Numerical Methods for the Chemical Master Equation

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von
M. Sc. Tudor Udrescu
aus
Constanţa

Tag der mündlichen Prüfung: 6. Juni 2012

Referent: Prof. Dr. Tobias Jahnke
Korreferent: Prof. Dr. Andreas Rieder

CONTENTS

ABSTRACT

This thesis is concerned with the construction and analysis of numerical methods for stochastic reaction networks, where the term *stochastic* means that the model contains a degree of randomness or unpredictability. The dynamics of such reaction networks are described by using a *Markov jump process* on large and usually high-dimensional state spaces, with the corresponding time-dependent probability distribution being the solution of the chemical master equation (CME). Adding an element of unpredictability in biological modeling is a relative new development, as only recently it has been recognized that biochemical kinetics, especially those at the intracellular level, are intrinsically stochastic. The solution of the CME provides the most accurate picture of the dynamics of such systems, but solving the equation numerically is hampered by the *curse of dimensionality*: the number of degrees of freedom scales exponentially with the number of species involved in the reaction network.

Consequently, we develop herein an approach that mitigates the effects of the curse of dimensionality by using *wavelet compression*. Adaptive wavelet-based numerical methods are devised for both the time-dependent and stationary CME. Reducing the number of degrees of freedom via wavelet compression is not the only challenge faced when investigating biochemical reaction networks via the CME: the *metastability* of many systems poses additional difficulties. Another objective of the thesis is to develop efficient numerical tools allowing the approximation of the committor probabilities - statistical objects that describe the progress of the transitions between subsets of the state space. Used within the framework of Transition Path Theory, the committor probabilities provide a detailed insight into the metastable dynamics of biological systems.

In order to exploit the multi-scale nature of many biological systems and achieve further reductions in the number of degrees of freedom required to approximate the stationary CME, an embedding of the adaptive wavelet method within a hybrid strategy is also explored. With a hybrid approach, significant reductions can be achieved by using the computationally intensive wavelet method only for the parts of the biochemical reaction networks composed of species with small concentrations and treating the remaining components in a deterministic setting. The methods are illustrated on multi-dimensional models with metastable solutions, which are defined on large state spaces.

ACKNOWLEDGMENTS

First and foremost, I would like to sincerely thank my advisor, Prof. Dr. Tobias Jahnke for his guidance, patience and support during the last four years. I am also grateful to Prof. Dr. Andreas Rieder for agreeing to be my second referee. Furthermore, I thank all members (past and present) of the IANM3 *Wissenschaftliches Rechnen* for creating such a nice atmosphere to work and sometimes to play: Dr. Martin Sauter, Dr. Wolfgang Müller, PD Dr. Nicolas Neuß, Daniel Maurer, Tim Kreutzmann, Prof. Dr. Christian Wieners, Andreas Arnold, Michael Kreim, Robert Winkler, Andreas Schulz, Ekkachai Thawinan, Nathalie Sonnefeld and Sonja Becker. Thanks also goes to Prof. Dr. Constantin Popa for initially drawing me to the field of numerical analysis and for offering encouraging words ever since.

I am also grateful to the Deutsche Forschungsgemeinschaft (DFG) which provided the financial support for my work under the priority programme SPP 1324 "Mathematische Methoden zur Extraktion quantifizierbarer Information aus komplexen Systemen". Moreover, as a member of the research group "Numerical methods for high-dimensional systems", I am also thankful for the financial support received from the "Concept for the Future" of Karlsruhe Institute of Technology (KIT) within the framework of the German Excellence Initiative.

Finally, I would like to express my deepest gratitude to my parents, Ing. Udrescu Viorel and Ing. Udrescu Steliana, whose support I could always count on.

CHAPTER

ONE

INTRODUCTION

Motivation

Recent decades have seen a tremendous level of activity in the field of molecular biology, where the use of new technologies enabled researchers to continuously expand the boundaries of knowledge on biological phenomena occurring at the cellular level. The large amounts of data being collected on individual cellular components and better understanding of the interactions also triggered the emergence of *systems biology* as a new interdisciplinary field that views biological processes as dynamical networks, thus expanding the toolbox of mathematical models and computer simulations available for the investigation of the complex relationships inside such systems. This *in-sillico* approach to molecular biology is playing an increasingly important role alongside conventional *in-vivo* methods, as it allows the quantitative assessment of various assumptions and hypotheses about the structure and internal mechanisms of biological networks, with significant time and cost savings compared to traditional laboratory methods.

However, while a large list of the "building blocks" of living organisms has already been assembled and their internal mechanisms are generally well understood (e.g. the seminal results from [GCC00, ESSL02, Pta04]), the integrated knowledge of how these pieces work together to influence phenotypic heterogeneity at higher levels is far from complete. The ultimate goal of *systems biology* is to enable the engineering of complex behavior in living organisms via changes that are robustly propagated either down-stream or up-stream of the location where they are added (see [Wil09] for a compelling argument on this subject), but achieving such ambitious goals necessitate further efforts in developing or adapting the mathematical and computational frameworks to handle the complexity of biological organization.

An important topic in computational systems biology has been the increasing awareness that stochasticity and discreteness play an important role in biological reaction networks at the cellular level [ADA09]. This is supported by experimental results [MA99, EBE01, EL00], which makes the study of stochastic fluctuations, although a challenging task due to the complexity of the dynamics involved, almost mandatory, a fact highlighted in many recent reviews [TSB04, KEBC05, RO04].

Consequently, questions related to the development of models that deliver the high resolutions needed to reveal important biological details like the effects of molecular noise have lately attracted a lot of interest. Any scientific research that has the goal of studying a "real" biological process, involves the question of how accurate the model used to depict reality should be, or to formulate the problem more precisely, how to construct a model using the available knowledge so that the discrepancy between the model and the real process is not too large and the model is simple enough to remain computationally tractable. Once such a model is constructed, the next question is whether this "exact" description can be used to investigate processes of genuine interest, or for practical reasons an "approximate" formulation is needed. Usually, the last choice is the only realistic option if the goal is to move away from studying the interactions of specific single molecules and describe instead the more complex behavior of biological systems that involve many components arranged in biological networks.

In the larger sense, this thesis is concerned with the construction, analysis and application of such "approximate" descriptions of complex biological processes. However, in order to keep things into perspective, it is important to note that the difference between the "exact" and "approximate" models is usually far less than that between the "exact" model and the real process, so choosing the appropriate modeling paradigm is of upmost importance.

Modeling choices

Ideally, describing a complex physical system would be done using some sort of *deterministic* model, meaning that given some past state we can completely characterize the future by employing some accurate evolution laws that obey the appropriate physics and keep track of the positions and speed of all the molecules involved, as well as their interactions.

Such *molecular dynamics* approaches can be very accurate, but the shear complexity of the interactions means they are usually too expensive from a computational point of view, especially if the model involves more than single molecules of each type and the dynamics are to be investigated over a longer time interval. A model on this scale is called *microscopic*, and employs Brownian dynamics for the movement of the molecules and the Smoluchowski model for their interactions. Not withstanding the challenges, advances in computational approaches like the Green's Function Reaction Dynamics (GFRD) algorithm proposed in [vZtW05], have enabled the application of such models to some biological systems, and their use will certainly increase in the future, especially in view of the development of hybrid approaches [HHL11]. However, biological complexity and the difficulty in formulating useful laws that take all effects into account, currently limit the use of *microscopic* models.

Much of the earlier mathematical modeling of cellular processes employed instead a *macroscopic* approach, that is, a deterministic model that assumes large population levels, discards the spatial dimension and is used to study the average behavior. Naturally, such simplifications can only be made under certain assumptions, namely that in addition to having large molecular copy-numbers that dampen the effects of molecular noise, the system is also *well-stirred*, meaning the molecules are uniformly spread within a container

of constant volume and the temperature is also constant. The time evolution of such a system can then be modeled via a system of ordinary differential equations (ODEs) representing the concentrations of the molecular populations involved, known as the *reaction rate equations* (RRE).

However, because biological processes at the cellular level such as gene regulatory networks, usually exhibit low copy numbers of participating molecules, this means that some of assumptions made in this classical deterministic setting are no longer valid. In order to obtain an accurate model for such systems, which is still reasonably simple to simulate despite the higher resolution, *randomness* has to be introduced into the mathematical model, while preserving the *well-stirred* characterization. Therefore, a *mesoscopic* model which lies between the very accurate but prohibitively costly *microscopic* scale and the coarse but from a computational point of view easily accessible *macroscopic* scale, has emerged as the the most popular choice for modeling stochastic effects, as it respects both the stochastic nature of biological processes and the discreteness of the population numbers. The model is based on the assumption that the process driving the evolution of the system is memoryless, i.e., depends only on the current state of the system and not the whole system history, with the mathematical formulation provided by a continuous time discrete space *Markov jump process* [Gil76].

In the *mesoscopic* formulation, the effects that are either too complex or too expensive to simulate are simply summarized in terms of random variables. Then, the future can no longer be unambiguously determined from the past and is described only in a probabilistic sense. This is suitable for most applications, because the questions being posed are of a quantitative nature, namely the time-evolution of the population numbers of the different interacting cellular components. From a computational point of view, realizations of the *Markov jump process* can be generated via the Stochastic Simulation Algorithm (SSA) also known as the Gillespie algorithm (see [Gil76]). Naturally, each run of a given model will produce a different result, but the probability distribution of the results for a certain time is determined by the underlying mathematical formulation and can be computed as the solution of the *Chemical Master equation* (CME). Thus, the CME provides an "exact" description of the stochastic model. However, the full probability distribution for the state of a biochemical system over time can only be computed in simple situations, which limits the direct use of CME. Numerical approximations of the solution are also not trivial to obtain as the CME is affected by the *curse of dimensionality*: the number of degrees of freedom needed for an accurate approximation grows exponentially with any increase in the number of components of the biochemical system.

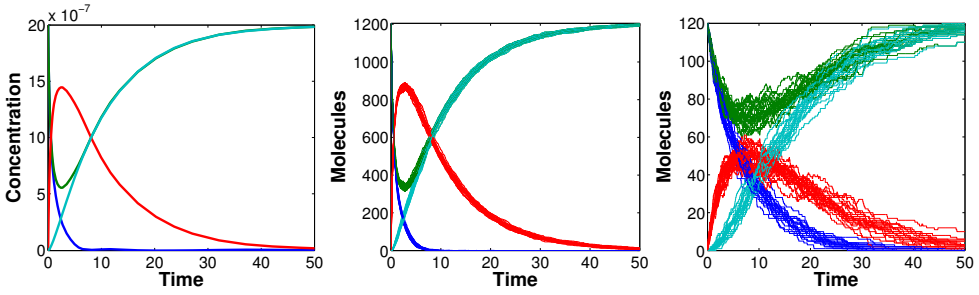
As the number of degrees of freedom present in most problems that merit investigation is tremendous, the usual computational approach in *mesoscopic* modeling has been based on Monte Carlo simulations using the SSA algorithm, either the original variant from [Gil76] or the many modifications that have been proposed since (see e.g., [GB00, Gil01, CGP05, CGP07]). In theory, the associated Monte Carlo error can be made arbitrarily small by increasing the number of simulations, but obtaining an accurate approximation of the probability distribution using stochastic simulations is usually not feasible because any change in the state of the system requires an update of the state vector. For systems with multiple time-scales, this can lead to high computational costs.

An alternative is to try to devise methods to solve CME directly, despite the challenges

posed by the *curse of dimensionality*. As both alternatives are computationally expensive, the question of the usefulness of stochastic modeling arises, and whether the incurred computational cost is justified. A comparison between the results obtained using the deterministic and stochastic approaches can thus shed light on why including molecular noise in the model is important, particularly in the case of gene regulatory networks.

Advantages of stochastic modeling

As stated before, system size is an important factor that contributes to stochasticity and larger copy numbers of molecules means that the influence of stochastic fluctuations on the dynamics of the system are less pronounced. This is illustrated in Figure 1.1, by comparing the deterministic and stochastic solutions of the Michelis-Menten model of enzyme kinetics (cf. [Hig08]). Plots 1.1a and 1.1b show the time evolution of each of the species, computed by using the deterministic reaction rate equations and SSA, respectively.



(a) Deterministic (b) Stochastic, large numbers (c) Stochastic, small numbers

Figure 1.1.: Stochastic fluctuations in the Michelis-Menten system

The initial number of molecules has been chosen to be in the range of 10^3, and the different scales are due to the fact that while the *mesoscopic* model delivers molecule numbers, the *macroscopic* model outputs concentrations. With the mention that plot 1.1b shows a number of superimposed independent SSA runs, it is evident that when large numbers of molecules are present, the stochastic solution looks like a slightly noisy solution of the corresponding differential equations. However, a 10-fold reduction of the initial copy numbers leads to an increase in fluctuations and their possible effects, as evident from the SSA runs plotted in 1.1c. In general, when N denotes the average number of molecules, a decrease in copy numbers will result in a $1/\sqrt{N}$ scaling of the noise [Wil09]. As a consequence, systems which exhibit *low-copy* numbers should be treated stochastically, as the deterministic model fails in such cases to capture the real dynamics.

Among the most important manifestations of stochasticity in cellular processes is the appearance of *multistability*. Multistable biological systems are also called *toggle switches*, and spend most of their life in two (or more) meta-states, until stochastic noise induces a sudden transition between these states. Using a well known model of a bistable toggle switch from [GCC00], which represents a synthetic gene regulatory network composed of a mutually repressible gene pair, we illustrate in Figure 1.2 how the stochastic model captures behavior which cannot be observed when using the deterministic model. For

any given initial conditions, the ODE solution depicted by the red line, will converge to one or the other stable state and remain there for all time. However, adding noise to the model via the stochastic description, leads to the observation that if the noise amplitude is sufficient, the solution will also visit the other stable state. The flexibility to switch between the stable states can therefore help explain the appearance of different behavior in isogenic cell populations, leading to the conclusion that stochasticity is in fact an evolutionary trait.

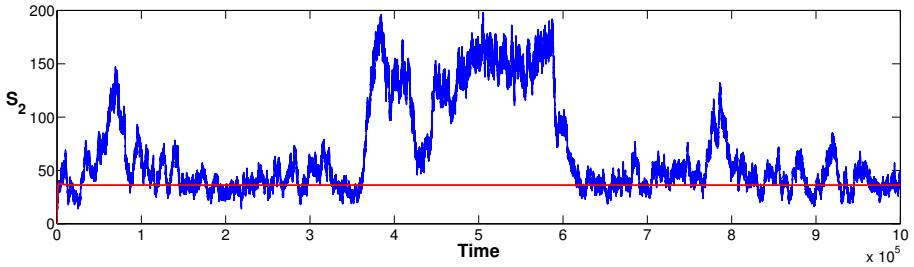


Figure 1.2.: Deterministic vs. stochastic solutions for a bistable toggle switch [GCC00], illustrating the effect of noise in multi-stable systems (figure adapted from [Eng08])

Another example where noise can impact the dynamics can be found in genetic oscillators, which are used by many living organisms as internal clocks for regulating behavior between day-time and night-time periods. A model of these oscillations can be found in [VKBL02], and Figure 1.3 showcases the manifestation of the *stochastic resonance* phenomenon, namely how noise can push the system out of a stable fixed point and start a new cycle. The occurrence of the oscillations can even be modulated by tuning the amount of noise. For all parameter sets, the stochastic model is sensitive enough to detect the fluctuations that send the system onto a new cycle, while the deterministic model is prone to settling into the stable state after the first oscillation for certain parameter values. These are just some of the effects of stochasticity in gene expression (for a comprehensive study on the subject see, e.g. [KEBC05] or the monograph [vK01]), which motivate a *mesoscopic* treatment of cellular processes.

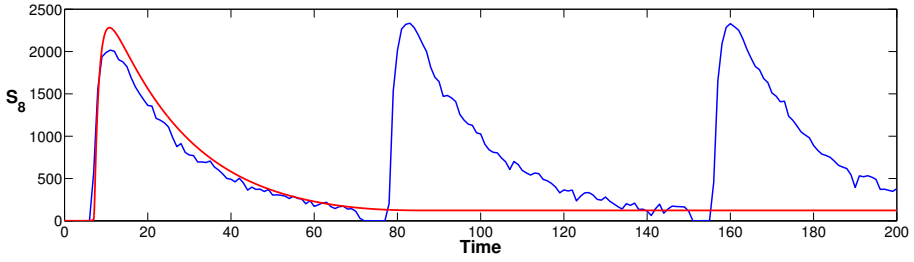


Figure 1.3.: Model of a *circadian oscillator* where the stochastic approach exhibits reliable oscillations while the deterministic model fails (figure adapted from [VKBL02])

State of art

As the CME can provide an accurate picture of the dynamics of intracellular networks similar to those presented in the previous section, a significant effort has been made in re-

cent years towards developing adequate numerical methods for this equation. The methods usually employ one of several computational approaches, either a discrete Galerkin method coupled with Rothe's method ([Eng09a, DHJW08, Jah10, JH08]), a finite state projection (FSP) algorithm ([MBBS08, MBS08, MK06]), an aggregation approach ([HBS+07]), sparse grids [HHL08] or adaptive lumping of states [FL09]. Explicit solution formulas for the special case of the monomolecular CME have also been derived in [JH07]. Sacrificing discreteness in the stochastic model is also possible, as stochastic effects can be modeled via the Chemical Langevin equation (CLE), a stochastic differential equation which extends the macroscopic model by appending a noise term. The corresponding probability density then evolves according to the Fokker-Planck equation, and delivers an approximation to the solution of the master equation [SLE09]. An important distinction between these stochastic approaches however, is that while the CME is fully discrete and thus faithful to biological reality, the CLE is instead continuous and molecular quantities are given as real numbers. An in-depth discussion on the relationships between different models and the conditions under which they can be employed can be found in, e.g. [Hig08, Gil07, Eng08].

With the CME describing the dynamics of the *mesoscopic* model and the reaction rate equations modeling the *macroscopic* model, there is also the possibility of choosing a model that fits somewhere between these upper and lower limits, respectively (see e.g. [FLH08]). This has motivated the idea of substituting non essential parts of the CME solution with results obtained with cheaper models, thus achieving significant model reduction for the price of lower accuracy. The construction of such *hybrid* models can be accomplished in many ways, and a number of promising approaches can be found in the literature ([FL07, HL07, FLH08, HHL08]).

Irrespective of the inner workings of each method, the central idea is always to reduce the number of degrees of freedom to more computationally manageable levels. Another possible approach to achieve this goal is through the use of wavelet compression, and the details of using adaptive wavelet methods for the CME represent the narrower focus of this thesis.

Focus of the thesis

Generally speaking, the efficiency of all the methods mentioned in the state of the art section depends on the compression ratio that can be achieved, i.e., the percentage of degrees of freedom required to obtain the desired accuracy. In a wavelet basis, the number of *essential* degrees of freedom represent only a small fraction of the total number of unknowns. This is due to the fact that wavelets decompose an input signal into a hierarchy of scales, and since smooth signals will contain relatively small amounts of detail information, many coefficients of the wavelet representation can be safely discarded with only a negligible effect on the approximation error. Because the solution of the CME evolves in time, not only the compression properties are important, but also determining which elements currently form the essential set. Additionally, it is advantageous to propagate the solution using an adaptive time-stepping strategy, as many biological systems exhibit a stiff behavior in an initial phase, and variable step-sizes can yield important savings for

simulations. The results concerning the construction of an adaptive wavelet method for the CME with adaptive-time stepping will be discussed at length in Chapter 4.

In some cases the transient behavior of biological systems is not as relevant as their behavior at equilibrium, which is given as the solution of the stationary CME. Because the stationary CME is a particular case of the time-dependent problem, the wavelet methods can also be modified to compute the stationary probability distribution of a biological system, and the proposed numerical method will be presented step by step in Chapter 5.

Reducing the number of degrees of freedom via wavelet compression is not the only challenge faced when investigating biochemical reaction networks via the CME: the bimodality and metastability of many systems pose additional difficulties. Another objective of the thesis is to develop numerical tools that allow the efficient computation of committor probabilities, mathematical objects that are used to model the mechanistic transitions between certain states of interest. Employed within the framework of Transition Path Theory [MSVE08, VE06], the committor probabilities provide a detailed insight into the metastable dynamics of biological systems. This is relevant particularly for gene regulatory networks as they contain toggle switches (cf. [HBS+07, MBS08]), leading to metastability in the solution of the CME. Transitions between metastable states are *rare events* and their analysis is of fundamental interest to biochemists looking for detailed insight into the kinetics of the system, such as the actual transition mechanisms involved. However, the problem of computing statistics for rare events is often not trivial, as using stochastic simulations in a brute force approach is impractical. This is due to the fact that the computationally affordable simulation times are usually insufficient to observe enough relevant events to compute probabilities. In some sense, the application of TPT can be seen in the context of expanding the knowledge about complex systems at equilibrium beyond what can be learned from the solution of the stationary CME. Because of the similarity of computing the stationary probability distribution and committor probability, it makes sense to apply wavelet compression to the TPT committor problem as well, with the details to be found also in Chapter 5.

In order to achieve further reductions in the number of degrees of freedom required to approximate the solution of the stationary CME, we also investigate the embedding of the adaptive wavelet method within a hybrid strategy. In many real-life applications, the number of species in the makeup of biochemical systems is far too big even for the capabilities of adaptive wavelet method. However, one is only interested in the behavior of a *few* species which due to their low copy numbers are considered as critical. This leads in a natural way to the idea of using the computationally intensive wavelet approach only for the parts of the biochemical system susceptible to stochastic fluctuations, and treat the rest of the components in a deterministic setting. In Chapter 6, we study the use of the wavelet method within a hybrid approach first proposed by A. Hellander and P. Lötsdedt in [HL07], and discuss both the potential and the limitations of this hybrid model.

CHAPTER

TWO

STOCHASTIC REACTION KINETICS

The kinetics of biological processes can be modeled using a network of reaction channels $R_1, \ldots, R_M$ that involve reactant and product molecules belonging to a set of d different species $S_1, \ldots, S_d$ with d and $M \in \mathbb{N}^+$. For example, we might know that when a molecule from the species S_1 encounters a molecule of type S_2 and certain microphysical conditions are met, the two molecules can combine into a new molecule of type S_3. Such an interaction "law" can be easily specified in a natural way by using the notation

$$R_1 : \; S_1 + S_2 \longrightarrow S_3. \tag{2.1}$$

Although such reaction channels R_j ($j = 1, \ldots, M$) capture the interactions between the species, they are not sufficient by themselves to describe the full dynamics of the biological process. This requires also knowledge of the "rates" at which the reaction channels fire and some initial conditions.

Such descriptions of biological processes naturally lead to the idea that the mathematical treatment should take into account that any changes induced by the reaction channels in the copy numbers of species S_i ($i = 1, \ldots, d$) are discrete. As already briefly discussed in Chapter 1, this intuition of using a discrete characterization is of course entirely correct, as it reproduces the intrinsic discreteness of nature. The purpose of this chapter is to review the mathematical formalisms that lead to the discrete stochastic approach to reaction kinetics.

2.1. Microphysical basis

The information required in most applications is represented by the copy numbers of the species S_i at time $t > 0$ or at chemical equilibrium, given that the initial amounts are known. As stated in Chapter 1, ideally this information would be extracted using a full deterministic model by keeping track of the positions, speed and interactions of all the participating molecules. However, because this *molecular dynamics* approach is usually not feasible, we are forced to stipulate a set of assumptions that simplify the problem,

namely that the system is in a *well-stirred* state within a container of constant volume V and additionally, it is at *thermal equilibrium*. As these two assumptions are crucial in allowing the probabilistic modeling of biological processes by converting the position and velocity components of the molecules into independent random variables, it is useful to spell them out in more detail.

Assumption 1. *A well-stirred system is one in which all the molecules are uniformly distributed inside a container H with volume V. If for example, we let P_1 and P_2 denote the positions of two randomly chosen molecules, with a subregion w of the container H having volume ΔV (cf. [Wil06]), we have*

$$\mathbb{P}(P_i \in w) = \frac{\Delta V}{V}, \quad i = 1, 2.$$

Assumption 2. *When a system is in thermal equilibrium, it means that the molecules have a Maxwell-Boltzmann velocity distribution, i.e., for a randomly selected molecule of mass m, the probability that its velocity lies in an infinitesimal region $d^3\mathbf{v}$ about $\mathbf{v}$ is given by $P_{MB}(\mathbf{v})d^3\mathbf{v}$ where*

$$P_{MB}(\mathbf{v}) = \Big(\frac{m^\star}{2\pi K_B T}\Big)^{3/2} \exp(-m^\star v^2/2K_B T),$$

with K_B denoting the Boltzmann constant, $m^\star$ the reduced mass of the two reactant molecules, T the temperature, and we have $\mathbf{v} = (v_x, v_y, v_z)$, $d^3\mathbf{v} = dv_x dv_y dv_z$, $v \equiv ||\mathbf{v}||$ (cf. [Gil92]). A cursory inspection of the expression for $P_{MB}(\mathbf{v})$ reveals that the velocity component is normally distributed with mean 0 and variance $K_B T/m^\star$.

2.2. Derivation of the chemical master equation

As the goal is to determine how the copy numbers of the species $S_1, \ldots, S_d$ evolve as time increases, we formally denote the state of the system by

$$X(t) = [X_1(t), X_2(t), \ldots, X_d(t)] \tag{2.2}$$

and stipulate the initial condition as $X(t_0) = \mathbf{x}_0 \in \mathbb{N}_0^d$ (from here on, the boldface notation $\mathbf{x} \equiv [x_1, \ldots, x_d]$ refers to vectors with d elements). The elements $X_i(t)$ of the state vector (2.2) represent random variables that encode the copy numbers x_i of the species S_i which are present within the container of volume V at time t. Each time one of the M reaction channels R_j fires, the state $X(t)$ changes. Without knowledge of the spatial movements of the molecules, the information required to determine the new state is which R_j reaction fired and when did this event occur. This makes $X(t)$ a stochastic process, as the firing time and the selection of reaction channel are both random events. Thus, the key in solving the problem is to specify the reaction channels R_j in terms of probabilities.

Under the Assumptions 1 and 2, namely that the system is *well-stirred* and at *thermal equilibrium*, it has been rigorously shown in [Gil76, Gil92] that for each reaction channel R_j ($j = 1, \ldots, M$), there exists a function α_j defined such that

$$\begin{aligned} \alpha_j(\mathbf{x})dt \quad = \quad & \text{the probability, given } X(t) = \mathbf{x} \in \mathbb{N}_0^d, \text{ that a randomly} \\ & \text{chosen reaction } R_j \text{ will fire inside the volume } V \text{ within} \\ & \text{the infinitesimal time interval } [t, t + dt), \text{ with } j = 1, \ldots, M, \end{aligned} \tag{2.3}$$

and a vector describing the corresponding state change, with components

$$\mu_i^j \quad = \quad \begin{array}{l}\text{change in the molecular count of species } S_i \text{ triggered} \\ \text{by the firing of reaction } R_j,\ i = 1, \ldots, d \text{ and } j = 1, \ldots, M.\end{array} \tag{2.4}$$

The function α_j is called *propensity function* and the vector μ^j is usually referred to as the *stoichiometric* vector, and together they completely specify the reaction channel R_j.

For example, in the case of the bimolecular reaction R_1 defined in (2.1), the stoichiometric vector encodes the decrease of the molecular counts for species S_1 and S_2 by one molecule, and the corresponding increase in the copy numbers of S_3 by the same number. Therefore, $X(t)$ changes to $X(t) + \mu^1$, with $\mu^1 = [-1,\ -1,\ 1]$, when assuming the whole system contains only three species.

The derivation of the propensity functions is more involved, using probability laws and molecular mechanics arguments and has a solid microphysical foundation. A comprehensive treatment of the subject can be found in, e.g. [Gil92], but for the sake of a self-contained exposition we will review the main ideas.

Generally speaking, the propensity functions have the form

$$\alpha_j(\mathbf{x}) = c_j h_j(\mathbf{x}) \tag{2.5}$$

with c_j being a *specific reaction rate constant*, defined such that $c_j dt$ is the probability that some random combination of suitable R_j reactant molecules will interact in the next infinitesimal time interval $[t, t + dt)$. We shall now take a closer look at the derivation of the two terms on the right hand side of (2.5) for the case of bimolecular reactions.

Let G be the event that a randomly selected pair of molecules collides in the infinitesimal time interval $[t, t + dt)$ and further, let $E_{\mathbf{v}}$ denote the event that the chosen pair has relative speed $\mathbf{v}$. We can use standard probability theory to write

$$\mathbb{P}(G) = \int_{\mathbf{v}} \mathbb{P}(E_{\mathbf{v}})\mathbb{P}(G|E_{\mathbf{v}}) \tag{2.6}$$

where $\mathbb{P}(G|E_{\mathbf{v}})$ is the conditional probability that the randomly selected pair will collide given that it has relative speed $\mathbf{v}$. Owing to Assumption 2, we have that

$$\mathbb{P}(E_{\mathbf{v}}) = P_{MB}(\mathbf{v})d^3\mathbf{v}, \tag{2.7}$$

which we remark is independent of the volume V of the container. For the conditional probability $\mathbb{P}(G|E_{\mathbf{v}})$ we use Assumption 1 and mechanical and structural arguments to stipulate that the volume of a suitable subregion w of the container is given by $\Delta V = (vdt)\pi r^2$, where r is the combined effective radius at which a collision can occur, implying

$$\mathbb{P}(G|E_{\mathbf{v}}) = \frac{\Delta V}{V} = \frac{vdt\pi r^2}{V}. \tag{2.8}$$

The derivation also makes use of the assumption that dt is infinitesimal and the effective radius is small compared with the dimensions of the container, which allows us to ignore interactions with other molecules.

Using (2.6), (2.7) and (2.8) we obtain that

$$\mathbb{P}(G) = \int_{\mathbf{v}} P_{MB}(\mathbf{v}) \frac{v dt \pi r^2}{V} d^3\mathbf{v}$$

and after integration

$$\mathbb{P}(G) = \frac{1}{V}\left(\frac{8K_BT}{\pi m^\star}\right)^{1/2} \pi r^2 dt. \tag{2.9}$$

Equation (2.9) basically means that the probability depends on the radii, the masses of the combined molecules and is inversely proportional to the volume V of the container. As it can not be expected that every collision between suitable reactant molecules leads to a reaction, we also need to compute another conditional probability, namely the *collision-conditioned reaction* probability. This depends on the impact energy of the specific reaction type and is computed as the probability that the energy will exceed a certain barrier $\Delta\varepsilon$. Leaving aside the technical details, it has been shown in [Gil92] that this probability has the exponential "Arrhenius" form $e^{-\frac{\Delta\varepsilon}{K_BT}}$, which does not depend on dt. Using now the probability multiplication law and (2.9), we conclude that the probability of a randomly selected combination of R_j reacting molecules to collide *and* react in the next infinitesimal time interval $[t, t+dt)$ has the form $c_j dt$ with

$$c_j = \frac{1}{V}\left(\frac{8K_BT}{\pi m^\star}\right)^{1/2} \pi r^2 e^{-\frac{\Delta\varepsilon}{K_BT}} \tag{2.10}$$

being independent of dt.

After establishing the formula (2.10) for the computation of the specific probability rate constant c_j, we proceed with the definition of the term $h_j(\mathbf{x})$ from (2.5), in order to complete the characterization of the propensity functions $\alpha_j(\mathbf{x})$. The function $h_j(\mathbf{x})$ is a combinatorial term that measures the number of distinct combinations of R_j reactants when exactly x_i molecules of species S_i are present. As is the case with the stoichiometric vectors μ^j, the functions $h_j(\mathbf{x})$ are based on the structure of the reaction channels themselves. For the bimolecular reaction R_1 used as an example, we have $h_1(\mathbf{x}) = x_1 x_2$. Considering now a generic reaction channel of the form

$$R_j : n_{j,1}S_1 + \cdots + n_{j,d}S_d \longrightarrow m_{j,1}S_1 + \cdots + m_{j,d}S_d, \tag{2.11}$$

with the i-th entry of the corresponding stoichiometric vector μ^j given by

$$\mu_i^j = m_{j,i} - n_{j,i}$$

we have

$$h_j(\mathbf{x}) = \binom{x_1}{x_1 - n_{j,1}} \cdot \ldots \cdot \binom{x_d}{x_d - n_{j,d}} = \prod_{i=1}^{d} \frac{x_i!}{n_{j,i}!(x_i - n_{j,i})!}. \tag{2.12}$$

The sum of all the stoichiometric coefficients on the reactants side of (2.11) denoted by $|s_j| = \sum_{i=1}^{d} n_{j,i}$, specifies the number of reactants and is called reaction order. Usually, in most reaction networks, only zero, first or second order reactions are considered.

Having defined the propensity functions and stoichiometric vectors, we can now use the results to describe how the evolution of $X(t)$ is driven by the reaction channels R_j ($j = 1, \ldots, M$).

In order to accomplish this task, we first have to establish what is the probability that given $X(t) = \mathbf{x}$, *exactly* one reaction of type R_j will occur in the next infinitesimal time interval $[t, t + dt)$, i.e., only one randomly selected pair of suitable molecules has collided and will react accordingly.

From Assumption 1 and (2.3) we know that each of the $h_j(\mathbf{x})$ pairs has probability $c_j dt$ of reacting in $[t, t + dt)$ and $1 - c_j dt$ probability of not reacting in the same interval. Consequently, by multiplying the independent probabilities, we obtain that the probability of a particular combination reacting while the rest will not is equal to

$$c_j dt(1 - c_j dt)^{(h_j(\mathbf{x})-1)} = c_j dt + \mathcal{O}(dt^2) .$$

Because the events involving the collision of molecule combinations are disjoint and exclusive, the probability that one combination will react according to R_j is the sum of the probabilities of the $h_j(\mathbf{x})$ pairs, implying

$$\begin{aligned} \mathbb{P}(\text{ exactly one reaction } R_j \text{ in } [t, t + dt)) &= h_j(\mathbf{x})\big(c_j dt + \mathcal{O}(dt^2)\big) \\ &= c_j h_j(\mathbf{x}) dt + \mathcal{O}(dt^2) . \end{aligned} \tag{2.13}$$

After computing the probability (2.13), we also need the probability that given $X(t) = \mathbf{x}$, *no* reaction channel fires in the infinitesimal time interval $[t, t + dt)$. The knowledge that $1 - c_j dt$ is the probability that a specific pair of molecules does not react according to R_j, and we have $h_j(\mathbf{x})$ combinations yields

$$(1 - c_j dt)^{h_j(\mathbf{x})} = 1 - c_j h_j(\mathbf{x}) dt + \mathcal{O}(dt^2) ,$$

and by taking the sum over all the reaction channels we get

$$\mathbb{P}(\text{ no reaction in } [t, t + dt)) = 1 - \sum_{j=1}^{M} c_j h_j(\mathbf{x}) dt + \mathcal{O}(dt^2) . \tag{2.14}$$

We only need to establish what is the probability that more than one reaction will occur. This is quickly derived from the observation that because the probability of exactly one reaction has the form $c_j dt$, this must be of order $\mathcal{O}(dt^2)$.

In short, under the assumption of a *well-stirred* system at *thermal equilibrium*, we now have available definitions for the probability that

- exactly one reaction R_j fires in $[t, t + dt)$ given by (2.13)
- no reactions fire in $[t, t + dt)$ given by (2.14)
- more than one reaction fires given as $\mathcal{O}(dt^2)$

and can proceed with the description in terms of these probabilities of the evolution of $X(t)$ conditioned on $X(t_0 = 0) = \mathbf{x}_0 \in \mathbb{N}^d$.

Let $\mathbb{P}(\mathbf{x}, t|\mathbf{x}_0, t_0)$ be the conditional probability that $X(t) = \mathbf{x}$, given that at time t_0 the system was in state $\mathbf{x}_0$. The goal is to determine $\mathbb{P}(\mathbf{x}, t + dt|\mathbf{x}_0, t_0)$. Intuitively, when starting in $X(t_0) = \mathbf{x}_0$, the state $X(t + dt) = \mathbf{x}$ can be reached using three mutually exclusive scenarios: either the system has already reached state $X(t) = \mathbf{x}$ and no other reaction will fire in the infinitesimal interval $[t, t+dt)$, or the system has reached a suitable

state $\mathbf{x} - \mu^j$ and will reach state $\mathbf{x}$ at $t + dt$ after exactly one reaction of type R_j fires, or finally, more than one reaction takes place in the time interval, in which case state $\mathbf{x}$ might not be reached.

Using the above argumentation together with (2.13) and (2.14), we conclude that

$$\begin{aligned}\mathbb{P}(\mathbf{x}, t+dt|\mathbf{x}_0, t_0) &= \mathbb{P}(\mathbf{x}, t|\mathbf{x}_0, t_0) \cdot \big(1 - \sum_{j=1}^{M} c_j h_j(\mathbf{x}) dt + \mathcal{O}(dt^2)\big) \\ &+ \sum_{j=1}^{M} \mathbb{P}(\mathbf{x} - \mu^j, t|\mathbf{x}_0, t_0) \cdot \big(c_j h_j(\mathbf{x} - \mu^j) dt + \mathcal{O}(dt^2)\big) \\ &+ \mathcal{O}(dt^2)\,.\end{aligned} \tag{2.15}$$

Subtracting $\mathbb{P}(\mathbf{x}, t|\mathbf{x}_0, t_0)$ from both sides of (2.15), using (2.5), dividing by dt and passing to the limit $dt \to 0$, finally leads to

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{P}(\mathbf{x}, t|\mathbf{x}_0, t_0) &= \sum_{j=1}^{M} \alpha_j(\mathbf{x} - \mu^j)\mathbb{P}(\mathbf{x} - \mu^j, t|\mathbf{x}_0, t_0) \\ &- \sum_{j=1}^{M} \alpha_j(\mathbf{x})\mathbb{P}(\mathbf{x}, t|\mathbf{x}_0, t_0)\end{aligned} \tag{2.16}$$

which is the *Chemical Master equation* (CME). This is a difference-differential equation that describes the probability flow responsible for creating and destroying any given state of the system under the condition of starting in state $\mathbf{x}_0$. The first term accounts for inflow into state $\mathbf{x}$ from neighboring states, while the second term represents the outflow from state $\mathbf{x}$.

At this stage, we must remark that because (2.16) is an exact consequence of the characterization of reaction channels purported by (2.3), which itself is grounded in sound microphysical arguments, solving the CME delivers the full picture of the dynamics of the process $X(t)$.

2.3. Stochastic simulation algorithm

Typically, solving the CME is not a trivial task, even if we assume a reduced state space, and regard (2.16) as a system of ODEs, one for each state. For example, a rather small system consisting of only three species where we limit the copy numbers to a maximum of 100 molecules per species, will contain 100^3 states and hence lead to 10^6 ODEs that have to be solved in order to compute the solution. Hence, most of the attempts have concentrated on Monte Carlo simulations using the *stochastic simulation algorithm* (SSA) proposed by Gillespie in his seminal paper [Gil76]. The SSA also uses the characterization of reaction channels given in (2.3), but the key aspect is that it circumvents computing the probability distribution, computing rather single realizations of the state vector $X(t)$. Assuming that the initial value $X(0) = \mathbf{x}_0$ is given, one time step of a naïve version of

such an algorithm would involve the following steps:

Algorithm 1: Naïve version of stochastic simulation
Step 1: Find the random time $t + dt$ at which the next reaction event will take place
Step 2: Determine the random index j of reaction channel R_j that will fire
Step 3: Update the value of $X(t + dt) = X(t) + \mu^j$ and the time $t = t + dt$

For an actual implementation however, we would need a way to sample the time to the next reaction and the appropriate reaction channel index from the underlying probability distribution. Another direct consequence of (2.3) is the existence of a function $p(\tau, j|\mathbf{x}, t)$ which is defined by

$$\begin{aligned} p(\tau, j|\mathbf{x}, t)d\tau &= \text{probability that the next reaction channel that} \\ &\quad \text{will fire in the infinitesimal interval } [t+\tau, t+\tau+dt) \\ &\quad \text{will be of } R_j \text{ type, } j = 1, \ldots, M. \end{aligned} \tag{2.17}$$

Using the same arguments as for the derivation of the CME, this probability is equal to the product between the probability of no reaction in the interval $[t, t+\tau)$ and the probability of the j-th reaction channel firing in the remaining time interval $[t+\tau, t+\tau+d\tau)$, quantities that have been defined in (2.13) and (2.14), respectively. By denoting now the probability of no reactions occurring in $[t, t+\tau)$ by $\mathbb{P}_0(\tau|\mathbf{x}, t)$ we can write

$$p(\tau, j|\mathbf{x}, t)d\tau = \mathbb{P}_0(\tau|\mathbf{x}, t) \cdot c_j h_j(\mathbf{x})d\tau + \mathcal{O}(\tau^2)\,. \tag{2.18}$$

For the purpose of deriving an explicit formula for the quantity $\mathbb{P}_0(\tau|\mathbf{x}, t)$, we divide the interval $[t, t+\tau)$ into N disjoint intervals with length $\varepsilon = \frac{\tau}{N}$ so that we have

$$[t+\tau) = \bigcup_{k=0}^{N-1} \Big[t + k\frac{\tau}{N}, t + (k+1)\frac{\tau}{N}\Big).$$

Next, using (2.14) and the multiplication of independent probabilities, we have

$$\mathbb{P}_0(\tau|\mathbf{x}, t) = \mathbb{P}_0(\varepsilon|\mathbf{x}, t)^N = \left(1 - \sum_{j=1}^{M} c_j h_j(\mathbf{x})\frac{\tau}{N} + \mathcal{O}\left(\frac{\tau^2}{N^2}\right)\right)^N.$$

Further, letting the number of subintervals N go to infinity and using the limit definition of the exponential function

$$e^{-\lambda} = \lim_{N\to\infty}\Big(1 - \frac{\lambda}{N}\Big)^N$$

yields that

$$\begin{aligned} \mathbb{P}_0(\tau|\mathbf{x}, t) &= \lim_{N\to\infty}\left(1 - \sum_{j=1}^{M} c_j h_j(\mathbf{x})\frac{\tau}{N} + \mathcal{O}\left(\frac{\tau^2}{N^2}\right)\right)^N \\ &= \exp\Big(-\sum_{j=1}^{M} c_j h_j(\mathbf{x})\tau\Big). \end{aligned} \tag{2.19}$$

By using (2.19) in (2.18), dividing by $d\tau$ and letting $d\tau \to 0$, we finally arrive at an explicit formula for the function

$$\begin{aligned} p(\tau, j|\mathbf{x}, t) &= c_j h_j(\mathbf{x}) \exp\Big(-\sum_{j=1}^{M} c_j h_j(\mathbf{x})\tau \Big) \\ &= \alpha_j(\mathbf{x}) \exp\Big(-\sum_{j=1}^{M} \alpha_j(\mathbf{x})\tau \Big), \end{aligned} \tag{2.20}$$

which represents a probability density. By denoting the sum of the propensities as

$$\gamma(\mathbf{x}) = \sum_{j=1}^{M} \alpha_j(\mathbf{x}), \tag{2.21}$$

we have

$$\int_0^\infty \sum_{j=1}^{M} p(\tau, j|\mathbf{x}, t) d\tau = \int_0^\infty \gamma(\mathbf{x}) \exp(-\gamma(\mathbf{x})\tau) d\tau = 1.$$

After computing the probability density $p(\tau, j|\mathbf{x}, t)$ we can revisit Algorithm 1 proposed earlier and qualify the first two steps as the process of generating two random numbers τ and j according to this joint probability density. Next, we set up the computational procedure for drawing the two random numbers. For this purpose, one can use *Bayes' formula*,

$$p(\tau, j|\mathbf{x}, t) = p_1(\tau|\mathbf{x}, t) \cdot p_2(j|\tau, \mathbf{x}, t) \tag{2.22}$$

to write $p(\tau, j|\mathbf{x}, t)$ as the product of two individual density functions. The first one, $p_1(\tau|\mathbf{x}, t)$ is computed by summing $p(\tau, j|\mathbf{x}, t)$ over the j random variable

$$\begin{aligned} p_1(\tau|\mathbf{x}, t) &= \sum_{j=1}^{M} p(\tau, j|\mathbf{x}, t) \\ &= \sum_{j=1}^{M} \alpha_j(\mathbf{x}) \exp\Big(-\sum_{j=1}^{M} \alpha_j(\mathbf{x})\tau \Big) \\ &= \gamma(\mathbf{x}) \exp(-\gamma(\mathbf{x})\tau), \end{aligned} \tag{2.23}$$

while for the second density $p_2(j|\tau, \mathbf{x}, t)$, using (2.22) and (2.23) we obtain

$$p_2(j|\tau, \mathbf{x}, t) = \frac{p(\tau, j|\mathbf{x}, t)}{p_1(\tau|\mathbf{x}, t)} = \frac{\alpha_j(\mathbf{x}) \cdot \exp(-\gamma(\mathbf{x})\tau)}{\gamma(\mathbf{x}) \cdot \exp(-\gamma(\mathbf{x})\tau)} = \frac{\alpha_j(\mathbf{x})}{\gamma(\mathbf{x})}. \tag{2.24}$$

Having established the two individual density functions (2.23) and (2.24), drawing the random numbers necessary can be accomplished in practice by using the inversion generating method which is based on the observation that the cumulative density function ranges uniformly over the interval $(0, 1)$.

First, let us remark that $p_1(\tau|\mathbf{x}, t) = \gamma(\mathbf{x}) \exp(-\gamma(\mathbf{x})\tau)$ is the density function of a random variable with the well-known exponential distribution. For continuous distributions we have that the cumulative density function is given as

$$F(x) = \int_{-\infty}^{x} P(y) dy$$

where $P(y)$ denotes a density function. Then, if u is a random drawn number from the uniform distribution $(0,1)$, by using $r = F^{-1}(u)$ we can generate a random number r from a continuous distribution with the specified density function. Choosing $p_1(\tau|\mathbf{x},t)$ as the density function we have

$$r_1 = \int_0^\tau p_1(\tau|\mathbf{x},t)d\tau = 1 - \exp(-\gamma(\mathbf{x})\tau). \tag{2.25}$$

Solving the equation for τ and replacing r_1 with the statistically equivalent random variable $1 - r_1$ we obtain that τ should be selected according to

$$\tau = \frac{1}{\gamma(\mathbf{x})} \ln\left(\frac{1}{r_1}\right).$$

The inversion method can also be used for discrete distributions. In such cases, the cumulative density function is related to the probability density function $P(y)$ by the formula

$$F(x) = \sum_{y \le x} P(y).$$

To generate a random number j_k according to the discrete density function $p_2(j|\tau,\mathbf{x},t)$, with $j_1 < ... < j_M$, a random number r_2 can be drawn from the uniform distribution such that

$$F(j-1) < r_2 \le F(j).$$

Using the formula obtained for the density function $p_2(j|\tau,\mathbf{x},t)$ in (2.24), we obtain

$$\sum_{k=1}^{j-1} \alpha_k(\mathbf{x}) < r_2\gamma(\mathbf{x}) \le \sum_{k=1}^{j} \alpha_k(\mathbf{x}).$$

Thus, we arrive at the "direct method" version of the *stochastic simulation algorithm* [Gil76], outlined below:

Algorithm 2: Gillespie's *direct method* (SSA)

0. Initialization: Set $t_0 = 0$ and fix initial value $X(t_0) = \mathbf{x}_0$

while $t < T_{final}$ **do**

1. Compute all propensities $\alpha_j(\mathbf{x})$ and their sum $\gamma(\mathbf{x}) = \sum_{j=1}^{M} \alpha_j(\mathbf{x})$

2. Draw random numbers r_1 and r_2 from the uniform distribution $(0,1)$

3. Let $\tau = \frac{1}{\gamma(\mathbf{x})} \ln\left(\frac{1}{r_1}\right)$

4. Determine the index j such that the inequality

$$\sum_{k=1}^{j-1} \alpha_k(\mathbf{x}) < r_2\gamma(\mathbf{x}) \le \sum_{k=1}^{j} \alpha_k(\mathbf{x}).$$

holds.

5. Update the state vector $X(t+\tau) = X(t) + \mu^j$ and let $t = t + \tau$

end

2.4. Markov jump process and the CME

In the last two sections, we have presented a derivation of the CME endorsed by microphysical arguments, and a computational procedure to simulate the dynamics induced by the underlying stochastic process, one realization at a time. However, as a rigorous definition of a stochastic process was not given, and moreover, another starting point for the derivation of the CME is provided by the theory of stochastic processes, we are motivated in taking a second look from this more abstract perspective.

2.4.1. Probability theory and stochastic processes

First, let us quickly compile the relevant theoretical tools from probability and stochastic process theory by adapting some definitions from [PS08, Chapter 3].

A *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ is defined as a triple composed of a sample space of outcomes $\Omega = \{w_1, w_2, \dots, \}$, a σ-algebra $\mathcal{F}$ over the subsets of Ω and a probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$, which satisfies the requirements $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$ and

$$\mathbb{P}(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \mathbb{P}(A_k)$$

for all sequences of pairwise disjoint sets $\{A_k\}_{k=1}^{\infty} \in \mathcal{F}$. Further, let $S \neq \emptyset$ be a finite or countable state set and $\mathcal{G}$ a σ-algebra over S, which together define a measurable space $(S, \mathcal{G})$.

Then, a *random variable* $X = X(w)$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ can be defined as a mapping

$$X : (\Omega, \mathcal{F}) \to (S, \mathcal{G})$$

between a *sample space* $(\Omega, \mathcal{F})$ and a *state space* $(S, \mathcal{G})$, both measurable, with the property that the events $\{w \in \Omega : X(w) \in A\} \in \mathcal{F}$ for any $A \in \mathcal{G}$. The *expectation* of the random variable X is defined by

$$\mathbb{E}X = \int_{\Omega} X(w) d\mathbb{P}(w)$$

as the weighted sum over all the possible outcomes that the random variable can take. Next, let $\mathcal{B}(U)$ denote the *Borel* σ-algebra of a topological space set U, in other words, the smallest σ-algebra containing all the open sets of U. Every random variable

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \to (S, \mathcal{B}(S))$$

induces then a probability measure on S,

$$\mathbb{P}_X(B) = \mathbb{P}X^{-1}(B) = \mathbb{P}(w \in \Omega; X(w) \in B), \quad B \in \mathcal{B}(S)$$

and we call P_X the *distribution* of X. For the case of $S = \mathbb{R}^d$, we can write

$$d\mathbb{P}_X(x) = p(x)dx$$

and refer to $p(x)$ as the *probability density function*.

We are now ready to define a *stochastic process* as a collection of random variables $X := \{X(t, w), w \in \Omega, t \in T\}$ with $T = \{t_0 \leq t_1 \leq \dots .\}$ an ordered set of time points. Fixing $w \in \Omega$ we obtain a realization or trajectory $X(t)$ of the process X, and by fixing t we get a random variable $X(w)$.

2.4.2. The Markov property

Speaking now in looser terms, we can think about a stochastic process as a system which evolves probabilistically in time, i.e., in which a certain time-dependent random variable exists. We can then measure its values $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots\}$ at certain times $\{t_0 \leq t_1 \leq \ldots. \leq t_n \leq \ldots.\}$ and assume that a joint probability density

$$p(\ldots; \mathbf{x}_n, t_n; \mathbf{x}_{n-1}, t_{n-1}; \ldots; \mathbf{x}_0, t_0) \tag{2.26}$$

exists, which describes the dynamics of the system completely [Gar09].

Next, we can use (2.26) to define the conditional probability density

$$p(\ldots; \mathbf{x}_n, t_n; \ldots; \mathbf{x}_{j+1}, t_{j+1} \mid \mathbf{x}_j, t_j; \ldots; \mathbf{x}_0, t_0) = \frac{p(\ldots; \mathbf{x}_n, t_n; \mathbf{x}_{n-1}, t_{n-1}; \ldots; \mathbf{x}_0, t_0)}{p(\mathbf{x}_j, t_j; \ldots; \mathbf{x}_0, t_0)} \tag{2.27}$$

with $0 \leq j < n$.

If all such conditional probabilities (2.27) would be available, this would also lead to a complete description of the dynamics. However, such a description would require a complete history of the system and thus be too complex. An effective idea to reduce the complexity is the *Markov assumption*. This stipulates that the conditional probability is entirely determined by the current state and not by the past, i.e.,

$$p(\mathbf{x}_n, t_n | \mathbf{x}_{n-1}, t_{n-1}, \ldots; \mathbf{x}_0, t_0) = p(\mathbf{x}_n, t_n | \mathbf{x}_{n-1}, t_{n-1}) \tag{2.28}$$

which is the *Markov property* (notice that in (2.28) we have used a finite set of measurements to simplify the notation). The *Markov property* has the important consequence that we can now express the joint probability density (2.26) in terms of simple conditional probabilities

$$\begin{aligned} p(\mathbf{x}_n, t_n; \ldots; \mathbf{x}_0, t_0) &= p(\mathbf{x}_n, t_n | \mathbf{x}_{n-1}, t_{n-1}) \cdot p(\mathbf{x}_{n-1}, t_{n-1} | \mathbf{x}_{n-2}, t_{n-2}) \cdot \ldots \\ &\quad \ldots. \cdot p(\mathbf{x}_1, t_1 | \mathbf{x}_0, t_0) \cdot p(\mathbf{x}_0, t_0) \end{aligned} \tag{2.29}$$

which means that any future state can be described given only an initial condition and the simple transition probability densities $p(\mathbf{x}_j, t_j | \mathbf{x}_{j-1}, t_{j-1})$, $1 \leq j \leq n$, thus simplifying the treatment of processes that exhibit property (2.28). Such processes are called *Markov processes* and are in effect *memoryless* because the future development of the process depends only on the current state and not on any of the past states.

The *Markov property* also has another important consequence. Starting from the addition law of probability for mutually exclusive events, and by eliminating one of the variables from the joint probability density by taking the sum over that variable, we have

$$p(\mathbf{x}_2, t_2 | \mathbf{x}_0, t_0) = \int p(\mathbf{x}_2, t_2; \mathbf{x}_1, t_1 | \mathbf{x}_0, t_0) d\mathbf{x}_1 \tag{2.30}$$

for three measurements taken at $t_0 \leq t_1 \leq t_2$. Using the definition (2.27) of the conditional probability density and the *Markov property* (2.28) we can write (2.30) as

$$\begin{aligned} p(\mathbf{x}_2, t_2 | \mathbf{x}_0, t_0) &= \int p(\mathbf{x}_2, t_2; \mathbf{x}_1, t_1 | \mathbf{x}_0, t_0) d\mathbf{x}_1 \\ &= \int p(\mathbf{x}_2, t_2 | \mathbf{x}_1, t_1; \mathbf{x}_0, t_0) \cdot p(\mathbf{x}_1, t_1 | \mathbf{x}_0, t_0) d\mathbf{x}_1 \\ &= \int p(\mathbf{x}_2, t_2 | \mathbf{x}_1, t_1) \cdot p(\mathbf{x}_1, t_1 | \mathbf{x}_0, t_0) d\mathbf{x}_1 \end{aligned} \tag{2.31}$$

which is the *Chapman-Kolmogorov* equation (cf. [Gar09]). In the case of discrete variables that take only integer values, the *Chapman-Kolmogorov* equation for discrete state spaces reads

$$\mathbb{P}(X(t_2) = \mathbf{x}_2 | X(t_0) = \mathbf{x}_0) = \sum_{\mathbf{x}_1} \mathbb{P}(X(t_2) = \mathbf{x}_2 | X(t_1) = \mathbf{x}_1) \cdot \mathbb{P}(X(t_1) = \mathbf{x}_1 | X(t_0) = \mathbf{x}_0). \tag{2.32}$$

Of course, before using consequence (2.29), the question is raised whether any natural process exists that actually observes the *Markov property* (2.28) exactly. If we assume a very fine time scale for observations, the answer is negative, because at the very least we would need the immediate history to predict the probabilistic future. Fortunately however, processes that have a relative short memory, meaning that their memory time is far smaller than the timescale used in recording the measurements, are common. Thus, it is reasonable to assume that a Markov process approximates such systems with sufficient accuracy and the popularity of Markovian models in many fields of science is evidence of this fact.

Another aspect of the current discussion about stochastic processes is whether the state space is discrete or continuous and whether the time evolution proceeds in a discrete or continuous way. Considering that the dynamics of biological processes evolve continuously in time and according to the arguments brought forward in Chapter 1, the quantities of interest take integer values, the focus in our case is predictably on the continuous-time Markov process with a discrete state space. In case the state space is finite or countable, and the time evolution discrete, the term Markov *chain* is sometimes employed. Without loss of generality we shall take the finite state space to be $S = \{1, \dots, N\} \subset \mathbb{N}$. Let us now present the construction of a continuous-time Markov process.

2.4.3. Continuous-time Markov process

The starting point for the construction of the continuous-time object is a discrete-time Markov *chain* which we proceed to define as in [PS08, Chapter 3].

Definition 2.1. *A random sequence $\{X_n\}_{n\geq 0}$ is a discrete-time Markov chain with initial distribution ρ_0 and transition matrix P, if it is a stochastic Markov process on the finite state space S with initial distribution ρ_0 (viewed as a column vector),*

$$(\rho_0)_i = \mathbb{P}(X_0 = i), i \in S$$

and transition probability from state i to state j given as

$$p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i), \quad i, j \in S,$$

for every $n \geq 0$ and $\mathbb{P}(X_n = i) > 0$.

If the transition probabilities are independent of n, then the process is said to be *homogeneous*. The transition probabilities $\{p_{ij}\}_{i,j\in S}$ can be assembled into a transition matrix $P \in \mathbb{R}^{N\times N}$, which satisfies

$$0 \leq p_{ij} \leq 1, \quad \forall\, i, j \in S \tag{2.33}$$

$$\sum_{j\in S} p_{ij} = 1. \tag{2.34}$$

Any matrix that satisfies the above conditions (2.33) and (2.34) is called a *stochastic* matrix.

Further, using the *Chapman-Kolmogorov* equation for discrete state spaces (2.32) and induction on n, it can be shown that the n-step transition probability from state i to state j, denoted by $p_{ij}^n = \mathbb{P}(X_n = j|X_0 = i)$ is equal to $(P^n)_{ij}$, and computing the probability that the Markov chain will be in state j at $n \geq 0$ will reduce to computing the corresponding power of the transition matrix. Consequently, for an initial distribution ρ_0 we have

$$\mathbb{P}(X_n = j) = \sum_{i\in S} \mathbb{P}(X_n = j|X_0 = i) \cdot \mathbb{P}(X_0 = i) = \sum_{i\in S} (\rho_0)_i (P^n)_{ij} = (\rho_0 P^n)_j. \tag{2.35}$$

Thus, if we know the initial distribution and the transition matrix we can determine the probability distribution at any later time point. Moreover, by using the notation introduced in Definition 2.1 for transition probabilities, we can write the general form of the *Chapman-Kolmogorov* equation as

$$p_{ij}^{(m+n)} = \sum_{k\in S} p_{ik}^{(m)} p_{kj}^{(n)} \tag{2.36}$$

which leads to

$$P^{m+n} = P^m P^n.$$

We turn now to the task of defining a continuous-time Markov process $\{X(t)\}_{t\in\mathbb{R}}$ with the same finite state space S as the discrete-time chain. In addition to observing the Markov property (2.28), we also want the process to be *time-homogenous*, i.e. to fulfill

$$\mathbb{P}(X(t) = j \mid X(s) = i) = \mathbb{P}(X(t-s) = j \mid X(0) = i) \tag{2.37}$$

for any states $i, j \in S$ and $s \leq t$. Intuitively, the main difference to the discrete-time setting discussed previously is that transitions can now occur at any time, so we need to establish how long the process will remain in a state $i \in S$ before performing a jump to a new state $j \in S$.

Let T_i denote the waiting time to the next jump while in state i. It can be shown by making use of the Markov property and the time-homogeneity requirement (2.37) that

$$\mathbb{P}(T_i > s + t \mid T_i > s) = \mathbb{P}(T_i > t). \tag{2.38}$$

Thus, T_i satisfies the memoryless requirement, as (2.38) basically says that the system forgets it has already waited for time s. This leads to the conclusion that T_i is exponentially distributed with a parameter $w(i)$, as the exponential distribution is the only continuous-time distribution that observes the Markov property (cf. [Ste09, Chapter 9.10]).

