}}(t_f)\right] \left[\frac{t}{t_f}\right]^{A_2} \tag{33}$$

where $A_1$and $A_2$ are parameters that define the curvature of the profiles. By combining these two profiles along the time horizon, many types of continuous trajectories are obtained. Profiles with prominent discontinuities or very difficult to implement in practise can be avoided. Only six parameters are required to connect the two curves and generate a continuous profile. These six control parameters, which will be the degrees of freedom in the optimization, are the initial and final value of the control function ($x_{2_{w_j}}(t_0)$ and $x_{2_{w_j}}(t_f)$), an intermediate time $t_{int}$ (time point in which the two curves connect) and the corresponding control function value $x_{2_{w_j}}(t_{int})$, and the exponential parameters $A_1$and $A_2$.

The problem is to find the optimal temperature profile that maximizes the exit conversion at a final time $t_f = 1.5$ subject to the dynamical system (27) with boundary conditions (28,29), initial conditions (30) and parameter values $Da = 0.1$, $Le = 1.0$, $Pe_1 = Pe_2 = 5.0$, $\gamma = 20.0$, $\beta = 1.50$ and $C = 12.0$. The PDEs are first discretized in 250 spatial nodes using central differences, resulting in a system of 500 ODEs. The dynamic system has been solved using a 4th order Runge-Kutta method. The resulting optimization problem has 500 state variables and 18 control parameters (6 variables per jacket). The numerical derivatives have been obtained according to the scheme (22) with $\varepsilon = 1 \times 10^{-6}$. We perform the dynamic optimization using $N = 5$ time subintervals with different dimensions of dominant eigenspace. With the

dominant subspace dimensions $m = 6$ and $m = 8$ the procedure is able to converge with almost identical optimal solutions. The values of the optimal control parameters are reported in Tables (1) and (2), respectively. For $m = 6$ 18 iterations are required, while 14 iterations are required for $m = 8$. In [38], we also observe that less iterations are required when the subspace is large enough to capture the slow action of the system. The same calculations were performed using a finer discretization of the time horizon ($N = 10$). For the three cases tested, the reduced optimisation procedure computes similar optimal parameters. In Table (3), we report the case for $m = 8$. Again, less iterations are required when increasing the size of the invariant subspace.

For the case with $N = 10$ and $m = 8$, the optimal control functions are shown in Fig. (2) For comparison purposes we perform dynamic optimization with a regular multiple shooting procedure. The results obtained show excellent good agreement with the ones obtained by reduced optimization. The total CPU time required by the standard multiple shooting lager by many days. Hence, a very significant computational speed-up is achieved. In [58] the performance of our procedure with respect to operating parameters (such as time horizon, subspace dimension, number of subintervals is extensively discussed.

Table 1: Optimal parameters that define the control function $x_{2_{w_j}}$ of each cooling jacket computed by reduced optimization, with $N = 5$ and $m = 6$, $fobj = 0.9997712$

| Parameter | Jacket $j = 1$ | Jacket $j = 2$ | Jacket $j = 3$ |
|---|---|---|---|
| $x_{2_{w_j}}(t_0)$ | 5.0 | 3.78261 | 1.88901 |
| $x_{2_{w_j}}(t_f)$ | 3.52167 | 3.79172 | 2.8607 |
| $t_{int}$ | 0.68721 | 0.276121 | 1.2567E-2 |
| $x_{2_{w_j}}(t_{int})$ | 3.62251E-3 | 8.1087E-4 | 0. |
| $A_1$ | 1. | 1. | 1. |
| $A_2$ | 1. | 1.000546 | 1. |

Table 2: Optimal parameters that define the control function $x_{2_{w_j}}$ of each cooling jacket computed by reduced optimization, with $N = 5$ and $m = 8$, $fobj = 0.9997892$

| Parameter | Jacket $j = 1$ | Jacket $j = 2$ | Jacket $j = 3$ |
|---|---|---|---|
| $x_{2_{w_j}}(t_0)$ | 5.0 | 3.76829 | 1.89432 |
| $x_{2_{w_j}}(t_f)$ | 3.51210 | 3.79352 | 2.83824 |
| $t_{int}$ | 0.69017 | 0.274873 | 1.1752E-2 |
| $x_{2_{w_j}}(t_{int})$ | 3.5243E-3 | 8.1972E-4 | 0. |
| $A_1$ | 1. | 1. | 1. |
| $A_2$ | 1. | 1.000476 | 1. |

