

Numerical Methods for the Chemical Master Equation

STEFAN ENGBLOM

UPPSALA UNIVERSITY
Department of Information Technology





UPPSALA
UNIVERSITET

**Numerical Methods for the Chemical Master
Equation**

BY
STEFAN ENGBLOM

August 2006

DIVISION OF SCIENTIFIC COMPUTING
DEPARTMENT OF INFORMATION TECHNOLOGY
UPPSALA UNIVERSITY
UPPSALA
SWEDEN

Dissertation for the degree of Licentiate of Technology in Scientific Computing
at Uppsala University 2006

Numerical Methods for the Chemical Master Equation

Stefan Engblom

Stefan.Engblom@it.uu.se

*Division of Scientific Computing
Department of Information Technology
Uppsala University
Box 337
SE-751 05 Uppsala
Sweden*

<http://www.it.uu.se/>

© Stefan Engblom 2006

ISSN 1404-5117

Printed by the Department of Information Technology, Uppsala University, Sweden

Abstract

The numerical solution of chemical reactions described at the meso-scale is the topic of this thesis. This description, the *master equation of chemical reactions*, is an accurate model of reactions where stochastic effects are crucial for explaining certain effects observed in real life. In particular, this general equation is needed when studying processes inside living cells where other macro-scale models fail to reproduce the actual behavior of the system considered.

The main contribution of the thesis is the numerical investigation of two different methods for obtaining numerical solutions of the master equation.

The first method produces statistical quantities of the solution and is a generalization of a frequently used macro-scale description. It is shown that the method is efficient while still being able to preserve stochastic effects.

By contrast, the other method obtains the full solution of the master equation and gains efficiency by an accurate representation of the state space.

The thesis contains necessary background material as well as directions for intended future research. An important conclusion of the thesis is that, depending on the setup of the problem, methods of highly different character are needed.

List of Papers

- A S. Engblom. Computing the moments of high dimensional solutions of the master equation. Technical Report 2005-020, Dept of Information Technology, Uppsala University, Uppsala, Sweden, 2005. Available at <http://www.it.uu.se/research>. To appear in *Appl. Math. Comput.*
- B S. Engblom. Gaussian quadratures with respect to discrete measures. Technical Report 2006-007, Dept of Information Technology, Uppsala University, Uppsala, Sweden, 2006. Available at <http://www.it.uu.se/research>.
- C S. Engblom. A discrete spectral method for the chemical master equation. Technical Report 2006-036, Dept of Information Technology, Uppsala University, Uppsala, Sweden, 2006. Available at <http://www.it.uu.se/research>.

Acknowledgments

I would like to thank my supervisor Per Lötstedt for time, patience and for many encouraging discussions. Various inputs by Paul Sjöberg, Måns Ehrenberg and Johan Elf have also been valuable. Financial support has been obtained from the Swedish National Graduate School in Mathematics and Computing.

On the personal side I would like to acknowledge the support and comfort provided for me by my family and especially that given by my wife Märta Cullhed.

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 2 | Chemical reactions at the meso-scale | 2 |
| 2.1 | Preys and predators: a motivating example | 2 |
| 2.2 | The Markov property and the master equation | 8 |
| 2.3 | Mathematical properties of the master operator | 11 |
| 2.4 | Model problems | 12 |
| 3 | Summary of papers | 16 |
| 3.1 | Paper A | 16 |
| 3.2 | Paper B | 17 |
| 3.3 | Paper C | 18 |
| 4 | Future work | 18 |
| 4.1 | Coupling of the macro-meso scales | 19 |
| 4.2 | Strong parameter estimation | 19 |
| 5 | Conclusions | 21 |

1 Introduction

Randomness enters in most descriptions of physical systems in some way. Under certain assumptions and in many, but certainly not in all cases, the effects of randomness can be captured accurately by *deterministic* models. Perhaps the most well-known such example is the diffusion equation which results from statistical considerations of random Brownian motion.

This thesis is concerned with the numerical solution of models of reality for which the effects of randomness cannot easily be taken into account in a deterministic way. To be specific, the dynamics of chemical reactions can be shown to satisfy a deterministic description provided that (i) the number of reacting molecules is sufficiently large to allow a statistical point of view and (ii) the systems state does not approach any critical points in phase-space. If these conditions are not met a more complete *stochastic* description of the system is necessary in order to obtain a realistic model.

The *master equation* is such a stochastic model and is derived from a basic and yet fundamental assumption on the dynamic properties of the underlying stochastic process — this is the Markov property.