We proceed now to study the transition probabilities. First, as $T_i \sim \exp(w(i))$ and satisfies (2.38), we infer that

$$\mathbb{P}(T_i < dt) = 1 - e^{-w(i)dt} = w(i)dt + \mathcal{O}\big(dt^2\big)$$

when $dt \to 0$. Next, using the notation from Definition 2.1, we write the probability that the process will jump to state j after leaving state i as

$$p_{ij} = \mathbb{P}(X(T_i) = j \mid X(0) = i).$$

The transition probability does not depend on the time spent by the process in i, because if it would do so, the Markov property will no longer be observed. By defining

$$w(i,j) = w(i) \cdot p_{ij} \tag{2.39}$$

as the transition intensity from state i to state j, we can write

$$\begin{aligned} \mathbb{P}(X(t+dt) = j \mid X(t) = i) &= \mathbb{P}(X(dt) = j \mid X(0) = i) \\ &= \mathbb{P}(T_i < dt, X(T_i) = j \mid X(0) = i) \\ &= w(i) \cdot p_{ij} dt + \mathcal{O}(dt^2) \\ &= w(i,j)dt + \mathcal{O}(dt^2) \end{aligned} \tag{2.40}$$

with $\mathcal{O}(dt^2)$ accounting for the probability of more than one jump in the interval $[t, t+dt)$. Because of the way we have defined the transition intensities (2.39), we also have for $i \in S$

$$\sum_{j \neq i} w(i,j) = \sum_{j \neq i} w(i) \cdot p_{ij} = w(i) \sum_{j \neq i} p_{ij} = w(i). \tag{2.41}$$

Taking (2.41) into account, we can now write the probability that no jump will take place in $[t, t+dt)$ as

$$\begin{aligned} \mathbb{P}(X(t+dt) = i \mid X(t) = i) &= \mathbb{P}(X(dt) = i \mid X(0) = i) \\ &= 1 - \sum_{j \neq i} \mathbb{P}(X(dt) = j \mid X(0) = i) \\ &= 1 - \sum_{j \neq i} w(i,j)dt + \mathcal{O}(dt^2) \\ &= 1 - w(i)dt + \mathcal{O}(dt^2)\,. \end{aligned} \tag{2.42}$$

Using (2.40) and (2.42) we are now ready to give a definition for a time-homogeneous continuous time Markov process with a finite state space S.

Definition 2.2. *A stochastic process $\{X(t)\}_{t \in \mathbb{R}}$ with a finite state space S is a time-homogeneous continuous time Markov process, if it satisfies*

$$\begin{aligned} \mathbb{P}(X(t+dt) = j \mid X(t) = i) &= w(i,j)dt + \mathcal{O}(dt^2) \\ \mathbb{P}(X(t+dt) = i \mid X(t) = i) &= 1 - w(i)dt + \mathcal{O}(dt^2) \end{aligned}$$

where $j \neq i$ and $w(i)$ is given by (2.41).

A classic (and arguably one of the most important) example of a continuous-time Markov process is the *Poisson process*, which is an integer valued counting process $N(t)$ of the number of jumps in the time interval $[0, t]$. The *Poisson process* satisfies

$$\begin{aligned} \mathbb{P}(N(t+dt) = i+1 \mid N(t) = i) &= wdt + \mathcal{O}(dt^2) \\ \mathbb{P}(N(t+dt) = i \mid N(t) = i) &= 1 - wdt + \mathcal{O}(dt^2) \end{aligned}$$

with $w > 0$ denoting the constant intensity of the process, which no longer depends on the state. Moreover, we have that the independent increments are exponentially distributed,

$$\mathbb{P}(N(t) - N(s) = k) = \frac{e^{-w(t-s)} \left(w(t-s)\right)^k}{k!},$$

and depend only on $t - s$ making the Poisson process *time-homogeneous*.

After these preparations, a recipe for the construction of a continuous-time Markov process $\{X(t)\}_{t\in\mathbb{R}}$ can be formulated (see also [PS08, Chapter 5]). The procedure involves two objects, the first ingredient being an independent and identically distributed sequence $\{\tau_n\}_{n\geq 0} \sim \exp(w)$ that will provide the transition times, with the second component represented by a discrete-time Markov *chain* $\{X_n\}_{n\geq 0}$ with transition matrix P defined as in (2.33 - 2.34), which provides the values for the states. We remark that $\{X_n\}_{n\geq 0}$ is sometimes called the *embedded* chain of the stochastic process $\{X(t)\}_{t\in\mathbb{R}}$. From an algorithmic viewpoint, first we set $X(0) = X_0$ and $t_0 = 0$ and let $t_{n+1} = t_n + \tau_n$ be the next jump time. Next, we define $X(t) = X_n$ for any $t \in [t_n, t_{n+1}) \, , \forall n \geq 0$. The process $X(t)$ thus obtained is called *Markov jump process*, and we note that the algorithm lightly sketched above is another formulation of the SSA algorithm presented in Section 2.3.

Next, we present a matrix characterization for the continuous-time Markov process. Similarly to the discrete case, we can assemble the transition probabilities of a Markov jump process into a matrix $P(t)$ with elements

$$p_{ij}(t) = \mathbb{P}(X(t) = j \mid X(0) = i). \tag{2.43}$$

Due to the exponential distribution of the jump times, we also have

$$\mathbb{P}(N(t) = k) = \frac{e^{-wt}\,(wt)^k}{k!}. \tag{2.44}$$

Combining the probability given in (2.44) with the k-step transition matrix of the embedded Markov chain leads to

$$p_{ij}(t) = \mathbb{P}(N(t) = k) \cdot \mathbb{P}(X_k = j \,|X_0 = i) = \sum_{k=0}^{\infty} \frac{e^{-wt}(wt)^k}{k!}\,(P^k)_{ij}.$$

Hence, in matrix form we have

$$P(t) = e^{-wt} \sum_{k=0}^{\infty} \frac{(wt)^k}{k!} P^k = e^{tw(P-I)} = e^{tL} \tag{2.45}$$

with $L = w(P - I)$ called the *generator* of the continuous-time Markov jump process. We remark that in case the state space is infinite, handling e^{tL} requires the operator theory of semigroups [PS08, Chapter 7.5]. Thus, given an intensity w and the transition matrix P of the *embedded* chain we can characterize the *Markov jump process*. Additionally, the generator L satisfies

$$L = \lim_{t\to 0} \frac{P(t) - I}{t} \tag{2.46}$$

and because P is a stochastic matrix, we have

$$\sum_{j\in S} l_{ij} = 0 \;\; \forall i \in S, \tag{2.47}$$

$$l_{ij} \in [0, \infty) \;\; \forall i, j \in S \text{ with } i \neq j \tag{2.48}$$

$$\text{and } l_{ii} \leq 0. \tag{2.49}$$

Summarizing (2.47), (2.48) and (2.49), the rows of L must sum up to zero, the off-diagonal elements are non-negative, while the diagonal elements are non-positive.

We are now finally in a position to bring the spotlight on the relationship between the time-continuous Markov chain, its generator and the CME derived in (2.16). As we have seen, the generator is built using the stochastic matrix $P(t)$ with elements defined by (2.43). The intention is to derive a set of differential equations that describe the evolution of the transition probabilities, or in other words a master equation. Hence, we begin by taking the time derivative

$$\begin{aligned}\frac{d}{dt}p_{ij}(t) &= \lim_{dt\to 0}\frac{p_{ij}(t+dt)-p_{ij}(t)}{dt} \\ &= \lim_{dt\to 0}\frac{1}{dt}\Big(\mathbb{P}(X(t+dt)=j \mid X(0)=i)-\mathbb{P}(X(t)=j \mid X(0)=i)\Big).\end{aligned} \tag{2.50}$$

Using now the *Chapman-Kolmogorov* equation (2.36), we introduce a new variable y in (2.50) and write

$$\begin{aligned}\frac{d}{dt}p_{ij}(t) = \lim_{dt\to 0}\frac{1}{dt}\Big(&\sum_{y\in S}\mathbb{P}(X(t+dt)=j \mid X(t)=y, X(0)=i) \\ &\cdot\mathbb{P}(X(t)=y \mid X(0)=i) \\ &-\mathbb{P}(X(t)=j \mid X(0)=i)\Big).\end{aligned} \tag{2.51}$$

Further, using (2.40) and (2.42) to expand the first term in (2.51), we have that

$$\begin{aligned}&\sum_{y\in S}\mathbb{P}(X(t+dt)=j \mid X(t)=y, X(0)=i)\cdot\mathbb{P}(X(t)=y \mid X(0)=i) \\ &\quad=\mathbb{P}(X(t+dt)=j \mid X(t)=j, X(0)=i)\cdot\mathbb{P}(X(t)=j \mid X(0)=i) \\ &\qquad+\sum_{y\neq j}\mathbb{P}(X(t+dt)=j \mid X(t)=y, X(0)=i)\cdot\mathbb{P}(X(t)=y \mid X(0)=i) \\ &\quad=\mathbb{P}(X(t+dt)=j \mid X(t)=j)\cdot\mathbb{P}(X(t)=j \mid X(0)=i) \\ &\qquad+\sum_{y\neq j}\mathbb{P}(X(t+dt)=j \mid X(t)=y)\cdot\mathbb{P}(X(t)=y \mid X(0)=i) \\ &\quad=\big(1-w(j)dt\big)p_{ij}(t)+\sum_{y\neq j}w(y,j)dt\cdot p_{iy}(t)+\mathcal{O}\big(dt^2\big)\,.\end{aligned} \tag{2.52}$$

Inserting (2.52) into (2.51), rearranging the terms and passing to the limit, yields via (2.41)

$$\begin{aligned}\frac{d}{dt}p_{ij}(t) &= \lim_{dt\to 0}\frac{1}{dt}\Big(\big(1-w(j)dt\big)p_{ij}(t)-p_{ij}(t) \\ &\qquad+\sum_{y\neq j}w(y,j)dt\cdot p_{iy}(t)+\mathcal{O}\big(dt^2\big)\Big) \\ &= -w(j)p_{ij}(t)+\sum_{y\neq j}w(y,j)p_{iy}(t) \\ &= \sum_{y\neq j}w(y,j)p_{iy}(t)-\Big(\sum_{y\neq j}w(j,y)\Big)p_{ij}(t)\end{aligned} \tag{2.53}$$

which are the *forward Kolmogorov equations* for the process $X(t)$. Comparing (2.53) with (2.16), we observe that the chemical master equation is a special case of the *forward Kolmogorov equation*, with the inflow and outflow terms readily recognizable.

Equation (2.53) can also be written in matrix form, by defining the matrix L as

$$L_{ij} = \begin{cases} -w(j), & \text{if } i = j \\ w(i,j), & \text{if } i \neq j. \end{cases} \tag{2.54}$$

Thus, we obtain

$$\frac{d}{dt}P(t) = P(t)L. \tag{2.55}$$

When the state space S is finite, and subject to the initial condition $P(0) = I$, equation (2.55) has the formal solution $P(t) = e^{tL}$. Comparing with (2.45), it is clear that by defining L as in (2.54), we have recovered the generator of the Markov jump process.

Besides the forward Kolmogorov equations, we can also obtain another set of differential equations called the *backward Kolmogorov* equations. Using again *Chapman-Kolmogorov* equations (2.31) to expand a transition matrix $Q(t+dt)$ this time as $Q(dt)Q(t)$ and taking the time derivative of $Q(t)$ at $t = 0$, we have

$$\begin{aligned} \frac{d}{dt}Q(t) &= \lim_{dt\to 0} \frac{Q(t+dt) - Q(t)}{dt} \\ &= \lim_{dt\to 0} \frac{Q(dt)Q(t) - Q(t)}{dt} \\ &= \lim_{dt\to 0} \frac{Q(dt) - I}{dt} Q(t) \\ &= L^{\star} Q(t). \end{aligned} \tag{2.56}$$

We conclude now this section by referring the readers interested in a more extensive treatment of stochastic processes to the monographs [CM65, Nor97]. For a viewpoint closer to the chemical master equation, [vK01, Gar09] are recommended.

2.5. The CME operator

Both guises of the CME introduced so far, either the form (2.16) derived from microphysical arguments, or (2.53) obtained using the *Chapman-Kolmogorov* equation, are somewhat unwieldy when it comes to presenting numerical methods, which is the aim of this thesis. Thus, the goal of this section is to introduce the CME operator, a more useful notation in the context of computing numerical approximations.

Recall that the goal is to compute the probability distribution

$$p(t, \mathbf{x}) = \mathbb{P}(X(t) = \mathbf{x} \mid X(0) = \mathbf{x}_0) \;\text{ with } \mathbf{x}, \mathbf{x}_0 \in \mathbb{N}_0^d, \tag{2.57}$$

i.e., the probability that at time t there are exactly x_i particles of the i-th species, given the initial copy numbers $X(0) = \mathbf{x}_0$. Rewriting (2.16) using (2.57), we get

$$\partial_t p(t, \mathbf{x}) = \sum_{j=1}^{M} \Big(\alpha_j(\mathbf{x} - \mu^j) p(t, \mathbf{x} - \mu^j) - \alpha_j(\mathbf{x}) p(t, \mathbf{x}) \Big). \tag{2.58}$$

The propensities α_j and stoichiometric vectors μ^j characterize the reaction channels R_j ($j = 1, \ldots, M$) completely, and it is reasonable to assume that only *feasible* reactions are

allowed when modeling biological processes. Therefore, as the inflow term $\mathbf{x}-\mu^j$ in (2.58) may have negative entries and this does not correspond to physical reality, we stipulate

$$\alpha_j(\mathbf{x}) = 0 \quad \text{and} \quad p(t,\mathbf{x}) = 0, \ \ \forall\, \mathbf{x} \notin \mathbb{N}_0^d. \tag{2.59}$$

Further, using the notation introduced in (2.57), the initial condition now reads

$$p(0,\mathbf{x}) = p_0(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{x}_0 \\ 0, & \text{else.} \end{cases} \tag{2.60}$$

We have shown in (2.55) that for a finite state space, the CME can be written as an initial value problem and the aim is to extend this formulation to the infinite case. Regarding now the CME (2.16) as a "discrete" PDE, where instead of partial derivatives with respect to the state $\mathbf{x} \in \mathbb{N}_0^d$ we have shifts, we reformulate the equation as an abstract Cauchy problem by defining a linear operator between function spaces. Let

$$l^1 = \{p : \mathbb{N}_0^d \to \mathbb{R} \text{ with } \sum_{\mathbf{x}\in\mathbb{N}^d} |p(\mathbf{x})| < \infty\}$$

the space of sequences whose series is absolutely convergent, equipped with the norm $||p||_1 = \sum_{\mathbf{x}\in\mathbb{N}_0^d} |p(\mathbf{x})|$. Consider now the operator $\mathcal{A}$ defined as

$$(\mathcal{A}p(t,\cdot))(\mathbf{x}) = \sum_{j=1}^{M} \Big(\alpha_j(\mathbf{x}-\mu^j)p(t,\mathbf{x}-\mu^j) - \alpha_j(\mathbf{x})p(t,\mathbf{x})\Big), \tag{2.61}$$

with domain

$$D(\mathcal{A}) = \{p \in l^1 \mid \alpha_j p \in l^1, \ \ \forall\, j = 1,\dots,M\}. \tag{2.62}$$

Using (2.61) in (2.58), as well as (2.60), we can now formulate the CME as an abstract Cauchy problem in the high dimensional sequence space $l^1(\mathbb{N}_0^d)$,

$$\begin{aligned} \partial_t p(t,\cdot) &= \mathcal{A}p(t,\cdot) \\ p(0,\cdot) &= p_0(\cdot). \end{aligned} \tag{2.63}$$

The CME is a linear equation, and if the operator $\mathcal{A}$ is *bounded*, i.e,

$$\exists\, C > 0 \text{ such that } ||\mathcal{A}p||_1 \le C||p||_1, \ \forall p \in D(\mathcal{A}),$$

the solution of (2.63) is given as

$$p(t,\cdot) = e^{t\mathcal{A}}p_0(\cdot), \text{ where } e^{t\mathcal{A}} = \sum_{k=0}^{\infty} \frac{(t\mathcal{A})^k}{k!}. \tag{2.64}$$

The boundness condition on $\mathcal{A}$ ensures convergence of the series $e^{t\mathcal{A}}$ in the l^1-sequence norm, leading to well-posedness for (2.63) in such a setting. However, for most applications of interest, the operator $\mathcal{A}$ is *unbounded*. Therefore, the term $e^{t\mathcal{A}}$ is no longer defined and the existence of the solution (2.64) is not clear.

Finding an analytical solution for the CME in the general case is usually impossible, although we remark that for networks consisting only of monomolecular reactions an

analytical form has been derived in [JH07]. Consequently, numerical approximation of the CME solution is usually the only option. However, as numerical approximations cannot be computed on an infinite state space, we proceed to define a finite subset of $\mathbb{N}_0^d$ by selecting a suitable truncation vector $\xi \in \mathbb{N}^d$ and considering

$$\Omega_\xi = \{\mathbf{x} \in \mathbb{N}_0^d \mid x_1 < \xi_1, \ldots, x_d < \xi_d\} \subseteq \mathbb{N}_0^d. \tag{2.65}$$

We also define the function space

$$V_\xi = \{p \in l^1 \mid p(\mathbf{x}) = 0 \text{ if } \mathbf{x} \notin \Omega_\xi\} \subseteq l^1$$

which is the set of all functions with support constrained to Ω_ξ. Further, let $\mathcal{P}_\xi : l^1 \to V_\xi$ be the projection from l^1 into the subset of the truncated functions, defined as

$$\Big(\mathcal{P}_\xi p(t,\cdot)\Big)(\mathbf{x}) = \begin{cases} p(t,\mathbf{x}) & \text{, if } \mathbf{x} \in \Omega_\xi \\ 0 & \text{, else.} \end{cases}$$

We can now define the "truncated" CME operator

$$\mathcal{A}_\xi : V_\xi \to V_\xi, \;\; \mathcal{A}_\xi = \sum_{j=1}^{M} \mathcal{A}_\xi^{(j)} \tag{2.66}$$

with

$$\big(\mathcal{A}_\xi^{(j)} p(t,\cdot)\big)(\mathbf{x}) = \begin{cases} \alpha_j(\mathbf{x}-\mu^j)p(t,\mathbf{x}-\mu^j) - \alpha_j(\mathbf{x})p(t,\mathbf{x}) & \text{if } \mathbf{x}, \mathbf{x}-\mu^j \in \Omega_\xi \\ 0 & \text{else} \end{cases}$$

denoting the CME operators for individual reaction channels R_j. The truncated counterpart to (2.63) is thus

$$\begin{aligned} \partial_t p_\xi(t,\cdot) &= \mathcal{A}_\xi p_\xi(t,\cdot) \\ p_\xi(0,\cdot) &= \mathcal{P}_\xi p_0(\cdot) \end{aligned} \tag{2.67}$$

with $p_\xi(t,\cdot) \in V_\xi$.

The question whether the solution of the truncated CME (2.67) converges to the solution of the CME (2.63) on $\mathbb{N}_0^d$ as $\xi = (\xi_1, \ldots, \xi_d) \to (\infty, \ldots, \infty)$ can be investigated by employing Trotter-Kato approximation theorems for operator semigroups [EN06, Chapter IV], which provide the following result.

Theorem 2.3. *If $range(\lambda_0 - \mathcal{A})$ is dense in l^1 for some $\lambda_0 > 0$ and $p_0 \in D(\mathcal{A})$, then the CME (2.63) on the unbounded space $\mathbb{N}_0^d$ has a unique solution which depends continuously on the initial data and furthermore, we have*

$$\lim_{\substack{\xi_i \to \infty \\ \forall i = 1,\ldots,d}} ||p_\xi(t,\cdot) - p(t,\cdot)|| = 0.$$

We note that Theorem 2.3 is a consequence of the Second Trotter-Kato approximation Theorem [EN06, Chapter IV, 1.9] and the following proof is only a rough sketch.

Proof. In order to apply the Second Trotter-Kato Theorem, we first need to show that the semigroups $(T_\xi(t))_{t\geq 0}$ generated by $\mathcal{A}_\xi$ satisfy the *stability condition*, i.e., there exist the constants $M \geq 1$ and $\omega \in \mathbb{R}$ such that

$$||T_\xi(t)||_1 \leq Me^{\omega t}$$

for all $t \geq 0$ and $\xi \in \mathbb{N}^d$. In other words, $||T_\xi(t)||$ has to remain bounded for all $t \geq 0$ as $\xi \to \infty$. Without loss of generality, we can choose $M = 1$ and $\omega = 0$ and show that $||T_\xi(t)|| \leq 1$ (the proof of this assertion is given in the subsequent Section 2.5.1). It can also be shown that the domain $D(\mathcal{A})$ from (2.62) is a dense subspace of l^1, and that $\mathcal{A}_\xi u_\xi \to \mathcal{A}u$ as $\xi \to \infty$ for all $u \in D(\mathcal{A})$. Then, the Second Trotter-Kato Theorem states that the closure of the operator $\mathcal{A}$ generates a strongly continous semigroup $(T(t))_{t\geq 0}$, and that

$$\lim_{\substack{\xi_i\to\infty \\ \forall i=1,\dots,d}} ||T_\xi(t)p_0(\cdot) - T(t)p_0(\cdot)|| \quad = \quad \lim_{\substack{\xi_i\to\infty \\ \forall i=1,\dots,d}} ||p_\xi(t,\cdot) - p(t,\cdot)|| = 0, \;\; \forall p_0 \in D(\mathcal{A}).$$

□

Because this thesis is concerned with numerical approximations, the discussion revolves around the truncated CME introduced in (2.67), and we proceed to expose its properties.

2.5.1. Properties of the truncated CME

The operator $\mathcal{A}_\xi$ is defined in (2.66) as a mapping between finite dimensional spaces V_ξ, and as such is isomorphic to a sparse matrix $A_\xi \in \mathbb{R}^{N\times N}$, where $N = \prod_{i=1}^d \xi_i$. If we perform a reshaping of the finite state space Ω_ξ into a column vector indexed on the states $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}\}$, we obtain the matrix form of (2.67)

$$\begin{aligned} \partial_t p(t) &= A_\xi p(t) \\ p(0) &= p_0 \end{aligned} \tag{2.68}$$

with $p(t) \in \mathbb{R}^N$ a vector where the i-th element is given by $p_i(t) = \mathbb{P}(X(t) = \mathbf{x}^{(i)})$ and

$$(A_\xi)_{ik} = \begin{cases} -\sum_{j=1}^M \alpha_j(\mathbf{x}^{(i)}) & \text{for } i = k \\ \sum_{j\in\mathcal{J}} \alpha_j(\mathbf{x}^{(i)}) & \text{for } \mathbf{x}^{(i)} = \mathbf{x}^{(k)} - s, \; \mathcal{J} = \{\text{all } j \text{ where } \; \mu^j = s\} \\ 0 & \text{otherwise.} \end{cases} \tag{2.69}$$

The set $\mathcal{J}$ denotes the indices of those reaction channels R_j which have the property that when they fire, they induce a jump into the corresponding new state on account of the old state and their specific stoichiometry. For brevity, we shall hereafter drop the superfluous parameter ξ, and proceed to use $A := A_\xi \in \mathbb{R}^{N\times N}$. Further, by a slight abuse of notation we will also use $\mathcal{A}$ instead of $\mathcal{A}_\xi$, but what is always meant is the truncated operator from (2.67). Similarly to (2.64), the solution of (2.68) is known to be given by

$$p(t) = e^{tA}p_0. \tag{2.70}$$

We now turn to the investigation of some of the properties induced by the particular structure of the matrix $A \in \mathbb{R}^{N\times N}$ defined in (2.69). Because of the positivity of the

propensity functions $\alpha_j(\mathbf{x})$ on $\mathbb{N}_0^d$ and implicitly on Ω_ξ, it follows that this matrix has non-positive diagonal entries, while its off-diagonal elements are non-negative, i.e.,

$$a_{ii} \leq 0, \ \ \forall i = \{1, \ldots, N\} \text{ and } a_{ik} \geq 0, \text{ for } i \neq k. \tag{2.71}$$

Using (2.59) ensured that the term $\mathbf{x} - \mu^j$ from (2.61) does not become negative, but when considering the truncated state space induced by $\xi \in \mathbb{N}^d$, we also need to impose another set of boundary conditions for the case in which the term lies outside the arbitrarily defined state space Ω_ξ from (2.65). As the choice of boundary conditions has important consequences, it warrants a detailed investigation.

Imposing discrete Neumann boundary conditions on the boundaries of Ω_ξ by setting

$$\alpha_j(\mathbf{x}) = 0, \ \text{ if } \ \mathbf{x} \in \Omega_\xi \ \text{ and } \ \mathbf{x} + \mu^j \notin \Omega_\xi, \tag{2.72}$$

leads to the suppression of all reactions that might trigger a jump from a state $\mathbf{x} \in \Omega_\xi$ into a state lying outside these boundaries. As a consequence of using (2.72) and considering the structure of $A \in \mathbb{R}^{N \times N}$, its elements a_{ik} now satisfy the condition

$$\sum_{i=1}^{N} a_{ik} = 0, \ \ \forall k \in \{1, \ldots, N\}, \tag{2.73}$$

in addition to having the properties given in (2.71). We remark that (2.73) and (2.71) are exactly the properties satisfied by the elements of the transposed generator of the continuous-time Markov process with finite state space given by (2.47), (2.48) and (2.49), i.e., $A \equiv L^T$. This is not a surprise, if we compare the definition of the generator (2.54) with that of A (2.69) and recall that the propensity functions give the probability of reaction channel R_j ($j = 1, \ldots, M$) firing next and triggering a transition to a new state.

Imposing the discrete Neumann boundary conditions (2.72) has several advantages, as it guarantees that the solution of the CME on the truncated state space remains a probability distribution if the initial data is a probability distribution, that a stationary distribution exists and all non-zero eigenvalues have negative real part. We proceed now to supply some proof of these assertions.

First, recall the indexing of the states $\mathbf{x} \in \Omega_\xi$ as $\{\mathbf{x}^{(i)}\}_{i=1,\ldots,N}$, and assume that for some particular state $i \in \{1, \ldots, N\}$ we have $p_i(t) = 0$ and $p_k(t) \geq 0$, $k \neq i$. Using (2.68) and taking (2.71) into account, we get that

$$\dot{p}_i(t) = a_{ii} p_i(t) + \sum_{k \neq i} a_{ik} p_k(t) \geq 0,$$

which means that $p_i(t)$ can not become negative, and we have $p(t) \geq 0, \ \ \forall t \geq 0$. Further, because

$$\sum_{i=1}^{N} \dot{p}_i(t) = (\mathbb{1}^T A) p(t) = 0,$$

we obtain $\sum_{i=1}^{N} p_i(t) = \sum_{i=1}^{N} (p_0)_i = 1$. This means that the solution of the CME on the truncated state space, obtained when imposing the discrete Neumann boundary conditions (2.72) and assuming $p_0 \in \mathbb{R}^N$ is a probability distribution, i.e.,

$$p_0 \geq 0 \ \ \text{ and } \ \ \sum_{i=1}^{N} (p_0)_i = 1,$$

has only non-negative elements and is also a probability distribution. Thus, discrete Neumann boundary conditions guarantee that the probability mass is preserved.

On the other hand, if we impose discrete Dirichlet boundary conditions by setting

$$p(t, \mathbf{x}) = 0, \quad \forall \mathbf{x} \in \mathbb{N}_0^d \setminus \Omega_\xi \tag{2.74}$$

this favorable property is lost. The reason is that imposing (2.74) might lead to

$$\sum_{i=1}^{N} \dot{p}_i(t) < 0$$

and further to $\sum_{i=1}^{N} p_i(t) \leq 1$ for some time $t > 0$, so probability mass would "leak out" of the truncated state space Ω_ξ. We remark that the truncation error incurred by imposing discrete Dirichlet boundary conditions has been investigated in [MK06] and is given as $1 - \sum_{i=1}^{N} p_i(t)$. For short time intervals, there is no difference between the two boundary conditions, especially if the probability mass is concentrated in states that are far from the artificially imposed borders of Ω_ξ, but for longer times it can be significant. Therefore, we have chosen to use Neumann boundary conditions (2.72) instead of (2.74).

For a matrix A with the properties (2.71) and (2.73), we also have that $||e^{tA}||_1 = 1$, $\forall t \geq 0$. To prove this assertion, let $v \in \mathbb{R}^N$, with $||v||_1 = 1$. Further, we define $v^+ \in \mathbb{R}^N$ and $v^- \in \mathbb{R}^N$ by

$$v_j^+ = \begin{cases} v_j & \text{if } v_j \geq 0 \\ 0 & \text{else} \end{cases} \qquad v_j^- = \begin{cases} v_j & \text{if } v_j < 0 \\ 0 & \text{else} \end{cases},$$

such that $v = v^+ + v^-$ and $v_j^+ \geq 0$ and $v_j^- \leq 0$ for all $j = 1, \ldots, N$. Then, $||v||_1 = \mathbb{1}^T v^+ - \mathbb{1}^T v^- = 1$ and we obtain,

$$\begin{aligned} ||e^{tA} v||_1 &\leq ||e^{tA} v^+||_1 + ||e^{tA} v^-||_1 \\ &= \mathbb{1}^T e^{tA} v^+ - \mathbb{1}^T e^{tA} v^- \\ &= \mathbb{1}^T v^+ - \mathbb{1}^T v^- = 1. \end{aligned} \tag{2.75}$$

In (2.75), we have used that

$$\sum_{i=1}^{N} p_i(t) = \mathbb{1}^T p(t) = \mathbb{1}^T e^{tA} p_0 = \mathbb{1}^T p_0 + \underbrace{\mathbb{1}^T \sum_{k=1}^{\infty} \frac{(tA)^k}{k!} p_0}_{=0} = \mathbb{1}^T p_0, \tag{2.76}$$

which yields $\mathbb{1}^T e^{tA} p_0 = \mathbb{1}^T p_0$.

Next, let $\sigma(A)$ denote the spectrum of A. Because of (2.73), we have $\mathbb{1}^T A = 0$, which yields that $0 \in \sigma(A)$ with trivial left eigenvector $\mathbb{1}^T = (1, \ldots, 1)$. Moreover, as the elements of A satisfy (2.71) and (2.73), we can apply Gerschgorin's Circle Theorem (see Appendix A for details) to show that all eigenvalues $\lambda \neq 0$, $\lambda \in \sigma(A)$ have strictly negative real part, i.e., $Re(\lambda) < 0$.

It is often the case that the transient behavior of the truncated CME (2.68) is of minor importance, with the focus being on the long-time dynamics of the system. This leads

to the question of convergence of the solution of (2.68) to the invariant distribution as $t \to \infty$.

Using the Perron-Frobenius theorem for primitive matrices it can be shown that under certain conditions an unique non-negative stationary distribution exists. Before the existence question is tackled however, we must first answer the question whether the system is *closed* or *open*. For *open* systems where particles are introduced into the system, there are many simple examples where we do not have a steady state. For a *closed* system, no particles may enter or leave, although they may combine in various ways inside the container which was considered for the system. *Closed* systems with a finite number of discrete states admit a unique steady-state solution if the matrix A is neither *decomposable* or of *splitting* type [vK01, Chapter V.3]. A matrix A with the properties (2.71) and (2.73) is called *decomposable* if it can be cast in the form

$$A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}$$

where A_{11} and A_{22} are matrices with the some properties of A but of smaller size. In such cases, the state space is decomposed into two subsets which are not connected and we have no unique steady-state solution. In case A is a *splitting* matrix, it can be written as

$$A = \begin{bmatrix} A_{11} & 0 & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix}$$

where A_{11} and A_{22} share the properties of A and at least some of the elements of A_{13}, A_{23} and A_{33} are non-zero. Again, the state space is decomposed into subsets that are not fully connected and we have at least two steady state solutions. These concepts can also be extended to the operator $\mathcal{A}$, with operators of decomposable or splitting type defined in a similar way (see [vK01, Chapter V.3]).

Next, following a similar result from [JH07] and using the Perron-Frobenius Theorem, we prove that if A is neither *decomposable* or of *splitting* type, an unique steady state solution exists. First, we give the definition of a primitive matrix.

Definition 2.4. *A matrix $M \in \mathbb{R}^{N\times N}$ with elements m_{ij} is called*

- *non-negative ($M \geq 0$) if $m_{ij} \geq 0, \forall i, j$.*
- *strictly non-negative ($M > 0$) if $m_{ij} > 0, \forall i, j$*
- *primitive if $M \geq 0$ and $M^k > 0$ for some $k \in \mathbb{N}$.*

Considering now the matrix A defined in (2.69), it is obvious that it does not satisfy the above requirements. However, as the solution of the CME on the truncated state space is given by $p(t) = e^{tA}p_0 \geq 0,\ \forall t > 0$, we get that the matrix $T(t) = e^{tA} \geq 0, \forall t \geq 0$ is primitive according to the definition 2.4 if A is neither of splitting or decomposable type. Next, we state the Perron-Frobenius theorem for primitive matrices [Sen81, Chapter 1].

Theorem 2.5 (Perron-Frobenius). *Let $M \geq 0$ be a primitive matrix. Then an eigenvalue $\mu \in \sigma(M)$ exists such that,*

(a) $\mu \in \mathbb{R}$, $\mu > 0$

(b) $\mu > |\lambda|$ *for any eigenvalue* $\lambda \neq \mu$

(c) μ *is a simple root for the characteristic polynomial*

(d) there are strictly positive right and left eigenvectors $v > 0$ *and* $w > 0$*, respectively, such that*

$$Mv = \mu v \text{ and } w^T M = \mu w^T$$

With $T(t) = e^{tA} \geq 0,\ \forall t \geq 0$ and A defined by (2.69), we then have the following

Corollary 2.6. *Suppose that* $T(t^\star) = e^{t^\star A} > 0$ *, for some* $t^\star > 0$*. Then,*

(i) there exists an unique invariant distribution $\rho \in \mathbb{R}^N$ $(\rho \geq 0,\ \sum_{i=1}^N \rho_i = 1)$*, i.e.,*

$$\begin{aligned} T(t)\rho &= \rho,\ \forall t \geq 0 \\ A\rho &= 0 \end{aligned}$$

(ii) we have

$$\lim_{t\to\infty} p(t) = \rho \quad \text{and} \quad \lim_{t\to\infty} T(t) = \underbrace{[\rho|\ldots|\rho]}_{N \text{ times}}$$

Proof. We have already established that every eigenvalue λ of $A \in \mathbb{R}^{N\times N}$ is either zero or has strictly negative real part (see Appendix A). Consequently, we have that $1 \in \sigma(T(t)),\ \forall t \geq 0$ and for any eigenvalue $\zeta \in \sigma(T(t)),\ \zeta \neq 1$ we have $|\zeta| < 1$ on account of

$$|e^{t\lambda}| = |e^{tRe(\lambda)}| < 1, \text{ for } Re(\lambda) < 0 \text{ and } t > 0.$$

Let $\hat{\rho}$ be the right eigenvector corresponding to eigenvalue $0 \in \sigma(A)$. It follows that $A\hat{\rho} = 0$ and also

$$T(t)\hat{\rho} = \sum_{k=0}^{\infty} \frac{(tA)^k}{k!}\hat{\rho} = \hat{\rho} + \sum_{k=1}^{\infty} \underbrace{\frac{(tA)^k}{k!}\hat{\rho}}_{=0} = \hat{\rho}.$$

Now, under the initial assumption $T(t^\star) > 0$, we can apply Perron-Frobenius as given in Theorem 2.5, which leads to the conclusion that the eigenvalue $1 \in \sigma(T(t))$ is simple, such that an unique (up to a constant $c > 0$) eigenvector $\rho = c\hat{\rho}$ exists, with $\rho > 0$ and $\sum_{i=0}^N \rho_i = 1$. Moreover, we have that

$$T(t)\rho = cT(t)\hat{\rho} = c\hat{\rho} = \rho,\ \ \forall t \geq 0 \text{ and } A\rho = cA\hat{\rho} = 0$$

which ends the proof of *(i)*.

To prove *(ii)*, we use the fact that there exists a decomposition of $A = SJS^{-1}$ with J the Jordan normal form, i.e., a block diagonal matrix

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_n \end{pmatrix}$$

where each block J_k of the square matrix J is of the form

$$J_k = \begin{pmatrix} \lambda_k & 1 & & \\ & \lambda_k & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{pmatrix}$$

and $S = [s_1|\dots|s_n]$ is an invertible matrix. We use now that $T(t) = e^{tA} = Se^{tJ}S^{-1}$ [EN06, Chapter I], and the fact that J has a block structure and each block can be treated separately. As the eigenvalue $\lambda_1 = 0 \in \sigma(A)$ is simple, the first Jordan block J_1 has only one entry, and we have

$$e^{tJ_1} = e^{t\lambda_1} = 1.$$

For the remaining Jordan blocks J_k,

$$\lim_{t\to\infty} e^{tJ_k} = \mathbf{0},\ k > 1$$

(where $\mathbf{0}$ is an appropriately sized zero matrix) holds, because all the other eigenvalues $\lambda_k,\ k > 1$ satisfy $Re(\lambda_k) < 0$ via Corollary A.2 from Appendix A. Hence,

$$T^\infty = \lim_{t\to\infty} T(t) = \lim_{t\to 0} Se^{tJ}S^{-1} = S\ \mathrm{diag}\{1, 0, \dots, 0\}\ S^{-1} = vw^T,$$

where v and w are the first column vector from S, and first row-vector from S^{-1} respectively. We know from *(i)* that ρ is the right eigenvector corresponding to the eigenvalue $\zeta = 1$, so we have

$$\lim_{t\to\infty} T(t)\rho = \rho = v(\underbrace{w^T\rho}_{=C>0})$$

which leads to $v = \frac{1}{C}\rho = c_1\rho$.

Further, we have that the vector $\mathbb{1}^T = (1, \dots, 1)$ is the left eigenvector corresponding to $\zeta = 1$, because

$$\mathbb{1}^T T(t) = \mathbb{1}^T e^{tA} = \mathbb{1}^T \sum_{k=0}^{\infty} \frac{(tA)^k}{k!} = \mathbb{1}^T\mathbf{I} + \sum_{k=1}^{\infty} \underbrace{\mathbb{1}^T \frac{(tA)^k}{k!}}_{=0} = \mathbb{1}^T$$

where we have used the fact that $\mathbb{1}^T$ is also the left eigenvector of A corresponding to eigenvalue 0 (a consequence of (2.73)). Passing now to the limit, we obtain

$$\lim_{t\to\infty} \mathbb{1}^T T(t) = \mathbb{1}^T = (\mathbb{1}^T v)w^T,$$

from which it follows that $w = c_2\mathbb{1}$ with some constant $c_2 > 0$. Without loss of generality, we can choose $c_1 = c_2 = 1$ such that $v = \rho$ and $w = \mathbb{1}$, respectively. This leads to

$$\lim_{t\to\infty} T(t) = vw^T = \rho\mathbb{1}^T = [\rho|\dots|\rho]$$

and additionally, we obtain that

$$\lim_{t\to\infty} p(t) = \lim_{t\to\infty} T(t)p_0 = \rho\mathbb{1}^T p_0 = \rho.$$

□

Drawing a line, corollary 2.6 shows that under certain conditions, and assuming the matrix $A \in \mathbb{R}^{N\times N}$ has the properties (2.71) and (2.73), there exists a unique (up to a constant) right eigenvector ρ to the eigenvalue 0, such that

$$A\rho = 0.$$

Further, if ρ is normalized and has only non-negative elements, then it represents the stationary probability distribution of the system.

Because in the following chapters we prefer to use the operator notation, with $\mathcal{A}$ denoting the operator restricted to Ω_ξ, we also define the stationary distribution as a discrete function $\pi : \Omega_\xi \to \mathbb{R}$ satisfying

$$\mathcal{A}\pi = 0, \;\; \pi(\mathbf{x}) \geq 0, \;\; \sum_{\mathbf{x}\in\Omega_\xi} \pi(\mathbf{x}) = 1. \tag{2.77}$$

2.5.2. The adjoint CME operator

In the same way that the CME defined with the help of operator $\mathcal{A}$ from (2.61) can be considered a special case of the *forward Kolmogorov equations* (2.55), it is sometimes useful to consider the *adjoint* operator $\mathcal{A}^\star$ which relates in a similar way to the *backward Kolmogorov equations* (2.56). The *adjoint* operator $\mathcal{A}^\star$ is defined such that it satisfies the property

$$\langle \mathcal{A}p, q\rangle = \langle p, \mathcal{A}^\star q\rangle \tag{2.78}$$

where $\langle\cdot,\cdot\rangle$ represents the Euclidean inner product between two functions $p, q \in l^1(\mathbb{N}^d)$. Substituting (2.61) in (2.78) we obtain

$$\begin{aligned}
\langle \mathcal{A}p, q\rangle &= \sum_{\mathbf{x}\in\Omega_\xi} \left(\sum_{j=1}^{M} \Big(\alpha_j(\mathbf{x}-\mu^j)p(t,\mathbf{x}-\mu^j) - \alpha_j(\mathbf{x})p(t,\mathbf{x}) \Big) \right) q(t,\mathbf{x}) \\
&= \sum_{\mathbf{x}\in\Omega_\xi} \left(\sum_{j=1}^{M} \alpha_j(\mathbf{x}) \Big(q(t,\mathbf{x}+\mu^j) - q(t,\mathbf{x}) \Big) \right) p(t,\mathbf{x})
\end{aligned}$$

where we have used that

$$\sum_{\mathbf{x}\in\Omega_\xi} \alpha_j(\mathbf{x}-\mu^j)p(t,\mathbf{x}-\mu^j)q(t,\mathbf{x}) = \sum_{\mathbf{x}\in\Omega_\xi} \alpha_j(\mathbf{x})p(t,\mathbf{x})q(t,\mathbf{x}+\mu^j)$$

because of (2.59). Thus, we obtain the following representation for the adjoint operator in terms of the already defined propensity functions α_j and stoichiometric vectors μ^j

$$(\mathcal{A}^\star q(t,\cdot))(\mathbf{x}) = \sum_{j=1}^{M} \alpha_j(\mathbf{x}) \Big(q(t,\mathbf{x}+\mu^j) - q(t,\mathbf{x}) \Big). \tag{2.79}$$

The adjoint CME equation takes the now familiar form

$$\begin{aligned}
\partial_t q(t,\cdot) &= \mathcal{A}^\star q(t,\cdot) \\
q(0,\cdot) &= q_0(\cdot),
\end{aligned} \tag{2.80}$$

and if we consider as before the truncated state space Ω_ξ, we get that the truncated adjoint CME operator $\mathcal{A}^\star_\xi$ is isomorphic to a sparse matrix $A^\star \in \mathbb{R}^{N\times N}$, with the corresponding matrix form of (2.80) having the formal solution $q(t) = e^{tA^\star}q_0$, with $q(t) \in \mathbb{R}^N$.

Finally, we have "detailed balance" if the operator $\mathcal{A}$ has the symmetry property, a condition that is rarely fulfilled by biological processes, as this would imply the reversibility of the underlying stochastic process. In matrix form, "detailed balance" can be expressed as

$$\mathrm{diag}(\pi)A = (\mathrm{diag}(\pi)A^\star)^T$$

where $\pi \in \mathbb{R}^N$ denotes the stationary probability distribution. This relation basically asserts the obvious fact that for a reversible process in steady state, the transitions between each pair of states must balance out.

The adjoint equation (2.80) is mostly used in connection with absorbing states and first-passage problems [vK01, Chapter XII], and in this thesis will make an appearance in Chapter 5 where methods for metastability analysis are discussed.

After deriving the CME from two perspectives and introducing the operator notation that will be central to the discussion in the later chapters, the aim of the next sections is to place the equation into context by quickly reviewing the main alternatives to the CME.

2.6. Macroscopic equations

In the discrete stochastic formulation we have investigated so far, the definition given for the propensity functions $\alpha_j(\mathbf{x})$ in (2.5), obscured somewhat the fact the propensities actually depend on the volume V of the space in which the particles are enclosed. However, the dependence on the volume is immediately clear, if for example, we inspect the expression (2.10) derived for the specific probability rate constant c_j of a bimolecular reaction. So far, we have simply kept the volume $V = 1$ fixed, but in order to investigate the behavior of the system when $V \to \infty$, let us redefine the propensities such that volume dependency is explicitly stated, as

$$\alpha_j(\mathbf{x}) = \frac{k_j}{\mathbf{\Omega}^{|s_j|-1}} \prod_{i=1}^{d} \binom{x_i}{x_i - n_{j,i}}. \tag{2.81}$$

In (2.81), the terms $n_{j,i}$ are the stoichiometric coefficients on the reactants side of (2.11), $|s_j| = \sum_{i=1}^{d} n_{j,i}$ is the reaction order, k_j denotes a macroscopic rate constant and $\mathbf{\Omega}$ is a scaling factor related to the conversion of the specific probability rate constant c_j in the macroscopic rate constant k_j. For example, $\mathbf{\Omega}$ can be the system volume V, *Avogadro's* constant $n_A = 6.02214179 \cdot 10^{23}\ mol^{-1}$, or their product $n_A \cdot V$.

Turning now to the *macroscopic* formulation of biochemical reaction kinetics, the state of the system is described by a deterministic process $y(t) : \mathbb{R} \to \mathbb{R}^d$ with $y(t) = [y_1(t), \ldots, y_d(t)]$. The variables $y_i(t) \in \mathbb{R}_0$ represent the concentrations of the species S_i at time t and are related to the copy numbers of the species from the stochastic formulation by

$$y_i(t) = X_i(t)/(n_A \cdot V).$$

Thus, using a vector notation, we have that the states $\mathbf{y} \in \mathbb{R}_+^d$ of the *macroscopic* model are related to the states $\mathbf{x} \in \mathbb{N}_0^d$ from the stochastic model via $\mathbf{y} = \mathbf{x}/\mathbf{\Omega}$, with $\mathbf{\Omega} = n_A \cdot V$. The time evolution of the deterministic process $y(t)$ is then given by a system of ordinary differential equations (ODEs)

$$\frac{d}{dt}y(t) = \sum_{j=1}^{M} \mu^j \tilde{\alpha}_j(y(t)), \tag{2.82}$$

where $\tilde{\alpha}_j(\mathbf{y}) = \alpha_j(\mathbf{x})/\mathbf{\Omega}$ are the $\mathbf{\Omega}$-scaled propensities of the reaction channel R_j, and we remark that the discrete characterization of nature has been replaced by a continuous point of view. The functions $\tilde{\alpha}_j$ are functionally similar to the propensity functions α_j from (2.81), and for elementary reactions we have

$$\begin{aligned}
\alpha_j(\mathbf{x}) &= \frac{k_j}{\mathbf{\Omega}^{|s_j|-1}} \prod_{i=1}^{d} \binom{x_i}{x_i - n_{j,i}} = \mathbf{\Omega} \frac{k_j}{\mathbf{\Omega}^{|s_j|}} \prod_{i=1}^{d} \binom{x_i}{x_i - n_{j,i}} \\
&= \mathbf{\Omega} k_j \prod_{i=1}^{d} \frac{1}{\mathbf{\Omega}^{n_{j,i}}} \cdot \frac{x_i!}{n_{j,i}!(x_i - n_{j,i})!} = \mathbf{\Omega} k_j \prod_{i=1}^{d} \frac{1}{\mathbf{\Omega}^{n_{j,i}} \cdot n_{j,i}!} \prod_{s=0}^{n_{j,i}-1} (x_i - s) \\
&= \mathbf{\Omega} k_j \prod_{i=1}^{d} \frac{1}{n_{j,i}!} \prod_{s=0}^{n_{j,i}-1} \left(y_i - \frac{s}{\mathbf{\Omega}}\right) = \mathbf{\Omega} \tilde{\alpha}_j(\mathbf{y}) \\
&= \mathbf{\Omega} \tilde{\alpha}_j(\mathbf{x}/\mathbf{\Omega}).
\end{aligned} \tag{2.83}$$

However, when modeling is based on the *law of mass action*, the evolution of the deterministic process $y(t)$ is described with the help of the classic reaction rates

$$a_j(\mathbf{y}) = k_j \prod_{i=1}^{d} \frac{y_i^{n_{j,i}}}{n_{j,i}!} \tag{2.84}$$

instead of the $\mathbf{\Omega}$-scaled propensity functions $\tilde{\alpha}_j$. For reactions of order zero and one, where $n_{j,i} \in \{0, 1\}$, the reaction rates $a_j(\mathbf{y})$ and the propensities $\tilde{\alpha}_j(\mathbf{y})$ coincide. For more complex reaction channels however, where we have $n_{j,i} > 1$, the reaction rate $a_j(\mathbf{y})$ only approximates $\tilde{\alpha}_j(\mathbf{y})$ for large $\mathbf{\Omega}$, because

$$\tilde{\alpha}_j(\mathbf{y}) = k_j \prod_{i=1}^{d} \frac{1}{n_{j,i}!} \prod_{s=0}^{n_{j,i}-1} \left(y_i - \frac{s}{\mathbf{\Omega}}\right) = k_j \prod_{i=1}^{d} \frac{y_i^{n_{j,i}}}{n_{j,i}!} + \mathcal{O}\left(\mathbf{\Omega}^{-1}\right). \tag{2.85}$$

Using (2.83), (2.84) and (2.85), it follows that

$$\frac{1}{\mathbf{\Omega}}\alpha_j(\mathbf{x}) - a_j\left(\frac{\mathbf{x}}{\mathbf{\Omega}}\right) = \begin{cases} 0 & \text{if } n_{j,i} \in \{0, 1\} \\ \mathcal{O}\left(\mathbf{\Omega}^{-2}\right) & \text{else} \end{cases} \tag{2.86}$$

which gives the relation between the volume-dependent propensities (2.81) of the CME and the classic reaction rates (2.84) of the *Reaction Rate Equations* (RRE). Further, we note that (2.82) represents the "concentrations" form of the RRE.

It has been shown by T.G. Kurtz in [Kur72], that in the thermodynamic limit, when the initial copy numbers of all the species $X(0)$ and $\mathbf{\Omega}$ approach infinity, and the initial species concentrations tend to some value

$$\lim_{\mathbf{\Omega} \to \infty} \frac{X(0)}{\mathbf{\Omega}} = \mathbf{y}_0,$$

the deterministic process $y(t)$ from (2.82) approaches the scaled discrete *Markov jump process* underlying the CME (2.63) for every finite time t, i.e.,

$$\lim_{\mathbf{\Omega}\to\infty} \mathbb{P}\Big(\sup_{t\in[0,T]} \Big| \frac{X(t)}{\mathbf{\Omega}} - y(t) \Big| > \varepsilon \Big) = 0, \quad \forall \varepsilon > 0.$$

This link between the stochastic and deterministic models is best illustrated by showing that the reaction-rate equations yield an approximation to the expected value $\mathbb{E}[X(t)] = \sum_{\mathbf{x}} \mathbf{x} \cdot p(t, \mathbf{x})$ of the stochastic process $X(t)$ at some specific time t. The derivation below follows that of [Gil00].

Multiplying the CME (2.58) by $\mathbf{x}$ and summing over all the states yields

$$\begin{aligned} \sum_{\mathbf{x}} \mathbf{x} \cdot \partial_t p(t, \mathbf{x}) &= \sum_{\mathbf{x}} \mathbf{x} \cdot \sum_{j=1}^{M} \Big(\alpha_j(\mathbf{x} - \mu^j) p(t, \mathbf{x} - \mu^j) - \alpha_j(\mathbf{x}) p(t, \mathbf{x}) \Big) \\ &= \sum_{j=1}^{M} \mu^j \sum_{\mathbf{x}} \alpha_j(\mathbf{x}) p(t, \mathbf{x}) \\ &= \sum_{j=1}^{M} \mu^j \mathbb{E}[\alpha_j(\mathbf{x})] \end{aligned} \tag{2.87}$$

where we have interchanged the sums and used a re-indexing of the term

$$\sum_{\mathbf{x}} \mathbf{x} \alpha_j(\mathbf{x} - \mu^j) p(t, \mathbf{x} - \mu^j) = \sum_{\mathbf{x}} (\mathbf{x} + \mu^j) \alpha_j(\mathbf{x}) p(t, \mathbf{x})$$

because of (2.59). Inserting the definition of the expectation on the left side of (2.87), we get

$$\frac{d}{dt} \mathbb{E}[X(t)] = \sum_{j=1}^{M} \mu^j \mathbb{E}[\alpha_j(X(t))]. \tag{2.88}$$

However, this is a closed differential equation only if we approximate the expectation of the propensity by the propensity of the expectation, i.e.,

$$\mathbb{E}[\alpha_j(X(t))] = \sum_{\mathbf{x}} \alpha_j(\mathbf{x}) p(t, \mathbf{x}) \approx \alpha_j\Big(\sum_{\mathbf{x}} \mathbf{x} p(t, \mathbf{x}) \Big) = \alpha_j(\mathbb{E}[X(t)]). \tag{2.89}$$

Under this condition we finally get

$$\frac{d}{dt} \mathbb{E}[X(t)] \approx \sum_{j=1}^{M} \mu^j \alpha_j(\mathbb{E}[X(t)]).$$

In case the propensities α_j are linear, in other words the corresponding reaction channels R_j ($j = 1, \ldots, M$) are of order zero or one, the approximation made in (2.89) is *exact*. For higher-order reaction channels however, the propensities are no longer linear, and we can expect the approximation to have a small error only close to the thermodynamic limit, while for the case when the species have low-copy numbers the error could be significant.

Therefore, for reaction networks with low-copy numbers or complex propensities, the CME is clearly the tool of choice for numerical treatment. We must remark, however, that

when the reaction network can be partitioned into two sets where some of the species have adequate copy numbers that allow their treatment by a deterministic model, while the others are present in small numbers and hence are treated stochastically, the computational complexity can be significantly reduced. Such hybrid models are increasingly popular, and in Chapter 6 we will discuss the advantages and also the pitfalls of using them.

2.7. Chemical Langevin equation

In the previous section, we have established a link between the CME and the much coarser RRE, and shown that the latter is a valid approximation only under certain conditions. Furthermore, a simple examination of the RRE (2.82) reveals that the *macroscopic* model is not equipped with the means of capturing stochastic effects which can be pronounced for models on the cellular level, as illustrated in Chapter 1. We now introduce another formulation, the *Chemical Langevin equation* (CLE), a stochastic differential equation that lies between the CME and the RRE. Making sense of the CLE can be done either from the perspective of the *macroscopic* model by considering this equation as an extension of the reaction rate equations via the addition of a term to deal with the inherent stochasticity, or from the perspective of the CME, whereby we trade discreteness for an easier numerical treatment.

However, solving the CLE is not the objective of this thesis, and the following discussion is aimed only at completing the overview of the main directions in simulating biochemical reaction networks. More details can be found in [vK01, Chapter IX] or the excellent review [Hig11].

The derivation of the CLE presented here, is again due to Gillespie and has the same microphysical basis as the CME. However, two rather strong dynamical conditions must be satisfied in order for the formulation to be valid. The following arguments closely follow Gillespie's original paper [Gil00].

As seen in the previous sections 2.3 and 2.4, the SSA algorithm is an exact reproduction of the *Markov jump process*. However, this means that every reaction event has to be considered separately, which is highly inefficient. If we assume now that in a certain time interval $[t, t+\tau)$ we have had more than one firing of some of the reaction channels R_j, we could just leap ahead and update the state of the process in one stroke. The problem with this approach however, is the propensity functions (2.3) which lie at the core of the derivation of CME and implicitly of the SSA algorithm, depend on the state of the system $X(t) = \mathbf{x}$ at the current time t. Therefore, in order for such a simplification to work, we must be able to assume that the propensities $\alpha_j(\mathbf{x})$ can be "frozen" for the specified time interval $[t, t+\tau)$. The number of times a reaction channel R_j fires can then be counted with an independent *Poisson* random variable, denoted by $\mathcal{P}_j(\alpha_j(\mathbf{x}), \tau)$. Thus, we arrive at the *tau*-leaping method [Gil01], one of the most notable improvements on the original SSA algorithm, which is given as

$$X(t+\tau) = \mathbf{x} + \sum_{j=1}^{M} \mu^j \mathcal{P}_j(\alpha_j(\mathbf{x}), \tau). \tag{2.90}$$

As stressed above, this approximation of the SSA is only valid if the propensities do not change too much in the time interval $[t, t+\tau)$, so we require that τ should be *small* enough for this condition to hold. On the other hand, if the time interval is too small and only relatively few firings of the reaction channels occur, then we would be better served by using the original version of SSA [Gil76]. Therefore, a second dynamical condition is that τ should be *large* enough, so an appreciable number of reaction channel firings occur in $[t, t+\tau)$, justifying the incurred error by a big increase in the computational efficiency. It is well known that the mean and variance of a *Poisson* random variable $\mathcal{P}_j(\alpha_j(\mathbf{x}), \tau)$ are both equal to $\alpha_j(\mathbf{x})\tau$. Thus, in order to have a sufficient number of reaction channel R_j firings in $[t, t+\tau)$ we must choose τ such that we have $\alpha_j(\mathbf{x})\tau \gg 1$ for all $j \in \{1, \ldots, M\}$. A consequence of choosing τ in such a manner would be that we can approximate a *Poisson* random variable with large variance and mean by a *normal* random variable with the same variance and mean $\mathcal{N}(\alpha_j(\mathbf{x})\tau, \alpha_j(\mathbf{x})\tau)$. Using now that a *normal* random variable with mean m and variance σ^2 can be replaced by

$$\mathcal{N}(m, \sigma^2) = m + \sigma\mathcal{N}(0, 1),$$

we can reformulate (2.90) as

$$y(t+\tau) = y(t) + \tau \sum_{j=1}^{M} \mu^j \alpha_j(y(t)) + \sqrt{\tau} \sum_{j=1}^{M} \mu^j \sqrt{\alpha_j(y(t))}\mathcal{N}_j(0, 1), \tag{2.91}$$

where we have marked the departure from the discrete representation of the state of the system by replacing the integer-valued $X(t)$ random variable with the continuous random variable $y(t)$ and denoted by $\mathcal{N}_j(0, 1)$ independent *normal* random variables. Equation (2.91) is the Euler-Maruyama method for stochastic differential equations (SDEs) [KP92], and hence for $\tau = dt$ and $dt \to 0$ it discretizes the SDE

$$dy(t) = \sum_{j=1}^{M} \mu^j \alpha_j(y(t))dt + \sum_{j=1}^{M} \mu^j \sqrt{\alpha_j(y(t))}dW_j(t) \tag{2.92}$$

where $W_j(t)$ are independent *Wiener* processes. The SDE given in (2.92) is the *Chemical Langevin equation*, and its solution is a continuous-time Markov process with a continuous state space. The derivation of the CLE implies that when we are able to approximate the number of firings of a reaction channels R_j ($j = 1, \ldots, M$) via *Poisson* random variables, and moreover, approximate these by *normal* random variables, we have that the *Markov jump process* underlying the CME (2.63) can be approximated by a continuous diffusion process described by the CLE (2.92). We also remark that by ignoring the second term on the right hand side of (2.92) we recover the *macroscopic* formulation (2.82). As is the case with the RRE however, the CLE is only valid under the conditions mentioned above, namely that τ is both *small* enough so the propensities do not change too much, and we have a sufficient number of firings given as $\alpha_j(\mathbf{x})\tau \gg 1, \forall j = 1, \ldots, M$, which are satisfied simultaneously only when we have large copy numbers for all the species. In [Gil00], Gillespie summarized these two conditions as requiring τ to have a "macroscopical infinitesimal" character, i.e., the reaction network must posses a domain of macroscopically infinitesimal time intervals where both dynamical conditions are fulfilled.

Note that the CLE (2.91) describes a continuous Markov process, and numerical simulations with the SDE form (2.92) lead to independent realizations of this process, which

is comparable to the way SSA simulations provide trajectories of the discrete Markov jump process underlying the CME. As previously discussed, the solution of the CME (2.16) gives the time evolution of the probability distribution $\mathbb{P}(\mathbf{x}, t|\mathbf{x}_0, t_0)$ of the *discrete* random variable $X(t)$ (2.2). Taking the same viewpoint and using continuous Markov process theory, yields that the evolution of the probability distribution $\mathbb{P}(\mathbf{y}, t|\mathbf{y}_0, t_0)$ of the *continuous* random variable $y(t)$ from (2.91) and (2.92) is the solution of the *Chemical Fokker-Planck equation* (CFPE) (cf. [Gil07]),

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{P}(\mathbf{y}, t|\mathbf{y}_0, t_0) &= -\sum_{j=1}^{M} \nabla\big(\alpha_j(\mathbf{y})\mathbb{P}(\mathbf{y}, t|\mathbf{y}_0, t_0)\big)\mu^j \\ &\quad + \frac{1}{2}\sum_{j=1}^{M}(\mu^j)^T\nabla^2\Big(\alpha_j(\mathbf{y})\mathbb{P}(\mathbf{y}, t|\mathbf{y}_0, t_0)\Big)\mu^j.\end{aligned} \tag{2.93}$$

The CFPE (2.93) is a partial differential equation (PDE), which can be viewed as the master equation companion to the CLE, and for a derivation the reader is referred to [Gil00, Gil96].