Table 3: Optimal parameters that define the control function $x_{2_{w_j}}$ of each cooling jacket computed by the reduced optimisation, with $N = 10$ and $m = 8$, $fobj = 0.99979205$

| Parameter | Jacket $j = 1$ | Jacket $j = 2$ | Jacket $j = 3$ |
|---|---|---|---|
| $x_{2_{w_j}}(t_0)$ | 5.0 | 3.78621 | 1.89431 |
| $x_{2_{w_j}}(t_f)$ | 3.55721 | 3.79513 | 2.8387 |
| $t_{int}$ | 0.68975 | 0.271234 | 1.1124E-2 |
| $x_{2_{w_j}}(t_{int})$ | 3.4711E-3 | 7.2523E-4 | 0. |
| $A_1$ | 1. | 1. | 1. |
| $A_2$ | 1. | 1.000782 | 1. |

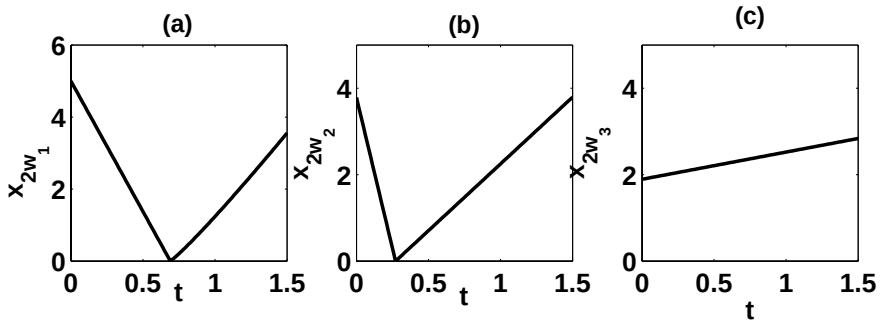



Fig. 2: (a) Control profile for jacket 1; (b) Control profile for jacket 2; (c) Control profile for jacket 3. This results were obtained using $N = 10$ and $m = 8$.

### 5.2 Parameter Identification of Chemical Reaction Kinetics

An important problem in dynamic systems is related to the accurate estimation of parameters of a mathematical model for a given set of time-varying measurements or observations of the state variables. These type of problems (inverse problems) appear in many applications and well developed solution techniques are available for deterministic systems (e.g.[61, 71, 72, 73]). For microscopic systems, where no closed form equations mathematical optimization is possible through the use of stochastic methods [70, 32] with high computational cost or after the extraction of a simplified model from the molecular data. When solving inverse problems, many mathematical issues arise and auxiliary tools are needed. For instance, it is usual that the inverse problem is ill-posed due to inexact measurements and/or partially observed data; then regularization techniques (e.g. [22, 65]) are required to well-pose the problem.

Also, in order to take advantage the special structure of the Hessian matrix and gradients of the least squares formulation, the optimization techniques applied are commonly based in Gauss-Newton [63] or Levenberg-Marquadt [69] methods which work acceptably for small to medium size problems. For large-scale systems, projected Hessian methods specially adapted to the problem of parameter identification can be readily used [68]. In this section we apply the methodology presented above for the optimisation of microscopic simulators. This can be done directly through the use of restricting and lifting operations [35].

We consider, the parameter estimation of a microscopic model of a set of biochemical reactions that synthesize inducible enzymes in bacterial cells [66]

$$\begin{gathered} X_2 \underset{k_{15}}{\overset{k_4}{\rightleftharpoons}} X_3 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} X_1 \overset{k_3}{\rightarrow} X_3 + X_6 \\ X_6 \overset{k_5}{\rightarrow} Y_1 \\ X_6 + X_7 \overset{k_{12}}{\rightarrow} X_4 \overset{k_7}{\rightarrow} X_7 + Y_2 \\ X_7 \overset{k_6}{\rightarrow} X_5 \overset{k_8}{\rightarrow} X_9 + X_7 \\ X_9 \overset{k_9}{\rightarrow} Y_3 \\ X_9 + X_8 \overset{k_{13}}{\rightarrow} X_{12} \overset{k_{14}}{\rightarrow} X_8 + Y_4 \\ X_{12} \overset{k_{10}}{\rightarrow} X_{11} \overset{k_{11}}{\rightarrow} X_{10} + X_{12} \end{gathered} \tag{34}$$

where $X_1$ is a complexing product of the regulator gene and a metabolic product of the repressing metabolite, $X_2$ is a complexing product of the regulator gene and a metabolic product of the inducer, $X_3$ is a regulator gene, $X_4$ is a repressed functional gene, $X_5$ is an assembly of mRNA precursors on the functional gene, $X_6$ is a repressor molecule, $X_7$ is a functional gene for mRNA synthesis, $X_8$ are ribosomes, $X_9$ is mRNA specific to the functional gene, $X_{10}$ enzyme produced and separated from the template, $X_{11}$ is an assembly of aminoacids on the template, $X_{12}$ is the template for the synthesis of the enzyme, $Y_1$ is a decomposition product of the repressor molecule, $Y_2$ is a decomposition product of the repressed functional gene, $Y_3$ is a decomposition product of mRNA and $Y_4$ is a decomposition product of the ribosomes. The above mechanism (34), can be also described by the following (macroscopic) ODE system