If a chemical system of D reacting species is described by counting the number of molecules of each kind, then the master equation is a differential-difference equation in D dimensions governing the dynamics of the probability distribution for the system. The description suffers from the well-known "curse of dimensionality"; — each species adds one dimension to the problem leading in many cases to a prohibitive computational complexity. Effective numerical methods for solving the master equation are of interest both in research and in practice as such solutions comes with a more accurate understanding of many interesting chemical processes.

The material in the thesis is organized as follows: in Section 2 we begin by looking closer at the master equation and its fundamentals. We give a motivation on the popular level for this type of descriptions and then highlights some mathematical prerequisites and implications. In Section 3 the contributed papers are summarized and possible directions for future research are indicated in Section 4. The conclusions of the thesis along with a short summary are finally found in Section 5.

2 Chemical reactions at the meso-scale

This section contains an overview of the physical background of the mesoscopic description of chemical reactions along with some mathematical considerations.

In Section 2.1 a very intuitive and yet interesting model of preys and predators is discussed. It motivates the need for randomness in certain descriptions of real-life systems. Despite the models popular setting in a population of animals, the found properties remain valid in similar descriptions of biochemical reactions found inside living cells.

In Section 2.2 some basic properties of stochastic processes are mentioned and a brief treatment of the critical Markov property is given. This material is then used to derive the master equation which is the focus of the present thesis. Further mathematical properties of this equation are discussed in Section 2.3 where the point of view is the Numerical Analyst's rather than the Physicist's.

Finally, Section 2.4 reviews a collection of model problems targeted in the contributing papers. The relevance of each model is discussed along with certain mathematical properties. Taken together, this gallery of models form the core of problems at which the contributed numerical methods aim.

2.1 Preys and predators: a motivating example

As a motivation for considering stochastic descriptions of physical phenomena we look at the behavior of an intuitive model involving preys and predators. The dynamics of the population in the fictitious world depicted in Figure 2.1 follows four simple rules. For each discrete time-step (i) the population increases by random immigration at a fixed probability, (ii) an animal dies of natural causes by a fixed probability, (iii) each prey eats and immediately reproduces itself and (iv) each predator that finds a prey eats it and immediately reproduces itself. Additionally, the animals move around in relatively high speed using periodicity at the boundaries.

It is a straightforward task to implement the above rules in a computer and simulate the dynamics of the population. Such a result is shown in Figure 2.2 and allows for a very intuitive explanation: once the number of preys reaches a certain level, the number of predators increases rapidly as a result of the increasing amount of food. This process evidently continues until the preys almost reach extinction. As a

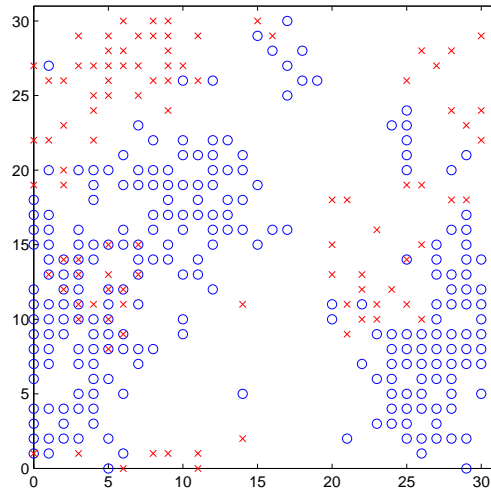


Figure 2.1: A 2-dimensional world with preys (circles) and predators (crosses) obeying simple rules.

result of starvation, the number of predators then rapidly decreases until the prey population has recovered. A new cycle of the system thus forms and the non-vanishing probability of immigration ensures that the system continues indefinitely.

How can we obtain qualitatively the result of this system without performing the expensive simulation of the whole world? The simplest idea is to form an ODE governing the population of the species. Let $x(t)$ and $y(t)$ denote the number of preys and predators, respectively. Then the above rules translates nicely into an ODE:

$$\left. \begin{aligned} \dot{x} &= k_0 - \mu x + k_1 x - k_2 xy \\ \dot{y} &= k_0 - \mu y + k_2 xy \end{aligned} \right\}, \quad (2.1)$$

where (k_0, μ) is the immigration/death-rate, k_1 the probability of preys finding food and where k_2 controls the probability of predators finding preys.

In Figure 2.3 data from a simulation using this model is shown. The parameters have been chosen so as to exactly match those used in the earlier simulation of the world. Clearly, the behavior of the ODE-model is quite different — apparently, stochasticity enters in more ways than merely as an "average".