2.8. A computational comparison

The overall aim of this chapter was to introduce the CME formally and discuss the properties of the truncated CME operator (2.67) which are relevant in the context of this thesis, specifically, the well-posedness on finite state spaces and the fact that under certain conditions, an unique stationary distribution exists. Additionally, the frameworks of Markov jump process, RRE and CLE have been reviewed in order to establish the links between them and the CME. A graphical overview of some of the connections between the frameworks (modified from [Gil07]) is given in Figure 2.1.

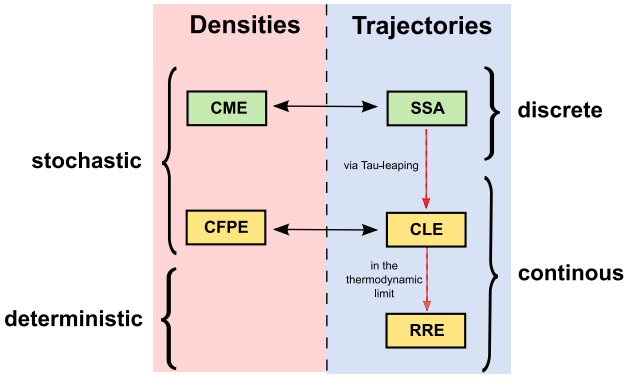


Figure 2.1.: Connections between frameworks of chemical kinetics. Red and blue strips divide the frameworks according to the type of solution obtained. Boxes in green indicate frameworks that provide exact results (CME and SSA), while yellow boxes denote frameworks which deliver approximate results (CLE, CFPE and RRE). The dashed red arrows show approximate inference routes between the frameworks, while solid black arrows indicate that frameworks operate with the same type of processes. We note that not all possible links between the frameworks are displayed (figure modified from [Gil07]).

We conclude the chapter with a computational investigation that illustrates how the different models compare with one another. As a model problem, we use the genetic toggle switch proposed in [GCC00], which was already featured in Figure 1.2 from Chapter 1. This synthetic bistable gene-regulatory network clearly exhibits stochastic behavior, and for certain parameter sets, the truncated state space is small enough to obtain a reference solution. Therefore, the model of the toggle switch will be used throughout this thesis to illustrate the performance of the numerical methods constructed. The toggle switch consists of a pair of mutually repressing genes, where the two competing species S_1 and S_2 each inhibits the transcription of its opponent. The reaction channels are

$$\begin{array}{rrcl|rcl|rcl} R_1: & \star & \longrightarrow & S_1 & \alpha_1 & = & c_{11}/(c_{12}+x_2^2) & \mu^1 & = & (1,0)^T \\ R_2: & \star & \longrightarrow & S_2 & \alpha_2 & = & c_{21}/(c_{22}+x_1^2) & \mu^2 & = & (0,1)^T \\ R_3: & S_1 & \longrightarrow & \star & \alpha_3 & = & c_3 x_1 & \mu^3 & = & (-1,0)^T \\ R_4: & S_2 & \longrightarrow & \star & \alpha_4 & = & c_4 x_2 & \mu^4 & = & (0,-1)^T \end{array} \tag{2.94}$$

and for the examples presented in this section we have used the parameters (modified from [Eng06])

$$c_{11} = c_{21} = 10^3, \quad c_{12} = c_{22} = 5.4 \cdot 10^3 \text{ and } c_3 = c_4 = 1.0005 \cdot 10^{-3}. \tag{2.95}$$

If copies of S_2 are present in abundance, then the propensity function for reaction R_1 almost vanishes, which inhibits the transcription of new copies of S_1. However, over sufficiently long time intervals, stochastic fluctuations can cause an increase in the copy-numbers of S_1, meaning that the production of S_2 will be inhibited instead, and leading to a *switch* in the roles of S_1 and S_2. Reactions R_3 and R_4 model the decay of the two competing species.

In the original paper [GCC00] where the toggle switch was proposed, the bistability of this simple two-gene regulatory network was investigated using the following deterministic model with $y_1(t)$ and $y_2(t)$ denoting the levels of S_1 and S_2 at time t.

$$\begin{aligned} \frac{dy_1}{dt} &= \frac{c_{11}}{c_{12}+y_2^2} - c_3 y_1 \\ \frac{dy_2}{dt} &= \frac{c_{21}}{c_{22}+y_1^2} - c_4 y_2 \end{aligned} \tag{2.96}$$

Plotting the phase portrait of the ODE system (2.96) together with the null-clines ($dy_1/dt = 0$ and $dy_2/dt = 0$) reveals the that the system posses two stable steady states at coordinates $(\sigma_{11}, \sigma_{12}) \approx (149, 36)$ and $(\sigma_{21}, \sigma_{22}) \approx (36, 149)$ as shown in Figure 2.2.

Although the deterministic model (2.96) can be used to reveal the bistability of the reaction network, it is unable to depict the dynamic switching between the steady states. This can only be done by using a stochastic description and the results are illustrated in Figure 2.3. Using the SSA algorithm and the Euler-Maruyama method we can obtain a trajectory of the Markov jump process underlying the CME as shown in Figure 2.3a, and a continuous path from the CLE as seen in Figure 2.3b, respectively. As these are independent realizations, they can not be compared directly, but we remark that both approaches capture the stochastic nature of the system, as apposed to the RRE solutions. Superficially, the CLE appears to deliver a comparable result to the Markov jump process, but

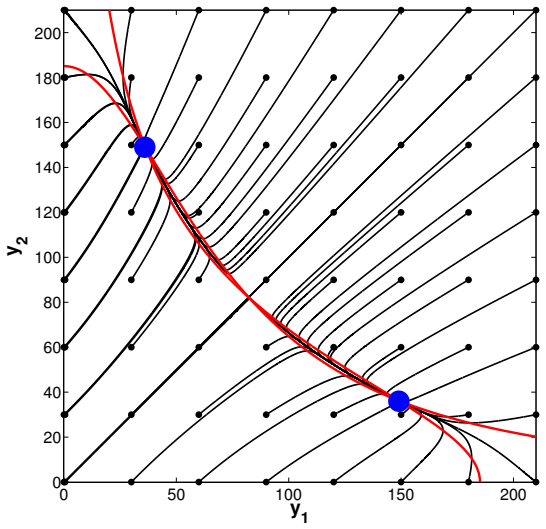
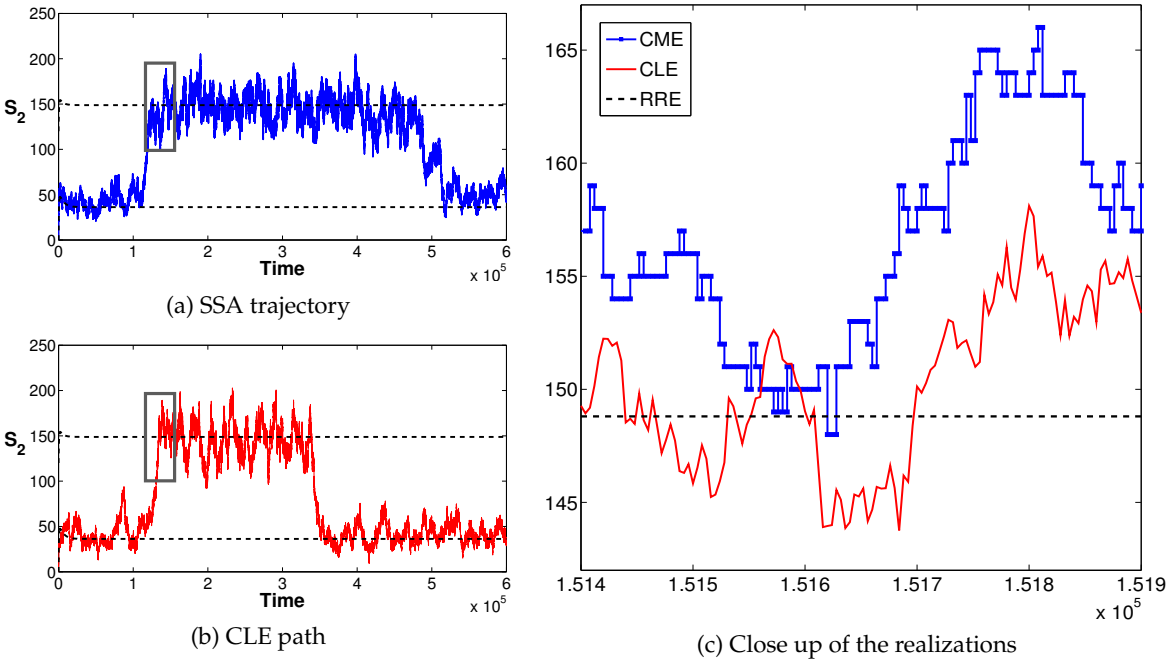


Figure 2.2.: Phase portrait of the deterministic reaction rate equations (2.96). Each black line represents a different trajectory of the deterministic process $y(t)$ in the y_1y_2-plane. The initial conditions are represented by small black dots and the two steady states are depicted by blue circles. The null-clines ($dy_1/dt = 0$ and $dy_2/dt = 0$) are depicted by red lines. Depending on the initial conditions, the deterministic process runs into one of the steady states.

examining a close up of the two independent realizations in Figure 2.3c, it is immediately clear that CME is the only one that respects the discrete character of biochemical reaction kinetics. Furthermore, if the molecule counts of the two genes are low, the assumptions on which the CLE rests are no longer valid, which means that the continuous stochastic model breaks down because the stochastic terms in (2.92) may involve negative square roots. This happens also in the case of the toggle switch model presented here, and requires the modification of the equation by taking the square roots of the absolute values in order to obtain a result (see e.g. [Hig11]).

Consequently, solving the CME represents the best available option when dealing with gene-regulatory networks or other systems where the RRE and CLE prove to be inadequate tools. Unfortunately, the CME is affected by the *curse of dimensionality* and even on a truncated state space it represents a massive ODE system that is too large to be tackled by standard methods. One possibility to approximate the solution of the CME is to generate independent realizations of the underlying Markov jump process via the SSA algorithm detailed in Section 2.3. In Figure 2.4a, such an approximation obtained by averaging $5 \cdot 10^5$ SSA runs is presented, while in Figure 2.4b the CME was solved directly. The SSA simulations were performed using the original version of the algorithm [Gil76] and illustrate the difficulties of estimating the probability distribution from independent realizations. Visually, it is evident that the probability computed using the CME solver is much smoother then the solution based on the SSA runs. Naturally, computing more realizations would improve the approximation, but this is computationally expensive. We remark that because of the discrete nature of the solution of the CME, the contour plots used in Figure 2.4 are somewhat misleading, since the probability distributions shown are only defined at discrete points $\mathbf{x} \in \mathbb{N}_0^d$. However, as such plots deliver a clear visualization of the solution profile, they will be used throughout this thesis.

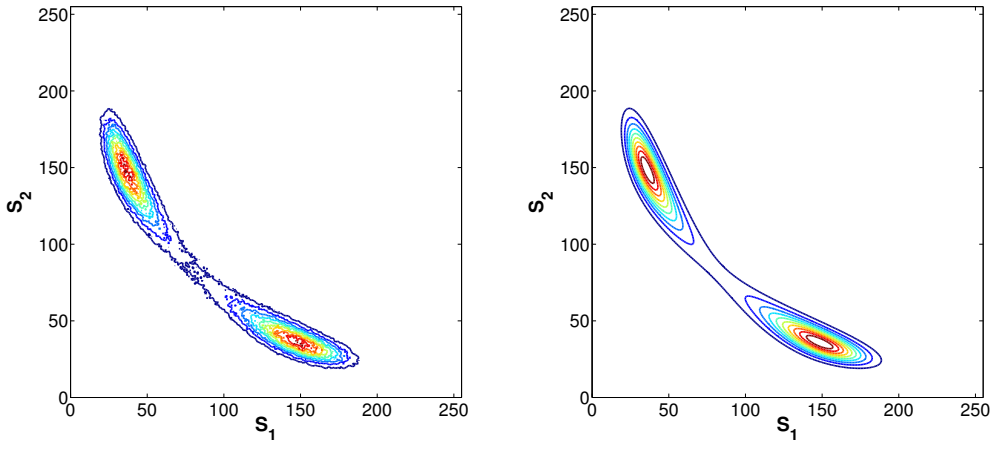


(a) SSA trajectory

(b) CLE path

(c) Close up of the realizations

Figure 2.3.: Comparison between the discrete (SSA) and continuous (CLE) stochastic models. Plot 2.3a shows a trajectory of the Markov jump process for one of the species (blue line), obtained using the SSA algorithm. After spending some time in the vicinity of one of the steady states, the system switches to the other steady state. The dotted black lines represent two solutions of the deterministic process with different initial conditions. Plot 2.3b depicts a path of the CLE (red line) with the dotted black lines again representing two solutions of the RRE. Figure 2.3c shows a close up of the two independent realizations taken from a time interval when both reside in the same attraction basin, with the approximate interval being marked by gray rectangles in 2.3a and 2.3b.



(a) SSA estimation

(b) CME solution

Figure 2.4.: Stationary probability distribution for the toggle switch model (2.94) with parameters (2.95). Left panel (2.4a): approximation of the probability distribution at $t = 2 \cdot 10^6$ (when the system has reached stationary state) obtained using $5 \cdot 10^5$ SSA trajectories. Right panel (2.4b): results obtained by approximating the solution of the stationary CME using a dedicated solver.

Besides the SSA approach, another option is to attempt to solve the truncated CME directly and some of the existing methods were mentioned in Chapter 1. The method that lies at the center of this thesis is based on wavelet compression, and we proceed with a short presentation of wavelet theory in order to set the stage for the construction of the wavelet-based numerical methods for the CME.

CHAPTER

THREE

WAVELET BASES

In the preceeding Chapters 1 and 2 we have introduced the stochastic modeling of chemical reaction kinetics and presented arguments that support a computational approach based on approximating the CME (2.67) directly. Moreover, we have established that the numerical treatment of the CME takes place by way of necessity on a high-dimensional state space Ω_ξ, defined by (2.65). Although finite, the truncated state space Ω_ξ is still huge, therefore special techniques are required to reduce the size of the CME problem to computationally manageable levels and an appealing idea is to use wavelet bases for this task. Accordingly, we concern ourselves in the present chapter with a brief overview of the properties and construction methods of some wavelet bases that can be employed for the numerical treatment of the CME on the truncated state space Ω_ξ. In doing so, we will forego the presentation of the more complex definitions from wavelet theory, and restrict the exposition to only a few well-known results that are needed to understand how our wavelet-based algorithms work. The finer points of wavelet theory and analysis, as well as the definitions presented in this chapter can be found in the monographs [Dau92, Coh03, Mal09, LMR98] or in the seminal articles [CDF92, Dah97, CDD01, DKU97].

The outline of the chapter is as follows. In Section 3.1 we present some of the basic properties common to all the wavelet constructions used in this thesis. Section 3.2 treats the concept of *multiresolution analysis* and presents the specific case of orthonormal wavelet bases. In Section 3.3 the multiresolution analysis concept is generalized to describe both the construction of biorthogonal wavelets on the real line and on bounded intervals. Finally, Section 3.4 deals with the use of wavelet bases on the high-dimensional state space Ω_ξ.

3.1. Basic properties

Let us begin with some basic properties shared by all wavelet bases presented in this chapter. In the following, we shall consider a separable Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and induced norm $|| \cdot ||_{\mathcal{H}}$. We remark that in the course of the exposition, $\mathcal{H}$ will

usually be taken to be the space $L_2(\Omega)$, with the domain Ω being subject to change. As a starting point we first consider Ω to be $\mathbb{R}$.

Translation and dilation. Traditionally, the elements of a wavelet basis for $L_2(\mathbb{R})$ have been constructed by applying translations and dilations onto two functions called *scaling function* $\varphi^{(m)} \in L^2(\mathbb{R})$, and *mother wavelet* $\psi^{(m)} \in L^2(\mathbb{R})$, respectively. The elements of the wavelet basis have then the following form,

$$\varphi_{j_0,k}^{(m)}(x) = 2^{j_0/2}\varphi^{(m)}(2^{j_0}x - k), \quad k \in \mathbb{Z} \tag{3.1}$$

$$\psi_{j,k}^{(m)}(x) = 2^{j/2}\psi^{(m)}(2^{j}x - k), \quad j = j_0, \ldots, j_{\max} - 1. \tag{3.2}$$

In the above equations (3.1) and (3.2), j refers to the resolution level or scale on which the basis element resides, and for all practical purposes ranges between a minimal resolution or "coarsest" scale j_0 and a "finest" scale denoted $j_{\max}$, while $k \in \mathbb{Z}$ is a translation parameter that reflects the location of the function on the domain Ω. For wavelets on the real line, all basis elements have the form (3.1)-(3.2), but as we shall see later, for wavelets on bounded domains this is no longer the case. Nevertheless, we remark that the *translation and dilation* property of the *scaling* and *mother wavelet* functions retains its usefulness also for wavelet bases on bounded domains, as most elements of such bases can still be constructed in this way. The exception is represented by the elements close to the boundaries of the domain. Next, using the notation introduced in (3.1)-(3.2) for the basis elements, we can define an (orthonormal) wavelet basis as the set of functions

$$\Psi = \left\{ \varphi_{j_0,k}^{(m)} \mid k \in \mathbb{Z} \right\} \cup \left\{ \psi_{j,k}^{(m)} \mid k \in \mathbb{Z}, \;\; j = j_0, \ldots, j_{\max} - 1 \right\} \subset L^2(\mathbb{R}) \tag{3.3}$$

with $j_0, j_{\max} \in \mathbb{N}$, and $m \in \mathbb{N} \setminus \{0\}$ denoting the order of the wavelet basis, i.e., all polynomials with a degree less than m can be represented exactly using this basis.

Local support. This is an important property of wavelet bases, especially in view of numerical realizations, which states that the support of the scaling function and the mother wavelet lies on a compact interval. This means that all basis elements have compact support that scales on a dyadic grid, i.e., $\operatorname{diam}(\operatorname{supp}(\psi_{j,k}^{(m)})) \sim 2^{-j}$ and $\operatorname{diam}(\operatorname{supp}(\varphi_{j,k}^{(m)})) \sim 2^{-j}$, respectively.

Vanishing moments. We have already mentioned this property, although not by name, having included the parameter m in the definition of the basis elements given in (3.1)-(3.2). For the *mother wavelet* $\psi^{(m)}(x)$ of order m, having vanishing moments means that

$$\int_{\mathbb{R}} x^n \psi^{(m)}(x) dx = 0$$

for all $n = 0, \ldots, m-1$. This is equivalent to saying that all polynomials with a degree less than m are exactly represented in the wavelet basis Ψ, by using the scaling functions $\varphi_{j_0,k}^{(m)}$. The *mother wavelet* $\tilde{\psi}^{(\tilde{m})}(x)$ of the dual basis $\tilde{\Psi}$ has the same property, but for an order $\tilde{m}$, possibly different from m. The dual basis $\tilde{\Psi}$ is defined as the set of dual functions

$$\tilde{\Psi} = \left\{ \tilde{\varphi}_{j_0,k}^{(\tilde{m})} \mid k \in \mathbb{Z} \right\} \cup \left\{ \tilde{\psi}_{j,k}^{(\tilde{m})} \mid k \in \mathbb{Z}, \;\; j = j_0, \ldots, j_{\max} - 1 \right\} \subset L^2(\mathbb{R}),$$

and forms a *biorthogonal system* with the basis Ψ (see (3.6)).

Wavelet Riesz bases. Before defining a *Riesz basis* for the Hilbert space $\mathcal{H}$, we generalize the concept of a wavelet basis as introduced in (3.3), and consider a countable family $\Psi = \{\psi_\lambda\}_{\lambda\in\mathcal{I}}$, with λ denoting a multi-index and $\mathcal{I}$ an ordered index set. For any such countable index set $\mathcal{I}$, let $\ell_2(\mathcal{I})$ be the space of sequences $c = \{c_\lambda\}_{\lambda\in\mathcal{I}}$ such that $||c||^2_{\ell_2(\mathcal{I})} := \sum_{\lambda\in\mathcal{I}} |c_\lambda|^2$ is finite. If the span of the family Ψ is dense in $\mathcal{H}$, and there exist two constants $C_1, C_2 \geq 0$ such that the norm equivalence

$$C_1||c||^2_{\ell_2(\mathcal{I})} \leq \Big\| \sum_\lambda c_\lambda \psi_\lambda \Big\|^2_{\mathcal{H}} \leq C_2||c||^2_{\ell_2(\mathcal{I})} \tag{3.4}$$

holds for all sequences $c = \{c_\lambda\}_{\lambda\in\mathcal{I}} \in \ell_2(\mathcal{I})$, then Ψ is a *Riesz* basis. We note that the constants $C_1,\ C_2$ are called *Riesz* constants and for the special case of an orthonormal basis we have $C_1 = C_2 = 1$. Moreover, sometimes we shall use the shorthand notation

$$||c||_{\ell_2(\mathcal{I})} \sim \Big\| \sum_\lambda c_\lambda \psi_\lambda \Big\|_{\mathcal{H}}$$

for (3.4). The significance of the *Riesz* basis property is the following: if Ψ is a *Riesz* basis for $\mathcal{H}$, then any $f \in \mathcal{H}$ has a unique representation as a linear combination of functions from the family Ψ, i.e.,

$$f = \sum_{\lambda\in\mathcal{I}} c_\lambda \psi_\lambda \tag{3.5}$$

with the coefficients in the expansion (3.5) being bounded on $\mathcal{H}$. Furthermore, for a *Riesz* basis Ψ there exists a unique family of dual functions $\tilde{\Psi} = \{\tilde{\psi}_\lambda\}_{\lambda\in\mathcal{I}}$, also a *Riesz* basis with constants $C_1^{-1},\ C_2^{-1}$, which is *biorthogonal* to Ψ, i.e., satisfies

$$\Big\langle \Psi, \tilde{\Psi} \Big\rangle_{\mathcal{H}} = I, \tag{3.6}$$

and moreover, we have $c_\lambda = \langle f, \tilde{\psi}_\lambda \rangle$. Consequently, we have that (3.5) can be reformulated as

$$f = \sum_{\lambda\in\mathcal{I}} \langle f, \tilde{\psi}_\lambda \rangle \psi_\lambda = \sum_{\lambda\in\mathcal{I}} \langle f, \psi_\lambda \rangle \tilde{\psi}_\lambda. \tag{3.7}$$

Note that in (3.6) we have used the short-hand notation

$$\Big\langle \Psi, \tilde{\Psi} \Big\rangle := \Big(\Big\langle \psi_{\lambda_i}, \tilde{\psi}_{\lambda_j} \Big\rangle \Big)_{\psi_{\lambda_i}\in\Psi, \tilde{\psi}_{\lambda_j}\in\tilde{\Psi}}, \quad \Big\langle \psi_{\lambda_i}, \tilde{\psi}_{\lambda_j} \Big\rangle = \delta_{i,j},$$

and if we have biorthogonality for the dual system Ψ and $\tilde{\Psi}$, then this can be used to show that Ψ is a *Riesz* basis (see [Coh03, Theorem 2.6.1]). For the orthonormal case we have of course $\psi_\lambda = \tilde{\psi}_\lambda$. The relevance of the Riesz basis property is that it provides a tight connection from the function norm to the discrete coefficient norm ℓ_2, meaning that small changes in the coefficients trigger small changes in the function and vice-versa (cf. [Dah97]), which is beneficial in terms of the stability of wavelet representations. Using the Riesz basis property, we can also distinguish between stability, uniform stability and stability over all levels, the last being equivalent to the Riesz basis property for Ψ. A family of vectors is stable if it has the *Riesz* basis property for its closed linear span, while a family $\{\psi_{j,k}\}_{j,k}$ is uniformly stable if each set $\{\psi_{j,k}\}_k$ is stable with constants that are independent of j.

At this point, it might be useful to place these properties into context. Given a function $f \in \mathcal{H}$ representing an object of interest and a basis Ψ, the hope is that most features of f can be captured by using as few of the coefficients $c = \{c_\lambda\}_{\lambda \in \mathcal{I}}$ from the expansion (3.5) as possible. As the elements of the wavelet basis Ψ are by construction *locally* supported, one advantage of using the wavelet representation (3.5) is that such a basis is better at reflecting the local behavior of an arbitrary function. Furthermore, a consequence of having vanishing moments is that the inner products between basis elements ψ_λ and sufficiently smooth functions decay exponentially fast when the scale j tends to infinity (for some estimates in terms of function smoothness see [Coh03, Chapter 3]). Thus, good spatial localization and *vanishing moments* combine to allow a high-accuracy representation of a function using only a reduced set of coefficients, by simply ignoring all coefficients that are smaller than a prescribed tolerance. Moreover, due to the *Riesz* basis property we have that the expansion in the chosen basis is stable. Last but not least, via the translation and dilation property, wavelet bases can be used to perform hierarchical decompositions by representing functions in terms of contributions on different scales. This leads directly to *adaptivity* in the sense that wavelet-based numerical approaches can concentrate the computational effort where it is mostly needed.

3.2. Multiresolution analysis

We proceed now to introduce an essential aspect of wavelet theory that clarifies the previous statement about hierarchical decomposition, namely the concept of *multiresolution analysis*. We adapt the following definition from [Mal09, Chapter 7], and remark that we treat first the orthonormal case, as our emphasis is on the construction of wavelet bases belonging to the Daubechies family [Dau92].

Definition 3.1. *A multiresolution analysis (MRA) is a sequence $\{\mathcal{S}_j\}_{j\in\mathbb{Z}}$ of closed subspaces of $L_2(\mathbb{R})$ with the following properties:*

(i) The spaces are nested and dense in $L_2(\mathbb{R})$, i.e,

$$\mathcal{S}_j \subset \mathcal{S}_{j+1}, \forall j \in \mathbb{Z} \text{ and } \operatorname{clos}_{L_2(\mathbb{R})}\left(\bigcup_{j\in\mathbb{Z}} \mathcal{S}_j\right) = L_2(\mathbb{R})$$

(ii) $\forall j \in \mathbb{Z}$, $f(x) \in \mathcal{S}_j \Leftrightarrow f(2x) \in \mathcal{S}_{j+1}$

(iii) There is a scaling function $\varphi \in \mathcal{S}_0$ such that $\{\varphi(x-k)\}_{k\in\mathbb{Z}}$ is a Riesz basis of $\mathcal{S}_0$.

(iv) The space $\mathcal{S}_j$ is translation invariant, i.e.,

$$\forall j \in \mathbb{Z}, f(x) \in \mathcal{S}_j \Leftrightarrow f(x - 2^j k) \in \mathcal{S}_j.$$

(v) The spaces $\{\mathcal{S}_j\}_{j\in\mathbb{Z}}$ satisfy

$$\bigcap_{j\in\mathbb{Z}} \mathcal{S}_j = \{0\}.$$

The nestedness property *(i)* of the sequence $\{\mathcal{S}_j\}_{j\in\mathbb{Z}}$ describes how the spaces $\mathcal{S}_j$ provide increasingly better approximations to a function f and the dense union ensures that pushing the resolution level $j \to \infty$ means that we can approximate arbitrarily well any function from $L_2(\mathbb{R})$. The approximation of a function at a specific resolution is formally defined as an orthogonal projection on the space $\mathcal{S}_j$. If we let $P_j : L_2(\mathbb{R}) \to \mathcal{S}_j$ denote such an orthogonal projector onto $\mathcal{S}_j$, then dense union yields

$$\lim_{j\to+\infty} ||f - \mathcal{P}_j f||_{L_2(\mathbb{R})} = 0.$$

Conversely, property *(v)* implies that if we let the resolution level go to 0, then we loose all the details from the approximation of f in $\mathcal{S}_j$, i.e.,

$$\lim_{j\to-\infty} ||\mathcal{P}_j f||_{L_2(\mathbb{R})} = 0.$$

Property *(ii)* relates the spaces $\mathcal{S}_j$ and $\mathcal{S}_{j+1}$ by stating that a function moves from $\mathcal{S}_j$ to $\mathcal{S}_{j+1}$ when rescaled on a dyadic grid. Further, for all $j \in \mathbb{Z}$, the spaces $\mathcal{S}_j$ are generated by the scaling function $\varphi^{(m)} \in L_2(\mathbb{R})$ via translations and dilations, such that

$$\Phi_j = \{\varphi_{j,k}^{(m)} := 2^{j/2}\varphi^{(m)}(2^j x - k) \mid k \in \mathbb{Z}\} \tag{3.8}$$

is a *Riesz* basis for $\mathcal{S}_j = \overline{\text{span}(\Phi_j)}$.

An important consequence of the nestedness of the spaces $\mathcal{S}_j$ is that the scaling function is *refinable*, which means that it satisfies

$$\varphi^{(m)}(x) = \sum_{k\in\mathbb{Z}} h_k \varphi^{(m)}(2x - k), \tag{3.9}$$

where the coefficients $\{h_k\}_{k\in\mathbb{Z}}$ are called *masks* or filter coefficients. The equation (3.9) is the *refinement equation* of the scaling function $\varphi^{(m)} \in L_2(\mathbb{R})$, and we remark that compact support for the scaling function translates into finite coefficient filters. For the Daubechies family of orthonormal wavelet bases we consider here, we have $h_k \neq 0$ for $k \in \{0, \ldots, 2m-1\}$, with m denoting the order of the scaling function. There are no formulas for computing the scaling functions if $m > 1$, but the values of $\varphi^{(m)}$ at any level can be computed with the *cascade algorithm*, which starting from known values at certain points, computes the desired values by iteratively applying the refinement equation (3.9).

By construction, the spaces $\mathcal{S}_j$ regroup all possible approximations of a function $f \in L_2(\mathbb{R})$ at the corresponding resolutions 2^{-j}. Assuming that we have an approximation at an arbitrary scale j, and we want to switch to an approximation at a finer scale $j+1$, it is advantageous to define a *complement* space $\mathcal{W}_j$ that describes the local differences between the two approximations in the nested spaces $\mathcal{S}_j$ and $\mathcal{S}_{j+1}$. We have then the decomposition

$$\mathcal{S}_{j+1} = \mathcal{S}_j \oplus \mathcal{W}_j, \tag{3.10}$$

which simply means that for any $f \in \mathcal{S}_{j+1}$, there exists an unique $u \in \mathcal{S}_j$ and $v \in \mathcal{W}_j$ such that $f = u + v$. Basically, the function f is split into a "coarse" part u representing the average of f, and a "detail" part v. Next, using the *mother wavelet* $\psi^{(m)}$ we construct a *Riesz* basis

$$\Psi_j := \{\psi_{j,k}^{(m)} := 2^{j/2}\psi^{(m)}(2^j x - k) \mid k \in \mathbb{Z}\} \tag{3.11}$$

for $\mathcal{W}_j = \overline{\text{span}(\Psi_j)}$.

Analogously to the scaling function, the *mother wavelet* $\psi^{(m)} \in L_2(\mathbb{R})$ also satisfies a refinement equation

$$\psi^{(m)}(x) \quad = \quad \sum_{k\in\mathbb{Z}} g_k \varphi^{(m)}(2x-k) \tag{3.12}$$

with a different *mask* $\{g_k\}_{k\in\mathbb{Z}}$. The compact support of wavelets is again equivalent with the mask having finitely many non-zero elements. For the Daubechies orthonormal wavelets we have $g_k = (-1)^k h_{1-k} \neq 0$, with $k = \{2-2m, \ldots, 1\}$.

Now, by applying the decomposition (3.10) recursively, we obtain a hierarchical decomposition of any space $\mathcal{S}_{j+1}$ as

$$\mathcal{S}_{j+1} = \mathcal{S}_j \oplus \mathcal{W}_j = \mathcal{S}_{j_0} \oplus \mathcal{W}_{j_0} \oplus \mathcal{W}_{j_0+1} \oplus \cdots \oplus \mathcal{W}_j,$$

which means that any function $f \in L_2(\mathbb{R})$ has an expansion in a multi-scale basis,

$$\Psi = \Phi_{j_0} \cup \bigcup_{j\in\mathbb{Z}} \Psi_j. \tag{3.13}$$

However, because numerical computations can only be performed on a bounded interval, we assume that the support of any function f is finite. Consequently, we can replace $j \in \mathbb{Z}$ in (3.13) with $j \in \{j_0, \ldots, j_{\max}-1\}$ where $j_{\max}$ denotes the original resolution of the function f. In doing so, equation (3.13) becomes a more compact version of (3.3). Using now the orthogonal projectors $\mathcal{P}_j$, it follows naturally that for every $f \in \mathcal{S}_{j_{\max}} \subset L_2(\mathbb{R})$, we have a representation in terms of contributions on different scales,

$$f = \mathcal{P}_{j_{\max}} f = \mathcal{P}_{j_0} f + \sum_{j=j_0}^{j_{\max}-1} (\mathcal{P}_{j+1} - \mathcal{P}_j) f. \tag{3.14}$$

The term

$$\mathcal{P}_{j_0} f \quad = \quad \sum_{k\in\mathbb{Z}} c^{(m)}_{j_0,k} \varphi^{(m)}_{j_0,k} \in \mathcal{S}_{j_0}, \tag{3.15}$$

approximates the function on the coarsest scale , whereas the terms

$$(\mathcal{P}_{j+1} - \mathcal{P}_j) f \quad = \quad \sum_{k\in\mathbb{Z}} d^{(m)}_{j,k} \psi^{(m)}_{j,k} \in \mathcal{W}_j \tag{3.16}$$

represent the detail information between successive spaces $\mathcal{S}_j$ and $\mathcal{S}_{j+1}$. A graphical representation of how multiresolution analysis works is given in Figure 3.1.

Using (3.15) and (3.16) in (3.14) we obtain that every $f \in \mathcal{S}_{j_{\max}} \subset L_2(\mathbb{R})$ has the representation

$$f \quad = \quad \sum_{k\in\mathbb{Z}} c^{(m)}_{j_0,k} \varphi^{(m)}_{j_0,k} + \sum_{j=j_0}^{j_{\max}-1} \sum_{k\in\mathbb{Z}} d^{(m)}_{j,k} \psi^{(m)}_{j,k} \tag{3.17}$$

with the coefficients given as $c^{(m)}_{j_0,k} = \langle f, \varphi^{(m)}_{j_0,k} \rangle_{L_2}$ and $d^{(m)}_{j,k} = \langle f, \psi^{(m)}_{j,k} \rangle_{L_2}$. The formulas for the coefficients are derived from the definitions of the orthogonal projectors.

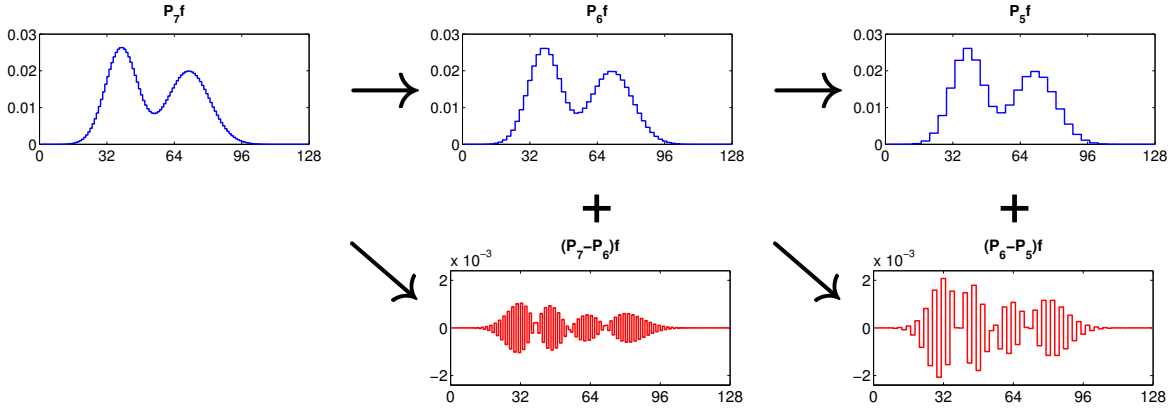


Figure 3.1.: Multi-scale decomposition via orthogonal projections using an orthonormal wavelet basis with $m = 1$ (Haar). The plots in the first row show successive approximations of the function f in increasingly coarser spaces $\mathcal{S}_j$, while in the second row the details from the complement spaces $\mathcal{W}_j$ are shown.

However, there is no need to compute the inner products because the coefficients of the wavelet expansion (3.17) can be efficiently obtained by exploiting the refinement equations (3.9) and (3.12) which give rise to a change of basis

$$\sum_k c_{j+1,k}^{(m)} \varphi_{j+1,k}^{(m)} = \sum_k c_{j,k}^{(m)} \varphi_{j,k}^{(m)} + \sum_k d_{j,k}^{(m)} \psi_{j,k}^{(m)}.$$

This mapping between the coefficients on different scales $\{c_{j+1,k}\}_k \mapsto \{c_{j,k}\}_k \cup \{d_{j,k}\}_k$, with

$$c_{j,k} = \sum_n h_n c_{j+1,n}, \quad d_{j,k} = \sum_n g_n c_{j+1,n},$$

is called a one-step wavelet transform and iterating this basic building block we obtain the *fast wavelet transform* (FWT). Conversely, reconstructing a function from its wavelet coefficients can be accomplished by using the inverse mapping $\{c_{j,k}\}_k \cup \{d_{j,k}\}_k \mapsto \{c_{j+1,k}\}_k$ to build the *(fast) inverse wavelet transform* (FIWT). We remark that both one-step wavelet transforms between successive levels j and $j+1$ perform 2^j operations, so the entire computational effort for the FWT and FIWT respectively, is $\sum_{j=j0}^{j_{\max}} 2^j = \mathcal{O}(N)$ where $N = 2^{j_{\max}}$ is the length of the original representation of the function f which we assumed had finite support (cf. [Coh03]).

As their name implies, the elements of orthonormal wavelet bases satisfy a number of orthogonality conditions. For all $i, j \in \{j_0, \ldots, j_{\max} - 1\}$ and $k, l \in \mathbb{Z}$, we have

$$\left\langle \varphi_{j_0,k}^{(m)}, \varphi_{j_0,l}^{(m)} \right\rangle_{L^2} = \delta_{k,l}, \quad \left\langle \psi_{j,k}^{(m)}, \psi_{i,l}^{(m)} \right\rangle_{L^2} = \delta_{j,i}\delta_{k,l}, \quad \left\langle \varphi_{j_0,k}^{(m)}, \psi_{i,l}^{(m)} \right\rangle_{L^2} = 0, \tag{3.18}$$

with $\delta_{k,l}$ denoting the Kronecker symbol. As a parenthesis, we remark that although a desirable property, the requirements to fulfill *orthonormality* by satisfying (3.18) are for some types of wavelet constructions on bounded intervals out of reach, therefore these constructions which will be discussed shortly satisfy only *biorthogonality* conditions as given in (3.6).

The reason why the representation given in (3.17) is more *compact* than the one in the canonical basis is because it exploits smoothness. If $\mathcal{P}_{j_{\max}} f$ is sufficiently smooth, the

function is already well approximated by coefficients on the coarsest scales, so many of the coefficients $d_{j,k}^{(m)}$ for large values of j corresponding to the detail information in the terms $(\mathcal{P}_{j+1} - \mathcal{P}_j)f$ are close to zero and thus can be dropped without incurring a major loss in accuracy for the approximation. A mathematically more rigorous estimate of how fast the wavelet coefficients decay in regions where f is smooth is given for the special case of the orthonormal wavelet basis with $m = 1$ in [Coh03, Remark 1.5.1] as

$$|d_{j,k}| \leq \sup_{x \in I_{j,k}} |f'(x)| 2^{-3j/2}. \tag{3.19}$$

In (3.19), f is taken to be differentiable with continuous derivative, i.e., $f \in C^1(I_{j,k})$ on the support $I_{j,k} = \text{supp}(\psi_{j,k})$. Using this result, a characterization of wavelet compression in terms of the smoothness of f can be given for the case $m = 1$ as

$$||f - \mathcal{P}_j f||_{L_2} \leq C 2^{-j} ||f'(x)||_{L_\infty} \tag{3.20}$$

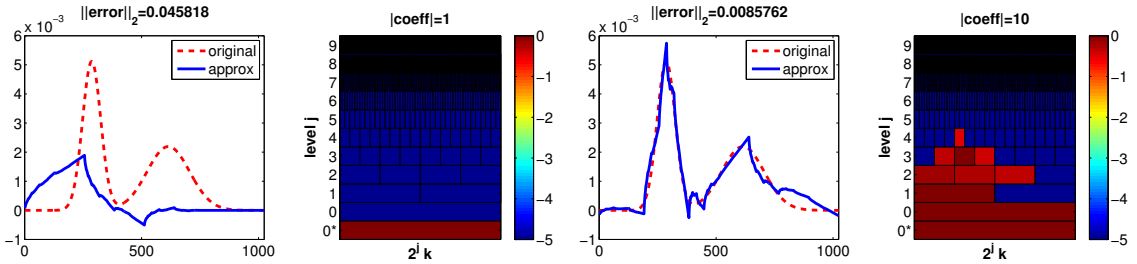
(cf. [Coh03, Remark 1.5.2]). For the general case $f \in C_0^n(\mathbb{R})$ with $n \in \{1, \ldots, m\}$, we have that the projection error is bounded for all $j \in \{j_0, \ldots, j_{\max}\}$ by

$$||f - \mathcal{P}_j f||_{L_2} \leq C 2^{-nj} ||f^{(n)}(x)||_{L_2}. \tag{3.21}$$

The estimate (3.21) depends on the smoothness of the function f, the truncation level j and the order of the wavelet m, as it holds only for $n \leq m$ (cf. [JU10]). Estimates of the type (3.21) are called *direct* or Jackson type estimates and they measure the compression power of the nested sequences of spaces $\mathcal{S}_j$ when $j \to \infty$. We also remark that similar estimates have been shown under weaker regularity assumptions (see [Coh03, Chapter 3]). In practical terms, these estimates give assurances that even if a relatively large number of coefficients are discarded, the resulting approximation is still reasonably accurate. This fact is illustrated in Figure 3.2, where we test wavelet compression on a smooth function f with finite support. We suppose that a wavelet representation of f which contains 1024 coefficients is available. The function is then reconstructed by using 1, 10 , 21 and 63 of the largest coefficients in absolute value. Each of the four panels is divided in a left plot showing the original function and its wavelet approximation and a right plot containing a wavelet scalogram used to visualize the location and magnitude of the wavelet coefficients. The scalogram is made up of building blocks that represent the scaling coefficients (level j_0*), and wavelet coefficients on level j_0 through $j_{\max} - 1$. Note that each wavelet level has twice as many coefficients as the previous "coarser" level and their magnitude is color coded according to the adjacent color-bar. For this particular example, an orthonormal Daubechies wavelet basis with $m = 2$ was used (see Section 3.2.1 for details).

3.2.1. Examples: orthonormal Daubechies wavelets

We present now some examples of orthonormal wavelet bases on $L_2(\mathbb{R})$ which will be used in some of the numerical examples presented later in this thesis. From a historical point of view, the first wavelet construction can be traced to the dissertation of Alfred Haar [Haa10]. The Haar wavelet system uses as scaling function $\varphi^{(m)}(x) = \chi_{[0,1)}(x)$, with χ_I denoting the characteristic function on the interval I. The oscillatory parts of a function f are represented with the help of the *Haar* mother wavelet $\psi^{(m)}(x) = \chi_{[0,1/2)} -$



(a) 1 largest coefficient (b) 10 largest coefficients

(c) 21 largest coefficients (d) 63 largest coefficients

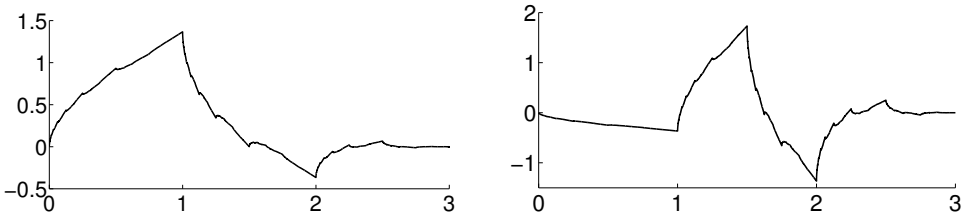
Figure 3.2.: Wavelet compression of a smooth function using different numbers of the largest wavelet coefficients. The approximation uses the *Daubechies db*2 orthonormal wavelet basis. There are a total of 1024 coefficients in the wavelet expansion of the original function.

$\chi_{[1/2,1)}$. The corresponding finite filter coefficients appearing in the refinement equations (3.9) and (3.12) are $h = \{1, 1\}$ and $g = \{1, -1\}$, respectively. The average spaces $\mathcal{S}_j$ spanned by the corresponding scaling functions Φ_j contain all piecewise constant functions on the intervals $2^{-j}k + [0, 2^j)$, $k \in \mathbb{Z}$ and we remark that the order of the Haar wavelet is $m = 1$. Because the *Haar* system is able to represent only piecewise constant functions exactly, its approximation power is somewhat limited. However, the compression properties can be arbitrarily improved by selecting a wavelet basis with $m > 1$, and the orthonormal wavelet bases we have defined in (3.3), are collectively known as Daubechies wavelets since their discovery in 1992 by Ingrid Daubechies [Dau92]. However, for these wavelet bases, the scaling function and the mother wavelet have no explicit formula and the filter coefficients are found indirectly, see [Dau92] for details. For the case $m = 2$, the scaling and mother wavelet functions are shown in Figure 3.3, and we remark that the compact support of both functions is larger than the compact support of the Haar system. For the $db2$ wavelet basis, the scaling function and mother wavelet are zero outside of the interval $[0, 3]$. As a rule, increasing the order m of the wavelet, also increases the compact support and the number of non-zero coefficients in the filter masks. The filter coefficients for $db2$ are

$$h_0 = \frac{1+\sqrt{3}}{4\sqrt{2}}, \quad h_1 = \frac{3+\sqrt{3}}{4\sqrt{2}}, \quad h_2 = \frac{3-\sqrt{3}}{4\sqrt{2}}, \quad h_3 = \frac{1-\sqrt{3}}{4\sqrt{2}} \quad \text{and} \quad g_n = (-1)^n h_{1-n}.$$

We remark that the Daubechies wavelet family is widely used, and the masks for the cases in which $m > 2$ can be looked up in the literature (see [Dau92, Mal09]).

The orthonormal Daubechies wavelet bases family described by (3.3) can be used successfully to solve the CME numerically, and examples in this sense will be presented in the next chapter. However, these wavelet bases have been defined on $\mathbb{R}$, and for numerical computation we need to represent functions on bounded intervals. There are two



(a) (b)

Figure 3.3.: Scaling function (3.3a) and mother wavelet (3.3b) of the *Daubechies db2* wavelet ($m = 2$)

ways to achieve this outcome. Either the target function is extended to $\mathbb{R}$ by periodic continuation and thus the wavelet bases described above can be employed directly, or special constructions that produce wavelet bases already adapted to the interval are used. Constructing compactly supported orthonormal wavelet bases on the interval is possible, and such a construction is detailed for example in [CDV93]. Another option is to renounce orthonormality and use biorthogonal wavelet bases. As some of the algorithms presented in Chapter 5 will use biorthogonal wavelets on the interval, we proceed to extend the concept of multiresolution analysis to the biorthogonal case.

3.3. Biorthogonal multiresolution analysis

As seen in the previous section, the construction of wavelet bases for a Hilbert space $\mathcal{H}$ uses the concept of multiresolution analysis. The orthonormal wavelet bases introduced so far have been defined for the space $L_2(\mathbb{R})$, and a similar viewpoint can also be taken with respect to the biorthogonal multiresolution analysis. Indeed, this was also the setting used in the original construction of biorthogonal wavelets by Cohen, Daubechies and Feauveau in [CDF92]. However, our purpose for the introduction of biorthogonal wavelets is to construct bases on bounded domains and such a setting clearly precludes the use of the techniques based on translation invariance that are employed on $\mathbb{R}$. Therefore, we adopt a presentation style also used in e.g. [Dah97, DKU97], that generalizes the multiresolution concept to the task of building wavelets on bounded domains $\Omega \subseteq \mathbb{R}$.

Despite the new setting, we remark that most of the basic principles already reviewed can still be employed. Analogously to the Definition 3.1, the starting point for a biorthogonal multiresolution analysis is to consider two sequences of nested subspaces $\{\mathcal{S}_j\}_{j\in\mathbb{Z}}$, $\{\tilde{\mathcal{S}}_j\}_{j\in\mathbb{Z}}$ dense in $L_2(\Omega)$,

$$\mathcal{S}_j \subset \mathcal{S}_{j+1}, \forall j \in \mathbb{Z} \text{ and } \operatorname{clos}_{L_2(\Omega)}\left(\bigcup_{j\in\mathbb{Z}} \mathcal{S}_j\right) = L_2(\Omega) \tag{3.22}$$

$$\tilde{\mathcal{S}}_j \subset \tilde{\mathcal{S}}_{j+1}, \forall j \in \mathbb{Z} \text{ and } \operatorname{clos}_{L_2(\Omega)}\left(\bigcup_{j\in\mathbb{Z}} \tilde{\mathcal{S}}_j\right) = L_2(\Omega). \tag{3.23}$$

Then, the spaces in (3.22) are generated by the bases $\Phi_j := \{\varphi_{j,k}^{(m)}, k \in \Delta_j\}$ such that we have $\mathcal{S}_j = \overline{\operatorname{span}(\Phi_j)}$ and likewise for the spaces in (3.23), we have dual bases $\tilde{\Phi}_j = \{\tilde{\varphi}_{j,k}^{(\tilde{m})}, k \in \Delta_j\}$ such that $\tilde{\mathcal{S}}_j = \overline{\operatorname{span}(\tilde{\Phi}_j)}$. We remark that $\Delta_j \subseteq \mathbb{Z}$ denotes the index set of

the basis elements on level j and for the case $\Omega = \mathbb{R}$, Δ_j will be an infinite set. The elements of the bases Φ_j and $\tilde{\Phi}_j$ are called *primal scaling functions*, and *dual scaling functions* respectively, and for the case $\Omega = \mathbb{R}$, are all generated using dilations and translations of a scaling function $\varphi^{(m)} \in L_2(\Omega)$ and a dual scaling function $\tilde{\varphi}^{(\tilde{m})} \in L_2(\Omega)$, both with compact support and approximation order m for the primal basis and $\tilde{m}$ for the dual basis, respectively. Finally, the sequences $\{\mathcal{S}_j\}_{j\in\mathbb{Z}}$, $\{\tilde{\mathcal{S}}_j\}_{j\in\mathbb{Z}}$ form a biorthogonal multiresolution of $L_2(\Omega)$ if the condition

$$\left\langle \varphi_{j,k}^{(m)}, \tilde{\varphi}_{j,l}^{(\tilde{m})} \right\rangle = \delta_{k,l}, \quad \forall \varphi_{j,k}^{(m)} \in \mathcal{S}_j, \quad \forall \tilde{\varphi}_{j,k}^{(\tilde{m})} \in \tilde{\mathcal{S}}_j \tag{3.24}$$

holds. For the case $\Delta_j = \mathbb{Z}$, we have via biorthogonality (3.24) and compact support, that the bases Φ_j are uniformly stable, i.e.,

$$||c||_{\ell_2(\Delta_j)} \sim || \sum_{k\in\Delta_j} c_{j,k}\varphi_{j,k}||_{L_2(\Omega)}, \quad \forall c \in l_2(\Delta_j),$$

with a similar relation being satisfied for the dual bases $\tilde{\Phi}_j$ (cf. [Dah97, Remark 5]). For bounded domains Ω however, the index sets Δ_j are finite and furthermore the bases Φ_j and $\tilde{\Phi}_j$ are no longer composed only of the translates and dilates of single functions. As a consequence, the uniform stability requires additional conditions (cf. [Dah97, Remark 7]).

Further, the nestedness of $\mathcal{S}_j \subset \mathcal{S}_{j+1}$ implies that the functions $\varphi_{j,k}$ are refineable and filter masks exist such that, viewing the set Φ_j as a column vector that contains the functions $\varphi_{j,k}$, we have

$$\Phi_j = M_{j,0}^T \Phi_{j+1}, \quad j \geq j_0. \tag{3.25}$$

Note that equation (3.25) is just a more compact notation of the refinement property from (3.9), and $M_{j,0} \in \mathbb{R}^{|\Delta_{j+1}|\times|\Delta_j|}$ is a refinement matrix with its k-th column containing the expansion coefficients $\{h_k\}_{k\in\Delta_j}$ of $\varphi_{j,k}$ with respect to functions on the next finer scale $j+1$. Because of compact support we have that filter masks have finitely many entries that are non-zero, and consequently the matrix $M_{j,0}$ is sparse. For the case in which $\Delta_j = \mathbb{Z}$, and we operate in the translation invariant setting, the matrix $M_{j,0}$ is bi-infinite, has a banded structure, and its size is independent of the level j. For wavelet constructions on bounded domains, the structure of the refinement matrices $M_{j,0}$ is more complicated. In such cases the refinement matrices feature two non-standard edge blocks M_L, M_R, which are independent of the refinement level, and away from the edges the matrix is again banded. The block structure of the refinement matrix is shown in Figure 3.4. The middle

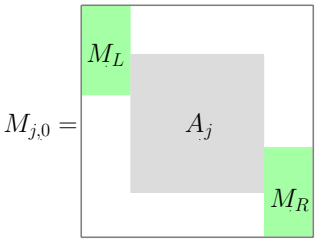
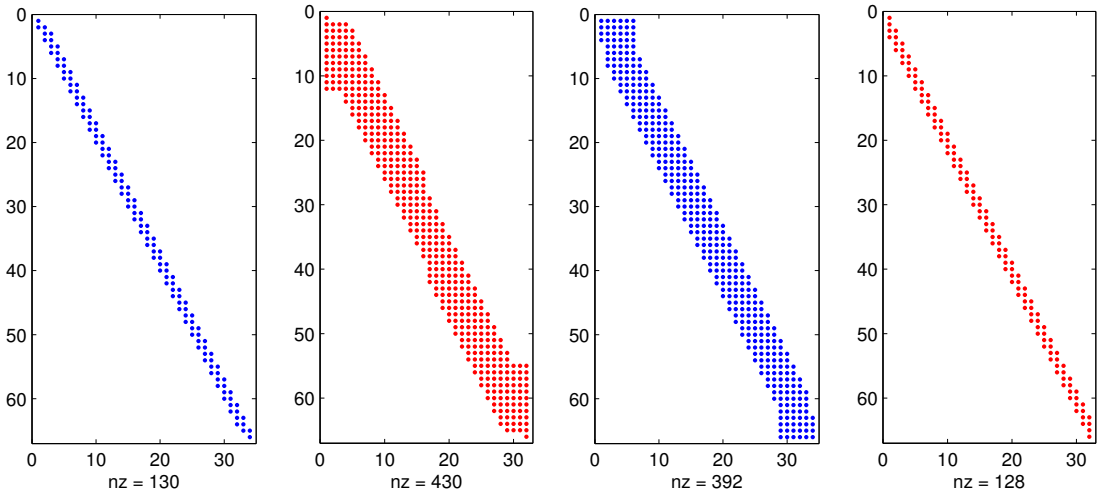


Figure 3.4.: Block structure of refinement matrix $M_{j,0}$

block A_j is a circular matrix that depends on the level j. The size of $M_{j,0}$ depends now

on the refinement level j, but the matrices are still sparse and the number on non-zero elements is uniformly bounded on j. In Figure (3.5a) we visualize the sparsity pattern for the particular case of the construction of biorthogonal wavelets on $[0, 1]$ from [Pri09].



(a) $M_{j,0}$ (b) $M_{j,1}$ (c) $\tilde{M}_{j,0}$ (d) $\tilde{M}_{j,1}$

Figure 3.5.: Sparsity pattern of the scaling and wavelet refinement matrices $M_{j,0}, M_{j,1}, \tilde{M}_{j,0}, \tilde{M}_{j,1}$ for the interval wavelet basis from [Pri09] with $m = 3$, $\tilde{m} = 5$ and $j = 5$. The refinement matrices for the scaling functions are colored blue, while the refinement matrices for wavelets use the red color.

Obviously, for the dual basis $\tilde{\Phi}_j$ with coefficient filter $\{\tilde{h}_k\}_{k\in\Delta_j}$ we can define a refinement matrix $\tilde{M}_{j,0} \in \mathbb{R}^{|\Delta_{j+1}|\times|\Delta_j|}$ such that

$$\tilde{\Phi}_j = \tilde{M}_{j,0}^T \tilde{\Phi}_{j+1}, \quad j \geq j_0. \tag{3.26}$$

From the biorthogonality condition (3.24) satisfied by the pair of bases Φ_j and $\tilde{\Phi}_j$, we have that the two refinement matrices $M_{j,0}$ and $\tilde{M}_{j,0}$ satisfy $M_{j,0}^T \tilde{M}_{j,0} = I$.

After constructing the biorthogonal bases Φ_j and $\tilde{\Phi}_j$ we are ready to introduce the next ingredient of multiscale decompositions, namely a pair of dual wavelet bases. This is accomplished by defining two complement spaces $\mathcal{W}_j$ and $\tilde{\mathcal{W}}_j$ of $\mathcal{S}_j$ and $\tilde{\mathcal{S}}_j$, such that we have the decompositions

$$\mathcal{S}_{j+1} = \mathcal{S}_j \oplus \mathcal{W}_j, \quad \tilde{\mathcal{S}}_{j+1} = \tilde{\mathcal{S}}_j \oplus \tilde{\mathcal{W}}_j, \quad j \geq j_0. \tag{3.27}$$

Additionally, the spaces $\mathcal{W}_j$ and $\tilde{\mathcal{W}}_j$ must satisfy the following orthogonality conditions

$$\tilde{\mathcal{W}}_j \perp \mathcal{S}_j, \quad \mathcal{W}_j \perp \tilde{\mathcal{S}}_j, \quad j \geq j_0 \tag{3.28}$$

where by $\perp$ we denote orthogonality with respect to the $L_2(\Omega)$ norm. From (3.27) and (3.28) we get that the two complement spaces $\mathcal{W}_j$ and $\tilde{\mathcal{W}}_j$ are uniquely determined and constructing the biorthogonal wavelets reduces to finding the bases

$$\Psi_j = \{\psi_{j,k} | k \in \nabla_j\}, \quad \tilde{\Psi}_j = \{\tilde{\psi}_{j,k} | k \in \nabla_j\},$$

with $\nabla_j = \Delta_{j+1} \setminus \Delta_j$ an appropriate index set. Ψ_j and $\tilde{\Psi}_j$ must be *Riesz* basis for $W_j = \overline{\operatorname{span}(\Psi_j)}$ and $\tilde{W}_j = \overline{\operatorname{span}(\tilde{\Psi}_j)}$, respectively. Additionally, they must fulfill the biorthogonality condition

$$\left\langle \Psi_j, \tilde{\Psi}_j \right\rangle = I. \tag{3.29}$$

The complete biorthogonal bases Ψ and $\tilde{\Psi}$ for the space $L_2(\Omega)$ are then obtained in the usual manner, by applying the space decompositions (3.27) recursively, which leads to the collections

$$\Psi = \{\psi_\lambda \mid \lambda \in \mathcal{I}\} = \Phi_{j_0} \cup \bigcup_{j=j_0}^{\infty} \Psi_j \quad \text{and} \quad \tilde{\Psi} = \{\tilde{\psi}_\lambda \mid \lambda \in \mathcal{I}\} = \tilde{\Phi}_{j_0} \cup \bigcup_{j=j_0}^{\infty} \tilde{\Psi}_j,$$

where $\lambda = (j, k, l)$ is a multi-index the aggregates the relevant information, i.e., level, location and type of basis element (0 for scaling functions on the coarsest level, and 1 for wavelets on levels $j \geq j_0$). By $\mathcal{I}$ we have denoted the corresponding index set, defined as $\mathcal{I} = j_0 \times \Delta_{j_0} \times \{0\} \cup \bigcup_{j=j_0}^{\infty} j \times \nabla_j \times \{1\}$.

It has been shown in [Dah97] that if the primal and dual scaling bases Φ_j and $\tilde{\Phi}_j$ have certain approximation properties, i.e., they allow the reproduction of polynomials of orders m and $\tilde{m}$, respectively and furthermore, biorthogonality, stability and compact support are also satisfied, then the bases Ψ and $\tilde{\Psi}$ are *Riesz* basis for the space $L_2(\Omega)$.