$$\frac{\mathrm{d}X_1}{\mathrm{d}t} = k_1 X_3 - k_2 X_1 - k_3 X_1 \tag{35a}$$
$$\frac{\mathrm{d}X_2}{\mathrm{d}t} = k_{15} X_3 - k_4 X_2 \tag{35b}$$
$$\frac{\mathrm{d}X_3}{\mathrm{d}t} = -k_1 X_3 + k_2 X_1 + k_3 X_1 - k_{15} X_3 + k_4 X_2 \tag{35c}$$
$$\frac{\mathrm{d}X_4}{\mathrm{d}t} = k_{12} X_7 X_6 - k_7 X_4 \tag{35d}$$
$$\frac{\mathrm{d}X_5}{\mathrm{d}t} = k_6 X_7 - k_8 X_5 \tag{35e}$$
$$\frac{\mathrm{d}X_6}{\mathrm{d}t} = k_3 X_1 - k_5 X_6 - k_{12} X_6 X_7 \tag{35f}$$
$$\frac{\mathrm{d}X_7}{\mathrm{d}t} = -k_{12} X_6 X_7 + k_7 X_4 - k_6 X_7 + k_8 X_5 \tag{35g}$$
$$\frac{\mathrm{d}X_8}{\mathrm{d}t} = k_{14} X_{12} - k_{13} X_8 X_9 \tag{35h}$$
$$\frac{\mathrm{d}X_9}{\mathrm{d}t} = k_8 X_5 - k_9 X_9 - k_{13} X_8 X_9 \tag{35i}$$
$$\frac{\mathrm{d}X_{10}}{\mathrm{d}t} = k_{11} X_{11} \tag{35j}$$
$$\frac{\mathrm{d}X_{11}}{\mathrm{d}t} = k_{10} X_{12} - k_{11} X_{11} \tag{35k}$$
$$\frac{\mathrm{d}X_{12}}{\mathrm{d}t} = k_{13} X_8 X_9 - k_{10} X_{12} + k_{11} X_{11} - k_{14} X_{12} \tag{35l}$$

where $X_1, \ldots, X_{12}, Y_1, \ldots, Y_4$ are concentrations of the biochemical species and $k_1, \ldots, k_{15}$ are the kinetic rate parameters.

The optimization problem consists of minimising the error between the experimental observations and the model predictions in time and can be formulated as

$$\min \ \frac{1}{2} \int_{t_0}^{t_f} \left( X^{obs}(t) - X(t,k) \right)^T \left( X^{obs}(t) - X(t,k) \right) \mathrm{d}t \tag{36a}$$
$$\text{s.t. microscopic model of (34)} \tag{36b}$$

Here the constraints (36b) are implemented in the form of a "stochastic simulation algorithm" [64] (Monte Carlo-based) that describe the stochastic time evolution of the reacting system (34). Briefly, this microscopic simulator consists of the following steps: Set a system with $L$ particles or molecules with $L_j$ is the number of molecules of species $j$; set simulation time $t = t_0$; calculate transition probabilities $R_k$ for the reaction $k$ using the current molecular distribution; sum of transition rates $R_{tot} = \sum_k R_k$; generate two random numbers uniformly distributed $ran1$ and $ran2$; use $ran2$ to select reaction $\mu$ that will occur so that $\sum_{k=1}^{\mu-1} r_k < ran2 \cdot R_{tot} \leq \sum_{k=1}^{\mu} r_k$; adjust the number of molecules participating in the reaction $\mu$ according to the stoichiometry;

set simulation time $t = t - \frac{\ln(ran1)}{R_{tot}}$; and repeat until a prescribed final time $t = t_f$ has been reached.

The inherent noise in the evolution profile of discrete and stochastic systems can complicate the numerical computation of sensitivities, the parameter perturbations have to be large enough to truly identify responses from the stochastic noise. In section 1 we discussed a few methodologies that have been used in the literature to address this problem In any case, a number of parallel realizations (simulations) need to be averaged in order to reduce the effect of the noise. Here, we have used central differences according to scheme (22) with $\varepsilon = 1 \times 10^{-3}$ to ensure that the perturbation captures the simulation noise and that the derivatives are acceptably accurate. Additionally, we use $L = 1 \times 10^6$ particles and average 10 realizations.

The set of "experimental" observations were generated from a single realization of the microscopic simulation at 8 time instants with the exact kinetic parameters. The initial conditions are: $X_3 = 1.0$, $X_7 = 1.0$ and $X_8 = 1.0$, the rest of the concentrations are zero. The experimental data were produced at the time instants: $t$ =[0.125, 0.3750, 0.5, 0.75, 1.0, 5.0, 10.0, 15.0]. The exact parameters are: $k$ =[5.0, 0.9, 0.1, 0.1, 0.035, 0.006, 0.1, 0.005, 0.1, 0.05, 0.2, 0.005, 0.02, 1.0]. We consider that the stochastic noise of the simulation is a random perturbation of the "true" dynamic trajectory

We perform parameter estimation computations using the microscopic simulator with $N = 5$ and $N = 10$ and $m = 2, 4, 10$. The reduced optimization method is able to compute the kinetic parameters with $N = 5$ and $m = 4$ and 10. For $m = 2$ there is no convergence. For $N = 10$ we encountered similar behavior. A comparison of some of the $k$ values estimated is given in Tables (4 and 5). A graphical comparison of the computed dynamic trajectories for some species is provided in Figs. (3, 4, 5, 6, 7, 8). The same calculations were performed with $L = 1 \times 10^5$ particles and averaging up to 50 realizations. Only the runs with $N = 10$ and $m = 10$ converged.

The estimated trajectories have good agreement with the exact dynamic trajectories at the beginning of the time horizon. This can be explained since half of the observations were chosen at the beginning of the time horizon.

Table 4: Estimated kinetic parameters using the microscopic model with $N = 5$ and $m = 4$. Note, that following [66] $k_{15} = 1.0$ and it is fixed in all calculations

| first | second | third | |
|---|---|---|---|
| $k_1$ | 5.4235 | $k_8$ | 0.09814 |
| $k_2$ | 0.83342 | $k_9$ | 0.00555 |
| $k_3$ | 0.20085 | $k_{10}$ | 0.1893 |
| $k_4$ | 0.1803 | $k_{11}$ | 0.00578 |
| $k_5$ | 0.03845 | $k_{12}$ | 0.2056 |
| $k_6$ | 0.00719 | $k_{13}$ | 0.00489 |
| $k_7$ | 0.1983 | $k_{14}$ | 0.02913 |

Table 5: Estimated kinetic parameters using the microscopic model with $N = 10$ and $m = 10$. Note, that following [66] $k_{15} = 1.0$ and it is fixed in all calculations

| | | | |
|---|---|---|---|
| $k_1$ | 4.9129 | $k_8$ | 0.10073 |
| $k_2$ | 0.9002 | $k_9$ | 0.0051 |
| $k_3$ | 0.1045 | $k_{10}$ | 0.1014 |
| $k_4$ | 0.1023 | $k_{11}$ | 0.00478 |
| $k_5$ | 0.02993 | $k_{12}$ | 0.2081 |
| $k_6$ | 0.00435 | $k_{13}$ | 0.0051 |
| $k_7$ | 0.1012 | $k_{14}$ | 0.01994 |

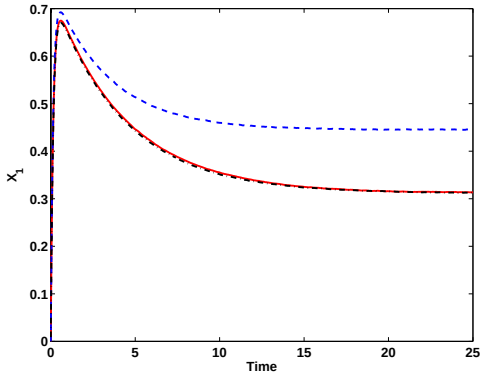



Fig. 3: Dynamic Trajectories using the microscopic simulator for species $X_1$ with exact kinetic parameters (solid line), estimated with $N = 5$ and $m = 4$ (broken line) and estimated with $N = 10$ and $m = 10$ (dashed-dotted line)

## 6 Conclusions

A model reduction-based computational framework has been developed to perform dynamic optimization using input/output macroscopic and microscopic simulators and can handle large multi-scale dynamic simulators. The algorithm is based on a multiple shooting discretization and it takes advantage of the low-dimensional dynamics that dissipative systems exhibit. It couples Newton-Picard methods with reduced Hessian techniques. Only very low-order block Jacobians are needed in each time subinterval. The low-order block Jacobian resulting from the multiple shooting discretization is used for a subsequent projection to the null-space of the system corresponding to the control parameters. All the essential gradients and Jacobian and Hessian matrices are efficiently computed numerically using only low-order projections.