3.3.1. Wavelet bases on the interval

Previously we have glossed over the issue of how exactly the wavelet bases Ψ_j and $\tilde{\Psi}_j$ are to be computed on the interval. For the construction of biorthogonal wavelet basis on $\mathbb{R}$, Fourier based techniques are available (see [CDF92]). However, in case of bounded domains, this strategy can no longer be used. Most of the works related to the construction of wavelets on the interval $[0, 1]$ (e.g. [DKU97, Pri09, Dij09]) follow another strategy. The first step is the construction of the scaling bases which preserve the properties required to obtain a *Riesz* basis, as outlined in the previous section. There are several choices available, but the guiding idea is to preserve as much as possible of the mechanisms that are used in the translation invariant setting on $\mathbb{R}$. Of course, this is not possible near the boundaries of the interval, where special scaling functions that preserve the order of polynomial replication need to be constructed. This has to be done for the primal and the dual scaling bases, after which the resulting bases must be biorthogonalized as they are no longer a dual system as a consequence of the modifications operated in the first step. After constructing these suitable biorthogonal bases Φ_j and $\tilde{\Phi}_j$, we can proceed with the construction of the wavelet bases Ψ_j and $\tilde{\Psi}_j$ proper.

Since the basis elements of Ψ_j are contained in $\mathcal{S}_{j+1}$ due to (3.27), we have by way of the refinement property that there exists a matrix $M_{j,1} \in \mathbb{R}^{|\Delta_{j+1}| \times |\nabla_j|}$ such that

$$\Psi_j = M_{j,1}^T \Phi_{j+1}, \tag{3.30}$$

and analogously for $\tilde{\Psi}_j$. The task of wavelet basis construction thus reduces to a matrix problem in the following way. Given $M_{j,0}$ and the composite mapping

$$M_j = (M_{j,0}, M_{j,1}) : l_2(\Delta_j) \oplus l_2(\nabla_j) \to l_2(\Delta_{j+1})$$

we have to determine $M_{j,1}$ which is called a *stable completion* [CDP96], such that M_j is invertible. The inverse of M_j is defined as the matrix $G_j = \begin{pmatrix} G_{j,0} \\ G_{j,1} \end{pmatrix}$ with $G_{j,0} \in \mathbb{R}^{|\Delta_j| \times |\Delta_{j+1}|}$ and $G_{j,1} \in \mathbb{R}^{|\nabla_j| \times |\Delta_{j+1}|}$, and we have

$$M_j G_j = G_j M_j = I.$$

We can qualify the remark made earlier about the special conditions required for the bases Φ_j and Ψ_j to be uniformly stable as the requirement that for all j

$$||M_j||_{\ell_2}, ||G_j||_{\ell_2} = \mathcal{O}(1)\,, \tag{3.31}$$

i.e., the norms remain uniformly bounded (cf. [Dah97, Remark 7]). We note that M_j and G_j depend on j, but only in the sense that once their values for a particular j are known, assembly for any other values can be done with little effort by expanding the middle section and using the same edge blocks (see Figure 3.5). However, the problem of computing the *stable completion* is non-trivial because we need to preserve sparsity in the matrix M_j and its inverse G_j. Even if an *initial* stable completion $\breve{M}_{j,1}$ is constructed, it might not be suitable for the purpose, for example would not satisfy biorthogonality. Fortunately, in [CDP96] a parameterization was given that made possible the modification of an initial stable completion into one that allows the construction of biorthogonal wavelets on intervals. Given an initial stable completion $\breve{M}_j$, a biorthogonal stable completion is computed as

$$M_{j,1} = (I - M_{j,0}\tilde{M}_{j,0}^T)\breve{M}_j, \tag{3.32}$$

and a corresponding result is given also for $\tilde{M}_{j,1}$. The new stable completions then satisfy the matrix equation $M_j M_j^T = I$ induced by (3.29), and we also have corresponding refinement relations

$$\Psi_j = M_{j,1}^T \Phi_{j+1}, \quad \tilde{\Psi}_j = \tilde{M}_{j,1}^T \tilde{\Phi}_{j+1}. \tag{3.33}$$

The implementation of the fast wavelet transforms in the setting we have discussed is trivial, as it involves multiplications of the coefficient vector with the sparse quadratic multiscale matrices $M_j := (M_{j,0}, M_{j,1})$, $\tilde{M}_j := (\tilde{M}_{j,0}, \tilde{M}_{j,1}) \in \mathbb{R}^{|\Delta_{j+1}| \times |\Delta_{j+1}|}$. Because the multiscale matrices are sparse (see Figure 3.5), the operations retain the usual $\mathcal{O}(N)$ complexity of fast wavelet transforms.

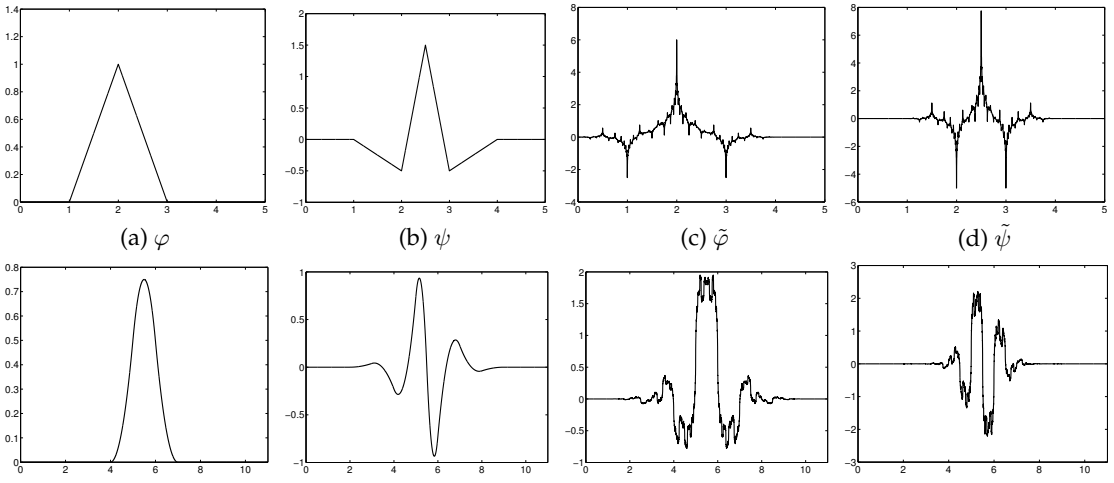
3.3.2. Examples: biorthogonal wavelet bases on the interval

The procedure sketched in the previous section is used by several constructions of wavelet bases on the bounded interval $\Omega = [0, 1]$, e.g. [DKU97, Dij09, Pri09] to name a few. The reason is that it is usually difficult to build a *Riesz* basis directly, so starting from a known primal basis one constructs an initial stable completion which is then refined into another one that is endowed with the desired properties, for example biorthogonality, or higher order vanishing moments. We remark that such methods of constructing wavelet bases can also be interpreted in special cases as the *lifting scheme* technique developed by Wim Sweldens (see [Swe98] for details).

In this thesis, we have opted to use the construction given in [Pri09]. There, the components of the primal basis that do not need to be modified to preserve the approximation properties near the boundaries, coincide up to a factor with dilates and translates of the cardinal *B-splines* of order m. A *B-spline* of degree m is defined recursively by

$$\begin{aligned} B_0(x) &= \chi_{[0,1)}(x) \\ B_m(x) &= \int_{\mathbb{R}} B_{m-1}(x-s)B_0(s)ds = \int_0^1 B_{m-1}(x-s)ds \end{aligned} \tag{3.34}$$

and they were used in [CDF92] in the role of primal scaling function $\varphi^{(m)} = B_m$ to construct biorthogonal *B-spline* wavelets on $L_2(\mathbb{R})$. Besides being refinable and replicating polynomials up to degree $m-1$, *B-splines* have also the advantage of an explicit formula and we remark that the filter coefficients for the refinement equations can be looked up in the literature (see e.g. [Mal09]). With increasing m, the support and smoothness also increases, and it has been shown in [CDF92] that for any $m \in \mathbb{N}$ there exists a *dual* function $\tilde{\varphi}^{(\tilde{m})} = \tilde{B}_{m,\tilde{m}}$ with $\tilde{m} \geq m$ and $m + \tilde{m}$ even, such that the dual scaling function is also refinable, locally supported and replicates polynomials up to degree $\tilde{m}-1$ exactly. There are no explicit formulas for these so called "dual B-splines" but we can compute their values using a cascade algorithm as finite filter coefficients are available. In Figure 3.6 we present two examples of scaling functions, wavelets and their duals on $L_2(\mathbb{R})$. The first row depicts the *B-spline* family with $m = 2,\ \tilde{m} = 2$ and in the second row the case $m = 3,\ \tilde{m} = 5$ is shown.



(a) φ (b) ψ (c) $\tilde{\varphi}$ (d) $\tilde{\psi}$

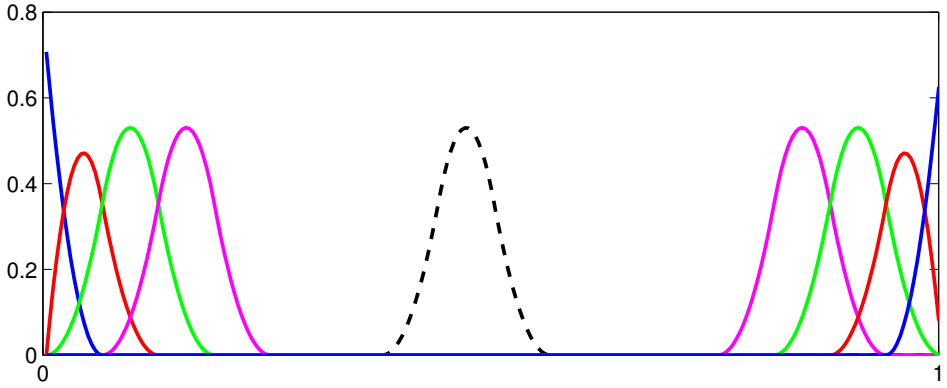
(e) φ (f) ψ (g) $\tilde{\varphi}$ (h) $\tilde{\psi}$

Figure 3.6.: *B-spline* scaling and wavelet functions on $L_2(\mathbb{R})$ with $m = 2,\ \tilde{m} = 2$ (top row) and $m = 3,\ \tilde{m} = 5$ (bottom row).

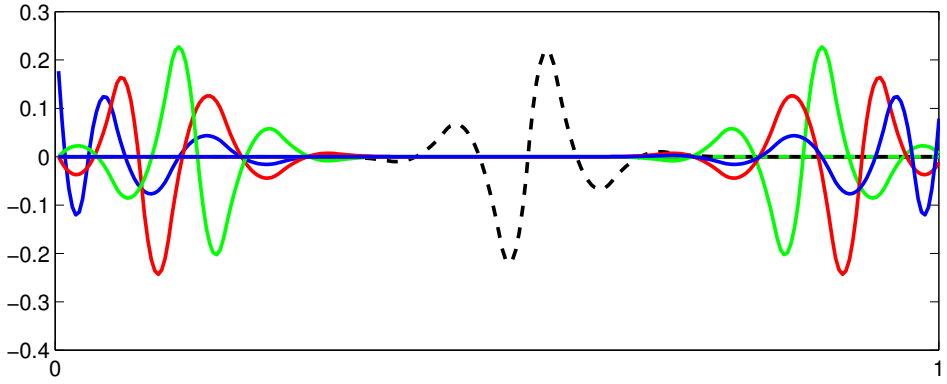
Like previously stated, modifications are required for those elements of the primal basis that are located near the boundaries. The construction given in [Pri09] solves this problem by employing for the primal basis elements of the *Schoenberg spline* space of order m corresponding to a knot sequence on $[0, 1]$, with boundary knots having multiplicity m. This means that no modifications have to be undertaken for the scaling functions near the boundaries, while the interior elements are just scaled *B-splines*.

An example of primal scaling functions $\varphi_{j,k}$ used in [Pri09] is shown in Figure 3.7 for the case $m = 3, \tilde{m} = 5$, and we remark that the plot contains the left and right bound-

ary scaling functions, with only one member of the set of interior basis functions being shown. The corresponding primal wavelets, again split into left, right and interior sets are shown in Figure 3.7b.



(a)



(b)

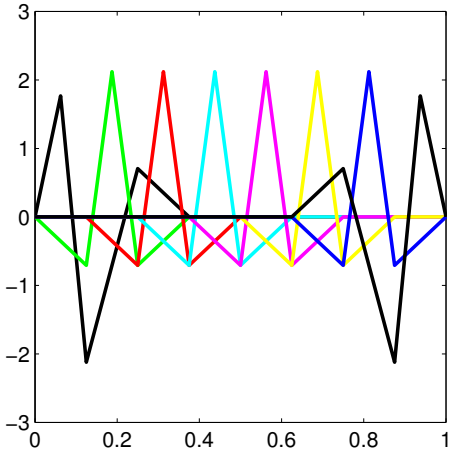
Figure 3.7.: Primal scaling basis (3.7a) and primal wavelet basis (3.7b) on the interval $[0, 1]$ from [Pri09]. The elements are shown for the case $m = 3$ and $\tilde{m} = 5$ on level $j = 4$. Left and right boundary elements are displayed in color, while a single representative from the translation-invariant set of interior elements is shown in black.

In the numerical examples presented in the following chapters, we also use the interval wavelet basis with $m = 2$ and $\tilde{m} = 2$, and the primal wavelets for $j = 3$ are shown in Figure 3.8.

Since the actual construction of the refinement matrices $M_{j,0}$ and $\tilde{M}_{j,0}$ is quite technical, we refer the reader to the original source [Pri06, Pri09] for specific details. We note however, that the coefficients of the j independent edge blocks of the refinement matrices are provided therein, so using these interval wavelet bases reduces to the implementation of the corresponding fast wavelet transforms.

3.4. Wavelets on Ω_ξ

The elements of the wavelet bases defined in this chapter are either functions on $\mathbb{R}$ or on the interval $[0, 1]$. Consequently, before using these bases for the approximation of the CME, they need to be transformed to wavelet bases on the bounded, discrete and



(a)

Figure 3.8.: Primal wavelet basis on the interval from [Pri09] with $m = 2$, $\tilde{m} = 2$ and level $j = 3$.

multidimensional domain $\Omega_\xi \subset \mathbb{N}_0^d$ defined in (2.65). So far we have only treated the univariate setting, and adapting wavelets to tensor product domains can be accomplished in two ways. Consider a multiscale decomposition

$$\mathcal{H}(\Omega_\xi) = \mathcal{S}_{j_0}(\Omega_\xi) \cup \bigcup_{j=j_0}^{j_{\max}-1} \mathcal{W}_j(\Omega_\xi),$$

where $\mathcal{W}_j = (\mathcal{S}_j \otimes \mathcal{W}_j) \oplus (\mathcal{W}_j \otimes \mathcal{S}_j) \oplus (\mathcal{W}_j \otimes \mathcal{W}_j)$, and a similar decomposition for the dual approximation spaces. Then, without loss of generality, we consider the case $d = 2$ and define the multivariate wavelet basis Ψ as being composed of the families

$$\{\varphi_{j_0,k_1} \otimes \varphi_{j_0,k_2}\}_{k_1,k_2 \in \Delta_{j_0}}, \quad \{\varphi_{j,k_1} \otimes \psi_{j,k_2}\}_{j \geq j_0, k_1 \in \Delta_j, k_2 \in \nabla_j},$$

$$\{\psi_{j,k_1} \otimes \varphi_{j,k_2}\}_{j \geq j_0, k_1 \in \nabla_j, k_2 \in \Delta_j} \quad \text{and} \quad \{\psi_{j,k_1} \otimes \psi_{j,k_2}\}_{j \geq j_0, k_1, k_2 \in \nabla_j}.$$

Similar families can be analogously defined for the dual basis $\tilde{\Psi}$ [CM97]. Such bases constructions Ψ and $\tilde{\Psi}$ are called *isotropic* as we have only products between basis elements with the same dyadic level in each direction.

A second choice for the realization of multivariate bases on tensor domains is to take straightforward tensor products of univariate basis elements, but in this case the dyadic levels can be different, and again for the case $d = 2$, the wavelet basis Ψ is given as

$$\Psi = \{\psi_{j_1,k_2} \otimes \psi_{j_2,k_2} \mid j_1, j_2 \geq j_0, k_1, k_2 \in \nabla_j\}.$$

Such bases are then called *anisotropic* (cf. [CM97]).

The interval wavelet bases can be used in a straightforward way for the bounded intervals defined by the truncation vector ξ from the definition of Ω_ξ given in (2.65). If the wavelets are defined on $\mathbb{R}$, the task of obtaining the wavelet representation of a function on bounded interval reduces to applying periodic continuation at the boundaries, in effect extending the target function on $\mathbb{R}$.

After the wavelets have been defined on the bounded intervals $[a, b]$, an equidistant grid $x_n = a + n(b-a)/2^r$ can be introduced with $r \in \mathbb{N}$ and $n = 0, \ldots, 2^r - 1$ such that

every function on the grid can be identified with a function on the discrete state space $\{0, \ldots, 2^r - 1\}$ via the relation $\tilde{f}(n) = f(x_n)$. This procedure then defines a wavelet basis for functions on Ω_ξ (cf. [JU10]).

In order to avoid any confusion for the reader, it is important to underline the fact that this is a *discrete* multi-dimensional wavelet basis built from tensor products of univariate wavelet bases, themselves discrete. The values of these discrete basis elements can be computed by using the *cascade algorithm* which starts from known values and computes intermediate values by applying the refinement equations iteratively. Considering the sequence of nested spaces $\{\mathcal{S}_j\}$ from Definition 3.1, we have that $j = \{j_0, \ldots, j_{\max} - 1\}$, and furthermore, $\mathcal{S}_{j_{\max}} = \Omega_\xi$. Thus, only a *finite* number of refinements exist, in contrast to the case of wavelets on $\mathbb{R}$ which was used in the presentation of the theory, where infinitely many refinements are possible (cf. [Jah10]). Moreover, we note that because of the discrete setting on which we operate, computing the scaling coefficients on the finest scale is simple, because there is no need to first project the original function onto a set of discrete points in order to apply the FWT. Thus, the fact that some scaling functions have no explicit representation presents no problems, as we can take the scaling coefficients on the finest scale to be the values of the original discrete function. Another point to be made is that different wavelet bases have different minimal resolution levels j_0 and cardinality. For example, the *B-spline* 3.5 basis on the interval has a minimal decomposition level $j_0 = 4$ due to the requirement that boundary elements should not overlap (see [Pri06] for details).

We conclude now the chapter dedicated to the construction of the wavelet bases, and proceed to describe the algorithms for the approximation of the CME.

CHAPTER

FOUR

NUMERICAL METHODS FOR THE CME

In Chapter 2 we have introduced a discrete stochastic formulation of biochemical reaction kinetics, where the time-evolution of the probability distribution $p(t,\cdot)$ for a system of interacting molecular species is governed by the *Chemical Master equation* (CME) on the infinite state space $\mathbb{N}_0^d$, given in (2.63). We present now the construction of an adaptive wavelet method for approximating the CME (2.67) on the truncated state $\Omega_\xi \subset \mathbb{N}_0^d$, which is equipped with an adaptive time-stepping strategy.

The main idea is to represent the solution of the CME in a thresholded sparse wavelet basis and propagate only the *essential* degrees of freedom in each time step, thus mitigating the effects of the *curse of dimensionality*. For the CME, this approach was first proposed in [Jah10] and then further developed in [JU10], which constitutes the basis of this chapter.

Conceptually, the adaptive wavelet method for the CME is closely related to similar methods for solving elliptic and parabolic equations (cf. [CDD01, CDD02, Dah97, Dah01]). Because wavelets are effective tools for data compression, as demonstrated in Chapter 3, by exploiting the favorable properties of a suitably chosen wavelet basis, such methods offer a good compression ratio and considerably reduce the number of degrees of freedom required to obtain reasonably accurate approximations. This is possible because for smooth data, the coefficients in the wavelet representation decay rapidly, and thus only a small subset of *essential* basis elements is required to represent the solution. For the CME, refining and propagating these *essential* basis elements is accomplished by an iterative procedure that combines Rothe's method with an adaptive Galerkin method: the problem is first discretized in time, followed by a projection into a suitably chosen low-dimensional space. The refinement of the approximation space is then guided by an *a posteriori* error analysis of the residual on the full state space.

We proceed now to describe the adaptive wavelet method for the CME in detail. Note that in order to make the notation simpler, we shall omit the spatial variable from the solution of the CME (2.67) with the operator $\mathcal{A}_\xi$ defined by (2.66), and use $p(t)$ instead of $p(t,\cdot)$. Moreover, we somewhat abuse our notation and refer to the operator $\mathcal{A}_\xi$ in the following sections as $\mathcal{A}$. As a precursory step, we begin by discussing the application of

wavelet compression to the CME solution and motivate the choice of Rothe's method as the integration strategy.

4.1. Using wavelet compression on the CME solution

Let $\mathcal{H}(\Omega_\xi)$ be the Hilbert space of all discrete functions $p : \Omega_\xi \longrightarrow \mathbb{R}$, equipped with the standard inner product

$$\langle p, q\rangle = \sum_{\mathbf{x}\in\Omega_\xi} p(\mathbf{x})q(\mathbf{x}) \quad \text{with} \quad p, q \in \mathcal{H}(\Omega_\xi),$$

and the norm $\|p\|_2 = \sqrt{\langle p, p\rangle}$. Let $N = \xi_1 \cdot \ldots \cdot \xi_d$ be the total number of states. Next, let $\{\psi_1^{(m)}, \ldots, \psi_N^{(m)}\}$ be a discrete orthonormal wavelet basis of order m for the space $\mathcal{H}(\Omega_\xi)$ and

$$p = \sum_{i=1}^{N} \beta_i^{(m)} \psi_i^{(m)}$$

the representation of the function p in this basis, with the wavelet coefficients given as $\beta_i^{(m)} = \langle p, \psi_i^{(m)}\rangle,\ i = 1, \ldots, N$. Because the domain $\Omega_\xi \subset \mathbb{N}_0^d$ defined in (2.65) is for most problems of interest high-dimensional, the corresponding multi-dimensional wavelet basis is built from tensor products of univariate wavelet bases obtained from a multiresolution analysis (as detailed in Chapter 3). Recall that there are two possible choices for constructing such wavelet bases on tensor product domains. Taking simply the tensor products of elements from one-dimensional wavelet bases leads to the *anisotropic* construction (cf. [Coh03, Section 2.2]). The *anisotropic* constructions are characterized by the fact that the resulting wavelets have different dyadic levels in each direction, and the construction does not generate a *multiresolution analysis* on the multi-dimensional domain. The second alternative is the *isotropic* construction [Coh03, Sections 1.4 and 2.12], which uses tensor products of wavelets with the same dyadic level in each direction, and generates a *multiresolution analysis* on Ω_ξ. Both the *anisotropic* and *isotropic* constructions can be used with our method. We remark that the numerical examples in this section use *isotropic* decompositions, if not otherwise specified.

In order to simplify the notation, the elements of the wavelet basis are indexed using a single index, instead of (multi-)indices for level and position. Assuming sufficient smoothness of the function p, a higher order (m) implies a faster decay of coefficients and consequently a better compression rate, because the proportion of coefficients that almost vanish increases, and a larger number can be discarded without impacting the quality of the approximation significantly.

Performing now a reordering of the wavelet coefficients in such a way that we have $|\beta_{j_1}^{(m)}| \geq \cdots \geq |\beta_{j_N}^{(m)}|$, we can write the best n-term approximation of p as

$$\tilde{p} = \sum_{i=1}^{n} \beta_{j_i}^{(m)} \psi_{j_i}^{(m)}.$$

As the name suggests, $\tilde{p}$ is the best approximation that can be obtained with a linear combination of n terms from the chosen wavelet basis, with $\{j_1, \ldots, j_n\}$ denoting the

corresponding indices. A comparison between an original function and its best n-term approximation is presented in Figure 4.1.

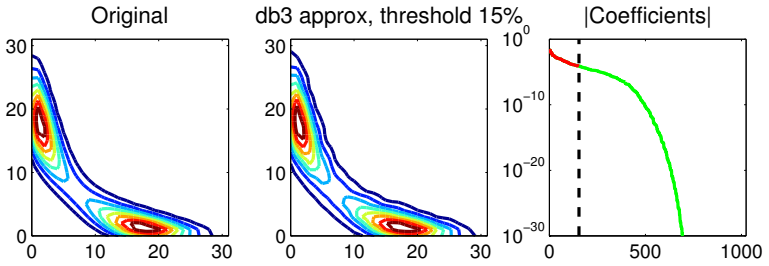


Figure 4.1.: Comparison between the exact solution p of the CME for the toggle switch model (2.94) and its best term approximation $\tilde{p}$ using 15% of the total degrees of freedom in Daubechies' *db3* orthonormal wavelet basis. In the right panel the values of the wavelet coefficients β_i ordered according to their absolute value are shown in logarithmic scaling. Only the coefficients on the left side of the dotted line were used for computing the best term approximation shown in the middle panel as a contour plot.

If the wavelet basis is orthogonal, the truncation error with respect to the $||\cdot||_2$ norm is the same as the 2-norm of the discarded coefficients, i.e.

$$||p - \tilde{p}||_2 = \sqrt{\sum_{i=n+1}^{N} |\beta_{j_i}^{(m)}|^2}.$$

Therefore, an efficient method for approximating the CME should strive to find a truncation index n, such that the approximation error remains below a certain prescribed tolerance. As evident from Figure 4.1, the number of degrees of freedom is significantly reduced by using the sparse wavelet representation, as $n \ll N$. However, this strategy can not be applied "as is" to the CME, because finding the best term approximation would require that all wavelet coefficients $(\beta_i^{(m)})_{i=1}^N$ are known. This is of course not a reasonable assumption, as the solution of the CME is *unknown*. It is possible to find the best term approximation for the initial probability distribution, but as the profile of the CME solution changes over time, and wavelets have local support, the set of best n terms at a later time would also change. Consequently, developing a method to propagate the set of *essential* basis elements is a prerequisite to the efficient use of wavelet compression for approximating the CME solution.

From a computational perspective, it is useful to remark that switching from the original representation of the function p in the canonical basis to the wavelet representation $p = \sum_{i=1}^{N} \beta_i^{(m)} \psi_i^{(m)}$ can be efficiently accomplished via a fast wavelet transform, requiring an $\mathcal{O}(N)$ effort. Conversely, reconstructing the function p from its wavelet coefficients $\beta_i^{(m)}$ can be done using a fast inverse wavelet transform, again at $\mathcal{O}(N)$ computational cost. Moreover, we remark at this point that the adaptive wavelet method for the CME is not restricted to any particular class of wavelets. The majority of the numerical examples shown in this chapter use Daubechies wavelets [Dau92], with periodic extensions at the boundaries, but in our implementation of the method also biorthogonal spline wavelets designed for bounded intervals or again using periodic continuation are available. However, in order not to complicate the exposition, we will restrict ourselves for the time being

to orthogonal wavelets, and discuss the biorthogonal case in the subsequent Chapter 5. The fast wavelet transforms can be regarded as black boxes that, given an input function or the wavelet coefficient vector return its counterpart. Because the method itself has a modular design, other wavelet bases besides those already available could be added in the future.

Usually, the basis elements $\psi_i^{(m)}$ have no explicit formulas, but are defined recursively via the discrete counterpart of the refinement equations. Sets of the appropriate one dimensional wavelet bases are computed in the initialization phase of the method, but the multi-dimensional basis elements are not explicitly computed. This is because in the adaptive wavelet method which will be presented in the next sections, only the one dimensional basis elements will be used in the computation of the Galerkin matrix given by (4.3).

4.2. Approximation with fixed step-size

Let $\{\psi_1^{(m)}, \ldots, \psi_N^{(m)}\}$ be the discrete wavelet basis for the space $\mathcal{H}(\Omega_\xi)$ introduced in the previous section. Because the polynomial order of the wavelet basis m is selected by the user and impacts only the compression properties and not the method itself, the superscript index (m) will be omitted from now on. We denote by

$$p_n = \sum_{i=1}^{\eta} \beta_i \psi_{j_i} \approx p(t_n) \tag{4.1}$$

the numerical approximation available at time $t_n = t_0 + nh$, with $h > 0$ fixed. Note that $p(t_n)$ represents the *exact* solution of the CME on the whole truncated state space Ω_ξ. Here, $\{j_1, \ldots, j_\eta\}$ is a small subset of the full index set $\{1, \ldots, N\}$, and $\beta = (\beta_1, \ldots, \beta_\eta)^T \in \mathbb{R}^\eta$ is the wavelet coefficient vector of p_n. As mentioned, the question is how to propagate the degrees of freedom required for the representation of the function p_n. Two strategies can be employed for this task, either the *method of lines*, or *Rothe's method*. The basic ideas behind these approaches are sketched in Figure 4.2.

In case the *method of lines* is used, the problem is first discretized in space and then in time. The space discretization is achieved in a straightforward manner, as the initial index set $\{j_1, \ldots, j_\eta\}$ of the *essential* elements can be easily obtained by performing a fast wavelet transform of the initial distribution $p(0)$ of the CME and selecting the best η coefficients such that the truncation error is below a prescribed tolerance,

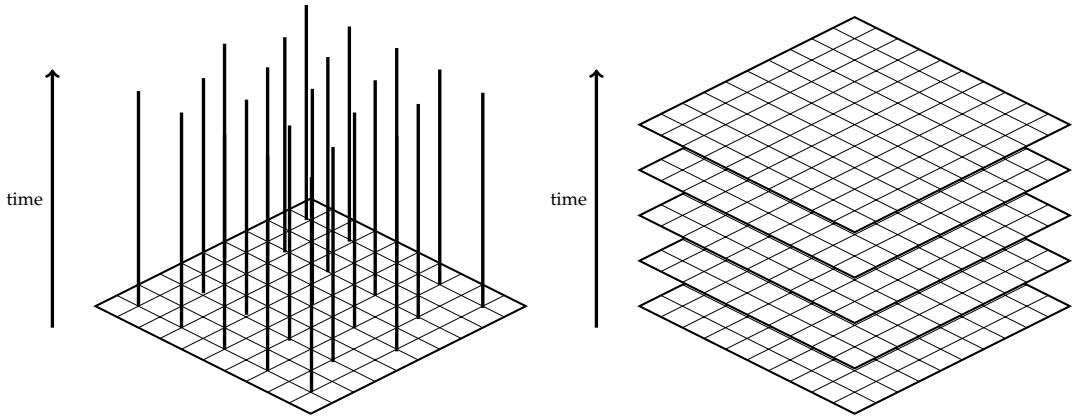
$$p(0) \approx q(0) = \sum_{i=1}^{\eta} \beta_i(0) \psi_{j_i}.$$

Then, the CME can be projected into the low-dimensional Galerkin space spanned by the elements $\{\psi_{j_1}, \ldots, \psi_{j_\eta}\}$ by imposing the Galerkin condition in (2.67) (cf. [Jah10]), i.e.,

$$\langle \psi_{j_i}, \partial_t q - \mathcal{A} q \rangle = 0 \text{ for all } i \in \{1, \ldots, \eta\}.$$

The evolution of the coefficient vector $\beta(t) = (\beta_1(t), \ldots, \beta_\eta(t))^T \in \mathbb{R}^\eta$ corresponding to the approximation is then given by the differential equation

$$\Gamma_\eta \frac{d}{dt} \beta(t) = M \beta(t) \tag{4.2}$$



(a) Method of Lines (b) Rothe's method

Figure 4.2.: Strategies for integrating the CME. In the left panel, the *method of lines* is shown (first space, then time discretization), while in the right panel *Rothe's method* is presented (first time, then space discretization) resulting in a series of stationary problems

where $M \in \mathbb{R}^{\eta\times\eta}$ is the Galerkin matrix defined by

$$M = (m_{ik})_{i,k=1}^{\eta}, \qquad m_{ik} = \langle \psi_{j_i}, \mathcal{A}\psi_{j_k} \rangle , \tag{4.3}$$

which is much smaller than the matrix $A \in \mathbb{R}^{N\times N}$ defined in (2.69) and used to represent the truncated operator $\mathcal{A}$. In case the wavelet basis is orthogonal we have that the Galerkin "mass" matrix $\Gamma_\eta = (\langle \psi_{j_i}, \psi_{j_k} \rangle) \in \mathbb{R}^{\eta\times\eta}$ is the identity matrix $I \in \mathbb{R}^{\eta\times\eta}$. In case a biorthogonal wavelet basis is used, the advantage of having an identity "mass" matrix is obtained only if a Petrov-Galerkin scheme is employed. Further details of such a scheme will be provided in Chapter 5, while in the present chapter only the Galerkin scheme will be discussed. The technical aspects related to the evaluation of the terms of the Galerkin matrix (4.3) will also be covered in Chapter 5, namely in Section 5.4.

Applying an ODE solver to (4.2) is now possible because the system contains only $\eta \ll N$ degrees of freedom. However, the approximation $||p_n - p(t_n)||_1$ deteriorates quickly as time increases, because the space discretization is fixed. As the elements of the wavelet basis are localized, and the solution of the CME will most likely occupy a different region of the state space Ω_ξ at a later time, the coefficients corresponding to the basis elements that have been discarded can become quite large and cannot be neglected anymore. Owing to these arguments, applying the *method of lines* and solving (4.2) constitutes a *non-adaptive* Galerkin method. Of course, a sequence of initial value problems could be solved, but this leads to a complicated procedure for adapting the space discretization during the integration of the CME.

A second strategy is *Rothe's method*, in which the system is first discretized in time and the spatial discretization is then adapted in each time step. Performing the time discretization first leads to a sequence of stationary problems and adapting the space discretization is therefore simplified. Examples of employing the adaptive *Rothe's method* ([Bor90, Bor91]) to the CME can be found in e.g., [DHJW08] (without the use of wavelets) and [Jah10] where wavelet compression was first proposed in this context.

Coupling *Rothe's method* with a Galerkin ansatz and using an iterative strategy to identify the *essential* degrees of freedom similar to the approach from [CDD01] constitutes the

computational core of the wavelet-based adaptive method for the CME which we present in this thesis. In our chosen setting, we can employ a variety of time integration methods. While in [Jah10], where the method was devised, the second order trapezoidal rule was used and Haar wavelet basis employed, in [JU10] the function p_n was propagated using a fourth-order integrator, namely the 2-stage Gauss-Runge-Kutta method and better wavelet bases were applied. Additionally, the method was endowed with adaptive time-stepping selection, and we proceed now to detail all these improvements. We remark that the results presented in the following sections have been already published in a slightly different form in [JU10], a co-authored paper.

For linear problems, the 2-stage Gauss-Runge-Kutta method is equivalent to the $(2,2)$-Padé approximation to the exponential function, and its order (order 4) is the highest possible among all integrators with two stages. Moreover, the method is A-stable, which is advantageous because the real parts of all eigenvalues of the operator $\mathcal{A}$ are non-positive (as shown in Appendix A) and the CME can be very stiff in the initial phase.

For a given approximation p_n, the new approximation $u_{n+1} \approx p(t_{n+1})$ is given as the solution of the linear equation

$$Q(h\mathcal{A})u_{n+1} \;=\; P(h\mathcal{A})p_n \tag{4.4}$$

with

$$Q(h\mathcal{A}) \;=\; I - \frac{h}{2}\mathcal{A} + \frac{h^2}{12}\mathcal{A}^2, \qquad P(h\mathcal{A}) \;=\; I + \frac{h}{2}\mathcal{A} + \frac{h^2}{12}\mathcal{A}^2. \tag{4.5}$$

Here and below, I denotes the identity operator/matrix. An equivalent formulation for the solution of (4.4) is

$$u_{n+1} \;=\; p_n + \frac{h}{2}(g_1 + g_2) \tag{4.6}$$

where (g_1, g_2) solves

$$\begin{pmatrix} I - \frac{h}{4}\mathcal{A} & -h\left(\frac{1}{4} - \frac{\sqrt{3}}{6}\right)\mathcal{A} \\ -h\left(\frac{1}{4} + \frac{\sqrt{3}}{6}\right)\mathcal{A} & I - \frac{h}{4}\mathcal{A} \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \;=\; \begin{pmatrix} \mathcal{A}p_n \\ \mathcal{A}p_n \end{pmatrix}. \tag{4.7}$$

Of course, both linear systems (4.4) or (4.7) contain all degrees of freedom and thus are usually far too large to be solved either directly of by some iterative scheme.

However, there is no need to solve either (4.4) or (4.7) *exactly*. It is enough to approximate the solution u_{n+1} of (4.4) up to a given tolerance that does not significantly increase the *local error* of the time integration.

Therefore, we can project (4.4) into a low-dimensional Galerkin space of $\mathcal{H}(\Omega_\xi)$ and approximate $u_{n+1} \approx p_{n+1}$. The main difference from the previously discussed *non-adaptive* Galerkin method (4.2) is that a refinement of the low-dimensional space is now performed, which adapts the spatial representation of p_{n+1}. As initial candidate for the new Galerkin space, we choose the space spanned by $\{\psi_{j_1}, \ldots, \psi_{j_\eta}\}$, i.e., the approximation space composed of the *essential* elements identified in the previous step. A first approximation

$$p_{n+1}^{(0)} = \sum_{i=1}^{\eta} \gamma_i^{(0)} \psi_{j_i}, \qquad p_{n+1}^{(0)} = p_n + \frac{h}{2}(g_1^{(0)} + g_2^{(0)}), \qquad g_s^{(0)} = \sum_{i=1}^{\eta} \zeta_{s,i}^{(0)} \psi_{j_i},\ s \in \{1,2\} \tag{4.8}$$

in such a space is obtained by imposing the Galerkin conditions in (4.7)

$$\begin{array}{rcrcl} \left\langle \psi_{j_i}, (I - \frac{h}{4}\mathcal{A})g_1^{(0)} \right\rangle & - & h\left(\frac{1}{4} - \frac{\sqrt{3}}{6}\right)\left\langle \psi_{j_i}, \mathcal{A}g_2^{(0)} \right\rangle & = & \langle \psi_{j_i}, \mathcal{A}p_n \rangle \\ -h\left(\frac{1}{4} + \frac{\sqrt{3}}{6}\right)\left\langle \psi_{j_i}, \mathcal{A}g_1^{(0)} \right\rangle & + & \left\langle \psi_{j_i}, (I - \frac{h}{4}\mathcal{A})g_2^{(0)} \right\rangle & = & \langle \psi_{j_i}, \mathcal{A}p_n \rangle \end{array} \tag{4.9}$$

for all $i = 1, \ldots, \eta$. Using the notation introduced in (4.3) we can rewrite (4.9) as

$$\begin{pmatrix} I - \frac{h}{4}M & -h\left(\frac{1}{4} - \frac{\sqrt{3}}{6}\right)M \\ -h\left(\frac{1}{4} + \frac{\sqrt{3}}{6}\right)M & I - \frac{h}{4}M \end{pmatrix} \begin{pmatrix} \zeta_1^{(0)} \\ \zeta_2^{(0)} \end{pmatrix} = \begin{pmatrix} M\beta \\ M\beta \end{pmatrix} \tag{4.10}$$

where $\zeta_s^{(0)} = \left(\zeta_{s,1}^{(0)}, \ldots, \zeta_{s,\eta}^{(0)}\right)^T$, $s \in \{1, 2\}$. The new linear system (4.10) has size $\mathbb{R}^{2\eta \times 2\eta}$, but is still considerably smaller than the problem on the state space Ω_ξ given in (4.4), and can be solved either directly, or by using GMRES or some other iterative method. The first approximation $p_{n+1}^{(0)}$ to u_{n+1} can then be easily obtained by performing a fast inverse wavelet transform of the new coefficient vector $\gamma^{(0)} = \beta + \frac{h}{2}(\zeta_1^{(0)} + \zeta_2^{(0)})$. Because (4.4) and (4.7) are equivalent, $\gamma^{(0)}$ also solves the equation

$$Q(hM)\gamma^{(0)} = P(hM)\beta. \tag{4.11}$$

However, the approximation $p_{n+1}^{(0)}$ might *not* be very close to the solution of the full problem (4.4) as the optimal low-dimensional space representation for p_{n+1} changes and we have used instead the space available for p_n. In order to guide the expansion of the approximation space we use a *posteriori* error analysis of the residual on the full space

$$r^{(0)} = Q(h\mathcal{A})p_{n+1}^{(0)} - P(h\mathcal{A})p_n.$$

Imposing the Galerkin condition (4.9) means that orthogonality between the residual and the approximation space is enforced. Hence, any "part" of the residual that is large indicates the specific areas which have been neglected by solving (4.11) instead of (4.4), i.e., where we need better coverage in the low-dimensional space in order to improve the approximation. In practical terms, the values $\langle r^{(0)}, \psi_k \rangle$ which denote the k-th coefficient of the residual in the chosen wavelet basis will encode how much the approximation needs the basis element ψ_k. If $|\langle r^{(0)}, \psi_k \rangle|$ is large, then the approximation will probably improve if ψ_k is added to the current basis selection, otherwise if the value is small, the gains will be negligible. If the basis element ψ_k is already contained in the approximation subspace, then we have by definition that $\langle r^{(0)}, \psi_k \rangle = 0$, so these elements must be excluded from the *a posteriori* analysis.

Now the adaptive wavelet method for the CME proceeds as follows. As a first step, the basis is enlarged by some number $\Delta\mu$ of new elements, which leads to a new candidate subspace spanned by $\psi_{j_1}, \ldots, \psi_{j_{\eta+\Delta\mu}}$. The new elements $\psi_{j_{\eta+1}}, \ldots, \psi_{j_{\eta+\Delta\mu}}$ are those which have the largest absolute values $|\langle r^{(0)}, \psi_k \rangle|$ for the coefficients of the residual in the wavelet basis. In practice, either a tolerance is selected, or a fixed number of elements is added. Next, the Galerkin matrix (4.3) is updated to include the newly selected basis elements by adding $\Delta\mu$ new lines and columns corresponding to $\psi_{j_{\eta+1}}, \ldots, \psi_{j_{\eta+\Delta\mu}}$, as sketched in Figure 4.3.

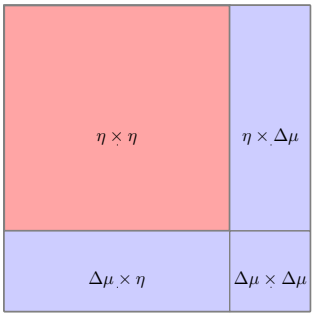


Figure 4.3.: Adding $\Delta\mu$ new elements to the Galerkin matrix $M \in \mathbb{R}^{\eta\times\eta}$ defined in (4.3)

Then, a new approximation $p_{n+1}^{(1)}$ is computed by solving (4.10) with the enlarged Galerkin matrix $\tilde{M} \in \mathbb{R}^{(\eta+\Delta\mu)\times(\eta+\Delta\mu)}$ and enlarged coefficient vector

$$\tilde{\beta} = \{\beta_1, \ldots, \beta_\eta, \underbrace{0, \ldots, 0}_{\Delta\mu}\} \in \mathbb{R}^{\eta+\Delta\mu}, \tag{4.12}$$

leading to a refined coefficient vector $\gamma^{(1)} = \tilde{\beta} + \frac{h}{2}(\zeta_1^{(1)} + \zeta_2^{(1)})$ solving (4.11). The improved approximation is then given by $p_{n+1}^{(1)} = \sum_{i=1}^{\eta^{(1)}} \gamma_i^{(1)} \psi_{j_i}$ with $\eta^{(1)} = \eta + \Delta\mu$ terms.

Iterating this basic procedure leads to a sequence of approximations $p_{n+1}^{(0)}, p_{n+1}^{(1)}, p_{n+1}^{(2)}, \ldots$ belonging to a hierarchy of increasingly larger approximation spaces. The iterative process stops at step (ℓ) if the 1-norm of residual $\|r^{(\ell)}\|_1$ is smaller than a prescribed tolerance, with the last approximation $p_{n+1}^{(\ell)}$ being accepted for time step t_{n+1}.

One drawback of the procedure described above is the growth of the approximation space with each iteration step. As a remedy, a thresholding of the current basis can be performed at the end of each step, or alternatively, every few steps. This post-processing step removes all dispensable basis elements from the representation of $p_{n+1}^{(\ell)} = \sum_{i=1}^{\eta^{(\ell)}} \gamma_i^{(\ell)} \psi_{j_i}$ that do not contribute significantly to the accuracy of the approximation. The thresholding is based on the fact that the truncation error, when measured in $||\cdot||_2$, is the 2-norm of the discarded coefficients, if an orthogonal wavelet basis is used. For other choices of wavelet basis, this is usually not the case, but other procedures can be applied to trim the basis, for example discarding all coefficients with values below a certain thresholding tolerance. We proceed now to describe the thresholding step in more detail.

If $\mathcal{I} \subset \{1, \ldots, \eta^{(\ell)}\}$ is a subset of the index set, and $p_{n+1}^{[\mathcal{I}]} = \sum_{i\in\mathcal{I}} \gamma_i^{(\ell)} \psi_{j_i}$ is the approximation obtained by deleting all terms with $i \notin \mathcal{I}$ from the wavelet representation, then for an orthogonal basis we have

$$\|p_{n+1}^{(\ell)} - p_{n+1}^{[\mathcal{I}]}\|_2 \quad = \quad \sqrt{\sum_{i\notin\mathcal{I}} \left(\gamma_i^{(\ell)}\right)^2}.$$

In order to reach an accuracy $\|p_{n+1}^{(\ell)} - p_{n+1}^{[\mathcal{I}]}\|_2 \leq \texttt{tol}_{trunc}$ with a minimal number of basis elements, we simply arrange the coefficients by magnitude and truncate the smallest coefficients as long as $\sum_{i\notin\mathcal{I}} \left(\gamma_i^{(\ell)}\right)^2 \leq \texttt{tol}_{trunc}^2$. However, the error of the adaptive wavelet method is measured with respect to the $\|\cdot\|_1$ rather than $\|\cdot\|_2$, so we choose

as $\texttt{tol}_{trunc} \leq c \cdot \texttt{tol}$, where c is a factor which accounts for the equivalence of the norms. The choice $c = 1/\sqrt{N}$ is correct, but often too pessimistic. A better choice is to replace N by the number of states where $p_{n+1}^{(\ell)}$ is essentially larger than zero. The function $p_{n+1} := p_{n+1}^{[\mathcal{I}]}$ obtained after this thresholding procedure is the final result of the entire time step.

The diagram 4.4 sketches the main steps of the algorithm, and the pseudocode for one single time step of the adaptive wavelet method with fixed step size is given in Algorithm 3. We note that the method does not store the sequence of approximations $p_{n+1}^{(0)}, p_{n+1}^{(1)}, p_{n+1}^{(2)}, \ldots$ but only one single approximation $\hat{p}_{n+1}$ which is overwritten in each iteration, and proceed to discuss some details that were glossed over in the presentation.

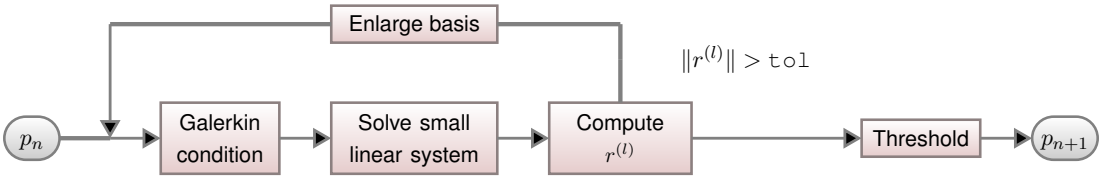


Figure 4.4.: Diagram illustrating the computational core of the adaptive wavelet method

Stopping criterion. As stated before, the iteration stops at some step (l) if the 1-norm of residual is smaller than a prescribed tolerance. However, it is advantageous to use as stopping criterion the condition $\|r\|_1 < C_{safe} \cdot \texttt{tol}$, where $C_{safe} \leq 1$ is a safety factor that is chosen based on the following argument. If $\|r\|_1 \leq C_{safe} \cdot \texttt{tol}$, then if we compare $Q(h\mathcal{A})u_{n+1} = P(h\mathcal{A})p_n$ (cf. (4.4)) with the approximation in the current step $Q(h\mathcal{A})\hat{p}_{n+1} = P(h\mathcal{A})p_n + r$, we get the error bound

$$\begin{aligned} \|u_{n+1} - \hat{p}_{n+1}\|_1 &\leq \|Q(h\mathcal{A})^{-1}P(h\mathcal{A})p_n - Q(h\mathcal{A})^{-1}\big(P(h\mathcal{A})p_n + r\big)\|_1 \\ &\leq \|Q(h\mathcal{A})^{-1}r\|_1 \\ &\leq \|Q(h\mathcal{A})^{-1}\|_1 \cdot C_{safe} \cdot \texttt{tol}. \end{aligned} \tag{4.13}$$

As the goal is to achieve $\|u_{n+1} - \hat{p}_{n+1}\|_1 \leq \texttt{tol}$, we have to choose $C_{safe} = 1/\|Q(h\mathcal{A})^{-1}\|_1$. Based on our numerical experiments, we conjecture that $\|Q(h\mathcal{A})^{-1}\|_1 = 1$, but unfortunately a rigorous proof is not available. However, since $(I - h\mathcal{A}/2)^{-1}$ is contractive (see Theorem 1 from [Jah10] for a proof) and $Q(h\mathcal{A})^{-1}$ is just a higher order perturbation, choosing $C_{safe} \lessapprox 1$ seems to be reasonable.

Additionally, sometimes it might be more convenient to prescribe *a priori* a maximum number of basis elements to be used, instead of just the accuracy of the approximation. This condition adds a degree of flexibility, because it is difficult to estimate *a priori* the needed accuracy for a given problem and besides, memory requirements impose anyway an upper bound for the maximal number of basis elements that can be used. Therefore, the number of basis elements μ that are kept *after* a time step (see step 6 of Algorithm 3) is chosen by the user, and if the thresholding procedure delivers a bigger value, the smaller of the two is selected. Moreover, a second parameter $\mu_{\max}$ which encodes the maximal number of basis elements *during* the time-step can be given. In practice, this means that the stopping criterion in Step 5 of Algorithm 3 is expanded to read "*If* $\|r\|_1 > C_r \cdot \texttt{tol}$ *and* $\hat{\mu} + \Delta\mu \leq \mu_{max}$".

Algorithm 3: One step of the adaptive wavelet method

Parameter : step-size $h > 0$, tolerance `tol`, safety factor C_{safe}

Input : index subset $\{j_1, \ldots, j_\eta\}$ and coefficients $\beta_1, \ldots, \beta_\eta$ of the current approximation $p_n = \sum_{i=1}^{\eta} \beta_i \psi_{j_i}$
Galerkin matrix M defined by (4.3)

Output : index subset $\{k_1, \ldots, k_\mu\}$ and coefficients $\gamma_1, \ldots, \gamma_\mu$ of the new approximation $p_{n+1} = \sum_{i=1}^{\mu} \gamma_i \psi_{k_i}$
updated Galerkin matrix

begin

1. Set $\hat{\mu} = \eta$.
2. Solve the linear system

$$\begin{pmatrix} I - \frac{h}{4}M & -h\left(\frac{1}{4} - \frac{\sqrt{3}}{6}\right)M \\ -h\left(\frac{1}{4} + \frac{\sqrt{3}}{6}\right)M & I - \frac{h}{4}M \end{pmatrix} \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} = \begin{pmatrix} M\hat{\beta} \\ M\hat{\beta} \end{pmatrix}$$

   and set $\hat{\gamma} = \hat{\beta} + \frac{h}{2}(\zeta_1 + \zeta_2)$. The vector $\hat{\beta}$ is an embedding of $\beta \in \mathbb{R}^\eta$ into $\mathbb{R}^{\hat{\mu}}$:

$$\hat{\beta} = (\beta_1, \ldots, \beta_\eta, \underbrace{0, \ldots, 0}_{\hat{\mu}-\eta})^T$$

3. Compute the new approximation $\hat{p}_{n+1} = \sum_{i=1}^{\hat{\mu}} \hat{\gamma}_i \psi_{j_i}$ by a fast inverse wavelet transform.
4. Compute the residual $r = Q(h\mathcal{A})\hat{p}_{n+1} - P(h\mathcal{A})p_n$ with Q and P defined by (4.5).
5. If $\|r\|_1 > C_r \cdot \mathtt{tol}$:
   a) Compute $\chi_l = \left| \langle \psi_l, r \rangle \right|$ for $l = 1, \ldots, N$ by a fast wavelet transform.
   b) Find the indices $j_{\hat{\mu}+1}, \ldots, j_{\hat{\mu}+\Delta\mu}$ of the $\Delta\mu$ largest entries of $(\chi_1, \ldots, \chi_N)$.
   c) Add $\psi_{j_{\hat{\mu}+1}}, \ldots, \psi_{j_{\hat{\mu}+\Delta\mu}}$ to the current selection of basis elements.
   d) Update the Galerkin matrix by adding new blocks corresponding to the new basis vectors:

$$M = (m_{ik})_{i,k=1}^{\hat{\mu}+\Delta\mu}, \qquad m_{ik} = \langle \psi_{j_i}, \mathcal{A}\psi_{j_k} \rangle .$$

   e) Set $\hat{\mu} \mapsto \hat{\mu} + \Delta\mu$.
   f) Go to step 2.
6. The result $p_{n+1} = \sum_{i=1}^{\mu} \gamma_i \psi_{k_i}$ is obtained by discarding all coefficients γ_i with $i \notin \mathcal{I}$, where $\mathcal{I}$ is the index set of the largest coefficients. The number of coefficients is chosen in such a way that $\|p_{n+1} - \hat{p}_{n+1}\|_1 \leq \mathtt{tol}$ (see above). The corresponding columns and lines are deleted from the Galerkin matrix M.

end

Minimum number of DOFs. If the fixed time step $h > 0$ is chosen too large, then the propagation of the approximation often requires more basis elements than representing the accepted approximation at the end of the time step. In such situations, sometimes a phenomenon which can be described as over-thresholding appears, i.e., basis elements that are discarded in the thresholding step need to be selected again in the next time step. Evidently, this decreases the efficiency of the algorithm as the corresponding entries in the Galerkin matrix M have to be computed once more. Therefore, it is better to give a *lower* bound for the number of DOFs that will be retained in step 6 of Algorithm 3. This has the effect that even if the thresholding procedure suggests discarding most of the coefficients, enough are retained to maintain computational efficiency. We remark however, that this situation mostly arises when choosing unreasonable parameter values.

Preserving positivity. An issue of upmost importance in the context of approximating the probability distribution representing the solution of CME is preserving the positivity of the numerical approximation p_{n+1}. The problem is not only related to the spatial discretization where the oscillatory nature of wavelet bases combined with the thresholding of elements can lead to negative values for the approximated function, but also affects the time integration scheme. Even in the case where the spatial representation is exact, meaning that all the basis elements are used to approximate the function, the numerical solution could exhibit negative entries if rather large-steps are used. This is because most Runge-Kutta methods only preserve positivity for sufficiently small time-steps. An exception is the implicit Euler method applied to $\dot{y} = -Ay$ where A is a M-matrix. In this case, the numerical solution is positive for every step-size $h > 0$ if $y_0 \geq 0$. Unfortunately, it is well known that the implicit Euler method has order 1, which is not sufficient for practical use. Moreover, methods based on other spatial approximations do not always preserve positivity either, so the issue is not singular to the wavelet method.

Alternative time integration scheme. So far, the adaptive wavelet method was built around the 4th-order, 2-stage Gauss-Runge-Kutta method which is equivalent to the $(2,2)$-Padé approximation to the exponential function. However, as detailed in (4.7), the algorithm uses in practice an equivalent formulation because it is not advisable to compute the coefficient vector by solving (4.11) directly. This is due to the fact that the matrix M^2 which would then occur in $Q(hM)$ typically has an adverse effect on the condition number of the linear system given by (4.11). The problem is circumvented by using the equivalent formulation (4.10). However, the price to be paid is that the linear system thus obtained in (4.10), has twice as many unknowns as (4.11). If the doubling of the size of the linear system obtained by projecting the problem in a low-dimensional approximation space needs to be avoided, the 2-stage Gauss-Runge-Kutta method could be replaced by another integration scheme. This is easily achieved, as the algorithm has a modular construction. Indeed, in [Jah10], the second-order *trapezoidal rule* was applied to the CME, meaning that instead of using (4.4), the approximation $u_{n+1} \approx p(t_{n+1})$ is given as the solution of the linear equation

$$\Big(I - \frac{h}{2}\mathcal{A}\Big)u_{n+1} \quad = \quad \Big(I + \frac{h}{2}\mathcal{A}\Big)p_n. \tag{4.14}$$

Analogously to (4.9), we then project the system (4.14) into a low-dimensional approximation space by imposing the Galerkin condition. Naturally, if another integration

method is used, a corresponding change must also be applied to the residual computation. Using (4.14) instead of (4.4), has however the disadvantage of employing only a method of order 2. To improve on this situation, another alternative is to use a singly diagonally implicit Runge-Kutta method (SDIRK) (cf. Section IV.6 in [HW96]). In this case, only linear systems with the same matrix $(I - chA) \in \mathbb{R}^{\eta\times\eta}$ but different right hand sides need to be solved. For a 2-stage SDIRK method, we would compute the new approximation $u_{n+1} \approx p_{n+1}$ of the full CME solution $p(t_{n+1})$ as

$$u_{n+1} = p_n + \frac{h}{2}(k_1 + k_2)$$

where k_1 and k_2 are the solutions of

$$\begin{aligned} \Big(I - ch\mathcal{A}\Big)k_1 &= \mathcal{A}p_n \\ \Big(I - ch\mathcal{A}\Big)k_2 &= \mathcal{A}(p_n + h(1-2c)k_1). \end{aligned} \tag{4.15}$$

If we choose $c = \frac{3+\sqrt{3}}{6}$ in (4.15), we obtain an A-stable method of order 3, thus less accurate then the 2-stage Gauss-Runge-Kutta method employed by the standard version of Algorithm 3. To obtain a SDIRK method of comparable order 4, three stages have to be used, which leads to the question whether the additional computational effort is justified compared to the standard approach (4.4). However, it is difficult to say *a priori* which method will be more efficient for a particular problem. In our numerical tests, no significant difference in efficiency was observed, but both methods are available in our implementation.

4.3. Adaptive step-size control

Up to now, we have only addressed the issue of space adaptivity for the wavelet method, and glossed over the question of time adaptivity. Solving the CME with a fixed step-size, however, can prove to be rather inefficient as a short stiff transient phase at the beginning of the time interval often leads to severe restrictions on the fixed step-size algorithm, whereas in reality much larger time steps can be taken later on, when the probability distribution slowly converges towards the stationary distribution.

The aim of this section is to introduce a viable strategy for selecting the step-size adaptively. We remark that *adaptive* in our context means the ability to control the global step-size and not a fully adaptive scheme where each degree of freedom is propagated with its own step-size. Although such a local time stepping method could increase performance, it is dependent on advanced knowledge of the problem, and our goal is to construct a method that is free of such assumptions. The main problem in devising an adaptive time-stepping strategy lies with the fact that two different types of errors have to be controlled, the error due to the spatial approximation via wavelet compression and the time approximation error. Because a classical strategy for adaptive step-size control based on embedded Runge-Kutta methods could not be applied in our case, a different approach was developed.

The overall goal is to select the step-size in such a way that the (local) approximation error remains under or close to the chosen tolerance `tol`. We must remark however, that

the error bounds given below only guarantee that the error is smaller than $C \cdot \texttt{tol}$ with some (moderate) constant $C > 1$. If the target is to keep the error always under a certain threshold, a good solution is to introduce an appropriate safety factor.

Constructing the adaptive time-step selection mechanism starts with the following bound for the local error.