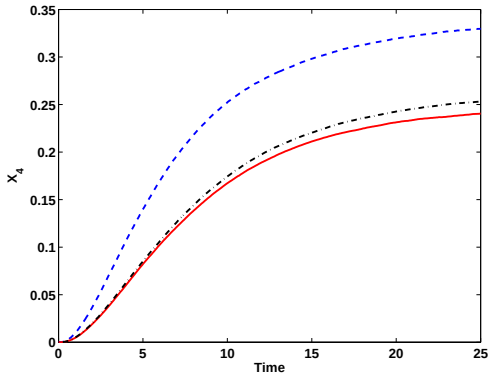



Fig. 4: Dynamic Trajectories using the microscopic simulator for species $X_4$ with exact kinetic parameters (solid line), estimated with $N = 5$ and $m = 4$ (broken line) and estimated with $N = 10$ and $m = 10$ (dashed-dotted line)

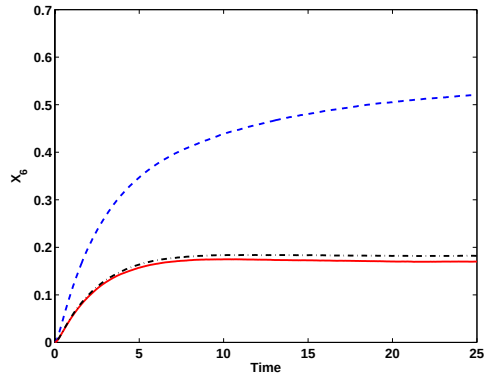



Fig. 5: Dynamic Trajectories using the microscopic simulator for species $X_6$ with exact kinetic parameters (solid line), estimated with $N = 5$ and $m = 4$ (broken line) and estimated with $N = 10$ and $m = 10$ (dashed-dotted line)

We have illustrated our methodology with the optimization of a tubular reactor simulated by a 4th Runge Kutta integrator. Excellent agreement with results from a *conventional* multiple shooting-based algorithm was obtained withe significant computational speed-up.

Several parametric studies have been performed by varying the invariant dominant subspace dimension ($m$) and number of time subintervals ($N$) of the multiple shooting discretization. If the number of time subintervals is increased the size of the subspace can be decreased. We also showed a parameter estimation problem using a MC-based simulator of a biochemical system. The results obtained from the dynamic optimization of the stochastic model were very good when compared with the exact kinetic parameters, showing that

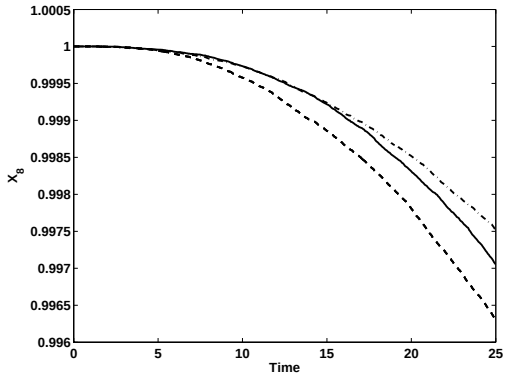



Fig. 6: Dynamic Trajectories using the microscopic simulator for species $X_8$ with exact kinetic parameters (solid line), estimated with $N = 5$ and $m = 4$ (broken line) and estimated with $N = 10$ and $m = 10$ (dashed-dotted line)

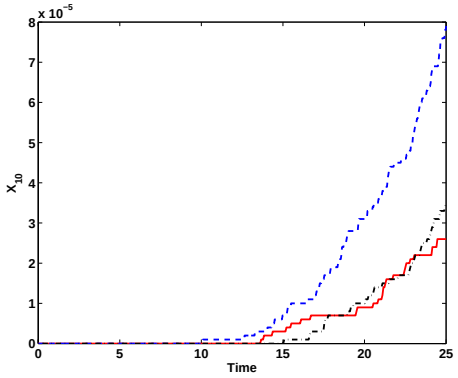



Fig. 7: Dynamic Trajectories using the microscopic simulator for species $X_{10}$ with exact kinetic parameters (solid line), estimated with $N = 5$ and $m = 4$ (broken line) and estimated with $N = 10$ and $m = 10$ (dashed-dotted line)

significant digit. the methodology has strong potential for the optimization of multi-scale systems.