Theorem 4.1 ([JU10]). *Let p_n be the approximation computed in the n-th time step with tolerance $\texttt{tol} > 0$ and let $p(t)$ be the* exact *solution of the CME*

$$\begin{aligned} \dot{p}(t) &= \mathcal{A}p(t) \quad \text{for } t \in [t_n, t_{n+1}] \\ p(t_n) &= p_n \end{aligned} \tag{4.16}$$

which starts from p_n at time t_n. Suppose that the representation of $\hat{p}_{n+1}$ before the truncation (step 6) is $\hat{p}_{n+1} = \sum_{i=1}^{\hat{\mu}} \hat{\gamma}_i \psi_{j_i}$ and let

$$V = span\{\psi_{j_1}, \ldots, \psi_{j_{\hat{\mu}}}\} \subset \mathcal{H}(\Omega_\xi)$$

*be the iteratively enlarged approximation space. Note that $p_n \in V$ because V is the approximation space **before** the truncation step. Let $q(t)$ be the solution of the* projected *CME*

$$\begin{aligned} \dot{q}(t) &= \mathcal{P}_V \mathcal{A} q(t) \quad \text{for } t \in [t_n, t_{n+1}] \\ q(t_n) &= p_n \end{aligned} \tag{4.17}$$

where

$$\mathcal{P}_V : \mathcal{H}(\Omega_\xi) \longrightarrow V, \qquad \mathcal{P}_V w = \sum_{i=1}^{\hat{\mu}} \langle w,\ \psi_{j_i} \rangle \, \psi_{j_i}$$

denotes the orthogonal projection from $\mathcal{H}(\Omega_\xi)$ onto V. Then, the local error $p_{n+1} - p(t_{n+1})$ is bounded by

$$\begin{aligned} \|p_{n+1} - p(t_{n+1})\|_1 \ \leq \ & \texttt{tol} \\ & + \frac{h^5}{720} \left\| (\mathcal{P}_V \mathcal{A})^5 p_n \right\|_1 + \mathcal{O}\left(h^6\right) \\ & + \int_{t_n}^{t_{n+1}} \|(\mathcal{P}_V - I)\mathcal{A} q(s)\|_1 ds. \end{aligned} \tag{4.18}$$

Proof. The error is split into the three parts

$$\begin{aligned} \|p_{n+1} - p(t_{n+1})\|_1 \ \leq \ & \|p_{n+1} - \hat{p}_{n+1}\|_1 \\ & + \|\hat{p}_{n+1} - q(t_{n+1})\|_1 \\ & + \|q(t_{n+1}) - p(t_{n+1})\|_1 . \end{aligned} \tag{4.19}$$

The error bound $\|p_{n+1} - \hat{p}_{n+1}\|_1 \leq \texttt{tol}$ follows directly from the definition of p_{n+1} in step 6 of Algorithm 3. Recall that steps 2 and 3 in Algorithm 3 are equivalent to applying the 2-stage Gauss method to the projected CME (4.17).

The local error of the Gauss method is bounded by

$$\|\hat{p}_{n+1} - q(t_{n+1})\| \leq \frac{h^5}{720} \left\|(\mathcal{P}_V \mathcal{A})^5 p_n\right\|_1 + \mathcal{O}(h^6)\,. \tag{4.20}$$

To show this, we use that in the equivalent $(2,2)$-Padé approximation, the problem can be formulated as

$$p_{n+1} = \Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1}\Big(I + \frac{h}{2}A + \frac{h^2}{12}A^2\Big)p_n \tag{4.21}$$

where we have used (4.5) and replaced $\mathcal{A}$ with the matrix $A \in \mathbb{R}^{N\times N}$. Next, from

$$\begin{aligned} I &= \Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1}\Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big) \\ &= \Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1} - \Big(\frac{h}{2}A - \frac{h^2}{12}A^2\Big)\Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1} \end{aligned}$$

we obtain that

$$\Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1} = I + \Big(\frac{h}{2}A - \frac{h^2}{12}A^2\Big)\Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1}. \tag{4.22}$$

Hence, using (4.22) we can now write

$$\begin{aligned} \Big(I + \frac{h}{2}A + \frac{h^2}{12}A^2\Big)\Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1} &= \Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1}\Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2 + hA\Big) \\ &= I + hA\Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1} \\ &= I + hA\left(I + \Big(\frac{h}{2}A - \frac{h^2}{12}A^2\Big)\Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1}\right) \\ &= I + hA + hA\Big(\frac{h}{2}A - \frac{h^2}{12}A^2\Big)\Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1}. \end{aligned}$$

If we multiply the last term in the last line in the above expression with

$$I = \Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big) + \Big(\frac{h}{2}A - \frac{h^2}{12}A^2\Big) \tag{4.23}$$

we get that

$$\begin{aligned} \Big(I + \frac{h}{2}A + \frac{h^2}{12}A^2\Big)\Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1} &= I + hA + hA\Big(\frac{h}{2}A - \frac{h^2}{12}A^2\Big) \\ &\quad + hA\Big(\frac{h}{2}A - \frac{h^2}{12}A^2\Big)^2\Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1}. \end{aligned}$$

Performing now the same procedure three more times, i.e., multiplying the last term of the previously obtained result by (4.23), expanding and grouping together the terms with the same powers, we finally obtain that

$$\begin{aligned} \Big(I + \frac{h}{2}A + \frac{h^2}{12}A^2\Big)\Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1} &= I + hA + \frac{h^2}{2}A^2 + \frac{h^3}{6}A^3 \\ &\quad + \frac{h^4}{24}A^4 + \frac{h^5}{144}A^5 + \mathcal{O}(h^6)\,. \end{aligned} \tag{4.24}$$

Comparing now the expansion of the $(2,2)$-Padé approximation to the exponential function obtained in (4.24) with the exponential series of the same order

$$p_{n+1} = \exp(hA)p_n = \Big(I + hA + \frac{h^2}{2}A^2 + \frac{h^3}{6}A^3 + \frac{h^4}{24}A^4 + \frac{h^5}{120}A^5\Big)p_n + \mathcal{O}\big(h^6\big)$$

we arrive at the expression for the *local error*

$$\begin{aligned}
\exp(hA)p_n - p_{n+1} &= \exp(hA)p_n - \Big(I - \frac{h}{2}A + \frac{h^2}{12}A^2\Big)^{-1}\Big(I + \frac{h}{2}A + \frac{h^2}{12}A^2\Big)p_n \\
&= h^5\Big(\frac{1}{120} - \frac{1}{144}\Big)A^5 p_n + \mathcal{O}\big(h^6\big) \\
&= \frac{h^5}{720}A^5 p_n + \mathcal{O}\big(h^6\big)
\end{aligned}$$

which we have used in (4.20).

In order to derive an error bound for the last term in (4.19), we state that $p(t) = q(t) - d(t)$ and obtain that $d(t)$ satisfies the equation

$$\begin{aligned}
\dot{d}(t) &= \dot{q}(t) - \dot{p}(t) \\
&= \mathcal{P}_V\mathcal{A}q(t) - \mathcal{A}p(t) \\
&= \mathcal{P}_V\mathcal{A}q(t) - \mathcal{A}q(t) + \mathcal{A}d(t) \\
&= \mathcal{A}d(t) + (\mathcal{P}_V - I)\mathcal{A}q(t).
\end{aligned} \tag{4.25}$$

Applying the variation-of-constants formula to (4.25) yields

$$d(t) = d(t_n) + \int_{t_n}^{t} \exp\big((t-s)\mathcal{A}\big)(\mathcal{P}_V - I)\mathcal{A}q(s)ds$$

where $\exp((t-t_n)\mathcal{A})$ denotes the flow of the CME (4.16). Since from (4.16) and (4.17) we know that $d(t_n) = q(t_n) - p(t_n) = 0$, and additionally we have from (2.75) that $\|\exp((t-t_n)\mathcal{A})\|_1 = 1$ for all $t \geq t_n$, it follows that

$$\|q(t_{n+1}) - p(t_{n+1})\|_1 = \|d(t_{n+1})\|_1 \leq \int_{t_n}^{t_{n+1}} \|(\mathcal{P}_V - I)\mathcal{A}q(s)\|_1 ds.$$

Substituting these bounds in (4.19) proves the assertion made in (4.18). ∎

We turn now to some computational issues related to the terms appearing in (4.18). First, the term $h^5 \left\|(\mathcal{P}_V\mathcal{A})^5 p_n\right\|_1 /720$ which arises from the time integration of the *projected* CME will be treated. Evaluating the expression $(\mathcal{P}_V\mathcal{A})^5 p_n$ in a straightforward way would imply five evaluations of $\mathcal{A}$ on the current numerical approximation p_n, but fortunately, this can easily be avoided by using the sparse wavelet representation $p_n = \sum_{i=1}^{\hat{\mu}} \beta_i \psi_{j_i}$. Then, the corresponding wavelet representation for the desired term is given by

$$(\mathcal{P}_V\mathcal{A})^5 p_n = \sum_{i=1}^{\hat{\mu}} \zeta_i \psi_{j_i}, \quad \text{where} \quad (\zeta_1, \ldots, \zeta_{\hat{\mu}})^T = M^5(\beta_1, \ldots, \beta_{\hat{\mu}})^T.$$

Hence, we have shifted the computation in the low-dimensional space, and only the relatively small Galerkin matrix M has to be applied five times, not the full operator $\mathcal{A}$.

Next, the integral term in (4.18) which represents the error caused by the spatial approximation has to be evaluated. The term can be understood in the sense that it describes how the solution of the *projected* CME differs from the solution of the *full* CME. However, one problem which immediately arises, is that since the function $q(t)$ is not computed within the adaptive wavelet method, an exact evaluation of this term is not available. Therefore, the term will be substituted with the following first order-approximation based on the rectangle method

$$\int_{t_n}^{t_{n+1}} \|(\mathcal{P}_V - I)\mathcal{A}q(s)\|_1 ds \;\approx\; (t_{n+1} - t_n)\|(\mathcal{P}_V - I)\mathcal{A}q(t_n)\|_1 \;=\; h\|(\mathcal{P}_V - I)\mathcal{A}p_n\|_1. \tag{4.26}$$

With (4.18) and (4.26) the condition $\|p(t_{n+1}) - p_{n+1}\|_1 \approx \texttt{tol}$ leads to the step-size selection formula

$$h \;=\; \min\left\{ \frac{\texttt{tol}}{\|(\mathcal{P}_V - I)\mathcal{A}p_n\|_1} \,,\; C_{safe} \cdot \left(\frac{720 \cdot \texttt{tol}}{\|(\mathcal{P}_V\mathcal{A})^5 p_n\|_1} \right)^{1/5} \right\} \tag{4.27}$$

with an optional safety factor $C_{safe} \leq 1$. However, we are faced now with the difficulty that the step-size h has to be chosen *before* the time step $p_n \mapsto p_{n+1}$ is carried out, but the space V is only known *after* the time step. At a particular time t_n, only the subspace

$$W = span\{\psi_{j_1}, \ldots, \psi_{j_\eta}\} \subset V \subset \mathcal{H}(\Omega_\xi)$$

which is spanned by the basis elements from the representation $p_n = \sum_{i=1}^{\eta} \beta_i \psi_{j_i} \approx p(t_n)$ is available. For the estimate (4.20), this difference is negligible, because this term estimates the error caused by the time integration. The problem is that simply replacing in the estimate of the spatial error V with W is far too pessimistic. We can however compute an estimate for the term $\|(I - \mathcal{P}_V)\mathcal{A}p_n\|_1$ by using a prediction of how many new basis elements are going to be added in the next time step. Let

$$\mathcal{A}p_n = \sum_{l=1}^{N} \theta_l \psi_l \tag{4.28}$$

be the representation of $\mathcal{A}p_n$ in the wavelet basis. First, we apply the projection $(I - \mathcal{P}_W)$ which removes all terms with index $l \in \{j_1, \ldots, j_\eta\}$ from (4.28). Next, from the remaining coefficients, the m largest coefficients in absolute value are discarded, because the corresponding basis elements are most likely of being selected during the enlargement process. The choice of the parameter m depends on how many basis elements are in the current active set which has cardinality (η), and on the maximal number of basis elements (μ_{max}), i.e. $m = s \cdot (\mu_{max} - \eta)$ with some safety factor $s \in [0, 1]$. In the numerical experiments we present later, we have used the value $s = 0.5$. When m is chosen large, then the new step-size h will also be large, but the downside is that this means that more

basis elements will be necessary for the approximation. We summarize now the above considerations in the following Algorithm 4 dedicated to step-size selection.

Algorithm 4: Adaptive step-size selection

Parameter : error tolerance $\texttt{tol} > 0$

Input : index subset $\{j_1, \ldots, j_\eta\}$ and coefficients $\beta_1, \ldots, \beta_\eta$ of the current approximation $p_n = \sum_{i=1}^{\eta} \beta_i \psi_{j_i}$
Galerkin matrix M defined by (4.3)

Output : step-size h for the step $t_n \mapsto t_{n+1} = t_n + h$

begin

1. Compute h_{space}:
   a) Compute $\mathcal{A}p_n$ and, via a fast wavelet transform, its representation (4.28).
   b) Set $\theta_l = 0$ for all $l = j_1, \ldots, j_\eta$.
   c) Put $m = s \cdot (\mu_{max} - \eta)$ and set the m largest (in modulus) coefficients to zero. With a fast inverse wavelet transform, compute $\varsigma = \sum_{l \notin \mathcal{D}} \theta_l v_l$ where $\mathcal{D}$ is the index set of the discarded terms.
   d) Set $h_{\text{space}} = \texttt{tol}/\|\varsigma\|_1$.
2. Compute h_{time}:
   a) Compute
   $$(\mathcal{P}_W \mathcal{A})^5 p_n = \sum_{i=1}^{\eta} \zeta_i \psi_{j_i}, \qquad (\zeta_1, \ldots, \zeta_\eta)^T = M^5 (\beta_1, \ldots, \beta_\eta)^T.$$
   b) Set $h_{\text{time}} = C_{safe} \cdot \left(\dfrac{720 \cdot \texttt{tol}}{\|(\mathcal{P}_W \mathcal{A})^5 p_n\|_1} \right)^{1/5}$
3. Choose $h = \min \left\{ h_{\text{space}}, h_{\text{time}} \right\}$.

end

Algorithm 4 is started at the beginning of every time step, i.e., before the main body of Algorithm 3 from Section 4.2. As a final remark, we note that the step-selection strategy *fixes* the step-size at the beginning of each iteration corresponding to a time-step. Changing the step-size during the iteration (steps 2 to 5 from Algorithm 3) by using the information gathered in the process of enlarging the approximation space was tested, but ultimately proved unsuccessful. The reason was that the decision which basis elements are selected for inclusion in the *essential* subset depends implicitly on the step-size. If the step-size is changed during the enlargement process, then basis elements that have been previously selected might no longer be suitable with a new step-size, which leads to strong oscillations in the step-size, and consequently to a decrease in the efficiency of the method.

4.4. Numerical examples

In this section we illustrate the potential of the adaptive wavelet method by numerically solving five model problems from molecular biology and epidemiology. The purpose of the first two examples is to investigate the accuracy of the method by comparing the results with reference solutions obtained by other means. The remaining three models are then used to showcase the ability of the wavelet method to deal with problems having large state spaces and metastable solution profiles.

4.4.1. Merging Modes

The first model is a toy problem that consists of two species S_1 and S_2 interacting via the reaction channels

$$\begin{array}{llllll|lll|lll} R_1: & S_1 & \longrightarrow & S_2 & & & \alpha_1 & = & c_1 x_1 & \mu^1 & = & (-1,1)^T \\ R_2: & S_2 & \longrightarrow & S_1 & & & \alpha_2 & = & c_2 x_2 & \mu^2 & = & (1,-1)^T \\ R_3: & S_1 & \longrightarrow & \star & & & \alpha_3 & = & c_3 x_1 & \mu^3 & = & (-1,0)^T \\ R_4: & S_2 & \longrightarrow & \star & & & \alpha_4 & = & c_4 x_2 & \mu^4 & = & (0,-1)^T \end{array} \tag{4.29}$$

and where the rate constants take the following values $c_1 = 1.5, c_2 = 0.7, c_3 = 0.7$ and $c_4 = 0.2$. Although the model and the parameters are not biologically relevant, this simple example has the advantage that an *exact* solution of the corresponding CME can be computed by using the analytical method proposed in [JH07] for *monomolecular* reaction systems. Thus, the model allows us to investigate the behavior of the error with respect to the user-defined tolerance. The same model, but with different parameters has been used for the same purpose in [Jah10].

We define now the multinomial distribution $\mathcal{M}(\mathbf{x}, N, r)$, which is a two-dimensional extension of the well known binomial distribution, as

$$\mathcal{M}(\mathbf{x}, N, r) = \begin{cases} N! \dfrac{r_1^{x_1}}{x_1!} \dfrac{r_2^{x_2}}{x_2!} \dfrac{(1-r_1-r_2)^{N-x_1-x_2}}{(N-x_1-x_2)!} & \text{if } x_1 + x_2 \leq N \\ 0 & \text{otherwise,} \end{cases}$$

for any $\mathbf{x} = (x_1, x_2) \in \mathbb{N}^2$, $N \in \mathbb{N}$ and any $r = (r_1, r_2)$ with $r_1, r_2 \in [0,1]$ and $r_1 + r_2 \leq 1$. As initial distribution of the CME problem (2.67),

$$\rho(\mathbf{x}) = 0.5 \cdot \mathcal{M}(\mathbf{x}, N, r^{(1)}) + 0.5 \cdot \mathcal{M}(\mathbf{x}, N, r^{(2)}) \tag{4.30}$$

was chosen, with $r^{(1)} = (0.7, 0.1)^T$, $r^{(2)} = (0.1, 0.7)^T$, and $N = 63$. Then, the *exact* CME solution of the (4.29) model is given by

$$p(t, \mathbf{x}) = 0.5 \cdot \mathcal{M}(\mathbf{x}, N, s^{(1)}(t)) + 0.5 \cdot \mathcal{M}(\mathbf{x}, N, s^{(2)}(t)), \tag{4.31}$$

with

$$s^{(i)}(t) = \exp(t\mathcal{C}) r^{(i)}, \quad \mathcal{C} = \begin{pmatrix} -(c_1 + c_3) & c_2 \\ c_1 & -(c_2 + c_4) \end{pmatrix}$$

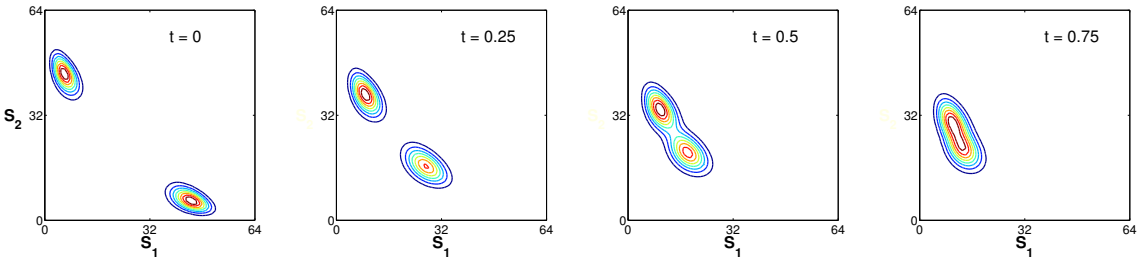


Figure 4.5.: Exact solution of the Merging Modes system at $t = 0, t = 0.25, t = 0.5$ and $t = 0.75$ (from left to right).

(cf. [JH07]). In Figure 4.5 snapshots of the time evolution of the CME solution are shown, depicting how the modes of $p(t, \mathbf{x})$ merge into one single peak.

The adaptive wavelet method was applied to the model (4.29) to obtain a numerical approximation of the CME solution on the time interval $[0, 1]$. As wavelet basis, an *isotropic* tensor product using *db2* wavelets was employed, and four independent runs using different tolerances for the 1-norm of the residual were performed. In Figure 4.6a, the error of the adaptive wavelet method for each of the four tolerances is shown. The errors are measured in the 1-norm and are obtained by comparing the approximations at the time steps chosen by the adaptive method with the corresponding explicitly derived solutions.

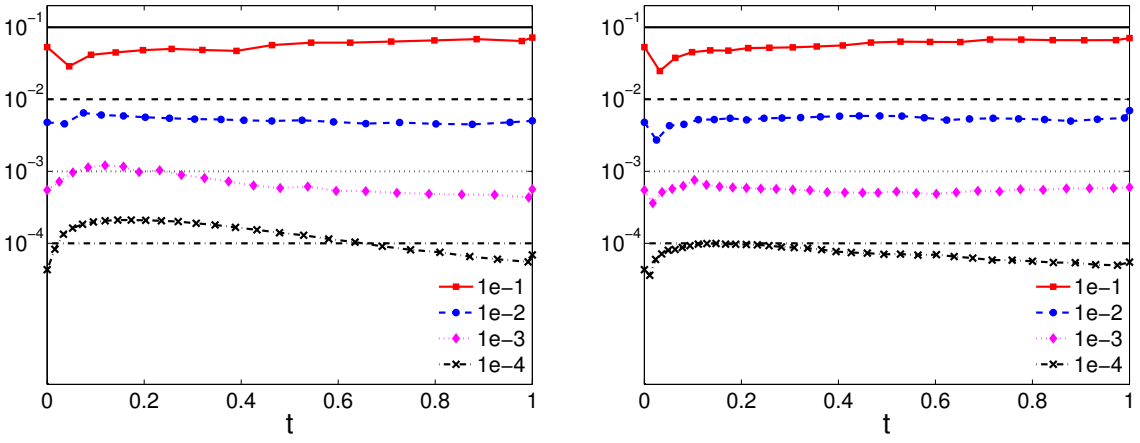
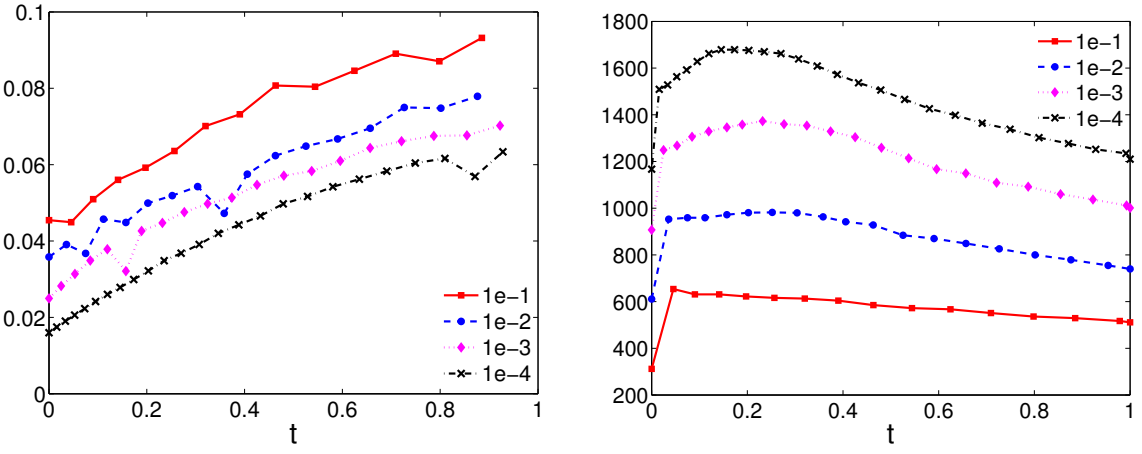


Figure 4.6.: Left panel (a): Error of the adaptive wavelet approximation of the Merging Modes problem (4.29) for $\texttt{tol}_1 = 10^{-1}$ (square), $\texttt{tol}_2 = 10^{-2}$ (circle), $\texttt{tol}_3 = 10^{-3}$(diamond) and $\texttt{tol}_4 = 10^{-4}$ (cross). The error was computed in the 1-norm by comparing each of the approximations with the exact solution. Right panel (b): Error of the adaptive wavelet approximation for $\texttt{tol} = 10^{-1}, 10^{-2}, 10^{-3}$ and 10^{-4} using a safety factor $C_{safe} = 0.7$ for h_{time}.

The results indicate that for tolerances up to $\texttt{tol} = 10^{-3}$ the error estimator given by (4.27) performs well, as the error is almost always below the chosen tolerance. In case smaller tolerances are used, however, some of the adaptively chosen step-sizes are slightly too optimistic, which translates into the error crossing the imposed barrier. We remark that this behavior appears only for tolerances that are going to be used for small problems. For problems with large state spaces, tolerances of 10^{-1} or 10^{-2} are sufficient

to provide an accurate approximation, because the 1-norm scales with the state space. There is also a simple countermeasure available to mitigate such error behavior, namely the use of a safety factor C_{safe} in the second term in (4.27), and the results obtained are presented in Figure 4.6b. In order to provide a comprehensive picture of the adaptive wavelet method, we also plot in Figure 4.7 the evolution of the step size and the number of basis elements used by the four runs without the safety factor.



(a) (b)

Figure 4.7.: Left panel (a): Evolution of the step-size h for the Merging Modes problem without the safety factor, using $\texttt{tol}_1 = 10^{-1}$ (solid), $\texttt{tol}_2 = 10^{-2}$ (dashed), $\texttt{tol}_3 = 10^{-3}$(dotted) and $\texttt{tol}_4 = 10^{-4}$ (dash-dot). Right panel (b): Number of basis elements used in each step to compute the approximation for $\texttt{tol}_1 = 10^{-1}$ (solid), $\texttt{tol}_2 = 10^{-2}$ (dashed), $\texttt{tol}_3 = 10^{-3}$ (dotted) and $\texttt{tol}_4 = 10^{-4}$ (dash-dot).

4.4.2. Genetic Toggle Switch

In this example, we revisit the genetic toggle switch described by the reaction network (2.94), but with a different set of parameters, namely

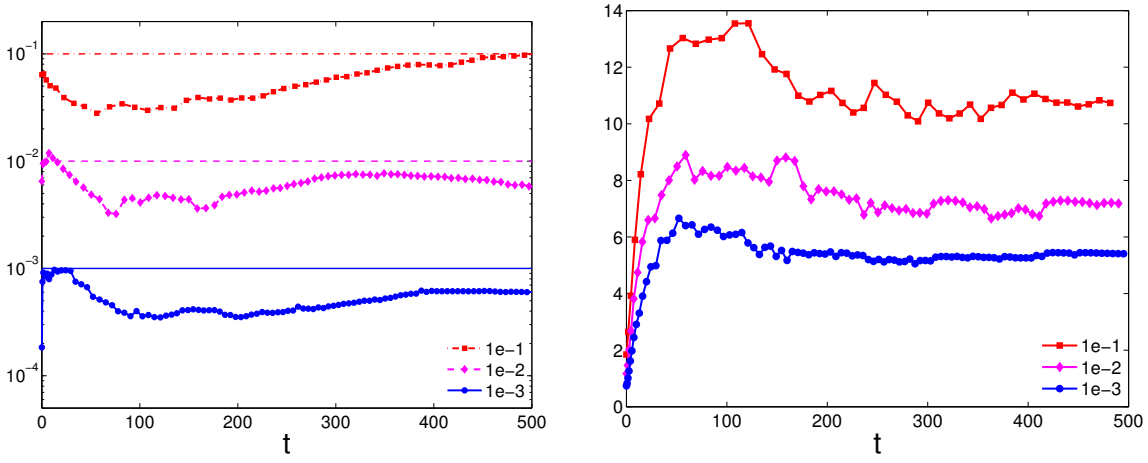
$$c_{11} = c_{21} = 10, \quad c_{12} = c_{22} = 30 \text{ and } c_3 = c_4 = 0.017. \tag{4.32}$$

Because the model (2.94) contains reaction channels with non-standard propensities, explicit solution formulas are no longer available. Fortunately, however, the truncated state space induced by the choice of parameters given in (4.32), contains only 32×32 total degrees of freedom and thus is small enough such that a "reference" solution can be obtained by solving the CME directly via the MATLAB routine $ode15s$.

The adaptive wavelet method was then used to obtain approximations of the CME on the time interval $[0, 500]$ using $db3$ wavelets. For an initial distribution, a "discrete Gaussian" centered at $\nu = (20, 18)$ and given by

$$\begin{aligned} p(0, \mathbf{x}) &= c_0 \cdot \exp(-(\mathbf{x} - \nu)^T C(\mathbf{x} - \nu)), \text{ for all } \mathbf{x} \in \Omega_\xi, \\ C &= \begin{pmatrix} 10000 & 0 \\ 0 & 10000 \end{pmatrix} \end{aligned}$$

was selected, with c_0 denoting a normalization constant obtained from the condition $\sum_{\mathbf{x}\in\Omega_\xi} p(0,\mathbf{x}) = 1$. In order to test the accuracy of the adaptive wavelet method for this small variant of the *toggle switch* model (2.94), three different runs using the same method parameters but different tolerances were performed, with results being shown in Figure 4.8. In the left panel 4.8a, the 1-norm error for each of the three tolerances is plotted, and we remark that the error was computed by comparing the approximations with results for the same time points provided by MATLAB's $ode15s$. Examining Figure 4.8a reveals that the error lies below the chosen tolerance, which means that the wavelet method provides in this case a very good approximation of the exact solution at the desired accuracy. In the right panel 4.8b, the time evolution of the step-size h corresponding to the different tolerances is shown. The plot supplies proof of the advantage of using adaptive step-size control, as it is clear that the *toggle switch* model requires small step sizes only for an initial stiff transient phase at the beginning of the time interval, while larger time steps can be used in a subsequent phase without a negative impact on the accuracy. As expected, the adaptive method selects larger time steps for low tolerances, while higher tolerances imply the use of smaller step-sizes.



(a) (b)

Figure 4.8.: Left panel (a): Error of the adaptive wavelet approximation of the toggle switch for $\texttt{tol}_1 = 0.1$ (square), $\texttt{tol}_2 = 0.01$ (diamond) and $\texttt{tol}_3 = 0.001$ (circle). The error was computed in the 1-norm by comparing each of the approximations with the reference solution. Right panel (b): Evolution of the step-size h for the toggle switch solved by the adaptive wavelet method with $\texttt{tol}_1 = 0.1$ (square), $\texttt{tol}_2 = 0.01$ (diamond) and $\texttt{tol}_3 = 0.001$ (circle).

4.4.3. Extended Toggle Switch

After using small test problems to investigate the accuracy of the method, we consider now a model with a far larger state space. This is another genetic toggle switch, which is obtained by extending the model from (2.94) by appending two more species and corresponding reaction channels. The resulting model then consists of two mutually repressing gene products, S_1 and S_2, with each gene expressing a different protein, denoted by S_3 and S_4, respectively.

The interactions between these four species ($d = 4$) are listed below:

$$\begin{array}{lcl|lcl|lcl} R_1: & \star & \longrightarrow S_1 & \alpha_1 & = & c_{11}/(c_{12}+x_2^2) & \mu^1 & = & (1,0,0,0)^T \\ R_2: & \star & \longrightarrow S_2 & \alpha_2 & = & c_{21}/(c_{22}+x_1^2) & \mu^2 & = & (0,1,0,0)^T \\ R_3: & S_1 & \longrightarrow \star & \alpha_3 & = & c_3 x_1 & \mu^3 & = & (-1,0,0,0)^T \\ R_4: & S_2 & \longrightarrow \star & \alpha_4 & = & c_4 x_2 & \mu^4 & = & (0,-1,0,0)^T \\ R_5: & S_1 & \longrightarrow S_1+S_3 & \alpha_5 & = & c_5 x_1 & \mu^5 & = & (0,0,1,0)^T \\ R_6: & S_2 & \longrightarrow S_2+S_4 & \alpha_6 & = & c_6 x_2 & \mu^6 & = & (0,0,0,1)^T \\ R_7: & S_3 & \longrightarrow \star & \alpha_7 & = & c_7 x_3 & \mu^7 & = & (0,0,-1,0)^T \\ R_8: & S_4 & \longrightarrow \star & \alpha_8 & = & c_8 x_4 & \mu^8 & = & (0,0,0,-1)^T \end{array} \tag{4.33}$$

We remark that reactions R_1 through R_4 are identical to the reaction network of the classic *toggle switch* from (2.94), with R_7 and R_8 modeling the decay of the added species S_3 and S_4. The expression of S_3 and S_4, which involves the genes S_1 and S_2 respectively, is described by reactions R_5 and R_6. The parameters for the reaction channels from (4.33) are

$$c_{11} = c_{21} = 10, \quad c_{12} = c_{22} = 30, \quad c_3 = c_4 = 0.017, \quad c_5 = c_6 = c_7 = c_8 = 0.01, \tag{4.34}$$

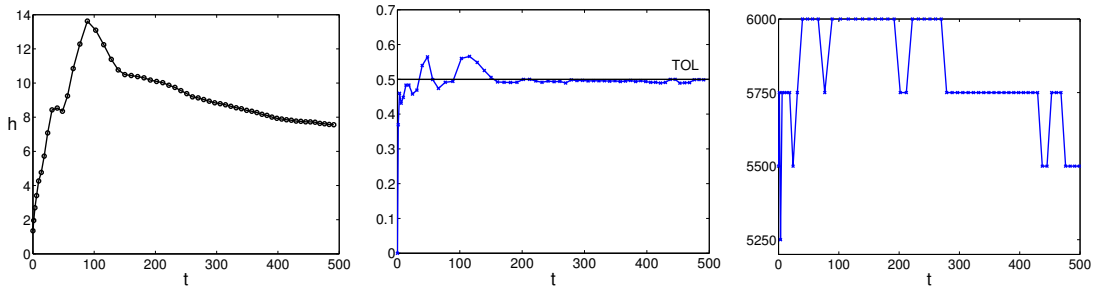
and as initial distribution of the CME problem, a $4D$-"discrete Gaussian" with a small variance, centered on the state $\nu = (20, 18, 22, 5)$ is used. This choice closely resembles a delta peak located at ν. The parameter set from (4.34) leads to a truncated state space of $32 \times 32 \times 32 \times 32 \approx 2^{20}$ total DOFs, which means that the corresponding CME can no longer be solved by traditional methods. However, we remark that by eliminating from the model the reactions involving the proteins S_3 and S_4, i.e., R_5 through R_8, we obtain the simplified $2D$ toggle switch (2.94) presented in Section 4.4.2. Because the solution of the original $2D$-model agrees with the marginal distribution for $S_1 - S_2$ of the larger $4D$-model, it follows that we can again use MATLAB's $ode15s$ routine to obtain a sort of reference solution by solving the CME for (2.94) on the truncated state space $\Omega_{32,32}$ directly, and comparing the result with the marginal of the approximation obtained by the adaptive wavelet method applied to the extended toggle switch (4.33).

The corresponding CME was approximated on the time interval $[0, 500]$, with the adaptive wavelet method being configured to use $\texttt{tol} = 0.5$ for the 1-norm of the residual. Although such a tolerance may seem exceedingly large, as the 1-norm scales with the state space and we have now 2^{20} total DOFs, this choice actually provides enough accuracy. As an example, considering an equally distributed error ε, a 1-norm measurement of $\|\varepsilon\|_1 = 0.5$ corresponds to $\|\varepsilon\|_\infty = 0.5/2^{20} \approx 4.77 \cdot 10^{-7}$ when the error is measured in the maximum norm. In the examples studied so far, all of which featured small state spaces, imposing limits for the number of DOFs in the *essential* set used in each time step did not play an important role. For the current model, however, using such limits brings an increase in computational efficiency. Thus, we have configured the method to keep a minimum of 5000 of the largest coefficients at the end of each time step, while the total number of elements that could be used within the algorithm was not allowed to exceed 6000. Hence, the solution was approximated using only 0.47% of the total number of $1,048,576$ degrees of freedom. New basis elements were proposed in batches of 250 elements each and the $db3$ wavelet basis was chosen to approximate the solution.

Approximations to the CME obtained with the adaptive wavelet method are shown in Figure 4.13. We remark that as the full distribution is a $4D$-object, only plots of the most

interesting $2D$-marginals at a succession of time points are given. The first two columns depict mesh and contour plots of the $2D$-marginal distribution of the two gene products $S_1 - S_2$, while in the third column, a contour plot for the $2D$-marginal distribution of the proteins $S_3 - S_4$ is provided. The two marginals clearly illustrate the multi-modal character of the solution profile at $t = 500$.

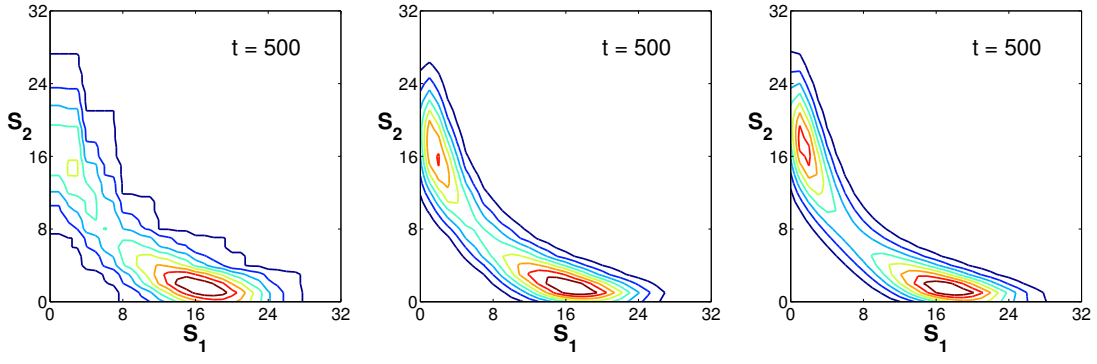
In the left panel (a) of Figure 4.9, the evolution of the step-size h for the wavelet integrator is shown. As was the case with the smaller model from which the current model is derived, small step-sizes are required in the initial stiff phase, with larger time-steps being selected as the approximation of the probability distribution approaches the invariant distribution. The middle panel (b) shows the evolution of the 1-norm of the residual, while the number of basis elements used by the adaptive wavelet method during the course of the simulation is given in the right panel (c).



(a) (b) (c)

Figure 4.9.: Left panel (a): Evolution of step-size h for the $4D$ toggle switch. Middle panel (b): Evolution of the 1-norm (scaled) of the residual for the same problem. Right panel (c): Number of basis elements used in each step to compute the approximation.

Further, in Figure 4.10, we use the previously discussed feature that the $2D$ marginal of the full $4D$ model can be compared with the solution of the small toggle switch from Section 4.4.2, and investigate the effects of using higher-order wavelet bases for problems with large state spaces. In the leftmost panel (a), the marginal distribution for species $S_1 - S_2$ obtained from an wavelet approximation at $t = 500$ for the $4D$ toggle switch using a multivariate Haar basis is shown.



(a) (b) (c)

Figure 4.10.: Comparison between approximations for the $4D$ toggle switch obtained using Haar basis (a), Daubechies $db3$ wavelet basis (b) and Matlab's $ode15s$ on the simplified $2D$ problem (c).

The middle panel displays the results obtained using the same parameters for the wavelet solver, but this time employing a $db3$ wavelet basis. The "reference" solution computed using MATLAB's $ode15s$ on the simplified $2D$ model is shown in the right panel. The results confirm that for problems of a certain size, the Haar wavelet basis used in [Jah10] can no longer cope, as the number of DOFs required to achieve a similar accuracy to the higher-order $db3$ basis would simply be too high. The increase in the size of the active DOFs set would then drive up the computational cost. Consequently, by providing the flexibility to choose from wavelets from the Daubechies wavelet familiy (including Haar), or biorthogonal wavelet bases, the improved wavelet method from [JU10] which was reviewed in this chapter allows the efficient numerical treatment of non-trivial problems.

4.4.4. Infectious diseases

In the following example, we temporarily leave gene regulatory networks to study the classic SEIR epidemic model that describes the spread of communicable diseases within a population (see [Het00] for an in-depth presentation). The SEIR model assumes that a population is split into four distinct classes ($d = 4$), namely individuals susceptible of becoming infected with a disease (S), exposed individuals (E) that are infected but not yet contagious, infectious individuals (I) and individuals that have recovered (R), and in the process have acquired immunity to the disease. The sub-populations of the model interact through seven reaction channels, as follows

$$\begin{array}{lcc|ccc|ccc} R_1: & S+I & \longrightarrow & E+I & \alpha_1 & = & c_1 x_1 x_3 & \mu^1 & = & (-1,1,0,0)^T \\ R_2: & E & \longrightarrow & I & \alpha_2 & = & c_2 x_2 & \mu^2 & = & (0,-1,1,0)^T \\ R_3: & I & \longrightarrow & S & \alpha_3 & = & c_3 x_3 & \mu^3 & = & (1,0,-1,0)^T \\ R_4: & S & \longrightarrow & \star & \alpha_4 & = & c_4 x_1 & \mu^4 & = & (-1,0,0,0)^T \\ R_5: & E & \longrightarrow & \star & \alpha_5 & = & c_5 x_2 & \mu^5 & = & (0,-1,0,0)^T \\ R_6: & I & \longrightarrow & R & \alpha_6 & = & c_6 x_3 & \mu^6 & = & (0,0,-1,1)^T \\ R_7: & \star & \longrightarrow & S & \alpha_7 & = & c_7 & \mu^7 & = & (1,0,0,0)^T \end{array} \tag{4.35}$$

The first reaction R_1 models the process through which susceptible individuals become infected by having contact with infectious ones. Individuals coming in contact with the disease first enter a latent phase and are assigned to the E sub-population. After an incubation period, they can become infectious themselves via reaction R_2 or can die of other causes, as described by reaction R_5. The temporary recovery of infected individuals can occur via reaction R_3, while by reaction R_6 the recovery process in which these individuals also acquire immunity to the disease is modeled. Reaction R_4 describes the death of susceptible individuals, whereas reaction R_7 represents new arrivals that are prone to becoming infected. We assume that the arrival of susceptible individuals via reaction R_7 is constant and is independent of the current size of the population. The model is studied for the case where the disease starts only with a few infected individuals, which means that a stochastic treatment is mandatory. As we shall see, the solution profile is multi-modal, as there are two possible scenarios: either the disease quickly spreads to a large section of the population or disappears at some early stage because the first few infectious individuals have already died. The parameters chosen for the simulation were

$$c_1 = 0.1, \quad c_2 = 0.5, \quad c_3 = 1, \quad c_4 = c_5 = c_6 = 0.01, \quad c_7 = 0.4,$$

and as initial distribution, a "discrete Gaussian" centered at $\nu = (50, 4, 0, 0)$ was considered.

In Figure 4.14 we plot snapshots of the evolution of the probability distribution of the SEIR model obtained by applying the adaptive wavelet method. For the marginal distribution in the S-E plane, both contour and mesh plots are shown (left and middle column, from top to bottom). The rightmost column shows a contour plot for the marginal distribution in the S-I plane. The time interval chosen was $[0, 7]$, and during the course of the simulation, the marginal distribution in the S-E plane splits up into two distinct peaks. The peak located at roughly $(50, 0)$ represents the scenario in which the first few infectious individuals have either died or recovered before their numbers reached a critical mass that could sustain the epidemic. Consequently, the disease disappears after some time. In the other scenario, the infection spreads quickly enough during the initial phase. With an increase in the number of carriers, the system will eventually reach a stage where the majority of the population is affected by the disease. The peak corresponding to this scenario is located around $(11, 27)$. The multi-modal character of the system is also evident in other marginal distributions, i.e., for the S-I species (left column) and for the E-I species (data not shown).

At this point, we remark that the multi-modal character of the solution, with peaks located far apart within the state space, poses no significant challenges to the wavelet method. The solution also exhibits a non-smooth character as it can be noticed in the last row of Figure 4.14. At the final integration time point $t = 7$, the solution vanishes close to the S-axis but does not vanish *on* the axis itself. Although wavelets are best suited for the approximation of sufficiently smooth signals, the method is also able to handle such difficult scenarios, where a certain degree of local non-smoothness is present.

Additionally, in the period from $t = 3$ to $t = 5$, the solution profile of the SEIR model is not parallel to any of the axes. Such behavior would pose challenges to methods where the solution is represented in terms of global tensor products (e.g., the method proposed in [JH08]), as they would need many more degrees of freedom to represent such profiles. The adaptive wavelet method overcomes such issues, as the elements of the multivariate wavelet bases are tensor products with *local* support.

The specific solver parameters used for this model are a $db3$ wavelet basis and a tolerance $\texttt{tol} = 0.61$ for the 1-norm of the residual. The inner iteration was stopped if the number of degrees of freedom exceeded 6500 and only a maximum of 6000 DOFs were kept at the end of each time step, which corresponds to the use of only 0.57% of the total 2^{20} degrees of freedom. New basis elements were proposed in batches of 250 elements each and solving the linear system (4.11) was accomplished using GMRES with restarts and a tolerance of $5 \cdot 1e^{-4}$. The method required 122 steps to simulate the evolution of the probability distribution for the chosen time interval $[0, 7]$.

4.4.5. Transcription regulation

As a last example, we present a simplified model of transcription regulation which was first introduced in [Gou05] and is usually referenced in the literature as the *Goutsias*

model. Although simplified, the model is biologically interesting as it represents a subsystem of the λ-phage, and can be used to model the transcription regulation of the CI protein which is responsible for maintaining the lysogenic cycle of viral reproduction in *E.coli* [Pta04, ARM98, Gou05]. Furthermore, this particular model has also been previously used to test other CME solvers like the Krylov FSP algorithm in [BHMS06]. The *Goutsias* model is a six species biochemical system, which is characterized by the following reaction channels

$$
\begin{array}{rrcl|rcl}
R_1: & \text{RNA} & \longrightarrow & \text{RNA} + \text{M} & \alpha_1 & = & c_1 x_3 \\
R_2: & \text{M} & \longrightarrow & \star & \alpha_2 & = & c_2 x_1 \\
R_3: & \text{DNA.D} & \longrightarrow & \text{RNA} + \text{DNA.D} & \alpha_3 & = & c_3 x_5 \\
R_4: & \text{RNA} & \longrightarrow & \star & \alpha_4 & = & c_4 x_3 \\
R_5: & \text{DNA} + \text{D} & \longrightarrow & \text{DNA.D} & \alpha_5 & = & c_5 x_2 x_4 \\
R_6: & \text{DNA.D} & \longrightarrow & \text{DNA} + \text{D} & \alpha_6 & = & c_6 x_5 \\
R_7: & \text{DNA.D} + \text{D} & \longrightarrow & \text{DNA.2D} & \alpha_7 & = & c_7 x_2 x_5 \\
R_8: & \text{DNA.2D} & \longrightarrow & \text{DNA.D} + \text{D} & \alpha_8 & = & c_8 x_6 \\
R_9: & \text{M} + \text{M} & \longrightarrow & \text{D} & \alpha_9 & = & c_9 (x_1 (x_1 - 1)/2) \\
R_{10}: & \text{D} & \longrightarrow & \text{M} + \text{M} & \alpha_{10} & = & c_{10} x_2
\end{array}
\tag{4.36}
$$

The model is best explained with the help of the schematic presented in Figure 4.11. There, we have the protein M (monomer) which is synthesized through the process of transcription of the gene into RNA (cf. reaction R_3) and subsequent translation of RNA into the protein through reaction R_1. Degradation of RNA and the monomer M are modeled by reactions R_4 and R_2, respectively. The monomers reversibly dimerize by reactions R_9 and R_{10} into transcription factors D (dimers), which then can reversibly bind to the operator sites O_1 and O_2 (processes modeled by reactions R_5 through R_8). The gene is labeled DNA, while DNA.D denotes the gene with a transcription factor D attached to the operator site O_1. Similarly, DNA.2D denotes the gene with both operator sites occupied by dimers, and the model assumes that D can bind at O_2 only if the site O_1 is already occupied. The gene produces RNA only if site O_1 is occupied, while binding of D at O_2 represses the transcription (cf. [Gou05]).

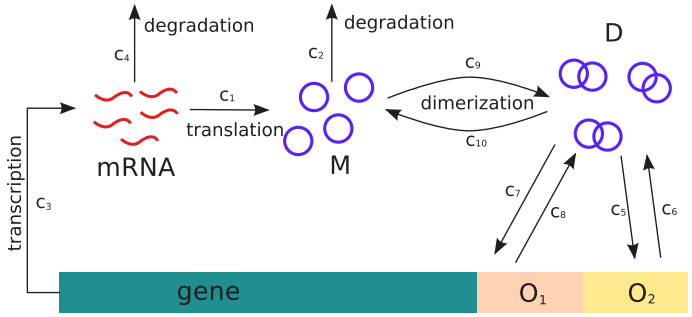


Figure 4.11.: Gene regulatory network for a subsystem of λ-phage (figure adapted from [Gou05])

In the description of the reaction system given in (4.36), we have retained the names used in [Gou05] in order to make the interpretation of the reaction channels easier, but to keep the propensities in the familiar description used throughout this thesis, we relabel the species as S_1 (M), S_2 (D), S_3 (RNA) , S_4 (DNA), S_5 (DNA.D) and S_6 (DNA.2D). The

reaction parameters are the same as those used in the original description by Goutsias and in [BHMS06], where the model was used for testing the Krylov FSP solver for the CME.

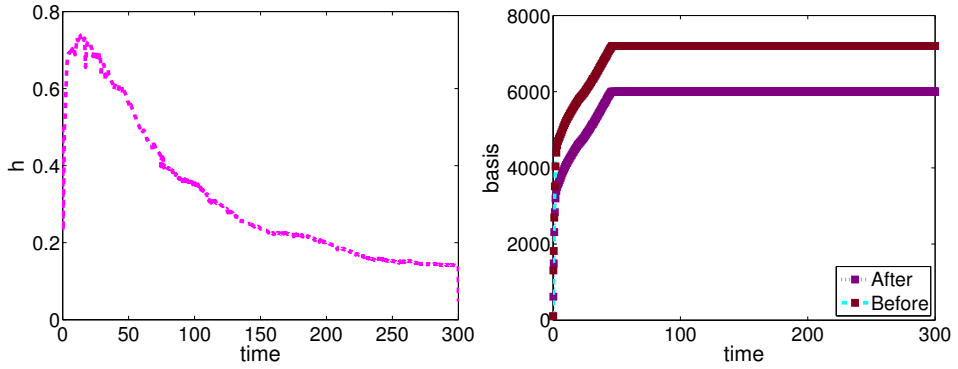
$$
\begin{aligned}
c_1 &= 0.043, & c_2 &= 0.0007, & c_3 &= 0.078,\\
c_4 &= 0.0039, & c_5 &= (0.012 \cdot 10^9)/n_A V, & c_6 &= 0.4791,\\
c_7 &= (0.00012 \cdot 10^9)/n_A V, & c_8 &= 0.8765 \cdot 10^{-11}, & c_9 &= (0.05 \cdot 10^9)/n_A V,\\
c_{10} &= 0.5
\end{aligned}
$$

with $n_A = 6.02214179 \cdot 10^{23}\ mol^{-1}$ being *Avogadro's* number and $V \approx 10^{-15} l$ the cell volume.

As initial distribution, a delta peak located at $\mathbf{x}_0 = (\, m,\, d,\, 0,\, g,\, 0,\, 0\,)^T$, with $m = 2$, $d = 6$ and $g = 2$, which are the biologically relevant parameters also used by Goutsias, was chosen. Because of the small number of genes, the number of possible configurations for some of the species is rather small, with S_4, S_5 and S_6 each having the possible configurations $\{0, 1, 2\}$. The species S_1, S_2 and S_3 are however endowed with a larger configuration space. Goutsias studied the model using stochastic simulations and observed how the configuration space for the dimers S_2 (D) explodes with increasing times, with the marginal distribution flattening and exhibiting an increasingly larger support.

Such models can be seen as being particularly challenging for all CME solvers, and difficulties with integrating the system for large times have also been reported in [BHMS06]. The model poses a challenge to the adaptive wavelet method as well, particularly because the state space has a combination of large and small directions, leading to a probability distribution which is highly non-smooth. The smoothness properties of the solution profile do not improve significantly with larger times either, also due to the presence of the small directions. Non withstanding these difficulties, we have applied the adaptive wavelet method to this model on a state space having $2^5 \times 2^6 \times 2^4 \times 2^2 \times 2^2 \times 2^2$ degrees of freedom. The time interval for integration was chosen as $[0, 300]$ and the method was configured to use a tolerance $\texttt{tol} = 0.2$ for the 1-norm of the residual. In contrast to the examples previously presented in this chapter, for the *Goutsias* model an *anisotropic* tensor product basis constructed from univariate *B-spline* 2.2 interval wavelet bases was chosen to approximate the solution. The switch to the biorthogonal wavelet bases on the interval also dictated a change to a Petrov-Galerkin scheme in order to preserve the identity matrices appearing in (4.11). We remark that a Galerkin ansatz could also have been used, but at the price of the evaluation of an extra Galerkin "mass" matrix, and that the Petrov-Galerkin scheme will be discussed in the next chapter. With respect to the other parameters, the maximum number of basis elements was not allowed to exceed 7000 inside the iteration dedicated to the refinement of the approximation space, with a maximum of 6000 basis elements being retained at the end of each time step. New basis elements were added in batches of 500 elements.

In Figure 4.15, marginal distributions at times $t \approx 10$, $t \approx 100$, $t \approx 200$ and $t = 300$ are shown which clearly illustrate how the number of active states explodes with increasing times. Integration beyond $t = 300$ is also possible, but as the maximum number of basis elements is kept fixed due to efficiency reasons, and the distribution enlarges its footprint, the adaptive time stepping strategy chooses increasingly small steps, which naturally



(a) (b)

Figure 4.12.: Evolution of the step size h for the *Goutsias* model and number of basis elements before/after truncation in each time step.

severely impacts the performance of the method. In Figure 4.12a the evolution of the time step, while in 4.12b the number of basis elements in each time step before and after truncation are plotted.

From a numerical point of view, the biggest limiting factor in computing the solution of the CME for all the non-trivial examples is the presence of huge state spaces with more than $1,000,000$ states. However, as it can be clearly seen in the panels of Figures 4.13, 4.14 or 4.15, most of these states are never populated throughout the time evolution of the corresponding probability distribution, which means that the subset of *essential* states is actually smaller. However, this information is of little practical use, because we only know which states can be ignored *a posteriori*. As the adaptive wavelet method is specifically designed to find the *essential* degrees of freedom, it is particularly suited to deal with these type of problems.

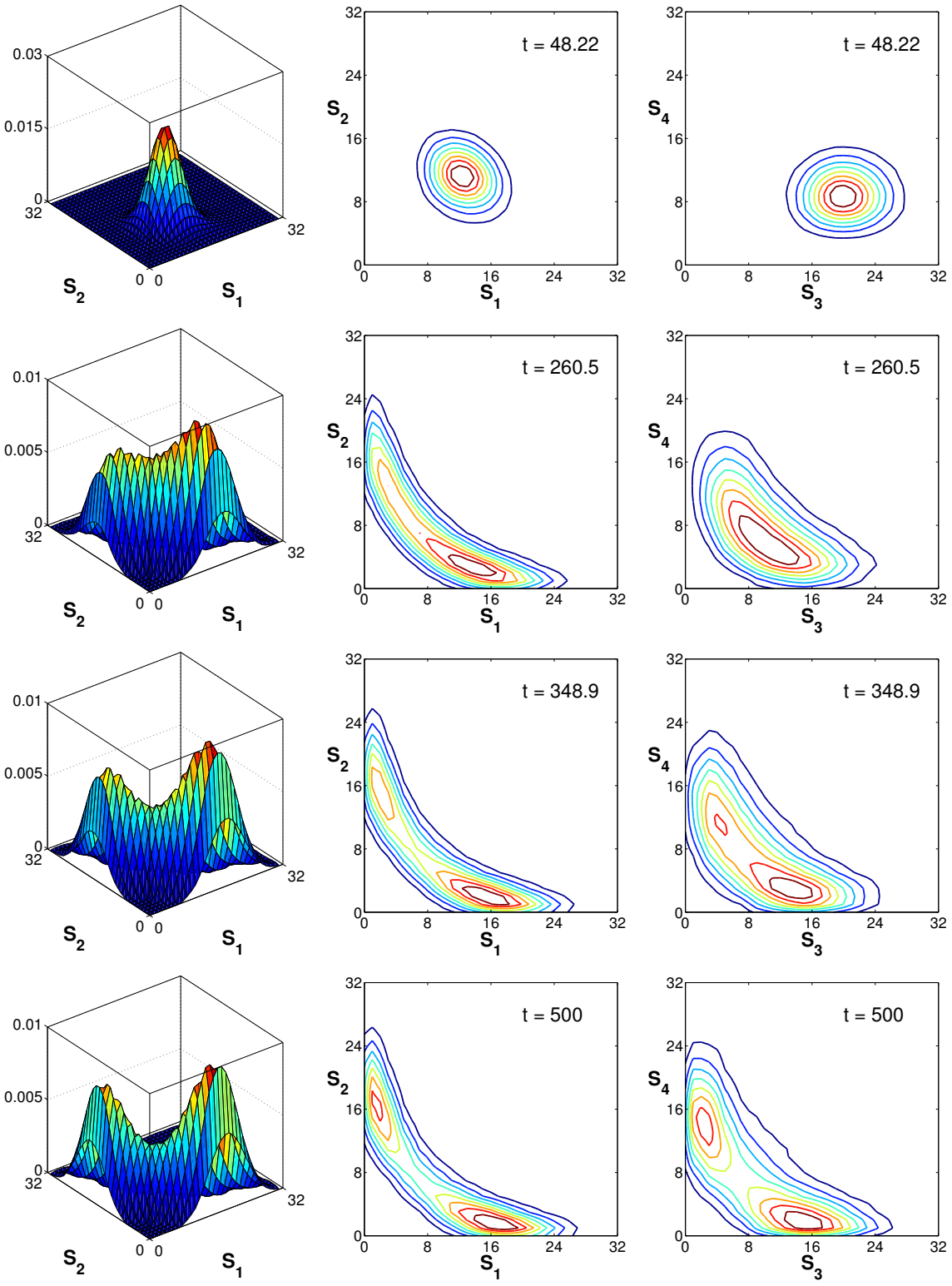


Figure 4.13.: Marginal distribution of the 4D toggle switch model (4.33) at different times. Surf plot (first column) and contour plots (second and third columns) of the approximation obtained with the adaptive wavelet method.

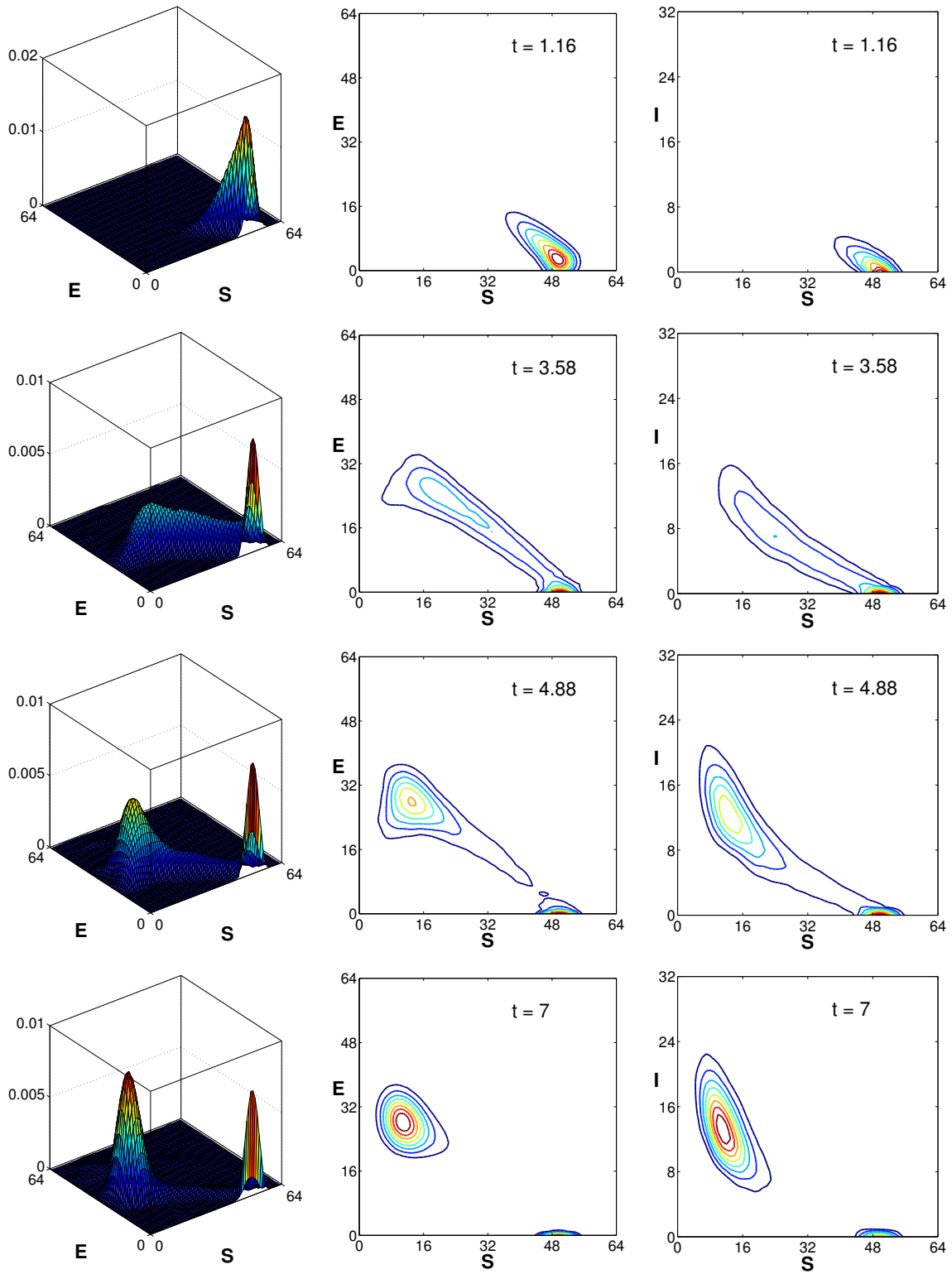


Figure 4.14.: Marginal distributions of the stochastic SEIR model (4.35) at different times. Surf plot (first column) and contour plots (second and third columns) of the approximation obtained with the adaptive wavelet method.