## References

1. P. Raghavan, S. Ghosh: Adaptive multi-scale computational Modeling of composite materials. *CMES-Computer Modeling Eng. Sci.* **5**, 151 (2004)
2. M. A. Gallivan, R. M. Murray: Reduction and identification methods for Markovian control systems, with application to thin film deposition *Int. J. Robust and Nonlin. Control* **14**, 113 (2004)

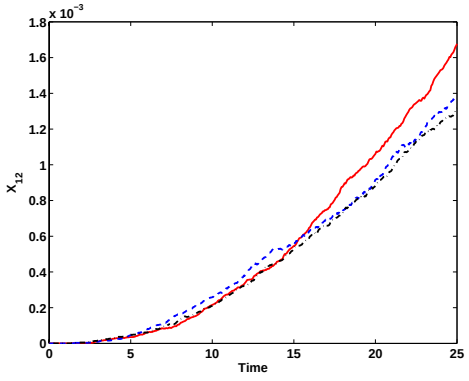



Fig. 8: Dynamic Trajectories using the microscopic simulator for species $X_{12}$ with exact kinetic parameters (solid line), estimated with $N = 5$ and $m = 4$ (broken line) and estimated with $N = 10$ and $m = 10$ (dashed-dotted line)


3. H. Schlichting, K. Gersten: *Boundary Layer Theory*, 8th edn (Springer, Berlin Heidelberg New York 2000)
4. O. I. Tureyen J. Caers: A parallel, multiscale approach to reservoir modeling. *Comput. Geosciences* **9**, 75 (2005)
5. Y. G. Yanovsky: Multiscale modeling of polymer composite properties. *Int. J. Multiscale Comput. Engineering* **3**, 199 (2005)
6. G. Csanyi, T. Albaret, G. Moras, M. C. Payne, A De Vita: Multiscale hybrid simulation methods for material systems. *J. Phys. Cond. Matter* **17**, R691 (2005)
7. G. S. D. Ayton, S. Bardenhagen, P. McMurry, D. Sulski et al: Interfacing molecular dynamics with continuum dynamics in computer simulation: Toward an application to biological membranes. *IBM J. Res. & Dev* **45**, 417 (2001)
8. G. S. D. Ayton, G. A. Voth: Simulation of Biomolecular Systems at Multiple Length and Time Scales. *Intl. J. Multiscale Comp. Eng.* **2**, 289 (2004)
9. R. D. Braatz, R. C. Alkire, E. Rusli et al: Multiscale systems engineering with applications to chemical reaction processes, *Chem. Eng. Sci.* **59**, 5623 (2004)
10. K. Burrage, T. Tian, P. Burrage: A multi-scaled approach for simulating chemical reaction systems. *Biophys. Mol. Biol.* **85**, 217 (2004)
11. J. Li, J. Zhang, W. Ge et al: Multi-scale methodology for complex systems. *Chem. Eng. Sci.* **59**, 1687 (2004)
12. H. H. McAdams, A. Arkin: It's a nosiy business! Genetic regulation at the nanomolar scale. *Trends Genet.* **15**, 65 (1999)
13. R. E. Miller: Direct Coupling of Atomistic and Continuum Mechanics in Computational Materials Science. *Intl. J. Multiscale Comp. Eng.* **1**, 5722 (2003)
14. E. Rusli, O. Drews, R. D. Braatz: System analysis and design of dynamically coupled multiscale reactor simulation codes. *Chem. Eng. Sci.* **59**, 22 (2004)
15. M. K. Gobbert, T. P. Merchant, L. J. Borucki, T. S. Cale: A multiscale simulator for low pressure chemical vapor deposition. *J. Electrochem. Soc.* **144**, 3945 (1997)
16. K. F. Jensen, S. T. Rodgers, R. Venkataramani: Multiscale modeling of thin film growth. *Curr. Opinion Solid State Mater. Sci.* **3**, 562 (1998)

17. S. Ogata, T. Igarashi: Concurrent coupling of electronic-density-functional, molecular dynamics, and coarse-grained particles schemes for multiscale simulation of nanostructured materials. *New Front. Proc. Engin. Advanced Mater. Sci. Forum* **502**, 33 (2005)
18. B. Protas, T. R. Bewley, G. Hagen: A computational framework for the regularization of adjoint analysis in multiscale PDE systems. *J. Comp. Phys.*, **195**, 49 (2004)
19. M. D. Gunzburger: *Perspectives in Flow Control and Optimization* (SIAM Philadelphia 2003)
20. S. S. Sritharan: *Optimal Control of Viscous flows* (SIAM Philadelphia 1998)
21. A. Griewank, C. Faure: Reduced functions, gradients and Hessians from fixed-point iterations for state equations. *Numer. Algorithms* **30**, 113 (2002)
22. H. Engl, M. Hanke, A. Neubauer: *Regularization of Inverse Problems* (Kluwer Dordrecht 1996)
23. C. R. Vogel: *Computational Methods for Inverse Problems* (SIAM Philadelphia 2002)
24. J. Fish, A. Ghouali: Multiscale analytical sensitivity analysis for composite materials. *Int. J. Numer. Methods Eng.* **50** 1501 (2001)
25. Y. Liu, B. S. Minsker: Full multiscale approach for optimal control of in situ bioremediation. *J. Water Res. Planning Management* **130**, 26 (2004)
26. T. Binder, L. Blank, W. Dahmen, W. Marquardt: Iterative algorithms for multiscale state estimation, part 1: Concepts. *J Optim. Theory Applic.* **111**, 501 (2001)
27. A. Lucia, P. A. DiMaggio, P. Depa: Funneling algorithms for multiscale optimization on rugged terrains. *Ind. Eng. Chem. Res.* **43**, 3770 (2004)
28. A. Lucia, F. Yang: Solving distillation problems by terrain methods. *Comput. Chem. Eng.* **28**, 2541 (2002)
29. B. Liu, H. Jiang, Y. Huang, S. Qu, M. F. Yu, K.C. Hwang: Atomic-scale finite element method in multiscale computation with applications to carbon nanotubes. *Phys. Rev. B* **72**, 035435 (2005)
30. A. Varshney, A. Armaou: Multiscale optimization using hybrid PDE/kMC process systems with application to thin film growth. *Chem. Eng. Sci.* **60**, 6780 (2005)
31. L. Sirovich: Turbulence and the dynamics of coherent structures. 1. Coherent structures. *Quart. Applied Math.* **XVL** 573 (1987)
32. S. Raimondeau, P. Aghalayam, A. B. Mhadeshwar, D.G. Vlachos: Parameter optimization of molecular models: Application to surface kinetics. *Ind. Eng. Chem. Res.* **42**, 1174 (2003)
33. R. Faber, T. Jockenhövel, G. Tsatsaronis: Dynamic optimization with simulated annealing *Comp. Chem. Eng.* **29**, 273 (2005)
34. E. H. L. Aarts and J. H. M. Korst: *Simulated Annealing and Boltzmann Machines: a Stochastic Approach to Combinatorial Optimization and Neural computing*, (Wiley, New York 1989)
35. I. G. Kevrekidis, C. W. Gear, J. M. Hyman, P. G. Kevrekidis, O. Runborg, C. Theodoropoulos: Equation-Free coarse-grained multiscale computation: Enabling microscopic simulators to perform system-level tasks. *Comm. Math. Sci.* **1**, 715 (2003)
36. C. Theodoropoulos, K. Sankaranarayanan, S. Sundaresan, I. G. Kevrekidis: Coarse bifurcation studies of bubble flow lattice Boltzmann simulations. *Chem. Eng. Sci.* **59** 2357 (2004)