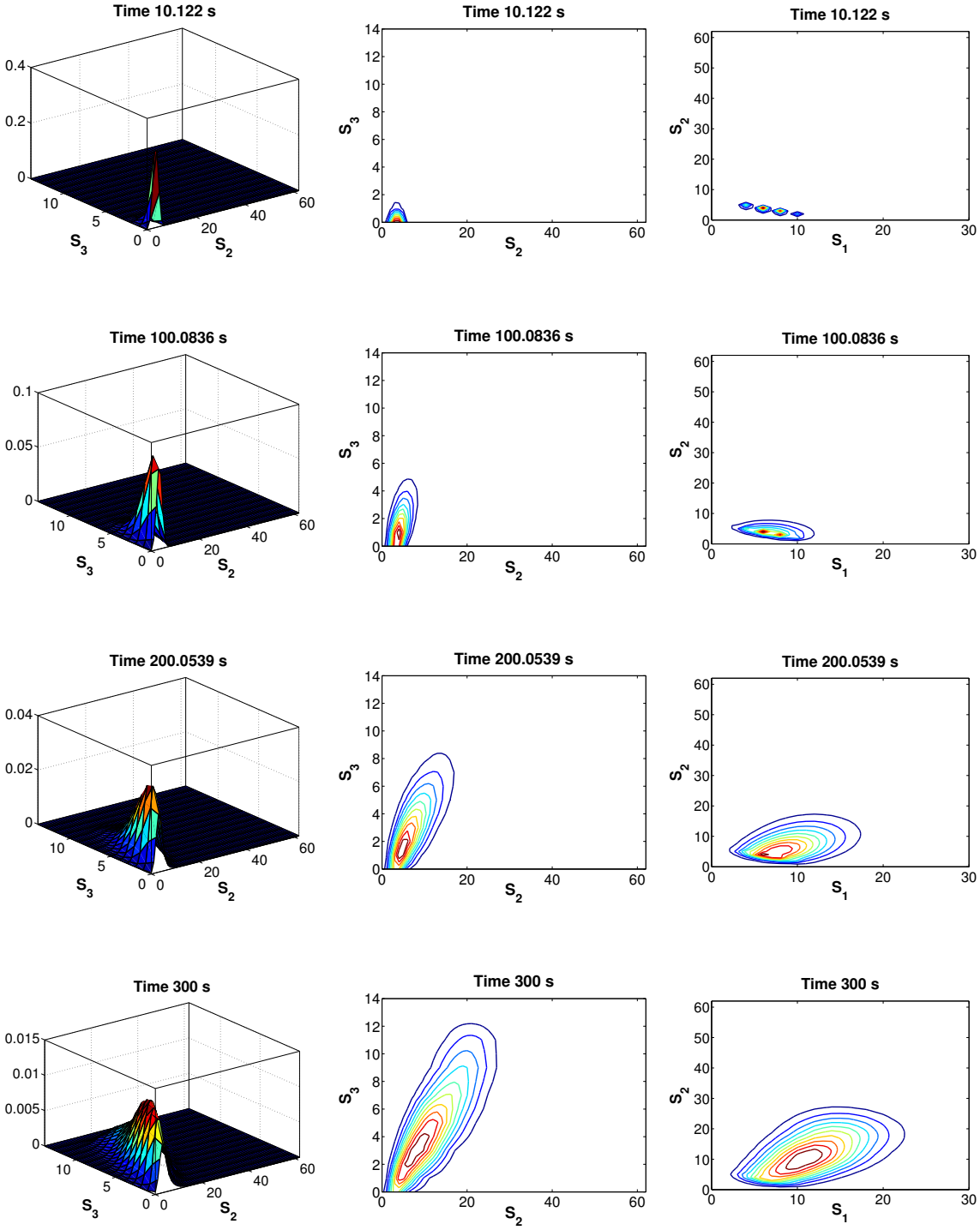


Figure 4.15.: Marginal distributions of the 6D Goutsias model (4.36) at different times. Surf plot (first column) and contour plots (second and third columns) of the approximation obtained with the adaptive wavelet method.

CHAPTER

FIVE

INVESTIGATING LONG-TIME DYNAMICS

The adaptive wavelet method presented in Chapter 4 demonstrated the use of wavelet compression to deal with the *curse of dimensionality* affecting the time-dependent CME on a finite, but high-dimensional state space $\Omega_\xi \subset \mathbb{N}_0^d$. In many applications, however, the transient behavior is of secondary importance, and the main goal is obtaining a characterization of the long-time dynamics of the system. Useful information in this regard can be obtained by approximating the invariant or stationary probability distribution, which is the solution of the *stationary* CME. While the invariant distribution does provide an insight into the long-time dynamics, it is not always sufficient by itself to investigate the qualitative behavior of the underlying Markov jump process at equilibrium. Information about the specific transition mechanisms between certain states are particularly important in biological systems exhibiting metastable dynamics, where *rare events* induce transitions between subsets of the state space. However, approximating the stationary distribution or gathering sufficient information to compute statistical objects related to the switching behavior are both non-trivial problems. Thus, in the present chapter, we are motivated to extend the wavelet method to the related tasks of approximating the *stationary* CME and the efficient computation of *committor probabilities*. The *committor probabilities* are objects that give a measure of the progress of transitions between two arbitrarily chosen subsets of the state space. Further, they can be used within the theoretical framework of Transition Path Theory (TPT) [VE06, MSVE08] to give a full statistical characterization of the switching mechanisms between metastable states.

5.1. Approximating the stationary distribution

5.1.1. Formulation as eigenvalue problem

As the *stationary* CME is a particular case of the time-dependent CME (2.67), and numerical computations are by necessity again restricted to the truncated state space Ω_ξ, the algorithm for the approximation of the stationary distribution uses the computational

core of the adaptive wavelet method presented in Algorithm 3, and we proceed to detail the required changes.

Recall that the truncated state space Ω_ξ contains a total of $N = \xi_1 \cdot \cdots \cdot \xi_d$ states, and Neumann boundary conditions (2.72) are imposed outside the boundaries defined by the truncation vector ξ. As stated in Chapter 2, when restricted to the state space Ω_ξ, the CME operator $\mathcal{A}$ is isomorphic to the generator matrix $A \in \mathbb{R}^{N\times N}$ of the underlying Markov jump process, with

$$a_{ik} \geq 0 \text{ for } i \neq k \text{ and } \sum_{i=1}^{N} a_{ik} = 0$$

leading to $a_{ii} = -\sum_{i\neq k} a_{ik}$, i.e., a usually non-symmetric matrix with non-positive diagonal elements, non-negative off-diagonal elements and zero columns sum. Assuming the matrix A is irreducible, we have by the Frobenius-Perron Theorem 2.5 that there exists a unique (non-negative) stationary distribution

$$\rho \in \mathbb{R}^N \text{ with } \rho_i \geq 0, \ \sum_{i=1}^{N} \rho_i = 1 \text{ and } A\rho = 0.$$

Moreover, we have already shown in Chapter 2 that the eigenvalue $\lambda_1 = 0$ of A is simple, and all the other eigenvalues have negative real part (see Appendix A). Consequently, the problem of computing the stationary distribution is equivalent to finding the eigenvector corresponding to the eigenvalue $\lambda_1 = 0$. A plethora of methods are available for such eigenvalue problems, among them direct methods, Krylov subspace techniques or single vector iterations [Saa92]. As the distinct eigenvalue $\lambda_1 = 0$ corresponding to the stationary distribution $\rho \in \mathbb{R}^N$ has the smallest absolute value of all eigenvalues belonging to the spectrum $\sigma(A)$, we use for this task the inverse power method with shift, also known as inverse iteration. The algorithm of the inverse iteration has the following form [GVL96]:

Algorithm 5: Inverse iteration

Parameter : s close to a distinct eigenvalue of A (in our case $s \approx 0$)

Input : initial guess $\rho^{(0)} \in \mathbb{R}^N$
matrix $A \in \mathbb{R}^{N\times N}$

Output : vector $\rho^{(k)}$ converging to eigenvector corresponding to eigenvalue close to s

for $k = 1, 2, \ldots,$ **do**

$$\text{solve } (A - sI)\hat{\rho}^{(k)} = \rho^{(k-1)} \tag{5.1}$$

$$\text{set } \rho^{(k)} = \hat{\rho}^{(k)}/||\hat{\rho}^{(k)}||$$

end

Naturally, Algorithm 5 operates on the entire truncated state space Ω_ξ and as such contains all the degrees of freedom $N = \prod_{i=1}^{d} \xi_i$, making efficient numerical approximation difficult, if not impossible. Therefore, the idea is to extend the inverse power method by once again using the favorable properties of a sparse wavelet representation and projecting the problem onto a low-dimensional approximation space.

5.1.2. Adaptive wavelet method for stationary CME

We recast now the problem of approximating the stationary distribution in the familiar operator notation, i.e., use $\mathcal{A}$ to denote the truncated operator and define the stationary distribution as a discrete function $\pi : \Omega_\xi \to \mathbb{R}$ satisfying

$$\mathcal{A}\pi = 0, \;\; \pi(\mathbf{x}) \geq 0, \;\; \sum_{\mathbf{x}\in\Omega_\xi} \pi(\mathbf{x}) = 1. \tag{5.2}$$

The task is to compute an approximation $p(\mathbf{x}) = \sum_{i=1}^{\eta} \gamma_i \psi_{j_i}(\mathbf{x}) \approx \pi(\mathbf{x})$ using a small subset $\{\psi_{j_1}, \ldots, \psi_{j_\eta}\}$ of a full wavelet basis $\{\psi_1, \ldots, \psi_N\}$, which satisfies $\mathcal{A}p \approx 0$ and $\sum_{\mathbf{x}\in\Omega_\xi} p(\mathbf{x}) = 1$. Thus, similarly with the time-dependent CME, we need a procedure to identify the suitable index subset $\{j_1, \ldots, j_\eta\}$. We use again the recipe from [CDD01], which means that given an approximation $p_k = \sum_{i=1}^{\eta} \gamma_i^{(k)} \psi_{j_i}$ at step k of an iterative procedure, we employ the residual on the whole state space Ω_ξ to guide the expansion of the approximation space in order to obtain the refined approximation $p_{k+1} = \sum_{i=1}^{\eta+\Delta\mu} \gamma_i^{(k+1)} \psi_{j_i}$.

In contrast to the time-dependent problem previously discussed in Chapter 4, however, we no longer have an initial distribution that allows the computation of the first set of basis elements for the refinement procedure. Recall that this set was obtained by performing a fast wavelet transform (FWT) of the initial distribution and retaining the best η-terms from the resulting wavelet coefficient vector $\gamma = (\gamma_i)_{i=1}^N$.

Consequently, the choice of wavelet basis plays now a far more important role. In Chapter 4, the use of an orthogonal wavelet basis $\Psi = \{\psi_1, \ldots, \psi_N\}$ with periodic extension at the boundaries of the domain Ω_ξ proved adequate for most numerical examples studied therein. That was because an initial set of coefficients was available and in each step this could be refined by the iterative procedure described in Algorithm 3. For the approximation of the *stationary* CME, however, we are forced to start with some arbitrary selection of basis elements and periodic wavelets have the disadvantage that they have high-amplitude coefficients in the neighborhood of the boundaries defined by the truncation vector ξ. This is because the boundary wavelets have separate components which no longer have vanishing moments and if the approximated function is non-periodic, as is the case in our application, the coefficients will behave as if discontinuities are present [Mal09]. Thus, these high-amplitude coefficients at the borders may cause problems with the residual-based expansion of the set of *essential* basis elements, especially for problems where the profile of the stationary distribution lies close to the boundaries. In our experience, better results for the approximation of the stationary distribution can be obtained by employing a boundary adapted wavelet basis that avoids the problems of periodic wavelets bases mentioned earlier. For the numerical tests that will be presented later, we have chosen *anisotropic* tensor products of wavelet bases on the interval. The specific univariate wavelet bases used are *B-spline* interval wavelet bases constructed using the procedure by Primbs from [Pri09], which was previously discussed in Chapter 3. We remark however, that other constructions of interval wavelet bases could also be adapted for use (e.g. [DKU97, Dij09]).

Let $\tilde{\Psi} = \{\tilde{\psi}_1, \ldots, \tilde{\psi}_N\}$ and $\Psi = \{\psi_1, \ldots, \psi_N\}$ be a pair of biorthogonal discrete wavelet bases on $\mathcal{H}(\Omega_\xi)$ constructed using the procedure outlined in Chapter 3. We shall refer to

Ψ as the *primal* basis, while $\tilde{\Psi}$ will be called *dual* basis. Then, the coefficients of a wavelet representation of a function $p = \sum_{i=1}^{N} \gamma_i \psi_i$ are given as $\gamma_i = \langle p, \tilde{\psi}_i \rangle$. We remark that the two bases are interchangeable, with the *primal* and *dual* basis switching places in the definitions above, i.e., using instead the *primal* basis for the decomposition and the *dual* basis for the reconstruction.

After establishing the choice of wavelet basis, the next step is projecting (5.1) into a low dimensional space. The use of a biorthogonal basis means that we have the option of a Petrov-Galerkin wavelet discretization in space. Using now the operator notation, we build a linear equation similar to (5.1) but with changed sign,

$$(sI - \mathcal{A})\hat{p}_k \quad = \quad -p_{k-1}. \tag{5.3}$$

Because Neumann boundary conditions (2.72) have been imposed, and zero is an eigenvalue with trivial left eigenvector $\mathbb{1}^T = (1, \ldots, 1)$ it follows that $(sI - \mathcal{A})$ is invertible for all $s > 0$ and all entries of the inverse are non-negative, because $-\mathcal{A} \in \mathbb{R}^{N \times N}$ is a M-matrix [Jah10]. Further, let us consider the wavelet expansions

$$\begin{aligned} \hat{p}_k &= \sum_{i=1}^{N} \hat{\gamma}_i \psi_i, \quad \hat{\gamma}_i = \langle \hat{p}_k, \tilde{\psi}_i \rangle \\ p_{k-1} &= \sum_{i=1}^{N} \beta_i \psi_i, \quad \beta_i = \langle p_{k-1}, \tilde{\psi}_i \rangle. \end{aligned}$$

Imposing now the Galerkin condition in (5.3) using a subset $\{\tilde{\psi}_{j_1}, \ldots, \tilde{\psi}_{j_\eta}\}$ of the basis $\tilde{\Psi}$, i.e.,

$$\langle \tilde{\psi}_{j_i}, (sI - \mathcal{A})\hat{p}_k \rangle = -\langle \tilde{\psi}_{j_i}, p_{k-1} \rangle,$$

for all $i = 1, \ldots, \eta$, we obtain the following algebraic formulation for our problem

$$(sI_\eta - \tilde{M})\hat{\gamma} = -\beta. \tag{5.4}$$

In equation (5.4), $\tilde{M} \in \mathbb{R}^{\eta \times \eta}$ is the Petrov-Galerkin matrix defined by

$$\tilde{M} = (\tilde{m}_{ik})_{i,k=1}^{\eta}, \qquad \tilde{m}_{ik} = \left\langle \tilde{\psi}_{j_i}, \mathcal{A}\psi_{j_k} \right\rangle, \tag{5.5}$$

and I_η is an $\eta \times \eta$ identity matrix, which follows from the biorthogonality conditions satisfied by the wavelet bases Ψ and $\tilde{\Psi}$. Further, $\beta = (\beta_1, \ldots, \beta_\eta)^T$ is the coefficient vector of the old approximation p_{k-1} and is obtained through a fast wavelet transform using the *dual* basis $\tilde{\Psi}$, while $\hat{\gamma}$ is the new coefficient vector, which after a normalization step $\gamma = \hat{\gamma}/||\hat{\gamma}||$, provides the new approximation p_k via an inverse wavelet transform using the *primal* basis Ψ.

Naturally, simply projecting the inverse iteration onto a low-dimensional approximation space does not constitute an adaptive numerical scheme, and this basic step for computing the new coefficient vector has to be coupled with residual-based basis refinement and coefficient thresholding strategies. It is also worth pointing out that simply using a naïve Galerkin approach instead of (5.3), i.e., projecting the stationary CME (5.2) directly and trying to solve a linear system of the form $\tilde{M}\gamma = 0$ is not a good idea. This is because

generally, the projected problem does not have a solution, as the simple zero eigenvalue is not preserved by the projection onto the approximation space.

Before detailing the main steps of the adaptive wavelet method for the stationary CME, let us first comment on the initialization phase of the method. As mentioned earlier, the choice of an initial set of basis elements is not as straightforward as in the time dependent case. For an initial approximation, we choose an uniform probability distribution $p_0 = c_0 \cdot \mathbb{1}$, where $c_0 = 1/\prod_{i=1}^{d} \xi_i$ is a normalization constant and $\mathbb{1} = (1, \ldots, 1)^T$. Next, the question how to select the initial index subset $\mathcal{J}_\eta = \{j_1, \ldots, j_\eta\}$ must be addressed. One idea is to compute a *few* long-time SSA trajectories, and after discarding the initial phase, use the data to compute a very coarse approximation of the stationary distribution. As the result is non-smooth, the next step would be to apply a smoothing procedure in multiple dimensions, followed by a fast-wavelet transform and selection of the best η-terms. The problem with such an initialization procedure is that the user must supply a multitude of problem dependent parameters, like the time interval for the SSA simulations, the smoothing parameter for computing the initial stationary approximation, the number of SSA trajectories to be computed, truncation index η, and so on. Moreover, because smoothing has to be applied, the approximation thus obtained could differ significantly from the true profile of the stationary distribution.

Therefore, in our numerical tests we have opted for another approach. Let $K(\xi) = I^{[0,\xi_1]} \otimes \cdots \otimes I^{[0,\xi_d]}$ where $I^{[0,\xi_i]} = \{k_i \in \mathbb{N}_0 \mid 0 \leq k_i \leq \xi_i\}$ is the local index interval for the i-th direction, denote the complete list of multi-indices $k^{(l)} := (k_1^{(l)}, \ldots, k_d^{(l)})$ assigned to the states of the multi-dimensional state space Ω_ξ. Moreover, there exists a bijective mapping between the single-indices j_l used for the enumeration of the wavelet bases and the multi-indices $k^{(l)}$, with $l = 1, \ldots, N$. Next, let $K(\varsigma) = I^{[0,\varsigma_1]} \otimes \cdots \otimes I^{[0,\varsigma_d]}$ be a subset of the complete multiple-indices set $K(\xi)$, where we have used the truncation vector ς with $0 < \varsigma_i \ll \xi_i$, and let $|K(\varsigma)| = \eta$. The initial index set for the adaptive wavelet method is then obtained via the mapping $\{k^{(1)} \ldots, k^{(\eta)}\} \mapsto \{j_1, \ldots, j_\eta\}$. This approach is quite natural in the context of using *anisotropic* tensor products of wavelet bases. Because of the way the univariate biorthogonal interval wavelet bases are constructed, the first indices belong to the scaling functions on the minimal resolution level. As such, taking the combinations of the first few basis elements in each direction translates into using parts of the nodal basis on an uniform grid covering the state space Ω_ξ. The approach can also be modified to use different local indexing intervals for each direction. We remark that although this initialization step is not optimal, the *a posteriori* analysis of the residual usually succeeds in identifying the correct coarse profile for the stationary distribution after a few iterations, which is then further improved. With the mention that the adaptive wavelet method for the stationary CME uses the same building blocks that can be found in Algorithm 3, we can proceed with the sketch of the new specialized Algorithm 6.

In Algorithm 6, steps (1) and (2) can be seen as one step of the inverse iteration in the low-dimensional space, and consequently they can be performed more than once, obtaining refined values for the wavelet coefficients for the currently active basis. However, as the critical issue is the expansion of the approximation space, performing more than one inverse iteration especially in the early stages does not bring significant advantages. The thresholding step (5) uses the same techniques as those used in step (6) from Algorithm 3 and can be either performed every step, or every few steps.

Algorithm 6: Adaptive wavelet method for stationary CME

Parameter : parameter $s \approx 0$, tolerance `tol`
$\Delta\mu$ (new basis elements per step), $\eta_{\max}$ (maximum allowed DOFs)
initial active set of coefficients $\mathcal{J}_\eta = \{j_1, \dots, j_\eta\}$

Input : coefficients $\gamma^{(0)} = (\gamma_1^{(0)}, \dots, \gamma_\eta^{(0)})^T$ via FWT of $p_0 = c_0 \cdot \mathbb{1}$ using $\tilde{\Psi}$
Petrov-Galerkin matrix $\tilde{M}$ defined as

$$\tilde{M} = (\tilde{m}_{ik})_{i,k=1}^{\eta}, \qquad \tilde{m}_{ik} = \left\langle \tilde{\psi}_{j_i}, \mathcal{A}\psi_{j_k} \right\rangle .$$

Output : approximation $p_k \approx \pi$ satisfying $\mathcal{A}p_k \approx 0$ and $\sum_{\mathbf{x}\in\Omega_\xi} p_k(\mathbf{x}) = 1$

Set $k = 1$

while *true* **do**

1. Solve the linear system

$$(sI - \tilde{M})\hat{\gamma}^{(k)} \quad = \quad -\gamma^{(k-1)}. \tag{5.6}$$

2. Set $\gamma^{(k)} = \hat{\gamma}^{(k)}/||\hat{\gamma}^{(k)}||$.

3. Compute approximation $\hat{p}_k = \sum_{i=1}^{\eta} \gamma_i^{(k)} \psi_{j_i}$ and normalize.

4. Compute the residual $r = \mathcal{A}\hat{p}_k$.

5. Apply optional thresholding step to $\gamma^{(k)}$ (analogously to step (6) of Algorithm 3), taking care to delete corresponding lines from the Petrov-Galerkin matrix $\tilde{M}$ and update the index subset $\mathcal{J}_\eta$, and value η, respectively. Then, the approximation for the current step is $p_k = \sum_{i=1}^{\eta} \gamma_i^{(k)} \psi_{j_i}$ with updated coefficient vector $\gamma^{(k)}$.

6. **if** $||r||_1 >$ `tol` *and* $\eta < \eta_{\max}$ **then**

   a) Compute $\chi_l = \left|\langle \tilde{\psi}_l, r\rangle\right|$ for $l = 1, \dots, N$ via fast wavelet transform.

   b) Find the new index subset $\mathcal{J}_{\Delta\mu} = \{j_{\eta+1}, \dots, j_{\eta+\Delta\mu}\}$ of the $\Delta\mu$ largest entries of $(\chi_1, \dots, \chi_N)$.

   c) Update the Petrov - Galerkin matrix by adding new blocks corresponding to the newly selected basis elements $\tilde{M} \in \mathbb{R}^{(\eta+\Delta\mu)\times(\eta+\Delta\mu)}$.

   d) Update active index subset $\mathcal{J}_\eta \mapsto \mathcal{J}_\eta + \mathcal{J}_{\Delta\mu}$ and set $\eta \mapsto \eta + \Delta\mu$.

   e) Update $\gamma^{(k)}$ with coefficients corresponding to the new index set $\mathcal{J}_\eta$.

   f) Set $k \mapsto k + 1$.

   else

   Exit while loop

   end

end

Step (6) implements the expansion of the state space and prepares the Petrov-Galerkin matrix $\tilde{M}$ and the right hand side of (5.6) for the next step of the adaptive wavelet method. Another option for substep (6e) is to reinitialize $\gamma^{(k)} = (\gamma_1^{(0)}, \ldots, \gamma_\eta^{(0)})^T$, i.e., use the coefficients computed via the fast wavelet transform of $p_0 = c_0 \cdot \mathbb{1}$ corresponding to the updated index subset.

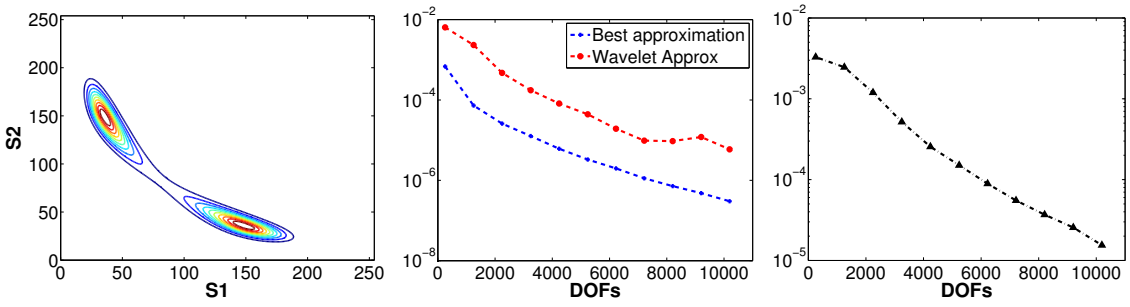
Before illustrating the adaptive wavelet method for the stationary CME with a few numerical examples, let us remark that if the we multiply (5.6) by $1/s$, the method can be viewed as using the implicit Euler version of the time-dependent adaptive wavelet method, with a fixed, very large step size $1/s$. Therefore, the stationary and time dependent versions are closely related and the Petrov-Galerkin ansatz can also be employed in Algorithm 3 in case a biorthogonal basis is used.

5.2. Numerical examples

In the first numerical example to be presented we shall revisit the *toggle switch* model with two competing species from (2.94). Due to the moderate size of its state space, this model allows us to investigate the accuracy of the proposed method. Further, two multi-dimensional models of toggle switches will be used to test the viability of the concept for larger state spaces.

5.2.1. Revisiting the 2D toggle switch

The state space of the *toggle switch* model from (2.94) with parameter set (2.95), is only $2^8 \times 2^8$. Thus, the total number of degrees of freedom $N = 65536$ is small enough to allow the computation of a reference solution by explicitly constructing the corresponding generator matrix $A \in \mathbb{R}^{N \times N}$ and solving the associated eigenvalue problem $A\rho = 0$ via the MATLAB routine *eigs*.



(a) Wavelet approximation (b) 2-norm Error (c) 1-norm of residual

Figure 5.1.: Accuracy of the adaptive wavelet method for the stationary CME applied to the *toggle switch* model (2.94). Panel 5.1a shows a contour plot of the approximation obtained using the wavelet method with 3.5% of the total degrees of freedom, panel 5.1b compares the 2-norm errors of the truncated reference solution and that of the wavelet approximation, while in panel 5.1c we plot the evolution of the residual for the wavelet method.

We can then use the adaptive wavelet method for the stationary CME on the same problem, and compare the error of the approximation obtained in each step of the wavelet

method, with the error of a best approximation. The best approximation in obtained by truncating the reference solution to the same number of coefficients as those used inside the wavelet method. The errors were computed in the 2-norm and are displayed in the Figure 5.1b. The adaptive wavelet method was configured to use *anisotropic* tensor products of *B-spline* 2.2 interval wavelet bases. The iterative procedure described by Algorithm 6 was stopped either if a tolerance in the 1-norm of $\texttt{tol} = 1e-5$ was reached by the residual, or the maximum number of allowed degrees of freedom $\eta_{\max} = 10000$ was exceeded. The results indicate that the 2-norm error of the approximation behaves similarly with the error of the truncated reference solution, and the wavelet approximation agrees well with the reference solution even after a few iterations. We remark that in Figure 5.1a the approximation after only 3 steps of the wavelet method is shown, which uses 3.5% of the total degrees of freedom.

As additional information, in Figure 5.2 the sets of basis elements used by the wavelet method are shown in the form of the sparsity pattern of the Petrov-Galerkin "stiffness" matrix for the full basis. Furthermore, mesh plots of the difference between the wavelet approximation and the reference solution obtained using MATLAB's *eigs* command for several iterations are shown in Figure 5.3. We remark that the scales of the plots change as the approximation is refined.

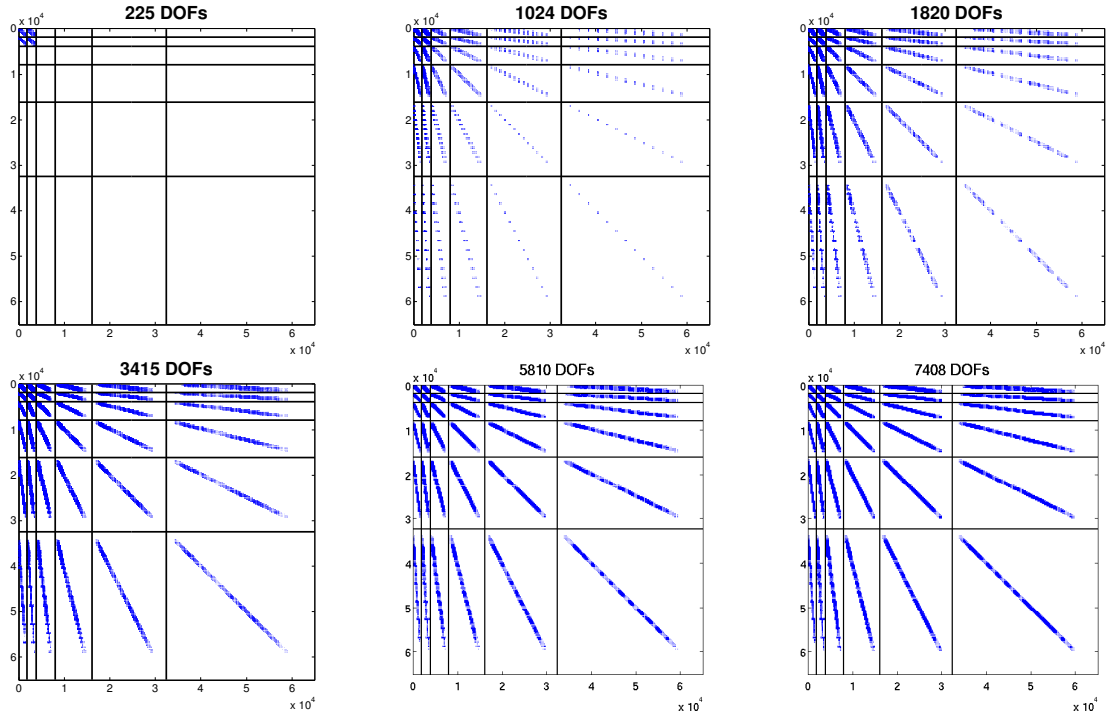


Figure 5.2.: Wavelet sets used by the adaptive wavelet method for the *toggle switch* model (2.94), plotted as sparsity pattern of the Petrov-Galerkin matrix for the full basis.

5.2.2. Multi-dimensional genetic toggle switches

After testing the accuracy, we now apply the adaptive wavelet method to more challenging examples with larger state spaces, where a reference solution is no longer available.

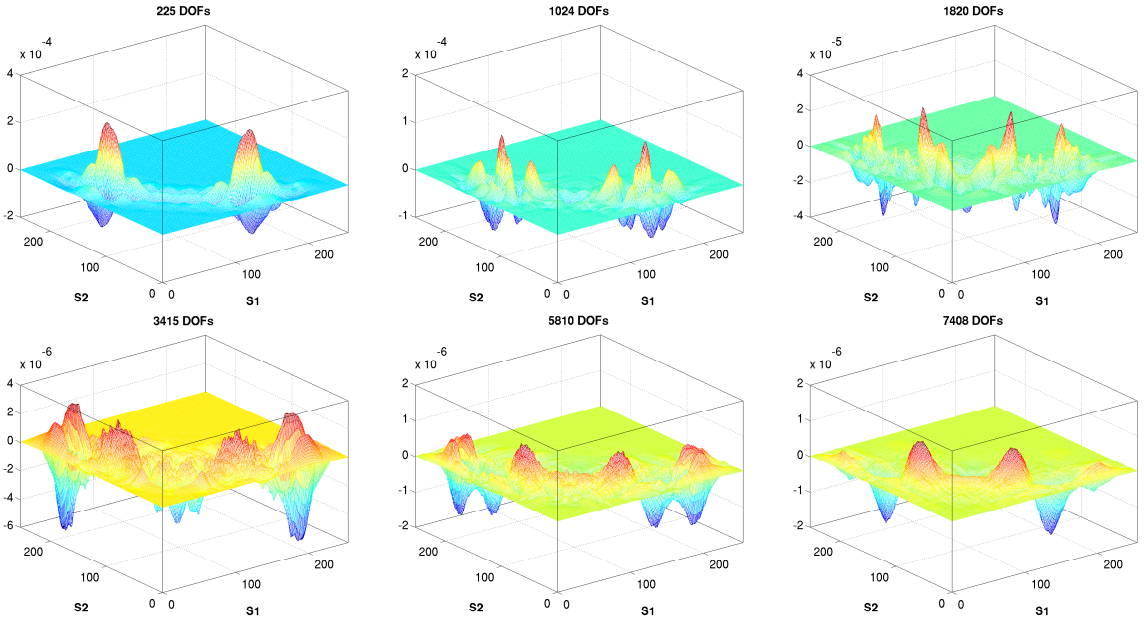


Figure 5.3.: Difference between the wavelet approximation and the reference solution for the *toggle switch* model (2.94), obtained with MATLAB's *eigs* command for increasingly refined active set of basis elements.

The first model is another genetic toggle switch, built around three mutually repressing gene products, denoted S_1, S_2 and S_3. The inhibition of each of the species by the other two components of the system is modeled by using non-standard propensities (reactions R_1 through R_3), while the other reaction channels are standard degradation reactions. The reaction channels and corresponding parameter set are summarized below

$$\begin{array}{lcccc|lcl|lcl} R_1: & \star & \longrightarrow & S_1 & & \alpha_1 & = & \frac{c_{11}}{(c_{12}+x_2^2)(c_{13}+x_3^2)} & \mu^1 & = & (1,0,0)^T \\ R_2: & \star & \longrightarrow & S_2 & & \alpha_2 & = & \frac{c_{21}}{(c_{22}+x_1^2)(c_{23}+x_3^2)} & \mu^2 & = & (0,1,0)^T \\ R_3: & \star & \longrightarrow & S_3 & & \alpha_3 & = & \frac{c_{31}}{(c_{32}+x_1^2)(c_{33}+x_2^2)} & \mu^3 & = & (0,0,1)^T \\ R_4: & S_1 & \longrightarrow & \star & & \alpha_4 & = & c_4 x_1 & \mu^4 & = & (-1,0,0)^T \\ R_5: & S_2 & \longrightarrow & \star & & \alpha_5 & = & c_5 x_2 & \mu^5 & = & (0,-1,0)^T \\ R_6: & S_2 & \longrightarrow & \star & & \alpha_6 & = & c_6 x_2 & \mu^6 & = & (0,0,-1)^T \end{array} \tag{5.7}$$

$$\begin{aligned} &c_{11} = 125000,\ c_{21} = 50000,\ c_{31} = 250000,\ c_{i2} = c_{i3} = 500,\ i = \{1,2,3\} \\ &c_4 = 0.005, \quad c_5 = 0.002, \quad c_6 = 0.01. \end{aligned} \tag{5.8}$$

The state space for the $3D$ *toggle switch* model (5.7) has $2^7 \times 2^7 \times 2^7$ total degrees of freedom. The stationary CME was solved using again *anisotropic* tensor products of *B-spline* 2.2 interval wavelets with the target tolerance for the 1-norm of the residual being set to $\texttt{tol} = 0.01$. Recall that the 1-norm scales with the size of the state space, so the choice provides sufficient accuracy. The method was configured to use a maximum of $\eta_{\max} = 30000$ basis elements, and in each step 1500 new coefficients were proposed. The method stopped before exhausting the maximum number of basis elements allowed, reporting $||r||_1 = 0.007$ after using 1.1% of the total degrees of freedom. In the first row of Figure 5.4 the approximation of the stationary distribution is shown in the form of $2D$

marginal plots for $S_1 - S_2$ and $S_2 - S_3$, respectively (the third $2D$ marginal which is not shown has a similar profile) and a $3D$ visualization of the profile of the multi-dimensional stationary probability distribution, with three metastable areas clearly identifiable. The second row contains results obtained by averaging the data from 1000 SSA simulations on the long time interval $[0, 10^6]$. The first 2% of the SSA data from each trajectory has been discarded before computing an approximation to the stationary distribution by averaging the remaining trajectory data. The SSA results confirm that the adaptive wavelet method delivers the correct qualitative description of the dynamics.

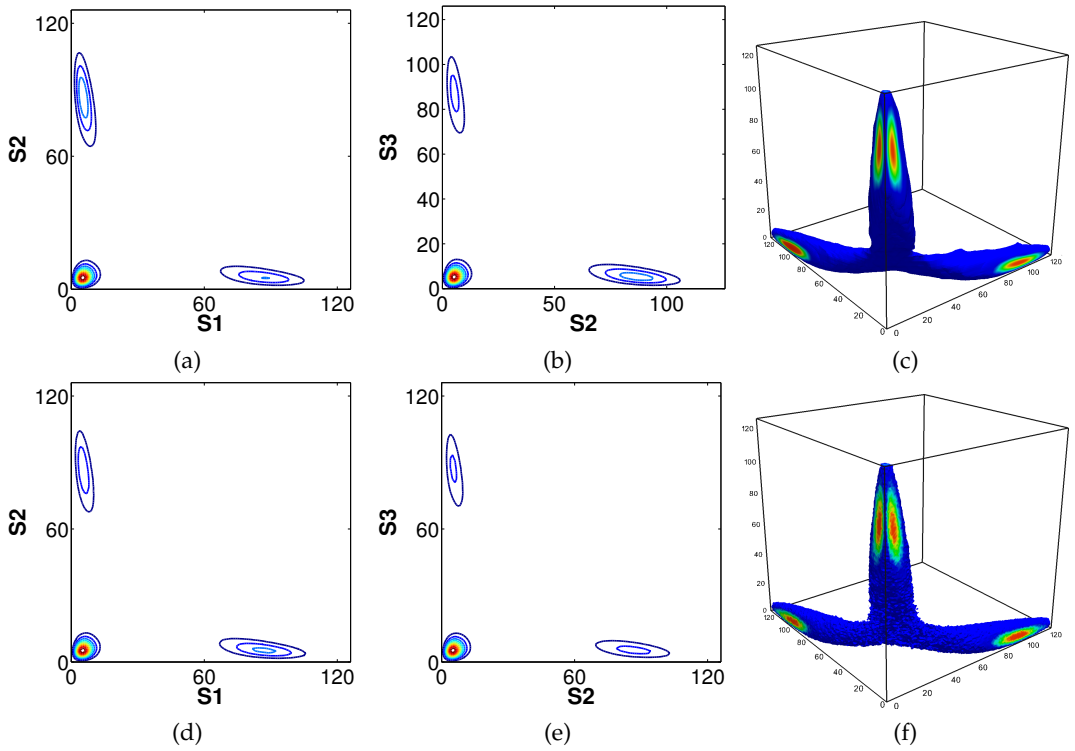


Figure 5.4.: $3D$ *toggle switch* model (5.7) with inhibition of each of the species by the other two species. First row (5.4a, 5.4b and 5.4c), shows results obtained with the wavelet method using 1.1% DOFs, while in the second row (5.4d, 5.4e and 5.4f) plots of the $2D$ marginals of an SSA-based approximation of the stationary distribution are shown. The right panel in each row depicts an iso-volume plot of the approximation to the stationary distribution where only values larger than $\varepsilon = 10^{-7}$ were used to construct the $3D$ visualization.

As a further example, a second large multi-dimensional genetic toggle switch ($d = 4$) is presented. The model actually contains six species, but we can perform a reduction based on algebraic arguments. The reduced species will be denoted by $\hat{S}_i,\ i \in \{1, 3\}$ in the description of the reaction channels given in (5.9). The first four reactions R_1, R_2, R_3 and R_4, model a reversible process by which two competing species S_1 and S_3 switch between their active and inactive state. This can take place for example by the binding of molecules belonging to other species to the S_1 and S_3 molecules, but the actual mechanisms are not explicitly modeled. As there are no degradation reactions for the two species S_1 and S_3, and assuming a *closed* system, their copy numbers remain constant.

$$
\begin{array}{lcl|lcl|lcl}
R_1: & S_1 & \longrightarrow \hat{S}_1 & \alpha_1 & = & c_1 x_1 & \mu^1 & = & (-1,0,0,0)^T \\
R_2: & S_3 & \longrightarrow \hat{S}_3 & \alpha_2 & = & c_2 x_3 & \mu^2 & = & (0,0,-1,0)^T \\
R_3: & \hat{S}_1 & \longrightarrow S_1 & \alpha_3 & = & \frac{c_{31}(K_1-x_1)}{(c_{32}+x_4^2)} & \mu^3 & = & (1,0,0,0)^T \\
R_4: & \hat{S}_3 & \longrightarrow S_3 & \alpha_4 & = & \frac{c_{41}(K_3-x_3)}{(c_{42}+x_2^2)} & \mu^4 & = & (0,0,1,0)^T \\
R_5: & S_1 & \longrightarrow S_1+S_2 & \alpha_5 & = & c_5 x_1 & \mu^5 & = & (0,1,0,0)^T \\
R_6: & S_3 & \longrightarrow S_3+S_4 & \alpha_6 & = & c_6 x_3 & \mu^6 & = & (0,0,0,1)^T \\
R_7: & S_2 & \longrightarrow \star & \alpha_7 & = & c_7 x_2 & \mu^7 & = & (0,-1,0,0)^T \\
R_8: & S_4 & \longrightarrow \star & \alpha_9 & = & c_8 x_4 & \mu^8 & = & (0,0,0,-1)^T
\end{array}
\tag{5.9}
$$

We can then omit the explicit modeling of the inactive state from the model, because we have $S_i + \hat{S}_i = K_i,\ i \in \{1,3\}$, with $K_1, K_3 \in \mathbb{N}$ constants denoting the total number of molecules of each of the species. The propensities of the reactions R_3 and R_4 are then modified by replacing the components of the reduced species with their equivalent formulation. Further, the two non-standard propensities belonging to reaction channels R_3 and R_4 describe how the species S_2 and S_4 act as inhibitors and repress the activation of the two competing species, S_3 and S_1, respectively. Finally, R_7 and R_8 are standard degradation reactions for S_2 and S_4.

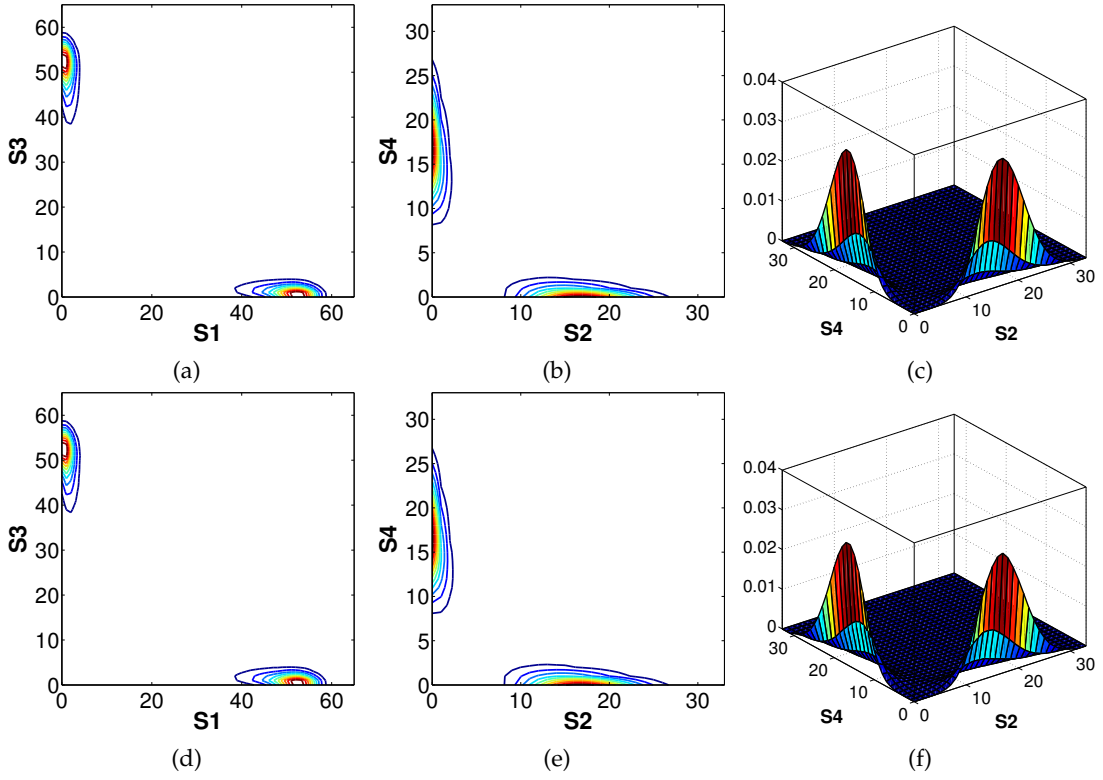


Figure 5.5.: $4D$ *toggle switch* given in (5.9). First row (5.5a, 5.5b and 5.5c), shows results obtained with the wavelet method using 0.53% DOFs, while in the second row (5.5d, 5.5e and 5.5f) contour and surf plots of the $2D$ marginals of an SSA-based approximation to the stationary distribution are shown.

Using the following parameter set

$$\begin{aligned} c_1 = c_2 &= 2, \quad c_{31} = c_{41} = 10, \quad c_{32} = c_{42} = 1, \\ c_5 = c_6 &= 1, \quad c_7 = c_8 = 3, \quad K_1 = K_3 = 63, \end{aligned} \tag{5.10}$$

leads to a truncated state space Ω_ξ of size $2^6 \times 2^5 \times 2^6 \times 2^5 \approx 4 \cdot 10^6$ total degrees of freedom, needed to capture the profile of the stationary distribution. Despite the large state space, the adaptive wavelet method was applied successfully to this $4D$ genetic toggle switch model, and the results are presented in Figure 5.5. *Anisotropic* tensor products of *B-spline* 3.5 interval wavelets were used, and the desired tolerance for the residual measured in the 1-norm was given as $\texttt{tol} = 0.7$. The method stopped after using 0.53% of the total number of degrees of freedom. Figure shows a comparison between the approximation obtained with wavelet compression, and from averaging 1000 long-time SSA trajectories for the interval $[0, 10^5]$. Again, the initial 2% of the data from each SSA trajectory was discarded before computing the SSA-based approximation of the stationary distribution.

We remark at this point that although the last two examples are not based on biological models in actual use, they do provide interesting test problems for the application of wavelet compression to the task of computing *committor probabilities*, as the stationary distributions of both models are non-trivial to compute and their profiles exhibit multiple metastable states.

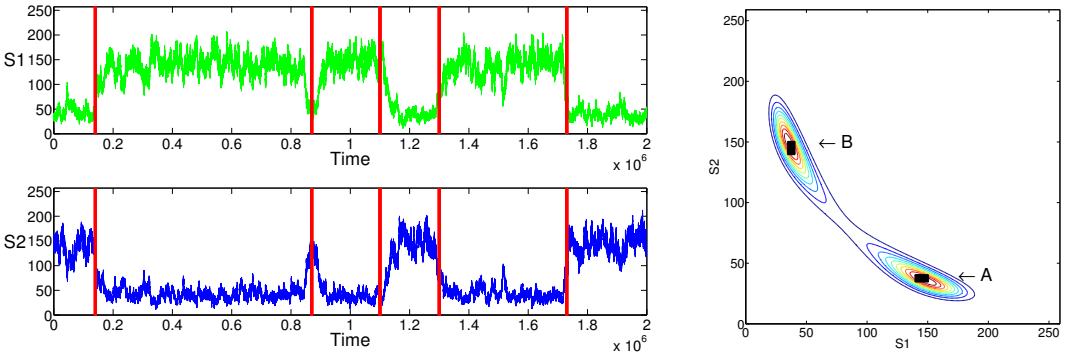
5.3. Transition Path Theory

Besides obtaining the solution of the stationary CME, computing other statistical properties of the underlying continuous-time Markov jump process, like characterizations of the mechanisms by which transitions between meta-states occur, might also prove relevant in the context of some applications. The transitions between metastable states are *rare events*, which are triggered by stochastic noise. This is not surprising, as we have already established in the previous chapters that the dynamics of biochemical systems are subject to random perturbations. Although stochastic simulations are capable of accurately modeling such dynamics, gathering sufficient statistics on the transition events between certain subsets of the state space using stochastic simulations is rarely practical. An alternative and more efficient strategy for solving such problems is given by Transition Path Theory (TPT) ([VE06, MSVE08]). Providing an elaborate description of TPT is far beyond the scope of this thesis, and indeed *not* actually required to understand how the adaptive wavelet method can be used to efficiently compute *committor probabilities*. These statistical objects are central to the theoretical framework of TPT, which uses them to compute the transition rates between meta-states or the dominant transition pathways, among other information. Obtaining the *committor probabilities* means solving large stationary problems closely related to the stationary CME, and consequently also affected by the *curse of dimensionality*. Based on the similarity of the committor problems to the stationary CME, it is only natural to try to extend the sparse wavelet compression concept to solving such problems as well. The purpose of the following short description of TPT is only to provide some context for the numerical results and introduce the notation required in presenting our algorithms. For further details, the reader is referred to the original sources

where TPT for Markov jump processes was introduced ([MSVE08, Met08]), which also form the basis of the next section.

5.3.1. Rare events and committor probabilities

The problem of investigating transition processes between metastable states can be best explained by making use once again of the *toggle switch* model from (2.94). In Figure 5.6a one SSA trajectory of the model is plotted, with the evolution of the copy numbers for each of the two species belonging to the model shown separately. We denote now by $\mathbb{A} = \{(x_1, x_2) : x_1 > x_2\}$ and $\mathbb{B} = \{(x_1, x_2) : x_1 < x_2\}$ two subsets of the truncated state space Ω_ξ of the model. Next, an examination of this single SSA trajectory leads to the observation that only 5 transitions between $\mathbb{A}$ and $\mathbb{B}$ occur in the time interval $[0,\ 2 \cdot 10^6]$, with the transitions being marked by vertical red lines in Figure 5.6a. Recall now that any change in the state vector means one step of the SSA algorithm, and from Figure 5.6a it is clear that the system spends most of its time in either one of the two metastable subsets of Ω_ξ. As a direct consequence, it is difficult to observe enough transitions as relatively few occur during computationally tractable stochastic simulation times. Moreover, the stationary distribution depicted in Figure 5.6b, does not represent a good reaction coordinate for investigating the transitions, where by the term of "good" reaction coordinate we refer to a quantity that can be used to describe the mechanisms by which the *rare events* occur. Although the metastable states are clearly visible in the profile of the stationary distribution, important information about the underlying dynamics of the system, like the preferred pathways for the transitions from one meta-state to the other, are somewhat concealed.



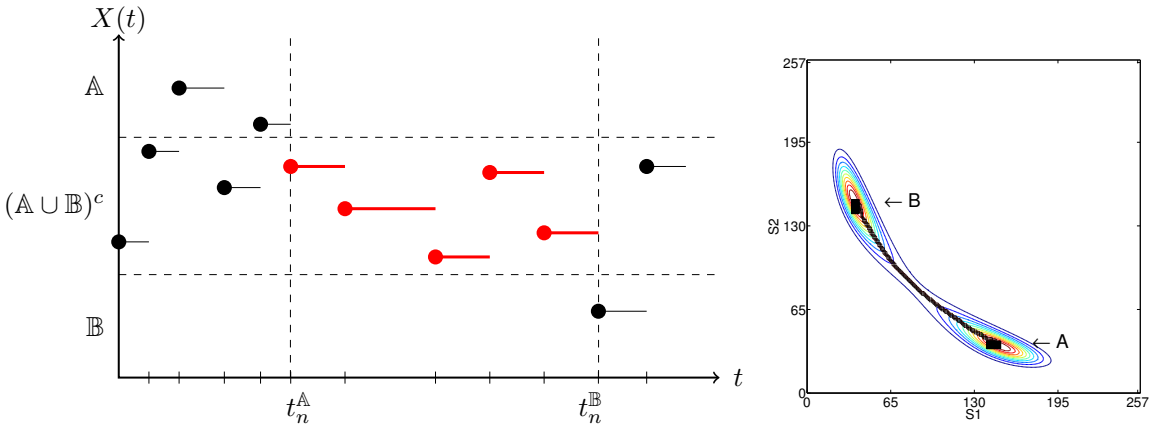
(a) (b)

Figure 5.6.: Investigating *rare events* with the help of SSA. Left panel: single SSA trajectory of the *toggle switch* model (2.94) shown for each of the two species in a separate plot. The transitions between two metastable sets $\mathbb{A}$ and $\mathbb{B}$ are marked by red lines. Right panel: Stationary distribution for the same model, with superimposed representation of the two sets $\mathbb{A}$ and $\mathbb{B}$.

Summarizing, in order to efficiently investigate *rare events*, we need a procedure for gathering the relevant statistics and a new reaction coordinate that is better suited for metastability analysis. Both these requirements can be fulfilled by using Transition Path Theory (TPT). Given a time-continuous Markov jump process on the finite multi - dimensional discrete state space Ω_ξ, and two non-empty disjoint sets $\mathbb{A}, \mathbb{B} \in \Omega_\xi$, TPT provides the means to describe the statistical properties of the transitions between these state space

subsets by analyzing in detail the associated *reactive trajectories*, i.e., those trajectories by which transitions from one set to another occur. By looking at the ensemble of all possible reactive trajectories, TPT computes then such statistical quantities as the *probability distribution of reactive trajectories*, or the *probability current* of reactive trajectories, i.e., what is the net amount of reactive trajectories going through a given state. It also computes the *rate of reaction* between a pair of arbitrarily selected sets and can single out the *dominant reaction pathways* used by the random walker to travel between the sets (cf. [Met08, MSVE08]). In a nutshell, TPT can lead to the full understanding of the discrete reaction mechanisms. A detailed derivation of TPT for Markov jump processes can be found in [MSVE08], so we mainly confine the presentation to the key ingredient needed for the computation of the various TPT objects, the *committor probability*.

In order to characterize the transition $\mathbb{A} \longrightarrow \mathbb{B}$, let us first consider an equilibrium path $\{X(t)\}_{t\in\mathbb{R}}$ of the jump process which oscillates infinitely many times between the set $\mathbb{A}$ and the set $\mathbb{B}$. We are only interested in the segments of the equilibrium path that leave $\mathbb{A}$ and go directly to $\mathbb{B}$, not the ones that return to $\mathbb{A}$ before proceeding to the destination or describe movements in the opposite direction. A schematic of a reactive trajectory is shown in the left panel of Figure 5.7 (adapted from [MSVE08]), with the right panel displaying an ensemble of reactive trajectories $\mathbb{A} \longrightarrow \mathbb{B}$ superimposed on the contour plot of the stationary distribution of the *toggle switch* model (2.94).



(a) (b)

Figure 5.7.: Schematic representation of a reactive trajectory (left panel, figure adapted from [MSVE08]) and ensemble of reactive trajectories between two sets $\mathbb{A}$, $\mathbb{B}$ for the *toggle switch* model (right panel).

Only the pieces involved in direct transitions $\mathbb{A} \longrightarrow \mathbb{B}$ are reactive trajectories, and by ignoring the other path segments we obtain an ensemble of transition pathways between the chosen subsets of the state space. We remark that pruning the trajectories of the non-reactive segments is achieved by defining the times $t_n^{\mathbb{A}}$ and $t_n^{\mathbb{B}}$, representing the last exit from the non-reactive set and entry into the reactive state space, and the first entry into the non-reactive set after exiting the reactive set, respectively. All the path segments lying outside these boundaries are then discarded, leaving us with an ordered sequence of path segments that give the successive states that were visited in the n-th transition between $\mathbb{A}$ and $\mathbb{B}$. All such finite sequences then build the ensemble of reactive trajectories.

As we are interested in the various statistical properties of this ensemble, objects must

be defined that quantify these properties. A natural first candidate is the probability distribution of reactive trajectories, i.e., the probability that when the system is in a state $\mathbf{x} \in \Omega_\xi$, it will first reach $\mathbb{B}$ rather than $\mathbb{A}$. Intuitively, this can be expressed as the probability that the process is arriving from $\mathbb{A}$ rather than $\mathbb{B}$, times the probability that it will reach $\mathbb{B}$ instead of $\mathbb{A}$ in the future. This leads to the operational definition of the two key objects of TPT, namely the *forward committor probability* and the *backward committor probability*, respectively.

The *discrete forward commitor* $q^+ : \Omega_\xi \to \mathbb{R}$ is defined for each state $\mathbf{x} \in \Omega_\xi$, as the probability that the Markov jump process starting in $\mathbf{x}$ will first reach $\mathbb{B}$ rather than $\mathbb{A}$. By this definition, the committor probability for all states $\mathbf{x} \in \mathbb{A}$ is $q^+(\mathbf{x}) = 0$ and similarly, for all $\mathbf{x} \in \mathbb{B}$ we have $q^+(\mathbf{x}) = 1$. For the other states $\mathbf{x} \in \Omega_\xi \setminus (\mathbb{A} \cup \mathbb{B})$, we use that $q^+(\mathbf{x})$ is the first entrance probability of the jump process $\{X(t), t \geq 0, X(0) = \mathbf{x}\}$ into set $\mathbb{B}$ *avoiding* set $\mathbb{A}$. Such entrance probabilities are handled by modifying the jump process such that these states become absorbing [MSVE08], and solving a *backward Kolmogorov* equation (2.56) with a modified generator matrix. Recall now that we operate in a setting characterized by the finite state space Ω_ξ, and the adjoint CME (2.80) with operator $\mathcal{A}^\star$ given by (2.79) is a special case of the *backward Kolmogorov* equation (2.56), as previously discussed in Section 2.5.2. Then, the adjoint operator $\mathcal{A}^\star$ is isomorphic to the large and sparse generator matrix of the Markov jump process given in (2.54), which by a slight misuse of notation we shall also denote by $\mathcal{A}^\star \in \mathbb{R}^{N \times N}$. The purpose of this rather convoluted argument is to be able to express the *discrete forward committor* as the solution of the following set of equations,

$$\begin{cases} (\mathcal{A}^\star q^+)(\mathbf{x}) = 0, & \text{for all } \mathbf{x} \in \Omega_\xi \setminus (\mathbb{A} \cup \mathbb{B}) \\ q^+(\mathbf{x}) = 0, & \text{for all } \mathbf{x} \in \mathbb{A} \\ q^+(\mathbf{x}) = 1, & \text{for all } \mathbf{x} \in \mathbb{B}. \end{cases} \tag{5.11}$$

We remark that (5.11) is just another way to write the committor problem in terms of the operator $\mathcal{A}^\star$, instead of the traditional description based on the generator of the Markov jump process, given in [MSVE08].

Analogously, the *discrete backward commitor* $q^- : \Omega_\xi \to \mathbb{R}$ for a state $\mathbf{x} \in \Omega_\xi$, is defined as the probability that the process arriving in state $\mathbf{x}$ is coming from the set $\mathbb{A}$ rather than from $\mathbb{B}$. For an operational definition we need the generator description of the time-reversed jump process, which is obtained from the "detailed balance" condition

$$\operatorname{diag}(\pi)\tilde{\mathcal{A}} = (\operatorname{diag}(\pi)\mathcal{A}^\star)^T. \tag{5.12}$$

Using (5.12), the generator of the time-reversed jump process is

$$\tilde{\mathcal{A}} = \operatorname{diag}(\pi^{-1})(\mathcal{A}^\star)^T \operatorname{diag}(\pi) = \operatorname{diag}(\pi^{-1})\mathcal{A} \operatorname{diag}(\pi), \tag{5.13}$$

where, again abusing the notation, $\mathcal{A} \in \mathbb{R}^{N \times N}$ denotes the matrix from (2.69) which is isomorphic to the truncated CME operator, and $\pi \in \mathbb{R}^N$ the unique (non-negative) stationary probability of the system.

Next, via similar arguments as those used for the *forward committor*, we have that the *backward committor* q^- is the solution of the following system of equations,

$$\begin{cases} (\tilde{\mathcal{A}} q^-)(\mathbf{x}) = 0, & \text{for all } \mathbf{x} \in \Omega_\xi \setminus (\mathbb{A} \cup \mathbb{B}) \\ q^-(\mathbf{x}) = 1, & \text{for all } \mathbf{x} \in \mathbb{A} \\ q^-(\mathbf{x}) = 0, & \text{for all } \mathbf{x} \in \mathbb{B}. \end{cases} \tag{5.14}$$

Note that as the stationary distribution π usually vanishes in large parts of the state space Ω_ξ, it follows that the *backward committor* (5.14) problem is not defined for states with $\pi(\mathbf{x}) \approx 0$, so the values of q^- are only computed for the subset of states where the corresponding stationary distribution values are non-zero.