37. A. Armaou, C. I. Siettos, I. G. Kevrekidis: Time-steppers and 'coarse' control of distributed microscopic processes. *Int J. Robust Nonlin. Control* **14** 89 (2004)
38. E. Luna-Ortiz, C. Theodoropoulos: An Input/Output Model Reduction-based optimization scheme for large-scale systems. *Multiscale Model. Simul.* **4**, 691 (2005)
39. A. Armaou, P. D. Christofides: Dynamic Optimization of dissipative PDE systems using nonlinear order reduction. *Chem. Eng. Sci.* **57**, 5083 (2002)
40. U. M. Ascher, R. M. M. Mattheij, R. D. Russell *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, (SIAM, Philadelphia 1995)
41. Z. Bai, G. W. Stewart: Algorithm 776: SRRIT: A Fortran Subroutine to Calculate the Dominant Invariant Subspace of a Nonsymmetric Matrix. *ACM Trans. Math. Softw.* **23**, 494 (1997)
42. A. Barclay, P. E. Gill, J. B. Rosen: SQP Methods and their application to numerical optimal control. In: *Variational Calculus, Optimal Control and Applications*, vol 124, ed by W. H. Schmidt, K. Heier, L. Bittner et al, pp 207–222 (Birkhäuser Basel 1998)
43. J. T. Betts: A Survey of Numerical Methods for Trajectory Optimization. *AIAA J. Guidance, Control and Dynamics* **21**, 193 (1998)
44. R. W. H. Sargent: Optimal Control. *J. Comp. Appl. Math.* **124**, 361 (2000)
45. O. von Stryk, R. Burlisch: Direct and Indirect Methods for Trajectory Optimization. *Annals Oper. Res.* **37**, 357 (1992)
46. L. T. Biegler: Solution of Dynamic Optimization Problems by Succesive Quadratic Programming and Orthogonal Collocation. *Comp. Chem. Eng.* **8**, 243 (1984)
47. T. H. Tsang, D. M. Himmelblau, T. F. Edgar: Optimal Control via collocation and nonlinear programming. *Intl. J. Control* **21**, 763 (1975)
48. K. Lust, D. Roose, A. Spence et al: An adaptive Newton-Picard algorithm with subspace iteration for computing periodic solutions. *SIAM J. Sci. Comput.* **19**, 1188 (1998)
49. P. Eberhard, C. Bischof: Automatic Differentiation of Numerical Integration Algorithms. *Math. Comp.* **68**, 717 (1999)
50. D. B. Leineweber. I. Bauer, H. G. Bock et al: An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization. Part I: theoretical aspects. *Comp. Chem. Eng.* **27**, 157 (2003)
51. K. Lust: Bifurcation Analysis of Periodic Solutions of Partial Differential Equations. PhD Thesis, Katholieke Universiteit Leuven, Belgium (1997)
52. C. Schmid, L. T. Biegler: Acceleration of Reduced Hessian methods for Large-Scale nonlinear programming. *Comp. Chem. Eng.* **17**, 451 (1993)
53. T. Maly, L. R. Petzold: Numerical Methods and software for sensitivity analysis of differential-algebraic systems. *Appl. Numer. Math.* **20**, 57 (1996)
54. L. T. Biegler, J. Nocedal, C. Schmid:A Reduced Hessian Method for Large-Scale Constrained Optimization. *SIAM J. Optim.* **5**, 314 (1995)
55. H. G. Bock, K. J. Plitt: A multiple shooting algorithm for direct solution of optimal control problems. In: *Proceedings of the 9th IFAC World Congress*, pp 431–439 (Pergamon Press 1984)
56. P. T. Boggs, J. W. Tolle: Sequential Quadratic Programming for large-scale nonlinear optimization. *J. Comp. Appl. Math.* **124**, 123 (2000)
57. O. Buchauer, P. Hiltman, M. Kiehl: Sensitivity analysis of initial-value problems with applications to shooting techniques. *Numer. Math.* **67**, 151 (1994)

58. E. Luna-Ortiz C. Theodoropoulos: *Manuscript in preparation*
59. K. L. Choong, R. Smith: Optimization of batch cooling crystallization. *Chem. Eng. Sci.* **59**, 313 (2004)
60. T. O. Drews, R. D. Braatz, R. C. Alkire: Parameter Sensitivity Analyisis of Monte Carlo Simulations of Copper Electrodeposition with Mutiple Additives. *J. Electrochem. Soc.* **150**, C807 (2003)
61. L. Edsberg, P. Wedin: Numerical tools for parameter estimation in ODE-systems. *Optim. Meth. Softw.* **6**, 193 (1995)
62. P. E. Gill, L. O. Jay, M. W. Leonard et al: An SQP method for the optimal control of dynamical systems. *J. Comp. Appl. Math.* **120**, 197 (2000)
63. P. E. Gill, W. Murray: Algorithms for the solution of the Nonlinear Least-Squares Problem. *SIAM J. Numer. Anal.* **15**, 977 (1978)
64. D. T. Gillespie: Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.* **81**, 2340 (1977)
65. M. Hanke, P. C. Hansen: Regularization methods for large-scale problems. *Surv. Math. Industry* **3**, 253 (1993)
66. F. Heinmets: Analog computer analysis of a model system for the induced enzyme synthesis, *J. Theoret. Biol.* **6**, 60 (1964)
67. K. F. Jensen, W. H. Ray: The bifurcation behavior of tubular reactors. *Chem. Eng. Sci.* **37**, 199 (1982)
68. K. Kunisch, E. W. Sachs: Reduced SQP Methods for Parameter Identification Problems. *SIAM J. Numer. Anal.* **29**, 1793 (1992)
69. D. W. Marquardt: An algorithm for least squares estimation of non-linear parameters. *SIAM J. Appl. Math.* **11**, 431 (1963)
70. K. Mosegaard, M. Sambridge: Monte Carlo analysis of inverse problems. *Inverse Problems* **18**, 29 (2002)
71. K. Schittkowski *Numerical Data Fitting in Dynamical Systems — A Practical Introduction with Applications and Software* (Kluwer, Dordrecht 2002)
72. W. Stortelder: Parameter Estimation in Nonlinear Dynamical Systems. PhD Thesis, CWI, Amsterdam, The Netherlands (1998)
73. C. R. Vogel: *Computational Methods for Inverse Problems* (SIAM, Philadelphia PA 2002)