We also remark that based on the definition of the *forward committor* q^+, i.e., the probability of reaching $\mathbb{B}$ before $\mathbb{A}$, we could in theory also compute q^+ via standard SSA simulations by starting multiple simulations from each state $\mathbf{x} \in \Omega_\xi$ and measuring the number of trajectories that have reached $\mathbb{B}$ first. Of course, due to the large number of simulations needed, this approach is not computationally efficient for large problems and provides at best only a very coarse approximation for the *forward committor* probability.

An example of *forward* and *backward committor* probabilities, respectively, is presented in Figure 5.8, again for the specific case of the $2D$ *toggle switch* model from (2.94). Transitions were investigated between the set $\mathbb{A} = \{(148, 38), (148, 39)\}$ and the set $\mathbb{B} = \{(38, 148), (39, 149)\}$ which are approximately located at the center of the two peaks marking the metastable sets. Examining Figure 5.8a, it becomes clear why the *committor probabilities* represent a better choice than the stationary distribution for investigating the transitions between metastable states. Knowing the current state of the system $\mathbf{x} \in \Omega_\xi$ and the corresponding value of the commitor $q^+(\mathbf{x})$ we can easily give an answer to the question how far the transition between $\mathbb{A}$ and $\mathbb{B}$ has progressed and what will occur next. Notice that multiple intermediate states along the possible route of a transition can have the same committor value, which means that the committor can be used to partition the state space into ensembles of transition states. In the example shown in Figure 5.8a, the states with $q^+(\mathbf{x}) = 0.5$ build the barrier between the two metastable states of the system. In Figure 5.8b a surf plot of the *backward committor* is shown. The unusual profile can be explained by the definition (5.14), which features a point-wise division by the elements of the stationary distribution vector $\pi \in \mathbb{R}^N$. Consequently, the subset of states where the stationary distribution almost vanishes is neglected when computing the values of q^-.

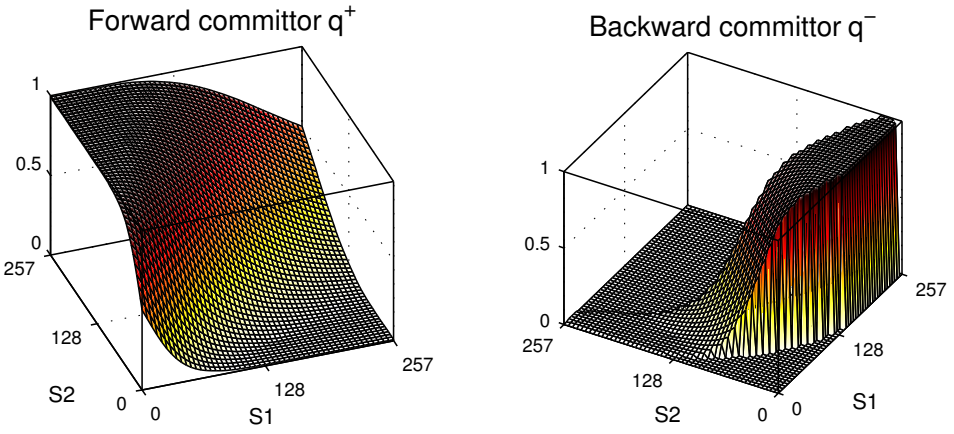


Figure 5.8.: Examples of committor probabilities for the $2D$ *toggle switch* model

5.3.2. TPT objects

Although they can provide useful information by themselves, the committor probabilities are better utilized within the framework of TPT to access various important aspects of the system dynamics through specific objects that rely on the committor values. One such object is the earlier mentioned probability distribution of *reactive trajectories*, $m^R = \{m^R(\mathbf{x})\}_{\mathbf{x}\in\Omega_\xi} \in \mathbb{R}^N$ which is defined for a state $\mathbf{x} \in \Omega_\xi$ as

$$m^R(\mathbf{x}) = q^+(\mathbf{x})\, \pi(\mathbf{x})\, q^-(\mathbf{x}). \tag{5.15}$$

If $\mathbf{x} \in \mathbb{A} \cup \mathbb{B}$, we have $m^R(\mathbf{x}) = 0$ and therefore m^R is not a normalized distribution. In order to normalize, we take $Z_{\mathbb{AB}} = \sum_{\mathbf{x}} m^R(\mathbf{x})$ and obtain

$$m^{\mathbb{AB}}(\mathbf{x}) = Z_{\mathbb{AB}}^{-1} q^+(\mathbf{x})\, \pi(\mathbf{x})\, q^-(\mathbf{x}), \tag{5.16}$$

which gives the probability of observing the system in state $\mathbf{x}$ and *in* a reactive trajectory, i.e., on its way from $\mathbb{A}$ to $\mathbb{B}$.

The definitions of the other TPT objects also rely on the expressions of the *forward* and *backward committor* functions, and their rigorous derivation can be found in [MSVE08]. We restrict ourselves to only informal definitions for these objects. One such object gives the amount of *discrete probability current* transported by the ensemble of reactive trajectories. For two distinct states $\mathbf{x}$ and $\mathbf{y}$, the *probability current* denoted $f_{\mathbf{xy}}$, is defined with the help of the transition probability between the two states, weighted by the committor probabilities such that only relevant contributions to the actual transition $\mathbb{A} \to \mathbb{B}$ are taken into account and contributions from trajectories returning to $\mathbb{A}$ before reaching $\mathbb{B}$ or belonging to transitions in the opposite direction $\mathbb{B} \to \mathbb{A}$ are ignored. The transition probability between two states of the state space is given by the corresponding entry into the generator matrix as defined by (2.54). Owing to the arguments made earlier about using the same notation for the adjoint operator $\mathcal{A}^\star$ restricted to Ω_ξ and the generator, the definition of the *discrete probability current* between two states $\mathbf{x}$ and $\mathbf{y}$ reads

$$f_{\mathbf{xy}}^{\mathbb{AB}} = \begin{cases} \pi(\mathbf{x}) q^-(\mathbf{x})\, \mathcal{A}^\star_{\mathbf{xy}}\, q^+(\mathbf{y}) & \text{if } \mathbf{x} \neq \mathbf{y} \\ 0 & \text{otherwise.} \end{cases} \tag{5.17}$$

We remark that the *probability current* can be interpreted as a very large sparse matrix $f^{\mathbb{AB}} \in \mathbb{R}^{N\times N}$. Furthermore, *probability current* is conserved, meaning that the amount of probability flux leaving subset $\mathbb{A}$ will enter subset $\mathbb{B}$ and this is also true for every state $\mathbf{x}$ along the way. This property leads to the computation of the *transition rate* between the sets $\mathbb{A}$ and $\mathbb{B}$ as

$$k_{\mathbb{AB}} = \sum_{\mathbf{x}\in\mathbb{A}, \mathbf{y}\in\Omega_\xi} f_{\mathbf{xy}}^{\mathbb{AB}} = \sum_{\mathbf{y}\in\Omega_\xi, \mathbf{z}\in\mathbb{B}} f_{\mathbf{yz}}^{\mathbb{AB}}, \tag{5.18}$$

As the definition (5.17) for $f_{\mathbf{xy}}^{\mathbb{AB}}$ might also contain "loops", meaning that a trajectory could pass through the states $\mathbf{x}$ or $\mathbf{y}$ more than once, an object called *effective current* is also needed. The *effective current* can be understood as the net amount of trajectories that perform a jump from $\mathbf{x}$ to $\mathbf{y}$ and is defined as the original *probability current* $f_{\mathbf{xy}}^{\mathbb{AB}}$ minus the contributions made by any loops, namely

$$f_{\mathbf{xy}}^{+} = \max\{f_{\mathbf{xy}}^{\mathbb{AB}} - f_{\mathbf{yx}}^{\mathbb{AB}}, 0\}. \tag{5.19}$$

$\{f^+_{\mathbf{xy}}\}_{\mathbf{x},\mathbf{y}\in\Omega_\xi}$ then defines a flow network from $\mathbb{A}$ to $\mathbb{B}$ and can be used to decompose the ensemble of reactive trajectories into single reaction pathways which are finite sequences

$$w = \left(\mathbf{x}^{(k_0)}, \ldots, \mathbf{x}^{(k_n)}\right)$$

with $\mathbf{x}^{(k_0)} \in \mathbb{A}$, $\mathbf{x}^{(k_n)} \in \mathbb{B}$ and $\mathbf{x}^{(k_1)}, \ldots, \mathbf{x}^{(k_{n-1})} \in \Omega_\xi \setminus \mathbb{A} \cup \mathbb{B}$. By $\{k_0, \ldots, k_n\}$ we have denoted a subset of the index set $\{1, \ldots, N\}$ used for an enumeration $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ of the states belonging to Ω_ξ. A pathway is simply a more concise representation for trajectories, as it forms a "tube" through which many trajectories travel depending on the size of the probability current allowed to pass through each section $(x^{(k_i)}, x^{(k_{i+1})})$ of the pathway. The flow through a specific pathway w is constrained by its *min-current*, given as

$$c(w) = \min_{m=(\mathbf{x},\mathbf{y})\in w} \{f^+_{\mathbf{xy}}\}. \tag{5.20}$$

The *dynamical bottleneck* of a reaction pathway is then defined as the pathway segment between adjacent states that can transport the smallest amount of *effective current*, i.e.,

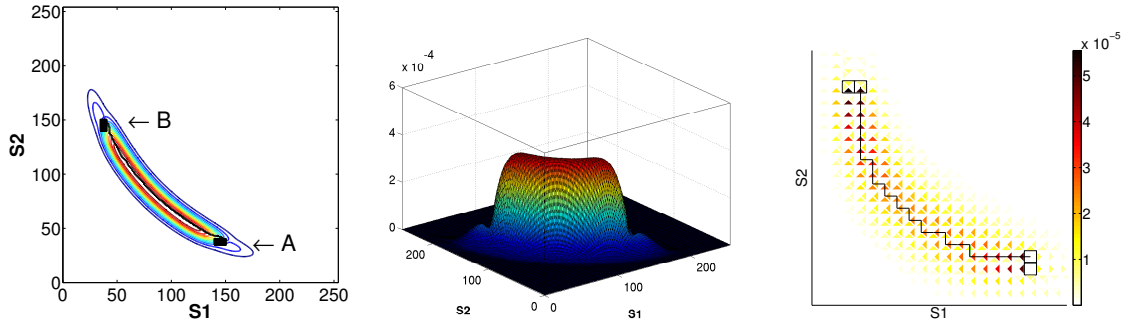
$$(b_1, b_2) = \operatorname*{argmin}_{m=(\mathbf{x},\mathbf{y})\in w} \{f^+_{\mathbf{xy}}\}. \tag{5.21}$$

Using the definitions (5.20) and (5.21), it is possible to perform a decomposition of the ensemble of reactive trajectories by identifying the "dominant" pathways, that is, those reaction pathways with the maximal *min-current* $c(w)$, or in other words, the pathways that incorporate the dynamical bottleneck with the largest throughput. By removing the *effective current* transported through the dominant pathway from the network, and repeating the process until there are no more pathways to be found between the sets $\mathbb{A}$ and $\mathbb{B}$, a non-unique decomposition of the flow network into individual pathways can be obtained. This decomposition might be useful to estimate how many pathways are necessary to carry a certain amount of the *effective current*. The algorithms for identifying bottlenecks and computing the dominant pathways come from the context of maximum flow problems, and further details can be found in [MSVE08]. An illustration of the probability distribution of reactive trajectories $m^{\mathbb{AB}}$, the *effective current* f^+ and the dominant pathways for a *toggle switch* model is provided in Figure 5.9.

Concluding this informal presentation, we note that from a numerical perspective, applying TPT to investigate a complex system can be viewed as a two-stage process. First, we need to approximate the solution of the stationary CME (5.2), which is then followed by approximating the *forward* q^+ and *backward* q^- committors, the solutions of (5.11) and (5.14), respectively. The second stage is actually a post-processing step, in which the computed approximations are used to obtain the different TPT objects presented in this section. As these computations involve algorithms that do not benefit from wavelet compression, our focus is solely on applying the wavelet method to the *committor* problems in order to enable TPT analysis for high-dimensional problems, and we proceed now to supply the specifics.

5.4. Wavelet approximation of the committors

Although similar to the stationary CME, the *forward* (5.11) and *backward* (5.14) committor problems also exhibit traits that require both a new formulation as well as a set of



(a) (b) (c)

Figure 5.9.: Left and middle panel: illustration of discrete probability distribution of reactive trajectories $m^{\mathbb{AB}}$ (contour (5.9a) and mesh plot (5.9b)) for the *toggle switch* model. Right panel: visualization of the effective current f^+ between the states of a *toggle switch*. An edge between two neighboring states $(\mathbf{x}^{(k_i)}, \mathbf{x}^{(k_{i+1})})$ with positive effective current is shown as a triangle oriented in the direction of the flux and color coded according to the intensity of the effective current. The dominant pathway that transports the largest amount of effective current is displayed in the left and right panels by the black line connecting the two sets $\mathbb{A}$ and $\mathbb{B}$ identified as a set of boxes.

changes to the adaptive wavelet method itself, before wavelet compression can be applied. We begin the discussion with the approximation of the *forward commitor* problem (5.11). Obtaining q^+ means solving a stationary *adjoint* CME on the truncated state space Ω_ξ with additional "boundary conditions", given by the requirements that $q^+(\mathbf{x}) = 0$ for all states $\mathbf{x} \in \mathbb{A}$ and $q^+(\mathbf{x}) = 1$ for all states $\mathbf{x} \in \mathbb{B}$, respectively. In order to simplify the presentation of the wavelet-based algorithms for the approximation of the committors, we first define a new operator $\mathcal{A}^\star_{bc}$ restricted to the state space Ω_ξ, given as

$$(\mathcal{A}^\star_{bc} q^+)(\mathbf{x}) = \begin{cases} (\mathcal{A}^\star q^+)(\mathbf{x}), & \text{if } \mathbf{x} \in \Omega_\xi \setminus (\mathbb{A} \cup \mathbb{B}) \\ q^+(\mathbf{x}), & \text{if } \mathbf{x} \in (\mathbb{A} \cup \mathbb{B}). \end{cases} \tag{5.22}$$

The notation (5.22) describes the effects of the "boundary conditions" from (5.11) by restricting the action of the *adjoint* operator $\mathcal{A}^\star$ to the state space $\Omega_\xi \setminus (\mathbb{A} \cup \mathbb{B})$, whereas for the states $\mathbb{A} \cup \mathbb{B}$, it leaves q^+ unchanged. Then, we end up with an equivalent formulation for the discrete *forward committor* problem (5.11), namely

$$\mathcal{A}^\star_{bc} q^+ \quad = \quad \chi_{\mathbb{B}}, \tag{5.23}$$

with $\chi_{\mathbb{B}}$ denoting the indicator function of the set $\mathbb{B}$. The next step is imposing the Galerkin condition on the equivalent formulation (5.23), thus projecting the problem onto a low-dimensional approximation space spanned by a subset $\{\psi_{j_1}, \ldots, \psi_{j_\eta}\}$ of a full wavelet basis $\Psi = \{\psi_1, \ldots, \psi_N\}$. For the committor problems, a Petrov-Galerkin scheme is no longer required if biorthogonal bases are used, because there is no need to preserve an identity in the low-dimensional setting, a situation we have encountered in the time-dependent and stationary CME algorithms. As a result, we can adapt the Galerkin scheme as described in [CDD01] to the task of solving the *forward committor* problem (5.23). Let $\Psi = \{\psi_1, \ldots, \psi_N\}$ and $\tilde{\Psi} = \{\tilde{\psi}_1, \ldots, \tilde{\psi}_N\}$ be a pair of biorthogonal wavelet

bases on $\mathcal{H}(\Omega_\xi)$, leading to the following wavelet representations

$$\begin{aligned} q^+ &= \sum_{i=1}^{N} \gamma_i \psi_i, \quad \gamma_i = \langle q^+, \tilde{\psi}_i \rangle \\ \chi_{\mathbb{B}} &= \sum_{i=1}^{N} \beta_i \tilde{\psi}_i, \quad \beta_i = \langle \chi_{\mathbb{B}}, \psi_i \rangle. \end{aligned}$$

We are interested in obtaining a numerical approximation $\tilde{q}^+ = \sum_{i=1}^{\eta} \gamma_i \psi_{j_i}$, and impose the Galerkin condition in (5.23), yielding

$$\langle \psi_{j_i}, \mathcal{A}^\star_{bc} \tilde{q}^+ \rangle = \langle \psi_{j_i}, \chi_{\mathbb{B}} \rangle, \tag{5.24}$$

for all $i = 1, \ldots, \eta$. Thus, we obtain the low dimensional linear system

$$G\gamma = \beta, \tag{5.25}$$

where, in the now customary manner, $G \in \mathbb{R}^{\eta \times \eta}$ represents a Galerkin matrix defined as,

$$G = (g_{ik})_{i,k=1}^{\eta}, \qquad g_{ik} = \langle \psi_{j_i}, \mathcal{A}^\star_{bc} \psi_{j_k} \rangle . \tag{5.26}$$

The approximation $\tilde{q}^+$ is then obtained by a fast inverse wavelet transform of the new coefficient vector $\gamma \in \mathbb{R}^\eta$.

Computing the elements of the Galerkin matrix (5.26) is however complicated by the inclusion of the "boundary conditions" in the definition of the operator $\mathcal{A}^\star_{bc}$. Before explaining how this evaluation is achieved, we use the opportunity to comment on the efficient computation of Galerkin matrix entries. The procedure that we describe next, is used *mutatis mutandis* by all the specialized versions of the adaptive wavelet method discussed in this thesis, as it is independent with respect to the choice of operator.

Two aspects make an efficient evaluation process of the Galerkin terms possible. The first one is the tensor-product approach used to build the elements of the multi - dimensional wavelet basis. For the sake of simplicity, it is always more convenient to avoid using multi-indices for the identification of the elements belonging to a wavelet basis Ψ, by using a suitable enumeration of the basis elements by single-indices. However, every element ψ_K of Ψ, with $K \in \{1, \ldots, N\}$, can also be identified via an equivalent multi-index notation $K = (k_1, \ldots, k_d)$. This leads to the representation

$$\psi_K = \psi^{(1)}_{k_1} \otimes \cdots \otimes \psi^{(d)}_{k_d}, \quad \psi_K(\mathbf{x}) = \psi^{(1)}_{k_1}(x_1) \cdot \ \ldots \ \cdot \psi^{(d)}_{k_d}(x_d), \tag{5.27}$$

where $\psi^{(i)}_{k_i}$ denotes the appropriate element from the i-th univariate wavelet basis involved in the tensor product. Of course, computing inner products of the type $\langle \psi_{j_i}, \mathcal{A}^\star \psi_{j_k} \rangle$, with j_i and j_k two arbitrarily selected single-indices from $\{j_1, \ldots, j_\eta\} \subset \{1, \ldots, N\}$, can use the multi-dimensional basis elements ψ_{j_i} and ψ_{j_k}, respectively. In principle, all we need to do is to apply the operator $\mathcal{A}^\star$ (or $\mathcal{A}$ if the problem requires it) to the basis function ψ_{j_k} and compute the inner product between the result and the element ψ_{j_i}. However, from an efficiency point of view, the adaptive wavelet method would certainly benefit if the reconstruction of these elements and the computations with the full tensor representations could be avoided. To achieve this goal, we need to make an assumption commonly

used in tensor-product approaches for solving the CME (cf. [DHJW08, JH08, Eng09a] among others), namely that the propensity functions $\alpha_j(\mathbf{x})$ given in (2.5), that appear in the definitions of both the CME operator $\mathcal{A}$ (2.61) and the *adjoint* operator $\mathcal{A}^\star$ (2.79), are separable and can be factorized as

$$\alpha_j(\mathbf{x}) = c_j \alpha_j^{(1)}(x_1) \cdot \alpha_j^{(2)}(x_2) \cdot \ \ldots \ \cdot \alpha_j^{(d)}(x_d), \tag{5.28}$$

for all reaction channels R_j ($j = 1, \ldots, M$). To illustrate how this works, we can use a simple example of a bimolecular reaction channel $S_1 + S_2 \xrightarrow{c_j} S_3$ using the standard propensity as defined in (2.12), i.e., $\alpha_j(\mathbf{x}) = c_j x_1 x_2$, $\mathbf{x} \in \mathbb{N}^d$. Applying now the factorization (5.28), we have $\alpha_j^{(1)}(x_1) = x_1$, $\alpha_j^{(2)}(x_2) = x_2$ and $\alpha_j^{(l)}(x_l) = 1$ for all $2 < l \leq d$, where d is the total number of species contained by the biochemical reaction network. Although some propensity functions cannot be factorized (e.g. a *toggle switch* model presented in [GRdO+11]), the standard propensities and even non-standard examples used in complicated reaction networks (e.g. models (2.94), (5.7) presented in this thesis or models found in literature [MK06, HHL08, MBS08, JH08, Eng09b, FL09]) can be written in the form (5.28).

Taken together, the product structure of the propensities and of the wavelet basis elements allows the evaluation of the inner products $\langle \psi_{j_i}, \mathcal{A}^\star \psi_{j_k} \rangle$ in the following way. Without loss of generality, and for the purpose of simplifying the notation, let us first denote by $u := \psi_{j_i}$ and $v := \psi_{j_k}$ two elements of a wavelet basis Ψ used to compute the entry $g_{ik}^\star$ of the Galerkin-type matrix $G^\star$ (see (5.32)). Then, using the definition of the *adjoint* operator $\mathcal{A}^\star$ (2.79), we have

$$\begin{aligned} \langle u, \mathcal{A}^\star v \rangle &= \sum_{j=1}^{M} \sum_{\mathbf{x} \in \Omega_\xi} \alpha_j(\mathbf{x}) v(\mathbf{x} + \mu^j) u(\mathbf{x}) \\ &\quad - \sum_{j=1}^{M} \sum_{\mathbf{x} \in \Omega_\xi} \alpha_j(\mathbf{x}) v(\mathbf{x}) u(\mathbf{x}) \\ &= \sum_{j=1}^{M} c_j \left(\prod_{i=1}^{d} \Theta_1(i, j, k_i, l_i) - \prod_{i=1}^{d} \Theta_0(i, j, k_i, l_i) \right) \end{aligned} \tag{5.29}$$

with

$$\begin{aligned} \Theta_1(i, j, k_i, l_i) &= \sum_{x_i=0}^{\xi_i} \alpha_j^{(i)}(x_i) v_{k_i}^{(i)}(x_i + \mu_i^j) u_{l_i}^{(i)}(x_i) \\ &= \left\langle \alpha_j^{(i)}(\cdot) v_{k_i}^{(i)}(\cdot + \mu_i^j), u_{l_i}^{(i)}(\cdot) \right\rangle \end{aligned} \tag{5.30}$$

$$\begin{aligned} \Theta_0(i, j, k_i, l_i) &= \sum_{x_i=0}^{\xi_i} \alpha_j^{(i)}(x_i) v_{k_i}^{(i)}(x_i) u_{l_i}^{(i)}(x_i) \\ &= \left\langle \alpha_j^{(i)}(\cdot) v_{k_i}^{(i)}(\cdot), u_{l_i}^{(i)}(\cdot) \right\rangle. \end{aligned} \tag{5.31}$$

In other words, by computing the inner products $\langle \psi_{j_i}, \mathcal{A}^\star \psi_{j_k} \rangle$ in the manner given by (5.29), we avoid the expensive evaluations with respect to the multi-dimensional state vector $\mathbf{x}$. These are then replaced with the inner products Θ_1 and Θ_0 with respect to every

single direction variable x_i, which are easier to compute. We remark at this point that the variable ξ_i appearing in the definitions of (5.30) and (5.31) represents the i-th entry of the truncation vector ξ. Furthermore, we can now easily take advantage of the available knowledge about the reaction network to increase the efficiency of the computation. By studying the propensity functions, we can establish whether the j-th reaction channel ($j = 1, \ldots, M$) affects the i-th spatial direction ($i = 1, \ldots, d$). If this is not the case, we have by definition $\mu_i^j = 0$ and $\alpha_j^{(i)} = 1$, which leads to $\Theta_1(i, j, k_i, l_i) = \Theta_0(i, j, k_i, l_i)$, i.e., the inner product for the "shifted" term of the operator is identical the inner product of the "non-shifted" term. Consequently, many of the terms in the expressions above vanish and do not need to be explicitly computed, which significantly reduces the computational workload. When evaluating the scalar products, improvements can also be made by using a sparse storage scheme for the elements of the univariate wavelet bases, and taking into consideration the support lengths of specific basis elements or whether the supports intersect.

After presenting the computation of the Galerkin terms in the general case, we proceed with the computation of the entries for the specific case of the committor problems. Usually, the number of the states assigned to the subsets $\mathbb{A}$ and $\mathbb{B}$, between which the various TPT objects are defined, is relatively small in comparison with the total number of states contained by the truncated state space Ω_ξ. As a result, the computation of the Galerkin matrix for the *forward committor* problem starts by computing the entries of a Galerkin matrix $G^\star \in \mathbb{R}^{\eta \times \eta}$ with

$$G^\star = (g_{ik}^\star)_{i,k=1}^{\eta}, \qquad g_{ik}^\star = \langle \psi_{j_i}, \mathcal{A}^\star \psi_{j_k} \rangle, \tag{5.32}$$

accomplished by using (5.29). Basically, instead of using the operator $\mathcal{A}_{bc}^\star$ from (5.23), we use the standard *adjoint* operator $\mathcal{A}^\star$ on the entire state space Ω_ξ ignoring for now the special conditions for the states assigned to $\mathbb{A}$ and $\mathbb{B}$. After computing $G^\star$, the next question is how to replace the entries of the matrix that are affected by the inclusion of the "boundary conditions" in the definition of the operator $\mathcal{A}_{bc}^\star$. A first step is to "cut out"' these states from the state space Ω_ξ. To this end, we give a more rigorous definition for a subset $\mathbb{A} \subset \Omega_\xi$ as a contiguous set of states forming a hypercube inside the larger hypercube defined by the truncation vector ξ, i.e.,

$$\mathbb{A} \;=\; \{\mathbf{x} \in \Omega_\xi \mid \varsigma_i^l \le x_i \le \varsigma_i^r,\;\; 0 \le \varsigma_i^l \le \varsigma_i^r \le \xi_i, \text{ for all } i = 1, \ldots, d\}. \tag{5.33}$$

The subset $\mathbb{A}$ is thus defined by the intervals $[\varsigma_i^l, \varsigma_i^r]$ giving the range for the values of the spatial variables x_i in each direction $i \in \{1, \ldots, d\}$. As the states $\mathbf{x} \in \Omega_\xi$ take integer values, we also have a natural mapping to a set of multi-indeces, i.e., the values of any state $\mathbf{x}$ represent also the values of the corresponding multi-index. Let $K(\xi) = I^{[0,\xi_1]} \otimes \cdots \otimes I^{[0,\xi_d]}$ denote the list of all such multi-indices associated with the states $\mathbf{x} \in \Omega_\xi$, with $I^{[0,\xi_i]} = \{k_i \in \mathbb{N} \mid 0 \le k_i \le \xi_i\}$ the local indices for the i-th direction. Further, let $K(\mathbb{A}) = I^{[\varsigma_1^l, \varsigma_1^r]} \otimes \cdots \otimes I^{[\varsigma_d^l, \varsigma_d^r]}$ describe the list of multi-indices for the states $\mathbf{x} \in \mathbb{A}$. Then, removing from the i-th direction the indices corresponding to the elements in subspace $\mathbb{A}$, means that the summations in (5.30) and (5.31) are computed only for the local index intervals $K(\xi_i, \varsigma_i^l, \varsigma_i^r) = [0, \varsigma_i^l - 1] \cup [\varsigma_i^r + 1, \xi_i]$ where $i \in \{1, \ldots, d\}$. However, from a computational point of view, rather than performing the summation using the index vectors $K(\xi_i, \varsigma_i^l, \varsigma_i^r)$, it is easier to perform the operation on the smaller index intervals

$[\varsigma_i^l, \varsigma_i^r]$. Additionally, we can take advantage of the localized support of the elements from the univariate wavelet bases to improve computational efficiency. Because for many elements we have $\mathrm{supp}(\psi_{k_i}^{(i)}) \cap [\varsigma_i^l, \varsigma_i^r] = \emptyset$, the corresponding terms vanish. Note that we use here $\mathrm{supp}(\psi_{k_i}^{(i)})$ in a loose sense, meant to identify the index values corresponding to the localized support of the discrete basis element. Thus, removing the entries of $G^\star$ that are affected by the "cut out" of $\mathbb{A}$ and $\mathbb{B}$, means selecting those elements from the index set $\{j_1, \ldots, j_\eta\}$ that belong to basis elements where at least one of the tensor components fulfills the condition $\mathrm{supp}(\psi_{k_i}^{(i)}) \cap [\varsigma_i^l, \varsigma_i^r] \neq \emptyset$. As the sets $\mathbb{A}$ and $\mathbb{B}$ are known beforehand, and the support lengths of the elements from the univariate basis elements as well, we can compute in a pre-processing step the index subset $\mathcal{M}_{\mathbb{A}\cup\mathbb{B}} = \{m_1, \ldots, m_\nu\} \subset \{1, \ldots, N\}$ of the elements satisfying these conditions. Naturally, only some of the elements identified by the indices in $\mathcal{M}_{\mathbb{A}\cup\mathbb{B}}$ are also contained in the active set $\mathcal{J}_\eta = \{j_1, \ldots, j_\eta\}$ which is used for computing the entries of the Galerkin matrix (5.32). Therefore, we only need to compute the entries of a smaller matrix $G^c \in \mathbb{R}^{\tilde{\eta}\times\tilde{\eta}}$, with $\tilde{\eta} \ll \eta$. For this task we use only the basis elements whose indices belong to the intersection set $\mathcal{J}_{\tilde{\eta}} = \mathcal{J}_\eta \cap \mathcal{M}_{\mathbb{A}\cup\mathbb{B}}$. Then, the entries of the matrix G^c are given as

$$G^c = (g_{ik}^c)_{i,k=1}^{\tilde{\eta}}, \qquad g_{ik}^c = \left\langle \psi_{j_i}, \mathcal{A}^\star \underline{\psi_{j_k}} \right\rangle. \tag{5.34}$$

In the expression (5.34) above, we have used the notation $\underline{\psi_{j_k}}$ to describe a multi - dimensional basis element ψ_{j_k}, with its tensor components containing non-zero elements only for the entries with local indices $[\varsigma_i^l, \varsigma_i^r] \cup [\varrho_i^l, \varrho_i^r]$ in the i-th spatial direction. Here, ϱ_i^l, ϱ_i^r are the minimum and maximum values for the spatial variables x_i in each direction that define the set $\mathbb{B} \subset \Omega_\xi$, analogously to (5.33).

In practical terms, the computation of the entries of G^c uses the same procedure as the one detailed in (5.29), but with the modifications that we have discussed above with respect to the summations in (5.30) and (5.31). Because the matrix G^c is smaller then $G^\star$, a final step is to perform an embedding of its values into an appropriately size matrix $G_\eta^c \in \mathbb{R}^{\eta\times\eta}$, by computing the corresponding locations in the larger matrix.

After computing the effects of removing the state subsets $\mathbb{A}$ and $\mathbb{B}$ from the state space Ω_ξ, the last step in the computation must fill in the missing values. For $\mathbf{x} \in \mathbb{A} \cup \mathbb{B}$, we have defined the operator $\mathcal{A}_{bc}^\star$ as leaving the function q^+ unchanged. For the Galerkin entries, this translates into computing the corresponding "mass" matrix $\Delta^c \in \mathbb{R}^{\tilde{\eta}\times\tilde{\eta}}$ with entries given by

$$\Delta^c = (\delta_{ik}^c)_{i,k=1}^{\tilde{\eta}}, \qquad \delta_{ik}^c = \left\langle \psi_{j_i}, \underline{\psi_{j_k}} \right\rangle. \tag{5.35}$$

After embedding the entries of the matrix Δ^c into an appropriately sized matrix $\Delta_\eta^c \in \mathbb{R}^{\eta\times\eta}$, we can recover the Galerkin matrix (5.26) for the *forward committor* problem by simply using the three previously computed matrices $G^\star$, G_η^c and Δ_η^c,

$$G = G^\star - G_\eta^c + \Delta_\eta^c. \tag{5.36}$$

We continue now the exposition with the details related to approximating the solution q^- of the *backward committor* (5.14) problem using wavelets. Computing q^- means solving a stationary-like CME equation with an operator $\tilde{\mathcal{A}}$ restricted to the state space Ω_ξ,

isomorphic to the generator matrix of the time-reversed process. Additionally we have the "boundary conditions" $q^-(\mathbf{x}) = 1$ for all states $\mathbf{x} \in \mathbb{A}$ and $q^-(\mathbf{x}) = 0$ for all states $\mathbf{x} \in \mathbb{B}$, respectively. The generator of the time-reversed process (5.13), and implicitly the operator $\tilde{\mathcal{A}}$, features divisions with the values of the stationary distribution π. This means that for the states where the stationary distribution almost vanishes, computing the entries of the *backward committor* q^- is not possible, a problem illustrated in Figure 5.8b. Consequently, the affected states must be omitted from the computation of the committor q^-. One possible solution would be to redefine the state space Ω_ξ such that it includes only the states $\mathbf{x}$ where $\pi(\mathbf{x}) > \texttt{tol}$, but this would entail first an exhaustive search of the high-dimensional space, and secondly and even more inconveniently, the simplicity of the domain would be lost, replaced by a complex multi-dimensional domain that conforms to the profile of the stationary distribution. Such a state space would then make the application of the adaptive wavelet method in its current form impossible. Fortunately, (5.14) can be replaced by the alternative formulation

$$\begin{cases} (\mathcal{A}\rho^-)(\mathbf{x}) = 0, & \text{for all } \mathbf{x} \in \Omega_\xi \setminus (\mathbb{A} \cup \mathbb{B}) \\ \rho^-(\mathbf{x}) = \pi(\mathbf{x}), & \text{for all } \mathbf{x} \in \mathbb{A} \\ \rho^-(\mathbf{x}) = 0, & \text{for all } \mathbf{x} \in \mathbb{B}. \end{cases} \tag{5.37}$$

In (5.37), instead of computing q^-, we compute ρ^- with $\rho^-(\mathbf{x}) = \pi(\mathbf{x})q^-(\mathbf{x})$ for all $\mathbf{x} \in \Omega_\xi \setminus (\mathbb{A} \cup \mathbb{B})$. The reasons for this simplification are two-fold. First, instead of the operator $\tilde{\mathcal{A}}$, we can use the more convenient CME operator $\mathcal{A}$ (2.61), and secondly, the computation of the *backward committor* q^- by itself is of limited practical interest. Its usefulness is much more related to its appearance in the operational definitions of the TPT objects (5.16) or (5.17). However, the TPT objects all feature terms of the type $\pi(\mathbf{x})q^-(\mathbf{x})$, so computing ρ^- instead of q^- is advantageous, as it avoids a post-processing step.

In order to have a wavelet "friendly" formulation, we proceed now in a similar way as in (5.23), and define a new operator $\mathcal{A}_{bc}$ as

$$(\mathcal{A}_{bc}\rho^-)(\mathbf{x}) = \begin{cases} (\mathcal{A}\rho^-)(\mathbf{x}), & \text{if } \mathbf{x} \in \Omega_\xi \setminus (\mathbb{A} \cup \mathbb{B}) \\ \rho^-(\mathbf{x}), & \text{if } \mathbf{x} \in (\mathbb{A} \cup \mathbb{B}). \end{cases} \tag{5.38}$$

Using the newly introduced notation (5.38) leads to the form of the *backward committor* problem to which the wavelet method will be applied, namely

$$\mathcal{A}_{bc}\rho^- \quad = \quad \chi_{\mathbb{A}}, \tag{5.39}$$

where $\chi_{\mathbb{A}}$ is a function defined as

$$\chi_{\mathbb{A}}(\mathbf{x}) = \begin{cases} \pi(\mathbf{x}), & \mathbf{x} \in \mathbb{A} \\ 0, & \mathbf{x} \notin \mathbb{A}. \end{cases}$$

Next, we can project (5.39) onto a low-dimensional space $\{\psi_{j_1}, \ldots, \psi_{j_\eta}\}$ by imposing the Galerkin condition in a manner similar to the *forward committor* case, which yields a low dimensional linear system featuring the Galerkin matrix $T \in \mathbb{R}^{\eta \times \eta}$ with entries

$$T = (t_{ik})_{i,k=1}^{\eta}, \qquad t_{ik} = \langle \psi_{j_i}, \mathcal{A}_{bc}\psi_{j_k} \rangle . \tag{5.40}$$

Computing the entries of the Galerkin matrix (5.40) then uses the same method as employed for the matrix (5.32) related to the *forward committor* problem, with minimal changes imposed by the use of operator $\mathcal{A}_{bc}$ instead of $\mathcal{A}_{bc}^\star$.

One last issue when using wavelet approximation for the committor problems is to ensure that the values of the numerical approximations for states $\mathbf{x} \in \mathbb{A} \cup \mathbb{B}$ are consistent with the definitions of the committors. Let

$$r = \mathcal{A}_{bc}^{\star} \tilde{q}^{+} - \chi_{\mathbb{B}} \tag{5.41}$$

be the residual of the *forward committor* problem (5.23). The goal is to have $r(\mathbf{x}) = 0$ for any state $\mathbf{x} \in \mathbb{A} \cup \mathbb{B}$. As we know that the Galerkin condition (5.24) implies that the residual is orthogonal to the approximation space, we just need to choose the elements of the approximation space such that we have $\chi_{\mathbb{A}} \in \operatorname{span}\{\psi_{j_1}, \ldots, \psi_{j_\eta}\}$ and $\chi_{\mathbb{B}} \in \operatorname{span}\{\psi_{j_1}, \ldots, \psi_{j_\eta}\}$, respectively. Here, $\chi_{\mathbb{A}}$ and $\chi_{\mathbb{B}}$ are the indicator functions for the sets $\mathbb{A}$ and $\mathbb{B}$. In other words, the approximation space must always contain the basis elements required to represent the states enclosed by the two subsets. Obtaining the corresponding indices is easily accomplished in a pre-processing step by using wavelet transforms of two multi-dimensional objects with the same size as Ω_ξ but non-zero entries only in the states $\mathbf{x} \in \mathbb{A}$ and $\mathbf{x} \in \mathbb{B}$, respectively. The indices corresponding to the wavelet coefficients that are larger than a prescribed tolerance are then saved in a set $\mathcal{R}_{\mathbb{A}\cup\mathbb{B}}$. The set $\mathcal{R}_{\mathbb{A}\cup\mathbb{B}}$ will then be retained throughout the residual-based refinement process of *essential* degrees of freedom in the active index set, with its elements being omitted from the thresholding procedure. Together with the Galerkin projection, these two building blocks constitute the computational core of the adaptive wavelet method for the committors. A sketch of the specialized method is given in Algorithm 7.

Algorithm 7: Adaptive wavelet method for the *forward* committor problem

Parameter : subsets $\mathbb{A}$ and $\mathbb{B} \subset \Omega_\xi$ with $\mathbb{A} \cap \mathbb{B} = \emptyset$, tolerance `tol`
index sets $\mathcal{M}_{\mathbb{A}\cup\mathbb{B}}$ and $\mathcal{R}_{\mathbb{A}\cup\mathbb{B}}$ (see comments above)
committor type: *forward*
$\Delta\mu$ (new basis elements per step), $\eta_{\max}$ (maximum allowed DOFs)
initial active set of coefficients $\mathcal{J}_\eta = \{j_1, \ldots, j_\eta\}$

Input : coefficient vector $\beta^{(0)} = (\beta_1^{(0)}, \ldots, \beta_\eta^{(0)})^T$ via FWT of $\chi_{\mathbb{B}}$
Galerkin matrix G defined as in (5.26)

Output : approximation $\tilde{q}^{+} \approx q^{+}$

Solve $G\gamma = \beta$ and recover approximation $\tilde{q}^{+}$, compute residual r using (5.41)

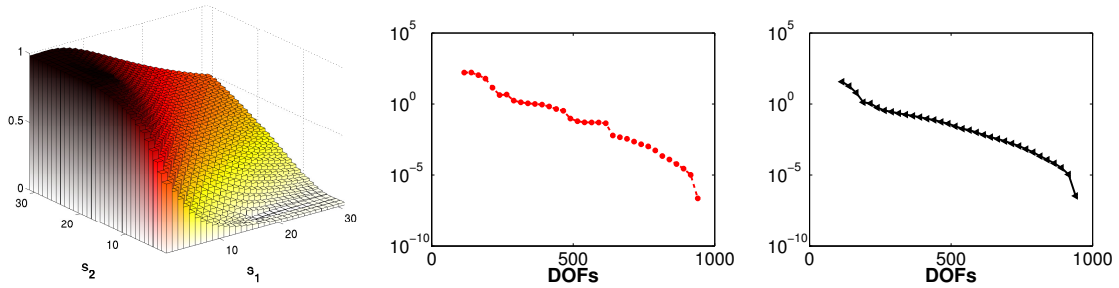
while $||r||_1 >$ `tol` *and* $\eta < \eta_{\max}$ **do**

1. Enlarge approximation space by $\Delta\mu$ elements using *a posteriori* analysis of r.
2. Solve the linear system
$$G\gamma = \beta$$
with enlarged Galerkin matrix $G \in \mathbb{R}^{(\eta+\Delta\mu)\times(\eta+\Delta\mu)}$ and vector $\beta \in \mathbb{R}^{\eta+\Delta\mu}$.
3. Compute approximation $\tilde{q}^{+} = \sum_{i=1}^{\eta+\Delta\mu} \gamma_i \psi_{j_i}$ and new residual r
4. Apply thresholding to coefficient vector γ, taking care to update the Galerkin matrix G and vector β. Thresholding is only applied to coefficients with indices in $(\mathcal{J}_\eta + \mathcal{J}_{\Delta\mu}) \setminus \mathcal{R}_{\mathbb{A}\cup\mathbb{B}}$.

end

We remark that because many substeps from Algorithm 7, like the enlargement of the approximation space or the thresholding strategy use the same mechanisms as their counterparts from Algorithms 3 and 6, some details have been omitted from the current description.

Next, we take a look at the feasibility of our approach by comparing the wavelet approximation for the *forward committor* to a reference solution. The model chosen is again the *toggle switch* given in (2.94) and using the parameter set (4.32) the corresponding state space $\Omega_{32\times 32}$ is sufficiently small, such that (5.23) can be solved directly. In the left panel of Figure 5.10, the committor function q^+ is shown, while in the middle panel we plot the error between the reference solution and the wavelet approximation computed by successive steps of the adaptive wavelet method. The right panel 5.10c displays the evolution of the 1-norm of the residual (5.41) for the wavelet approximation. For this test problem, we used the sets $\mathbb{A} = \{(18, 3), (18, 4)\}$ and $\mathbb{B} = \{(3, 18), (4, 18)\}$.



(a) Committor q^+ (b) 1-norm of error (c) 1-norm of residual

Figure 5.10.: Comparing the wavelet approximation of the forward committor function q^+ for the model (2.94) with parameter set (4.32) with a reference solution obtained by solving the linear system (5.23) directly. Both the error and the residual of the wavelet approximation are measured in the 1-norm.

Drawing a line, the specialized version of the wavelet method for the committors uses the same numerical recipe as the method for the stationary CME discussed earlier in this chapter. The main differences are related to the use of a non-standard state space $\Omega_\xi/\mathbb{A}\cup\mathbb{B}$ which leads to an increase in the complexity of the bookkeeping required to project the system in the low-dimensional space. We remark that although many of the issues related to approximating the committor equations have been resolved, additional technical difficulties remain, for example making the computation of the specific Galerkin matrix entries more efficient. Moreover, after approximating the committor functions, a post-processing step must be applied in order to compute the dominant transition pathways and the corresponding transition rates. For models with large state spaces, this is also a non-trivial problem that must be solved before utilizing the current method at its full potential.

Before presenting some examples of using the committor approximations for metastability analysis, we comment on another newly developed approach for the computation of committor probabilities. Instead of solving the system of linear equations from (5.23), an alternative formulation for the committors problem is to express the committors in terms of the dominant eigenvectors of a modified operator, and based on this reformulation a method for their efficient computation has been recently proposed in [PHSN11].

Instead of the master operator description of system dynamics, the authors use a description in terms of the time-discrete transition matrix $T(\tau) \in \mathbb{R}^{N\times N}$, which appears in the *Chapman-Kolmogorov* equation

$$p(k\tau) = p(0)T^k(\tau).$$

However, they also provide a transformation procedure between the matrix T and the generator matrix underlying our formulation. The approach then uses a modified transition matrix and reduces the computation of the committor to finding one largest nontrivial right eigenvector (cf. [PHSN11]), with the help of the power method [GVL96]. Consequently, we remark that wavelet compression can also be used for the eigenvalue formulation of the committor problem, by modifying the method designed for the stationary CME to include the changes related to the computation of the Galerkin entries detailed in section 5.4.

5.5. Metastability analysis with TPT

We illustrate now the usefulness of approximating the committors by visualizing some TPT objects for the multi-dimensional genetic toggle switches given in (5.7) and (5.9). Figure 5.11 depicts a selection of reaction pathways between two of the three metastable states of the $3D$ *toggle switch* model (5.7).

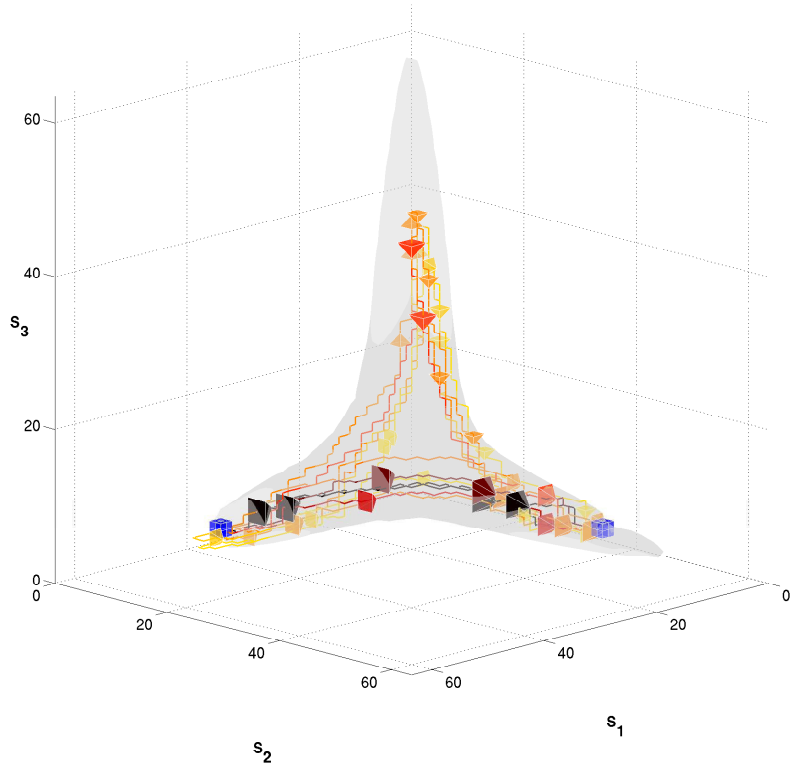


Figure 5.11.: Selection of pathways for the 3D toggle switch model (5.7)

The two sets $\mathbb{A}$ and $\mathbb{B}$ were defined by the limits

$$\mathbb{A} = \{\mathbf{x} \in \Omega_\xi |\ 36 \leq x_1 \leq 37,\quad 2 \leq x_2 \leq 3,\quad 2 \leq x_3 \leq 3\}$$
$$\mathbb{B} = \{\mathbf{x} \in \Omega_\xi |\ 2 \leq x_1 \leq 3,\ 36 \leq x_2 \leq 37,\quad 2 \leq x_3 \leq 3\}$$

and are identified in Figure 5.11 by blue voxels. Each pathway is color-coded according to its *min-current* $c(w)$ (darker colors indicate larger values), and at selected points along the pathways, pyramids oriented in the direction of the flux and also color-coded with respect to the intensity of the effective current flowing through the corresponding pathway segments are shown. The size of the pyramid-shaped markers also indicate the intensity of the effective current, and we note that the featured pathways are only an arbitrary selection of the full decomposition in single pathways. Showing all the individual pathways however, would render the visualization ineffective. Moreover, the normalized probability distribution of reactive trajectories $m^{\mathbb{AB}}$ is shown as a transparent gray isosurface, which conforms roughly to the profile of the stationary distribution presented in Figure 5.4c. Note that compared to the 3D visualization of the stationary distribution profile from Figure 5.4c, a different view point is used. We also remark that for visualization purposes, in Figure 5.11 results for the model (5.7) using the scaled parameter set

$$\begin{aligned} &c_{11} = 4225 \cdot 0.5,\quad c_{21} = 4225 \cdot 0.2,\quad c_{31} = 4225 \cdot 1,\quad c_{i2} = c_{i3} = 65,\ \ i = \{1,2,3\} \\ &c_4 = 0.025 \cdot 0.5,\quad c_5 = 0.025 \cdot 0.2,\quad c_6 = 0.025 \cdot 1. \end{aligned} \tag{5.42}$$

are shown. As it can be observed, TPT allows a precise analysis of the underlying dynamics of the system. While most of the probability current flows using the direct route between the two states $\mathbb{A}$ and $\mathbb{B}$ as indicated by the dark-colored pathways, some of the pathways also make a detour towards the other metastable state, before continuing to the set $\mathbb{B}$. However, these pathways have a lower *effective current* throughput, which is illustrated by their lighter color and the corresponding smaller size of the markers. Thus, TPT can potentially be used to identify those degrees of freedom that describe the essential dynamics, and therefore enable the creation of reduced models that discard the unused DOFs.

As a last example, we show the probability distribution of reactive trajectories and the dominant pathway, i.e., the pathway which contains the dynamical bottleneck with the highest intensity, for the $4D$ *toggle switch* model from (5.9). Because we are now trying to visualize a four-dimensional object, the subplots in Figure 5.12 show only $3D$ marginals of both the probability distribution of reactive trajectories and the dominant pathway. The 4th dimension of the pathway segments is visualized by color-coding the pathway markers using a heat map. The two sets $\mathbb{A}$ and $\mathbb{B}$ were defined by the limits

$$\mathbb{A} = \{\mathbf{x} \in \Omega_\xi |\ 51 \leq x_1 \leq 55,\ 16 \leq x_2 \leq 20,\quad 1 \leq x_3 \leq 3,\quad 1 \leq x_4 \leq 2\}$$
$$\mathbb{B} = \{\mathbf{x} \in \Omega_\xi |\ 1 \leq x_1 \leq 3,\ 1 \leq x_2 \leq 2,\quad 51 \leq x_3 \leq 55,\ 16 \leq x_4 \leq 20\}$$

and are shown using blue voxels for set $\mathbb{A}$ and red voxels for set $\mathbb{B}$, respectively.

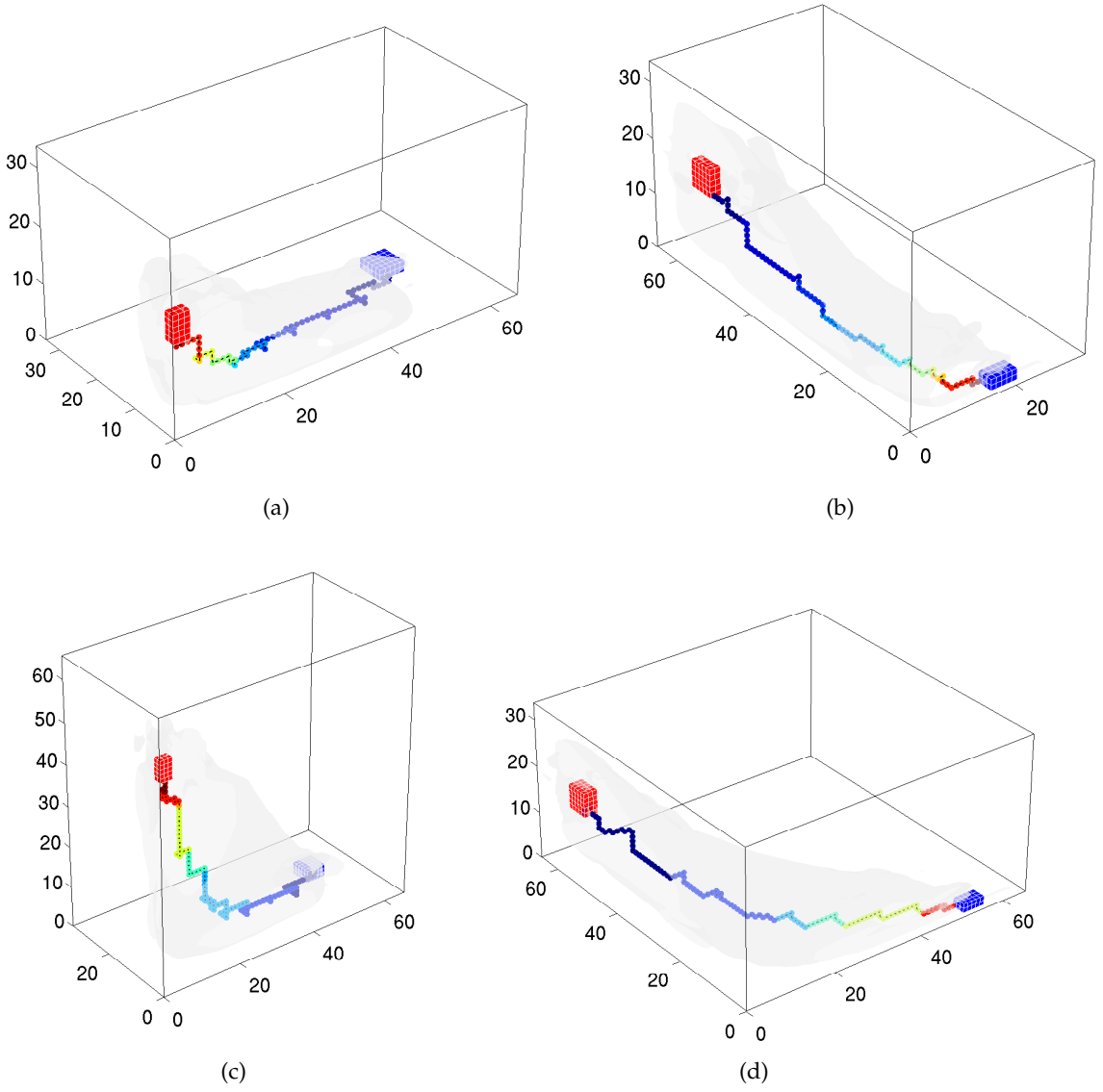


Figure 5.12.: Examples of dominant pathway for the $4D$ *toggle switch* model (5.9) visualized using $3D$ projections: $S_1 - S_2 - S_3$ (5.12a), $S_1 - S_2 - S_4$ (5.12b), $S_1 - S_3 - S_4$ (5.12c) and $S_2 - S_3 - S_4$ (5.12d).

CHAPTER

SIX

HYBRID DETERMINISTIC-STOCHASTIC MODELS

The previous Chapters 4 and 5, demonstrated that using wavelet compression is a valid way of mitigating the effects the *curse of dimensionality* has on both the time-dependent CME (2.67), and its stationary version given by (5.2). In order to achieve further reductions in the numbers of degrees of freedom required to approximate the solution of the stationary CME, we present in this chapter how the adaptive wavelet method can be embedded within a hybrid strategy. As is customary with hybrid approaches (see e.g. [FL07, HL07, FLH08, HHL08]), the main idea is to split the model into stochastic and deterministic parts. The part that contains the species with low copy-numbers and hence more susceptible to stochastic fluctuations is modeled with the computationally expensive CME, while the part that includes the remaining system components with higher molecular counts is dealt with in a deterministic setting. Then, the stationary solution of the CME for the full system is approximated by alternately calling the adaptive wavelet method for the stochastic part, and a Newton method for the deterministic description. This hybrid approach allows the numerical treatment of systems with larger state spaces by employing the adaptive wavelet method more efficiently, i.e., only for the critical subparts of the biochemical system, and the potential will be illustrated by a numerical example. However, the advantages brought by the reduction in the size of the CME state space when using hybrid methods must be weighted against the loss of accuracy inherent when solving the modified equation. Additionally, biochemical systems do not always exhibit a clear separation of scales, so partitioning the systems must be handled with care. We proceed now with the derivation of a hybrid model for the CME first proposed by A. Hellander and P. Lötstedt in [HL07], followed by a presentation of the hybrid adaptive wavelet method for the stationary CME based on this approach. However, numerical results and the detailed error analysis of the Hellander-Lötstedt hybrid model by Jahnke [Jah11], showed that this hybrid model cannot *always* deliver a reasonably accurate and qualitatively correct description of the system dynamics. We conclude the chapter with two numerical examples, a non-trivial problem where the hybrid model can be successfully applied, and a second example with a bimodal solution profile where the approach faces difficulties. Finally, possible refinements to the hybrid adaptive wavelet method allowing the treatment of problems having multi-modal stationary distributions are mentioned.

6.1. Derivation of the Hellander-Lötstedt hybrid model

6.1.1. Splitting the model

The derivation of the hybrid model starts by way of necessity with the partition of the full model which contains d different species $S_1, \ldots, S_d$, into two groups $S_1, \ldots, S_{d_1}$ and $S_{d_1+1}, \ldots, S_{d_1+d_2}$, where $d_1 < d$ and $d = d_1 + d_2$. The first group $S_1, \ldots, S_d$ is made up of a *few* critical species while the second one encompasses the other species from the biochemical reaction network that interact with the critical species, but are less prone to stochastic fluctuations due to their larger copy-numbers. In accordance with this partition of the species, the state vector $\mathbf{x} \in \mathbb{N}_0^d$ of the full system and the stoichiometric vectors μ^j of the reaction channels R_j are also decomposed into

$$\mathbf{x} = (\mathbf{y}, \mathbf{z}) \in \mathbb{N}_0^d, \text{ and } \mu^j = (\nu^j, \zeta^j) \in \mathbb{Z}^d \quad \text{for all } j \in 1, \ldots, M \tag{6.1}$$

where

$$\begin{aligned} \mathbf{y} &= (y_1, \ldots, y_{d_1}) &&= (x_1, \ldots, x_{d_1}) \in \mathbb{N}_0^{d_1} \\ \mathbf{z} &= (z_1, \ldots, z_{d_2}) &&= (x_{d_1+1}, \ldots, x_{d_1+d_2}) \in \mathbb{N}_0^{d_2}. \end{aligned}$$

Moreover, we have $\nu^j = (\nu_i^j)_{i=1}^{d_1} = (\mu_i^j)_{i=1}^{d_1} \in \mathbb{Z}^{d_1}$ and $\zeta^j = (\zeta_k^j)_{k=1}^{d_2} = (\mu_k^j)_{k=d_1+1}^{d_2} \in \mathbb{Z}^{d_2}$ for every reaction index $j = 1, \ldots, M$. We also assume that the propensity functions $\alpha_j(\mathbf{x})$ defined in (2.5) are separable, i.e., for every j there exist two functions $\beta_j : \mathbb{N}_0^{d_1} \to \mathbb{R}$ and $\gamma_j : \mathbb{N}_0^{d_2} \to \mathbb{R}$ such that

$$\alpha_j(\mathbf{x}) = \alpha_j(\mathbf{y}, \mathbf{z}) = \beta_j(\mathbf{y})\gamma_j(\mathbf{z}). \tag{6.2}$$

The propensities with the general form given by (2.5) naturally satisfy this assumption, but also non-standard propensities like the ones used to model inhibition by competing species encountered for example in the model of the Gardner *toggle switch* listed in (2.94) (reactions R_1 and R_2), are amenable to a separation along the lines described by (6.2). We remark however, that the decomposition of the propensities is not unique in the sense that the reaction constant c_j appearing in the general form (2.5) of the propensity functions can be inserted either in the definition of the function β_j or that of γ_j.

Substituting the state vector $\mathbf{x} \in \mathbb{N}_0^d$ with its decomposition (6.1) into the CME given in (2.58), and making use of the assumption (6.2), we have

$$\begin{aligned} \partial_t p(t, \mathbf{y}, \mathbf{z}) &= (\mathcal{A}p)(t, \mathbf{y}, \mathbf{z}) \\ &= \sum_{j=1}^{M} \Big(\alpha_j(\mathbf{y} - \nu^j, \mathbf{z} - \zeta^j) p(t, \mathbf{y} - \nu^j, \mathbf{z} - \zeta^j) - \alpha_j(\mathbf{y}, \mathbf{z}) p(t, \mathbf{y}, \mathbf{z}) \Big) \\ &= \sum_{j=1}^{M} \Big(\beta_j(\mathbf{y} - \nu^j)\gamma_j(\mathbf{z} - \zeta^j) p(t, \mathbf{y} - \nu^j, \mathbf{z} - \zeta^j) - \beta_j(\mathbf{y})\gamma_j(\mathbf{z}) p(t, \mathbf{y}, \mathbf{z}) \Big). \end{aligned} \tag{6.3}$$

As was already shown in section 2.5, restricting the CME operator (2.61) to a finite state space Ω_ξ (2.65), leads to the truncated operator being isomorphic to a large sparse matrix, which by a slight abuse of notation we also denote by $\mathcal{A} \in \mathbb{R}^{N \times N}$, with $N = \prod_{i=1}^{d} \xi_i$ and ξ denoting a suitably chosen truncation vector. Furthermore, in case Neumann boundary

conditions are imposed and the assumptions detailed in Chapter 2 are satisfied, we have that if $p(0, \mathbf{y}, \mathbf{z})$ is a probability distribution, the solution of the CME will converge to a stationary distribution $\rho = \rho(\mathbf{y}, \mathbf{z})$ with $\mathcal{A}\rho = 0$.

Consequently, we extend now the definition of the truncation of the infinite state space $\mathbb{N}_0^d$ given by (2.65) to the partitioned model by writing $\Omega_\xi = \Omega_{\xi_{|\mathbf{y}}} \times \Omega_{\xi_{|\mathbf{z}}}$, where

$$\begin{aligned} \Omega_{\xi_{|\mathbf{y}}} &= \{\mathbf{y} \in \mathbb{N}_0^{d_1} \mid \mathbf{y} < \xi_{|\mathbf{y}}\ ,\ \xi_{|\mathbf{y}} = (\xi_1, \ldots, \xi_{d_1})\} \\ \Omega_{\xi_{|\mathbf{z}}} &= \{\mathbf{z} \in \mathbb{N}_0^{d_2} \mid \mathbf{z} < \xi_{|\mathbf{z}}\ ,\ \xi_{|\mathbf{z}} = (\xi_{d_1+1}, \ldots, \xi_{d_1+d_2})\}. \end{aligned}$$

Similarly to the conditions imposed in (2.59), we further stipulate that

$$\beta_j(\mathbf{y}) = 0 \text{ if } \mathbf{y} \notin \Omega_{\xi_{|\mathbf{y}}},\ \ \gamma_j(\mathbf{z}) = 0 \text{ if } \mathbf{z} \notin \Omega_{\xi_{|\mathbf{z}}},\ \ p(\cdot, \mathbf{y}, \mathbf{z}) = 0 \text{ if } (\mathbf{y}, \mathbf{z}) \notin \Omega_{\xi_{|\mathbf{y}}} \times \Omega_{\xi_{|\mathbf{z}}}, \tag{6.4}$$

and analogously to (2.72), impose discrete Neumann boundary conditions, i.e.,

$$\beta_j(\mathbf{y}) = 0 \text{ if } \mathbf{y} + \nu^j \notin \Omega_{\xi_{|\mathbf{y}}},\ \ \gamma_j(\mathbf{z}) = 0 \text{ if } \mathbf{z} + \zeta^j \notin \Omega_{\xi_{|\mathbf{z}}}. \tag{6.5}$$

Condition (6.5) leads to the suppression of all reaction channels $j \in \{1, \ldots, M\}$ that could cause a jump from a state $(\mathbf{y}, \mathbf{z}) \in \Omega_{\xi_{|\mathbf{y}}} \times \Omega_{\xi_{|\mathbf{z}}}$ to a state $(\mathbf{y} + \nu^j, \mathbf{z} + \zeta^j) \notin \Omega_{\xi_{|\mathbf{y}}} \times \Omega_{\xi_{|\mathbf{z}}}$ lying outside the truncated state space.

6.1.2. Model reduction by product approximation

The next step in the derivation of the hybrid model is to approximate the solution $\rho(\mathbf{y}, \mathbf{z})$ of the stationary CME via a direct product

$$\rho(\mathbf{y}, \mathbf{z}) \approx (u \otimes q)(\mathbf{y}, \mathbf{z}) = u(\mathbf{y})q(\mathbf{z}) \tag{6.6}$$

of two stationary marginal distributions $u(\mathbf{y})$ and $q(\mathbf{z})$ which depend only on $\mathbf{y}$ and $\mathbf{z}$, respectively. Naturally, for most cases this product approximation is very coarse.

With the aim of deriving a set of coupled equations for the two marginal distributions, we impose the conditions

$$0 = \sum_{\mathbf{z} \in \Omega_{\xi_{|\mathbf{z}}}} \mathcal{A}(u \otimes q)(\mathbf{y}, \mathbf{z}), \text{ for all } \mathbf{y} \in \Omega_{\xi_{|\mathbf{y}}} \tag{6.7}$$

$$0 = \sum_{\mathbf{y} \in \Omega_{\xi_{|\mathbf{y}}}} \mathcal{A}(u \otimes q)(\mathbf{y}, \mathbf{z}), \text{ for all } \mathbf{z} \in \Omega_{\xi_{|\mathbf{z}}} \tag{6.8}$$

For brevity, we shall drop from now on the superfluous indicator of the state space from the notation and write $\sum_{\mathbf{z}}$ and $\sum_{\mathbf{y}}$ instead of $\sum_{\mathbf{z} \in \Omega_{\xi_{|\mathbf{z}}}}$ and $\sum_{\mathbf{y} \in \Omega_{\xi_{|\mathbf{y}}}}$, respectively.

Expanding now the right-hand side of condition (6.7) by using (6.3) together with the product ansatz made in (6.6), and grouping the terms depending on $\mathbf{z}$, we have

$$\begin{aligned} \sum_{\mathbf{z}} \mathcal{A}(u \otimes q)(\mathbf{y}, \mathbf{z}) &= \sum_{j=1}^{M} \Big(\sum_{\mathbf{z}} \gamma_j(\mathbf{z} - \zeta^j) q(\mathbf{z} - \zeta^j) \Big) \beta_j(\mathbf{y} - \nu^j) u(\mathbf{y} - \nu^j) \\ &\quad - \sum_{j=1}^{M} \Big(\sum_{\mathbf{z}} \gamma_j(\mathbf{z}) q(\mathbf{z}) \Big) \beta_j(\mathbf{y}) u(\mathbf{y}). \end{aligned} \tag{6.9}$$

From the conditions (6.4) and (6.5), and introducing the notation $\tilde{\mathbf{z}} = \mathbf{z} - \zeta^j$, we obtain that

$$\sum_{\mathbf{z}} \gamma_j(\mathbf{z} - \zeta^j) q(\mathbf{z} - \zeta^j) = \sum_{\tilde{\mathbf{z}} + \zeta^j} \gamma_j(\tilde{\mathbf{z}}) q(\tilde{\mathbf{z}}) = \sum_{\mathbf{z}} \gamma_j(\mathbf{z}) q(\mathbf{z}). \tag{6.10}$$

Using now relation (6.10) in (6.9) and applying similar arguments also to the second equation (6.8), yields the desired system of coupled equations for the marginal distributions $u(\mathbf{y})$ and $q(\mathbf{z})$,

$$0 = \sum_{j=1}^{M} \Big(\sum_{\mathbf{z}} \gamma_j(\mathbf{z}) q(\mathbf{z}) \Big) \Big(\beta_j(\mathbf{y} - \nu^j) u(\mathbf{y} - \nu^j) \; - \; \beta_j(\mathbf{y}) u(\mathbf{y}) \Big) \tag{6.11}$$

$$0 = \sum_{j=1}^{M} \Big(\sum_{\mathbf{y}} \beta_j(\mathbf{y}) u(\mathbf{y}) \Big) \Big(\gamma_j(\mathbf{z} - \zeta^j) q(\mathbf{z} - \zeta^j) \; - \; \gamma_j(\mathbf{z}) q(\mathbf{z}) \Big). \tag{6.12}$$

Notice that the product approximation presented above has significantly reduced the *full* CME on the truncated space $\Omega_\xi = \Omega_{\xi_{|\mathbf{y}}} \times \Omega_{\xi_{|\mathbf{z}}}$ which has a total number of degrees of freedom given by $N = \prod_{i=1}^{d_1} (\xi_{|\mathbf{y}})_i \cdot \prod_{k=1}^{d_2} (\xi_{|\mathbf{z}})_k$, to the CME model (6.11)-(6.12) which has only $\tilde{N} = \prod_{i=1}^{d_1} (\xi_{|\mathbf{y}})_i + \prod_{k=1}^{d_2} (\xi_{|\mathbf{z}})_k$ degrees of freedom. Basically, the *linear* full CME has now been replaced with two CME-like equations, where the corresponding propensities are multiplied by factors that depend on the other marginal distribution, i.e., $\sum_{\mathbf{z}} \gamma_j(\mathbf{z}) q(\mathbf{z})$ for (6.11), and $\sum_{\mathbf{y}} \beta_j(\mathbf{y}) u(\mathbf{y})$ for (6.12). By *fixing* $u(\mathbf{y})$ and $q(\mathbf{z})$, we obtain two lower-dimensional stationary CMEs, which are obviously easier to solve than the full equation.

6.1.3. Hellander-Lötstedt hybrid model

The last step in the derivation of the Hellander-Lötstedt model is to replace the marginal $q(\mathbf{z})$ from the coupled system of equations (6.11)-(6.12) with an approximation, thus further reducing the number of degrees of freedom.

To this end, we make the assumption that we can replace $q(\mathbf{z})$ with the approximate expectation denoted by

$$\eta \approx \sum_{\mathbf{z}} \mathbf{z} q(\mathbf{z}). \tag{6.13}$$

Because the coupling term in (6.11) is actually $\sum_{\mathbf{z}} \gamma_j(\mathbf{z}) q(\mathbf{z})$, we introduce an approximation in terms of η,

$$\sum_{\mathbf{z}} \gamma_j(\mathbf{z}) q(\mathbf{z}) \approx \gamma_j \Big(\sum_{\mathbf{z}} \mathbf{z} q(\mathbf{z}) \Big) \approx \gamma_j(\eta), \tag{6.14}$$

i.e., the expectation of the propensity is approximated by the propensity of the expectation. For the second equation (6.12), we again use conditions (6.4) and (6.5) to obtain that

$$\sum_{\mathbf{z}} \mathbf{z} \Big(\gamma_j(\mathbf{z} - \zeta^j) q(\mathbf{z} - \zeta^j) \Big) = \sum_{\mathbf{z}} (\mathbf{z} + \zeta^j) \gamma_j(\mathbf{z}) q(\mathbf{z}). \tag{6.15}$$

Inserting (6.15) into (6.12), where we have first taken the sum over $\mathbf{z}$, yields

$$\begin{aligned} 0 &= \sum_{j=1}^{M} \Big(\sum_{\mathbf{y}} \beta_j(\mathbf{y}) u(\mathbf{y}) \Big) \sum_{\mathbf{z}} \mathbf{z} \Big(\gamma_j(\mathbf{z}-\zeta^j) q(\mathbf{z}-\zeta^j) \; - \; \gamma_j(\mathbf{z}) q(\mathbf{z}) \Big) \qquad (6.16) \\ &= \sum_{j=1}^{M} \Big(\sum_{\mathbf{y}} \beta_j(\mathbf{y}) u(\mathbf{y}) \Big) \zeta^j \gamma_j(\eta). \end{aligned}$$

Substituting (6.14) into (6.11) and using the result from (6.16), we arrive at the final form of the coupled system of equations which forms the Hellander-Lötstedt model, given as

$$0 = \sum_{j=1}^{M} \Big(\gamma_j(\eta) \Big) \Big(\beta_j(\mathbf{y}-\nu^j) u(\mathbf{y}-\nu^j) \; - \; \beta_j(\mathbf{y}) u(\mathbf{y}) \Big) \; := \tilde{\mathcal{A}}(\eta) u(\mathbf{y}) \qquad (6.17)$$

$$0 = \sum_{j=1}^{M} \Big(\sum_{\mathbf{y}} \beta_j(\mathbf{y}) u(\mathbf{y}) \Big) \zeta^j \gamma_j(\eta) \qquad := F(\eta, u(\mathbf{y})). \qquad (6.18)$$

The model presented above was first proposed for the case of the time-dependent CME in [HL07], using a different derivation. We also remark that similar models have also appeared in the context of other applications like molecular dynamics (see [Bor91] or [GJ08]). In the Hellander-Lötstedt approach, the solution consists of two components, on one hand $u(\mathbf{y})$ which solves a reduced *linear* CME where the operator $\tilde{\mathcal{A}}(\eta)$ has propensities that depend on the values of η, and on the other hand, the vector of approximate expectations for the non-critical species η, which is the solution of a *non-linear* equation that includes factors that depend on the marginal distribution of the critical species $u(\mathbf{y})$. Compared with the product based model reduction, the number of degrees of freedom present in this hybrid model is $\prod_{i=1}^{d_1} (\xi_{|\mathbf{y}})_i + d_2$. At this point, we note that the derivation of the hybrid model for the stationary CME which was presented above follows the arguments for the time-dependent case to be found in [Jah11], where a detailed analysis of the modeling error for the Hellander-Lötstedt was carried out and an extension to the model was introduced. We proceed now to give some algorithmic details about the embedding of the adaptive wavelet method within the hybrid approach (6.17)-(6.18).

6.2. Hybrid algorithm for stationary CME using wavelets

Solving the two subproblems that make up the hybrid model (6.17)-(6.18) is done alternately. For the stochastic part, the adaptive wavelet method for the stationary CME is used, with the mention that the propensities of the CME operator $\tilde{\mathcal{A}}(\eta)$ now depend on η, therefore the routine that evaluates $(\tilde{A}q)(\mathbf{y})$ for all $\mathbf{y} \in \Omega_{\xi_{|\mathbf{y}}}$ uses the values of η computed in the previous step. The deterministic section of the approximation is obtained by using an d_2-dimensional Newton method that uses the approximation of the marginal distribution $u(\mathbf{y})$ computed in a preceding step performed with the wavelet method. Now, we briefly discuss Newton's method for multi-dimensional problems [Ste08].

The task is to solve the non-linear subproblem from (6.18), i.e.

$$F(\eta, u(\mathbf{y})) = 0.$$

Let $F : \mathbb{R}^{d_2} \times \Omega_{\xi_{|\mathbf{y}}} \to \mathbb{R}^{d_2}$ be given by $F(\eta, u(\mathbf{y})) = (F_i(\eta, u(\mathbf{y})))_{i=1:d_2}$, where

$$F_i(\eta, u(\mathbf{y})) = \sum_{j=1}^{M} \Big(\sum_{\mathbf{y}} \beta_j(\mathbf{y}) u(\mathbf{y}) \Big) \zeta_i^j \gamma_j(\eta),$$

and $\eta = (\eta_i)_{i=1:d_2}$. Additionally, the marginal distribution $u(\mathbf{y})$ is known. We denote now the gradient of $F(\eta, u(\mathbf{y}))$ by $DF(\eta, u(\mathbf{y})) \in \mathbb{R}^{d_2 \times d_2}$,

$$DF(\eta, u(\mathbf{y})) = \begin{pmatrix} \frac{\partial F_1(\eta, u(\mathbf{y}))}{\partial \eta_1} & \cdots & \frac{\partial F_1(\eta, u(\mathbf{y}))}{\partial \eta_{d_2}} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_{d_2}(\eta, u(\mathbf{y}))}{\partial \eta_1} & \cdots & \frac{\partial F_{d_2}(\eta, u(\mathbf{y}))}{\partial \eta_{d_2}} \end{pmatrix}$$

and by standard arguments, obtain the following recursion for Newton's method for multi-dimensional problems,

$$\eta^{k+1} = \eta^k - (DF(\eta^k, u(\mathbf{y})))^{-1} F(\eta^k, u(\mathbf{y})), \forall k \geq 0. \tag{6.19}$$

At each step k of the Newton method, we need to compute the vector

$$v^k = (DF(\eta^k, u(\mathbf{y})))^{-1} F(\eta^k, u(\mathbf{y})) \in \mathbb{R}^{d_2}.$$

Naturally, we can avoid the need to explicitly compute the matrix $(DF(\eta^k, u(\mathbf{y})))^{-1}$, by solving instead the linear system

$$DF(\eta^k, u(\mathbf{y}))v^k = F(\eta^k, u(\mathbf{y})).$$

We remark that as the size of the matrix $DF(\eta^k, u(\mathbf{y})) \in \mathbb{R}^{d_2 \times d_2}$ is usually small, the linear system can be easily solved. The multi-dimensional Newton method is stopped and convergence to a solution of the problem $F(\eta, u(\mathbf{y})) = 0$ is declared when $||F(\eta^{new}, u(\mathbf{y}))||_2 \leq$ `tol`, with `tol` a prescribed tolerance, usually chosen as 10^{-10}. We summarize now the multi-dimensional Newton method used to approximate η, in the following Algorithm 8.

Algorithm 8: Multi-dimensional Newton Method

Parameter : tolerance `tol` for the largest admissible value of $||F(\eta^{new}, u(\mathbf{y}))||_2$

Input : initial guess $\eta^0 \in \mathbb{R}^{d_2}$
current approximation of marginal distribution $u(\mathbf{y})$
function $F(\eta, u(\mathbf{y}))$

Output : approximate solution $\eta^{new} \in \mathbb{R}^{d_2}$ of $F(\eta, u(\mathbf{y})) = 0$

$\eta^{new} = \eta^0$
while $||F(\eta^{new}, u(\mathbf{y}))||_2 >$ `tol` **do**
  $\eta^{old} = \eta^{new}$
  compute $DF(\eta^{old}, u(\mathbf{y}))$
  solve the linear system $DF(\eta^{old}, u(\mathbf{y}))v = F(\eta^{old}, u(\mathbf{y}))$
  update $\eta^{new} = \eta^{old} - v$
end

Combining the multi-dimensional Newton method with the adaptive wavelet solver for the stationary CME can now be easily accomplished by employing the two solvers alternately and the pseudocode for the hybrid method is listed in Algorithm 9.

Algorithm 9: Hybrid algorithm

Parameter : tolerance $\texttt{tol}_{stochastic}$ for wavelet solver
tolerance $\texttt{tol}_{Newton}$ for Newton method
Maximal number of basis elements to be used $\max_{basis}$

Input : vector $\eta^{(0)} \in \mathbb{R}^{d_2}$ for initial step of hybrid method
initial index subset $\{j_1, \ldots, j_\delta\}$ for adaptive wavelet method (AWM)

Output : approximate solutions η and $u(\mathbf{y})$ of hybrid model (6.17)-(6.18)

Set $k = 1$
while *size current basis* $< \max_{basis}$ **do**

Solve $\tilde{\mathcal{A}}(\eta^{(k-1)})u^{(k)} = 0$ with **AWM**$(\texttt{tol}_{stochastic}) \implies$ obtain $u^{(k)}$
Solve $F(\eta^{(k)}, u^{(k)}) = 0$ with **Newton**$(\texttt{tol}_{Newton}) \implies$ obtain $\eta^{(k)}$
Set $k = k + 1$

end

Consequently, employing Algorithm 9 leads to a sequence of approximations for the two subproblems

$$\eta^{(0)} \longrightarrow u^{(1)} \longrightarrow \eta^{(1)} \longrightarrow u^{(2)} \longrightarrow \ldots,$$

and we remark that, also it is possible to start with an approximation $u^{(0)}$ of the marginal distribution for the critical species and then compute $\eta^{(1)}$, we prefer to use as starting point an initial guess $\eta^{(0)} \in \mathbb{R}^{d_2}$ supplied by the user. For example, an easy way to obtain good values for $\eta^{(0)}$ would be to perform a few runs of the SSA algorithm on the full model. Furthermore, if the deterministic subproblem is linear with respect to η, then the Newton method will yield the exact result in only one step.

6.3. Numerical example: *lac* Operon

We illustrate now the potential of the hybrid wavelet method for the stationary CME by using a classic example of prokaryotic gene regulation - the *lac* operon. The source of this stochastic *lac* operon model, together with most of the parameter values, is [Wil06]. The biological relevance of this example is derived from the fact that operons, which are clusters of coregulated genes which can be turned on and off together, provide a mechanism for bacteria to quickly adapt to environments where nutrient availability may vary greatly [Ral08]. The *lac* operon can be found in *E.coli* and encompasses the genes required for lactose metabolism [JM61].

The stochastic model from [Wil06] contains 11 species and 16 reaction channels and is best explained with the help of the schematic in Figure 6.1. The *lac* operon contains three genes, denoted by *z*, *y* and *a*, that encode three enzymes *Z*, *Y* and *A*, which act together in

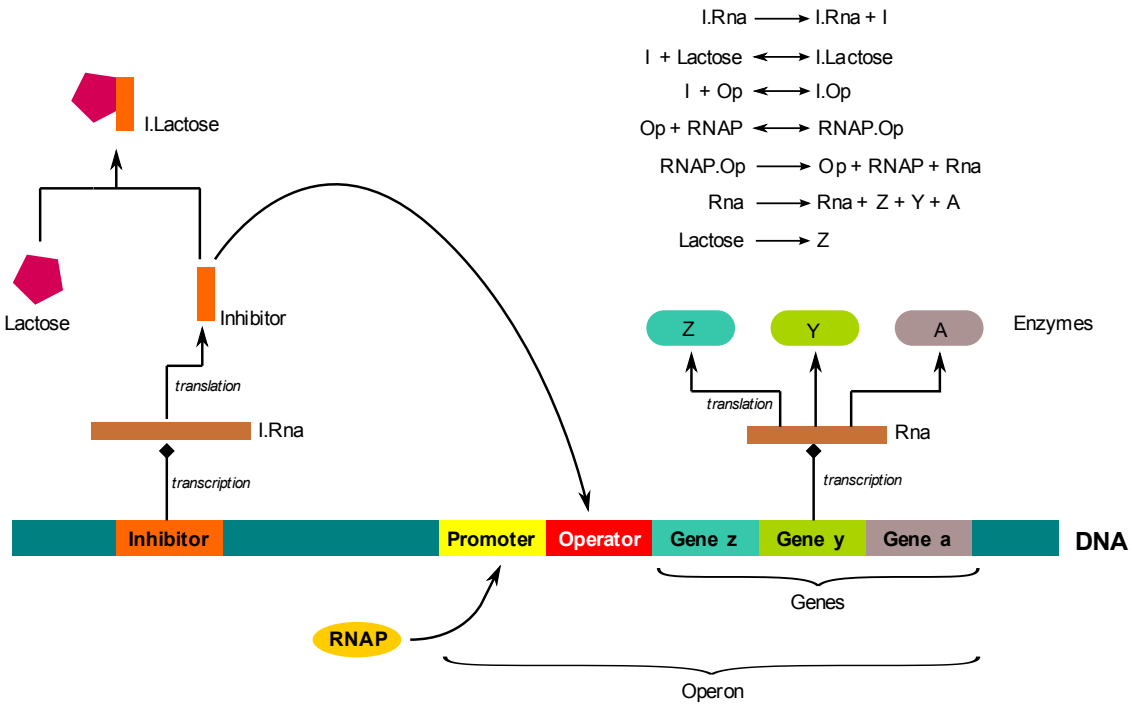


Figure 6.1.: *lac* operon structure and control mechanisms (Figure adapted from [Wil06]).

the process of lactose transport and conversion to glucose for use as food source within the *E.coli* cell. However, these enzymes are only required if lactose is present in abundance in the environment, so their transcription is regulated by the presence of lactose itself. The structure of the operon is such that the three genes are adjacent, and located downstream of a promoter that marks the binding site for the enzyme RNA polymerase. After binding, the RNAP transcribes all the genes into a single mRNA molecule. Upstream of the operon in the DNA strand, there is also a separate inhibitor gene which encodes a protein that represses transcription of the operon genes by binding to the DNA just downstream of the promoter site for RNAP binding. Under normal conditions, i.e., in the absence of lactose, the transcription of the genes is turned off, as the inhibitor binds downstream of the promoter site. However, when lactose is present in the environment, the inhibitor binds preferentially to the lactose molecules, and thus transcription is turned on (cf. [Wil06, Ral08]).

As previously stated, the original model contains 11 different species, involved in 16 reaction channels. We slightly modify the model by adding another reaction channel ($R17$) that constantly supplies Lactose, which is different from the original model from [Wil06] where Lactose is inserted in a single burst. Additionally, a reduction of three species is performed based on the observation that copy numbers respect the following algebraic relations

$$\begin{aligned} \text{I.dna} &= 1 \\ \text{Op} + \text{I.Op} + \text{Rnap.Op} &= K_1 = 1 \\ \text{Rnap} + \text{Rnap.Op} &= K_2 = 100. \end{aligned}$$

Consequently, the propensities can be modified, and the resulting model becomes $8D$,

with only six of species being treated stochastically and the remaining two handled deterministically. The reaction channels of the reduced model are listed below, with table 6.1 providing the correspondence between the original name and the new labeling, in addition to the choice of numerical treatment for each of the species.

R_1 :	$\star$	$\longrightarrow$	S_1	α_1	$=$	c_1	Inhibitor Transcription
R_2 :	S_1	$\longrightarrow$	$S_1 + S_2$	α_2	$=$	$c_2 x_1$	Inhibitor Translation
R_3 :	$S_2 + S_6$	$\longrightarrow$	S_7	α_3	$=$	$c_3 x_2 x_6$	Lactose Inhibitor Binding
R_4 :	S_7	$\longrightarrow$	$S_2 + S_6$	α_4	$=$	$c_4 x_7$	Lactose Inhibitor Dissociation
R_5 :	$S_2 + S_3$	$\longrightarrow$	$\star$	α_5	$=$	$c_5 x_2 x_3$	Inhibitor Binding
R_6 :	$\star$	$\longrightarrow$	$S_2 + S_3$	α_6	$=$	$c_6(1 - x_3)(1 - x_8)$	Inhibitor Dissociation
R_7 :	S_3	$\longrightarrow$	S_8	α_7	$=$	$c_7 x_3(K_2 - x_8)$	Rnap Binding
R_8 :	S_8	$\longrightarrow$	S_3	α_8	$=$	$c_8 x_8$	Rnap Dissociation
R_9 :	S_8	$\longrightarrow$	$S_3 + S_4$	α_9	$=$	$c_9 x_8$	Transcription
R_{10} :	S_4	$\longrightarrow$	$S_4 + S_5$	α_{10}	$=$	$c_{10} x_4$	Translation
R_{11} :	$S_6 + S_5$	$\longrightarrow$	S_5	α_{11}	$=$	$c_{11} x_6 x_5$	Conversion
R_{12} :	S_1	$\longrightarrow$	$\star$	α_{12}	$=$	$c_{12} x_1$	Inhibitor Rna degradation
R_{13} :	S_2	$\longrightarrow$	$\star$	α_{13}	$=$	$c_{13} x_2$	Inhibitor Degradation
R_{14} :	S_7	$\longrightarrow$	S_6	α_{14}	$=$	$c_{14} x_7$	Lactose Inhibitor Degradation
R_{15} :	S_4	$\longrightarrow$	$\star$	α_{15}	$=$	$c_{15} x_4$	Rna Degradation
R_{16} :	S_5	$\longrightarrow$	$\star$	α_{16}	$=$	$c_{16} x_5$	Z Degradation
R_{17} :	$\star$	$\longrightarrow$	S_6	α_{17}	$=$	c_{17}	Lactose Production

$c_1 = 0.02$, $c_2 = 0.1$, $c_3 = 0.005$, $c_4 = 0.1$, $c_5 = 1$,
$c_6 = 0.01$, $c_7 = 0.1$, $c_8 = 0.01$, $c_9 = 0.03$, $c_{10} = 0.1$,
$c_{11} = 1e-5$, $c_{12} = 0.01$, $c_{13} = c_{14} = 0.002$, $c_{15} = 0.01$, $c_{16} = 0.001$,
$c_{17} = 10$, $\mathbf{K_2} = 100$

Species	Name	Initial amount	Treatment	Description
	I.dna	1	reduced	inhibitor gene
S_1	I.Rna	0	stochastic	associated mRNA
S_2	I	50	stochastic	repressor protein
S_3	Op	1	stochastic	lac operon
	Rnap	100	reduced	RNAP complex
S_4	Rna	0	stochastic	mRNA
S_5	Z	0	stochastic	enzyme
S_6	Lactose	20	deterministic	lactose molecule
S_7	I.Lactose	0	deterministic	complex
	I.Op	0	reduced	complex
S_8	Rnap.Op	0	stochastic	complex

Table 6.1.: *lac* operon species

Even after performing the reductions and assigning S_6 and S_7 to the set of species that are handled deterministically, the state space for the reduced CME of the *lac* operon model has more than $2^3 \times 2^3 \times 2^1 \times 2^4 \times 2^9 \times 2^1 \approx 2 \cdot 10^6$ degrees of freedom. We remark that the choice of the species that are treated deterministically is made relying on

an analysis of a few stochastic simulations via Gillespie's SSA algorithm, which reveal a clear separation of scales.

A comparison between the marginal distributions obtained using the hybrid method and marginal distributions computed by averaging the data from 10^7 SSA runs on the interval $[0, 10^6]$, is presented in Figure 6.2, and shows that for this model, the dynamics for the critical species are well approximated by the hybrid method. The initial amounts used in the SSA algorithm are given in Table 6.1, and we note that the first 1% of the SSA data was discarded before computing the approximations of the marginal distributions.

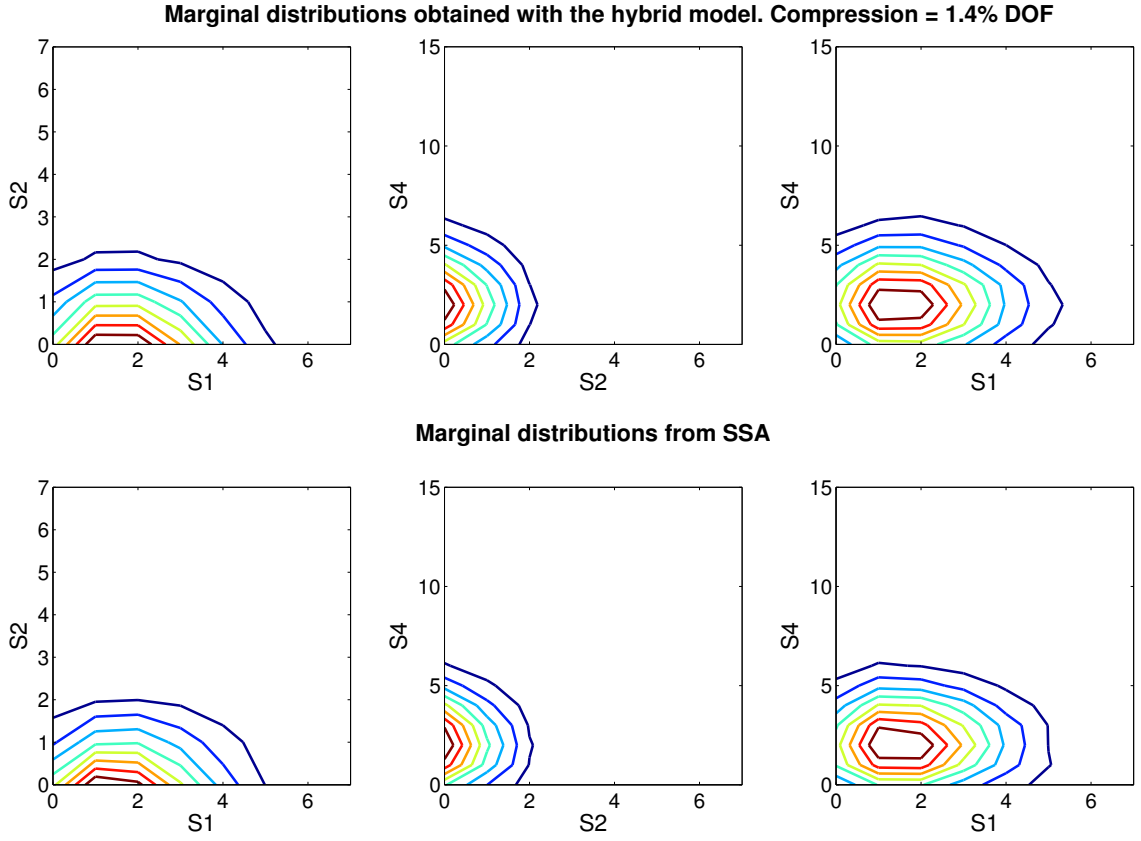


Figure 6.2.: Numerical results for the *lac* operon hybrid model. First row shows marginal distributions obtained with the hybrid wavelet method using a total of 1.4% of the total number of degrees of freedom, while the second row depicts marginal distributions obtained by averaging 10^7 SSA simulations.

The parameters used in the stochastic part of the hybrid solver were a multivariate anisotropic *Haar* basis, $\texttt{tol}_{stochastic} = 0.01$ and $\max_{basis} = 30000$. The choice of the Haar wavelet system was dictated by the small directions present in the reduced stochastic model. The solver reached the maximum number of DOFs allowed without reporting a 1-norm of the residual smaller than $\texttt{tol}_{stochastic}$, and in Figure 6.2 the results obtained with the hybrid wavelet method using 1.4% of the total number of degrees of freedom are shown. For the deterministic part, $\eta^{(0)} = \{4000, 100\}$ was chosen as initial approximation and $\texttt{tol}_{Newton}$ was set to $1e-10$. We remark however that as the deterministic subproblem is linear with respect to η, Newton's method yields the exact result after one single step.

6.4. Discussion about the modeling error of the hybrid model

As previously stated, the hybrid model can not be applied with the same success to every model. A detailed analysis of the modeling error of the Hellander-Lötstedt model can be found in [Jah11], and the results presented therein explain why for some classes of models where the solution exhibits a multi-modal profile, the hybrid model (6.17)-(6.18) no longer provides the correct qualitative description of the underlying dynamics. This assertion can be checked by applying the hybrid model to a small test case, namely the *toggle switch* model from (2.94) with a different set of initial parameters

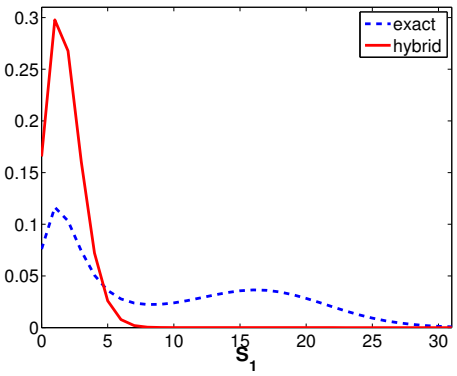
$$c_{11} = c_{21} = 10, \quad c_{12} = c_{22} = 30 \text{ and } c_3 = c_4 = 0.017. \tag{6.20}$$

The parameter set (6.20) leads to a 2D system with a state space of $\Omega_{32\times 32}$ where we treat one of the species stochastically and the other deterministically. As the system is symmetric, the precise assignment of the species to the two groups does not matter, and S_1 is placed in the stochastic group, with S_2 assigned to the deterministic set. Because of the small size of the system, we can eliminate the errors that might be caused by the wavelet discretization, and solve the stochastic part exactly. In (6.21) we summarize the partition of the system, namely the form of the propensity functions β_j and γ_j for each reaction.

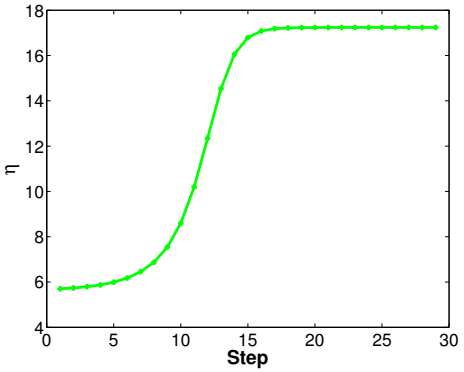
$$\begin{array}{llll|lll|lll} R_1: & \star & \longrightarrow & S_1 & \beta_1 & = & 1 & \gamma_1 & = & c_{11}/(c_{12}+\eta^2) \\ R_2: & \star & \longrightarrow & S_2 & \beta_2 & = & c_{21}/(c_{22}+x_1^2) & \gamma_2 & = & 1 \\ R_3: & S_1 & \longrightarrow & \star & \beta_3 & = & c_3 x_1 & \gamma_3 & = & 1 \\ R_4: & S_2 & \longrightarrow & \star & \beta_4 & = & 1 & \gamma_4 & = & c_4\eta \end{array} \tag{6.21}$$

The corresponding stoichiometric vectors for the two parts are $\nu = (1,-1)^T$ and $\zeta = (1,-1)^T$, respectively. We can now apply the hybrid method presented in Algorithm 9, and the results are presented in Figure 6.3.

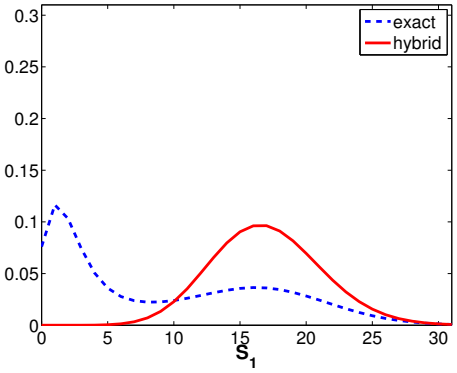
Two different runs of the hybrid algorithm were performed, one with the initial guess $\eta^{(0)} = 5.5$, while the other was launched with $\eta^{(0)} = 5.7$. As it can be seen in the left panels, the approximation for the marginal distribution u converges to one of the steady states, and no longer approximates the true dynamics that were obtained by computing the marginal from the exact CME solution of the full system. Consequently, the hybrid model is ill-suited for such problems. A solution would be to couple the adaptive wavelet method to the refined hybrid model proposed in [Jah11] based on conditional expectations, which proved capable of dealing with such bimodal solution profiles, and this represents one of the future aims in the development of the adaptive wavelet method.



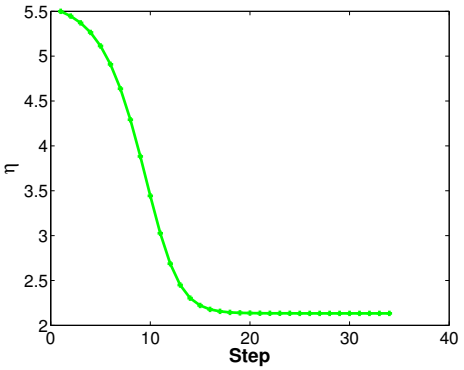
(a) Marginal pdf (S_1) when $\eta^{(0)} = 5.7$



(b) Values of $\eta^{(i)}$ for (S_2)



(c) Marginal pdf (S_1) when $\eta^{(0)} = 5.5$



(d) Values of $\eta^{(i)}$ for (S_2)

Figure 6.3.: Results of the application of the hybrid method to the toggle switch model (2.94) with parameter set (6.20). In the first row, the results obtained with an initial guess $\eta^{(0)} = 5.7$ are plotted (marginal distribution for S_1 obtained from full CME solution in blue, and marginal distribution from hybrid method in red), while in the second row, $\eta^{(0)} = 5.5$ was chosen.

CHAPTER

SEVEN

CONCLUSIONS AND OUTLOOK

In this thesis the adaptive wavelet method for the CME proposed in [Jah10], was further refined to include a better fourth-order time integrator instead of the second-order trapezoidal rule used initially, and far more importantly, the use of a larger spectrum of wavelet bases significantly improved the performance of the original method. Moreover, the method has now gained also adaptivity in time, alongside its spatial adaptivity, via an adaptive time-stepping strategy that markedly improved its computational efficiency. This is because in most cases the time evolution of a system starts with a fast (stiff) transient phase that requires small steps, before slowly converging to the stationary distribution, at which point larger step-sizes are possible. The biggest obstacle in endowing the method with adaptive time-step control, was to successfully balance the two errors caused by the spatial approximation using wavelet compression on one hand, and the error caused by time integration on the other. The results obtained in this respect have been disseminated in a journal article [JU10] and a conference proceedings [UJ09]. In conclusion, the adaptive wavelet method for the time-dependent CME problem has gained new features that make it a feaseable alternative to other approaches aimed at solving the CME directly.

Another direction that was investigated herein is related to obtaining approximations of the stationary CME. For this task the computational core of the adaptive wavelet method has been combined with the inverse power method with shift to produce a stationary solver that shows promising results when applied to non-trivial problems, as seen in Chapter 5. However, further improvements are necessary, particularly with respect to the selection of an initial set of active basis elements. Another open problem is the absence of a theoretical result concerning the convergence of the method. Although our numerical tests suggest that the method performs as designed, a rigorous proof is still not available, mostly because the CME operator is neither symmetric nor elliptic, such that the norm equivalences usually at hand for applications of adaptive wavelet methods for PDEs (e.g. [CDD01]) can not be used.

A further result that was discussed at length in Chapter 5 was the development of a numerical method for approximating the solution of the discrete committor problems.

Although improvements still need to be made before the method can be used in real-life applications, the technical difficulties related to the approximation of the committors using wavelets have largely been resolved. Steps have also been made in the direction of developing the necessary visualization protocols that are indispensable to the valorification of the information provided by the committors in high-dimensions.

By far the most promising research goal that remains to be investigated is the coupling of the adaptive wavelet method with the hybrid method recently proposed in [Jah11]. The embedding of the wavelet-based stationary CME solver in a hybrid strategy has been explored in Chapter 6, and the results showed the potential of the approach on one hand, but also highlighted the need for an enhanced hybrid model if the solver is to be succesfuly applied to model problems with metastable solution profiles.

Although not explicitly discussed in the thesis, work on parallelizing the software codes has begun. The computation of the Galerkin matrices which appear in all the algorithms discussed so far, is accomplished in parallel by splitting the workload between several processors. As the evaluation of the entries in the Galerkin matrix is the most computationally expensive part of the adaptive wavelet method, parallelization has brought important benefits with regard to running times. However, the potential for parallelization has not been exhausted, and future work in this direction could lead to substantial progress.

APPENDIX

A

PROPERTIES OF THE TRUNCATED CME

This appendix provides some details that were glossed over in (2.5). We begin by recalling that the truncated CME operator (2.67) is isomorphic to a large sparse matrix $A \in \mathbb{R}^{N\times N}$ with non-positive diagonal, non-negative off-diagonal entries and the property that the sum over each of its columns is zero. The following arguments are adapted from [And83, Section 12].

Theorem A.1 (Gerschgorin circle theorem). *Let $M \in \mathbb{R}^{N\times N}$ a complex matrix with entries m_{ij}. For $i \in \{1,\dots,N\}$ let*

$$R_i = \sum_{\substack{i=1\\ j\neq i}}^{N} |m_{ij}|$$

be the sum of the off-diagonal entries in column j. Further, let $D(m_{ii}, R_i)$ denote the Gerschgorin circle centered on m_{ii} with radius R_i. Then, every eigenvalue of M lies within at least one of the Gerschgorin circles $D(m_{ii}, R_i), i = \{1,\dots,N\}$.

Proof. Let λ be an eigenvalue of M and $v = (v_i)$ the corresponding eigenvector, such that

$$Mv = \lambda v. \tag{A.1}$$

Let $i \in \{1,\dots,N\}$ be chosen such that $|v_i| = \max_j |v_j|$, i.e., i is selected so that v_i is the largest entry in absolute value of v. Then $v_i > 0$ otherwise, $v = 0$. Next, we write the j-th line of the system (A.1)

$$\sum_{j=1}^{N} m_{ij}v_j = \lambda v_i.$$

By splitting the above sum

$$\sum_{j\neq i} m_{ij}v_j = (\lambda - m_{ii})v_i$$

we obtain that

$$|\lambda - m_{ii}| \leq \frac{1}{|v_i|}\sum_{j\neq i} |m_{ij}||v_j| \leq \sum_{j\neq i} |m_{ij}|$$

where the last part of the inequality is derived from the fact that $\frac{|v_j|}{|v_i|} \leq 1,\ \forall j \neq i$. □

Corollary A.2. *Let $A \in \mathbb{R}^{N\times N}$ be a matrix with elements satisfying (2.71) and (2.73), and denote by $\sigma(A)$ its spectrum. Then,*

1. $0 \in \sigma(A)$
2. *If $\lambda \neq 0$, $\lambda \in \sigma(A)$, we have $Re(\lambda) < 0$.*

Proof. The first assertion is trivial, as from (2.73), we have that $\mathbb{1}^T A = 0$ leading to $\mathbb{1}^T = (1, \dots, 1)$ being a left eigenvector of A with eigenvalue 0. The proof of the second point also starts from (2.73). By splitting the sum we have that

$$-a_{ii} = \sum_{j\neq i} a_{ij} = \sum_{j\neq i} |a_{ij}| = R_i.$$

Applying now (A.1) we get

$$R_i^2 \geq |\lambda - a_{ii}|^2 = |\lambda + R_i|^2 = (Re(\lambda) + R_i)^2 + Im(\lambda)^2.$$

Hence, $Re(\lambda) \leq 0$. Moreover, if $Re(\lambda) = 0$ we have that $Im(\lambda) = 0$, which concludes the proof. □

BIBLIOGRAPHY

[ADA09] S. S. Andrews, T. Dinh, and A. P. Arkin, *Stochastic Models of Biological Processes*, Encyclopedia of Complexity and System Science (Robert Meyers, ed.), vol. 9, Springer New-York, 2009, pp. 8730–8749.

[And83] D. H. Anderson, *Compartmental Modeling and Tracer Kinetics*, Springer, New York - Heidelberg - Berlin, 1983.

[ARM98] A. P. Arkin, J. Ross, and H. H. McAdams, *Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected Escherichia coli cells*, Genetics **149** (1998), 1633–1648.

[BHMS06] K. Burrage, M. Hegland, S. MacNamara, and R. B. Sidje, *A Krylov-based finite state projection algorithm for solving the chemical master equation arising in the discrete modelling of biological systems*, Markov Anniversary Meeting: An international conference to celebrate the 150th anniversary of the birth of A.A. Markov (A. N. Langville and W. J. Stewart, eds.), Boston Books, 2006, pp. 21 – 38.

[Bor90] F. A. Bornemann, *An adaptive multilevel approach to parabolic equations I. General theory and 1D implementation*, IMPACT of Computing in Science and Engineering **2** (1990), no. 4, 279 – 317.

[Bor91] _____, *An Adaptive Multilevel Approach to Parabolic Equations in Two space Dimensions*, Ph.D. thesis, Freie Universität Berlin, 1991.

[CDD01] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations - convergence rates*, Math. Comp. **70** (2001), no. 233, 27–75.

[CDD02] _____, *Adaptive wavelet methods II: beyond the elliptic case*, Found. Comput. Math. **2** (2002), no. 3, 203–245.

[CDF92] A. Cohen, I. Daubechies, and J.-C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure and Appl. Math. **45** (1992), no. 5, 485–560.

[CDP96] J. M. Carnicer, W. Dahmen, and J. M. Peña, *Local Decomposition of Refinable Spaces and Wavelets*, Appl. Comp. Harm. Anal. **3** (1996), 127–153.

[CDV93] A. Cohen, I. Daubechies, and P. Vial, *Wavelets on the Interval and Fast Wavelet Transforms*, Applied and Computational Harmonic Analysis **1** (1993), no. 1, 54 – 81.

[CGP05] Y. Cao, D. T. Gillespie, and L. R. Petzold, *The slow-scale stochastic simulation algorithm*, J. Chem. Phys. **122** (2005), no. 1, 014116.

[CGP07] ———, *The adaptive explicit-implicit tau-leaping method with automatic tau selection*, J. Chem. Phys. **126** (2007), no. 22, 224101.

[CM65] D.R. Cox and H.D. Miller, *The theory of stochastic processes*, Wiley publications in statistics, Wiley, 1965.

[CM97] A. Cohen and R. Masson, *Wavelet methods for second order elliptic problems, preconditioning and adaptivity*, SIAM J. Sci. Comp **21** (1997), 1006–1026.

[Coh03] A. Cohen, *Numerical analysis of wavelet methods*, Studies in Mathematics and Its Applications, vol. 32, Elsevier, 2003.

[Dah97] W. Dahmen, *Wavelet and multiscale methods for operator equations*, Acta Numerica **6** (1997), 55–228.

[Dah01] ———, *Wavelet methods for PDEs - some recent developments*, J. Comput. Appl. Math. **128** (2001), 133–185.

[Dau92] I. Daubechies, *Ten lectures on wavelets*, vol. 61, SIAM, Philadelphia, 1992.

[DHJW08] P. Deuflhard, W. Huisinga, T. Jahnke, and M. Wulkow, *Adaptive discrete Galerkin methods applied to the chemical master equation*, SIAM J. Sci. Comput. **30** (2008), no. 6, 2990–3011.

[Dij09] T. J. Dijkema, *Adaptive tensor product wavelet methods for solving PDEs*, Ph.D. thesis, Universitaet Utrecht, 2009.

[DKU97] W. Dahmen, A. Kunoth, and K. Urban, *Biorthogonal Spline-Wavelets on the Interval - Stability and Moment Conditions*, Appl. Comp. Harm. Anal. **6** (1997), 132–196.

[EBE01] J. Elf, O. G. Berg, and M. Ehrenberg, *Comparison of repressor and transcriptional attenuator systems for control of amino acid biosynthetic operons*, J. Mol. Biol. **313** (2001), 941–954.

[EL00] M. B. Elowitz and S. Leibler, *A synthetic oscillatory network of transcriptional regulators*, Nature **403** (2000), no. 6767, 335–338.

[EN06] K.-J. Engel and R. Nagel, *A short course on operator semigroups*, Springer, New-York, 2006.

[Eng06] S. Engblom, *Numerical methods for the chemical master equation*, Licentiate thesis, Department of Information Technology, Uppsala University, September 2006.

[Eng08] ———, *Numerical Solution Methods in Stochastic Chemical Kinetics*, Ph.D. thesis, Upsalla University, 2008.

[Eng09a] ———, *Galerkin spectral method applied to the chemical master equation*, Commun. Comput. Phys. **5** (2009), no. 5, 871–896.

[Eng09b] ———, *Spectral approximation of solutions to the chemical master equation*, J. Comput. Appl. Math. **229** (2009), no. 1, 208–221.

[ESSL02] M. B. Elowitz, E. D. Siggia, P. S. Swain, and A. J. Levine, *Stochastic gene expression in a single cell*, Science **297** (2002), 1183–1186.

[FL07] L. Ferm and P. Lötstedt, *Numerical method for coupling the macro and meso scales in stochastic chemical kinetics*, BIT Numerical Mathematics **47** (2007), 735–762.

[FL09] ———, *Adaptive solution of the master equation in low dimensions*, Appl. Numer. Math. **59** (2009), no. 1, 187–204.

[FLH08] L. Ferm, P. Lötstedt, and A. Hellander, *A hierarchy of approximations of the master equation scaled by a size parameter*, SIAM J. Sci. Comput. **34** (2008), 127–151.

[Gar09] C. W. Gardiner, *Handbook of stochastic methods*, 4th revised and augmented ed., Springer Series in Synergetics, Springer, Berlin, 2009.

[GB00] M. A. Gibson and J. Bruck, *Efficient exact stochastic simulation of chemical systems with many species and many channels*, J. Phys. Chem. A **104** (2000), no. 9, 1876–1889.

[GCC00] T. S. Gardner, C. R. Cantor, and J. J. Collins, *Construction of a genetic toggle switch in Escherichia coli*, Nature **403** (2000), no. 6767, 339–342.

[Gil76] D. T. Gillespie, *A general method for numerically simulating the stochastic time evolution of coupled chemical reactions*, J. Comput. Phys. **22** (1976), 403–434.

[Gil92] ———, *A rigorous derivation of the chemical master equation*, Physica A **188** (1992), 404–425.

[Gil96] ———, *The multivariate Langevin and Fokker–Planck equations*, American Journal of Physics **64** (1996), no. 10, 1246–1257.

[Gil00] ———, *The chemical Langevin equation*, J. Chem. Phys. **113** (2000), 297–306.

[Gil01] ———, *Approximate accelerated stochastic simulation of chemically reacting systems*, J. Chem. Phys. **115** (2001), no. 4, 1716–1733.

[Gil07] ———, *Stochastic simulation of chemical kinetics*, Annual Review of Physical Chemistry **58** (2007), no. 1, 35–55.

[GJ08] M. Griebel and L. Jager, *The BGY3dM model for the approximation of solvent densities*, The Journal of Chemical Physics **129** (2008), no. 17, 511–525.

[Gou05] J. Goutsias, *Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems*, The Journal of Chemical Physics **122** (2005), no. 18, 184102.

[GRdO+11] E. Giampieri, D. Remondini, L. de Oliveira, G. Castellani, and P. Lio, *Stochastic analysis of a miRNA-protein toggle switch*, Mol. BioSyst. **7** (2011), 2796–2803.

[GVL96] G. H. Golub and C. F. Van Loan, *Matrix computations*, third ed., Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, 1996.

[Haa10] A. Haar, *Zur Theorie der orthogonalen Funktionensysteme*, Mathematische Annalen **69** (1910), no. 3, 331–371.

[HBS+07] M. Hegland, C. Burden, L. Santoso, S. MacNamara, and H. Booth, *A solver for the stochastic master equation applied to gene regulatory networks*, J. Comput. Appl. Math. **205** (2007), 708–724.

[Het00] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Review **42** (2000), no. 4, 599–653.

[HHL08] M. Hegland, A. Hellander, and P. Lötstedt, *Sparse grids and hybrid methods for the chemical master equation*, BIT **48** (2008), 265–284.

[HHL11] A. Hellander, S. Hellander, and P. Lötstedt, *Coupled mesoscopic and microscopic simulation of stochastic reaction-diffusion processes in mixed dimensions*, Tech. Report 2011-005, Department of Information Technology, Uppsala University, 2011.

[Hig08] D. J. Higham, *Modeling and simulating chemical reactions*, SIAM Rev. **50** (2008), no. 2, 347–368.

[Hig11] ———, *Stochastic ordinary differential equations in applied and computational mathematics*, IMA Journal of Applied Mathematics **76** (2011), no. 3, 449–474.

[HL07] A. Hellander and P. Lötstedt, *Hybrid method for the chemical master equation*, J. Comput. Phys. **227** (2007), 100–122.

[HW96] E. Hairer and G. Wanner, *Solving ordinary differential equations II: Stiff and differential-algebraic problems*, 2nd rev. ed., vol. 14, Springer, Berlin, 1996.

[Jah10] T. Jahnke, *An adaptive wavelet method for the chemical master equation*, SIAM J. Sci. Comput. **31** (2010), no. 6, 4373–4394.

[Jah11] ———, *On reduced models for the chemical master equation*, Multiscale Modeling & Simulation **9** (2011), no. 4, 1646–1676.

[JH07] T. Jahnke and W. Huisinga, *Solving the chemical master equation for monomolecular reaction systems analytically*, J. Math. Biol. **54** (2007), no. 1, 1–26.

[JH08] ———, *A dynamical low-rank approach to the chemical master equation*, Bull. Math. Biol. **70** (2008), no. 8, 2283–2302.

[JM61] F. Jacob and J. Monod, *On the regulation of gene activity*, Cold Spring Harbor Symposia on Quantitative Biology **26** (1961), 193–211.

[JU10] T. Jahnke and T. Udrescu, *Solving chemical master equations by adaptive wavelet compression*, Journal of Computational Physics **229** (2010), no. 16, 5724–5741.

[KEBC05] M. Kærn, T. C. Elston, W. J. Blake, and J. J. Collins, *Stochasticity in gene expression: from theories to phenotypes*, Nature Reviews Genetics **6** (2005), no. 6, 451 – 464.

[KP92] P. E. Kloeden and E. Platen, *Numerical solution of stochastic differential equations*, Springer-Verlag, Berlin, 1992.

[Kur72] T. G. Kurtz, *The relationship between stochastic and deterministic models of chemical reactions*, J. Chem. Phys. (1972), no. 57, 2976–2978.

[LMR98] A. K. Louis, P. Maass, and A. Rieder, *Wavelets: Theorie und Anwendungen*, Teubner Studienbücher: Mathematik, Teubner, 1998.

[MA99] H. H. McAdams and A. P. Arkin, *It's a noisy business! Genetic regulation at the nanomolar scale*, Trends in Genetics **15** (1999), no. 2, 65 – 69.

[Mal09] S. Mallat, *A wavelet tour of signal processing*, 3rd ed., Elsevier, Amsterdam, 2009.

[MBBS08] S. MacNamara, A. M. Bersani, K. Burrage, and R. B. Sidje, *Stochastic chemical kinetics and the total quasi-steady-state assumption: Application to the stochastic simulation algorithm and chemical master equation*, The Journal of Chemical Physics **129** (2008), no. 9, 095105.

[MBS08] S. MacNamara, K. Burrage, and R. B. Sidje, *Multiscale modeling of chemical kinetics via the master equation*, Multiscale Model. Simul. **6** (2008), no. 4, 1146–1168.

[Met08] P. Metzner, *Transition Path Theory for Markov Processes*, Ph.D. thesis, Freie Universität Berlin, 2008.

[MK06] B. Munsky and M. Khammash, *The finite state projection algorithm for the solution of the chemical master equation*, J. Chem. Phys. **124** (2006), no. 4, 044104.

[MSVE08] P. Metzner, C. Schütte, and E. Vanden-Eijnden, *Transition Path Theory for Markov Jump Processes*, Multiscale Model. Simul. **7** (2008), no. 3, 1192–1219.

[Nor97] J. R. Norris, *Markov chains*, Cambridge University Press, 1997.

[PHSN11] J.-H. Prinz, M. Held, J. C. Smith, and F. Noe, *Efficient computation, sensitivity, and error analysis of committor probabilities for complex dynamical processes*, Multiscale Modeling & Simulation **9** (2011), no. 2, 545–567.

[Pri06] M. Primbs, *Stabile biorthogonale Spline-Waveletbasen auf dem Intervall*, Ph.D. thesis, Universität Duisburg-Essen, 2006.

[Pri09] _____, *The method of stable completion in the classical wavelet context*, Results in Mathematics **53** (2009), no. 3-4, 391–398.

[PS08] G. A. Pavliotis and A. M. Stuart, *Multiscale methods: Averaging and homogenization*, Springer, 2008.

[Pta04] M. Ptashne, *A genetic switch: Phage lambda revisited*, Cold Spring Harbor Laboratory Press, 2004.

[Ral08] A. Ralston, *Operons and prokaryotic gene regulation*, Nature Education, vol. 1(1), 2008.

[RO04] J. M. Raser and E. K. O'Shea, *Control of stochasticity in eukaryotic gene expression*, Science **304** (2004), 1811–1814.

[Saa92] Y. Saad, *Numerical methods for large eigenvalue problems*, Algorithms and architectures for advanced scientific computing, Manchester University Press, 1992.

[Sen81] E. Seneta, *Non-negative matrices and markov chains*, Springer Verlag, New York - Heidelberg - Berlin, 1981.

[SLE09] P. Sjöberg, P. Lötstedt, and J. Elf, *Fokker-Planck approximation of the master equation in molecular biology*, Comput. Visual. Sci. **12** (2009), no. 1, 37–50.

[Ste08] D. Stefanica, *A primer for the mathematics of financial engineering*, Financial engineering advanced background series, FE Press, 2008.

[Ste09] W.J. Stewart, *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling*, Princeton University Press, 2009.

[Swe98] W. Sweldens, *The lifting scheme: A construction of second generation wavelets*, SIAM J. Math. Anal. **29** (1998), 511–546.

[TSB04] T. E. Turner, S. Schnell, and K. Burrage, *Stochastic approaches for modelling in vivo reactions*, Comput. Biol. Chem. **28** (2004), 165–178.

[UJ09] T. Udrescu and T. Jahnke, *An adaptive method for solving chemical master equations using a sparse wavelet basis*, AIP Conference Proceedings **1168** (2009), no. 1, 489–492.

[VE06] E. Vanden-Eijnden, *Transition path theory*, Computer Simulations in Condensed Matter: From Materials to Chemical Biology **2** (2006), no. 3, 439–478.

[vK01] N. G. van Kampen, *Stochastic processes in physics and chemistry*, 3rd ed., North-Holland Personal Library, Amsterdam: North-Holland, 2001.

[VKBL02] J. M. G. Vilar, H. Y. Kueh, N. Barkai, and S. Leibler, *Mechanisms of noise-resistance in genetic oscillators*, Proceedings of the National Academy of Sciences **99** (2002), no. 9, 5988–5992.

[vZtW05] J. S. van Zon and P. R. ten Wolde, *Simulating biochemical networks at the particle level and in time and space: Green's Function Reaction Dynamics*, Phys. Rev. Lett. **94** (2005), 8103.

[Wil06] D. J. Wilkinson, *Stochastic Modelling for Systems Biology*, Chapman and Hall/CRC, Mathematical & Computational Biology, 2006.

[Wil09] _____, *Stochastic modelling for quantitative description of heterogeneous biological systems*, Nature Reviews Genetics **10** (2009), no. 2, 122–133.