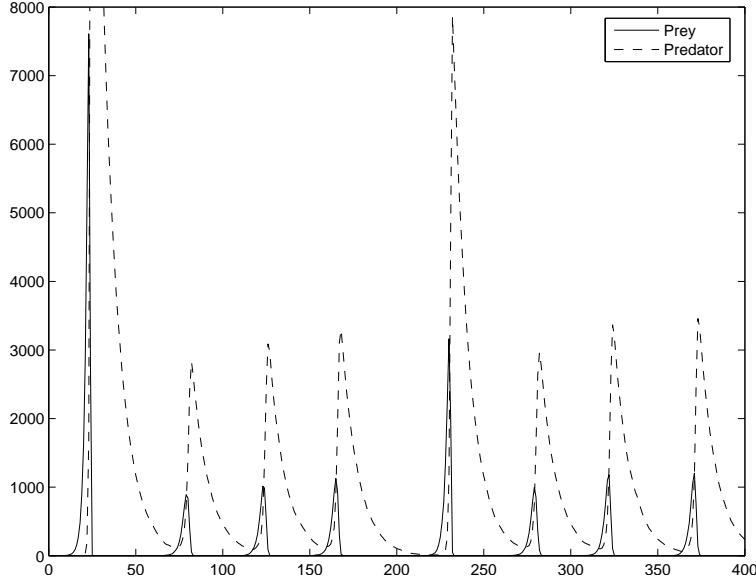
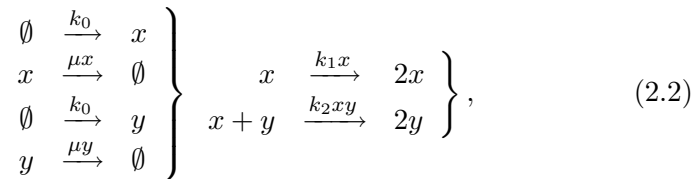


Figure 2.2: Number of individuals as a function of time from a direct simulation of the world in Figure 2.1.

As the *macroscopic* description (2.1) does not accurately captures the rules formulated at the *microscopic* level, one wonders whether a better formalism exists. This model must take stochasticity into account in some manageable way, yet being computationally tractable.

A partial answer is provided by the master equation where the rules are translated into the formalism of chemical reactions:



where all parameters have the same meaning as before. Under a certain model for the stochasticity of the reactions, the master equation to be derived in the next section *exactly* governs the time-dependent probability for the number of individuals of each species in the system (2.2):

$$\begin{aligned}
 \frac{\partial p(x, y, t)}{\partial t} = & -k_0 \nabla_x p + \mu \Delta_x [xp] - k_0 \nabla_y p + \mu \Delta_y [yp] \\
 & - k_1 \nabla_x [xp] - k_2 (\Delta_x \nabla_y - \Delta_x + \nabla_y) [xyp], \quad (2.3)
 \end{aligned}$$

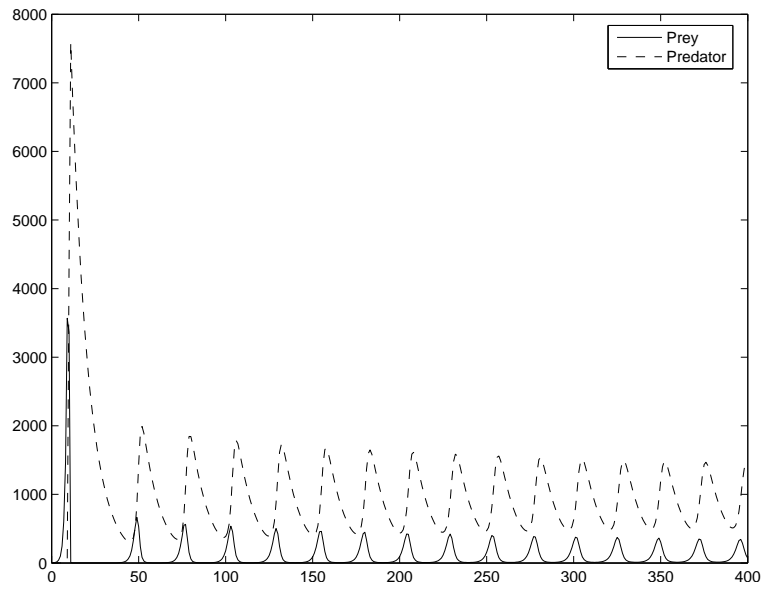


Figure 2.3: Solution of the prey-predator macroscopic model (2.1). The model behaves quite differently compared to the experiment in Figure 2.2. The continuous model is smoother with less sharp features and the oscillation seems to be of higher frequency.

where $\nabla q = q(x) - q(x - 1)$ and $\Delta q = q(x + 1) - q(x)$ and where the subscript indicates the target coordinate.

There are stochastic simulation techniques that allows the simulation of sample trajectories of the master equation. One such technique is *Gillespie's SSA* method [6] and the result of such a simulation is shown in Figure 2.4. Clearly, the *mesoscopic* model in the form of the master equation does a much better job in capturing the actual behavior of the prey-predator system.

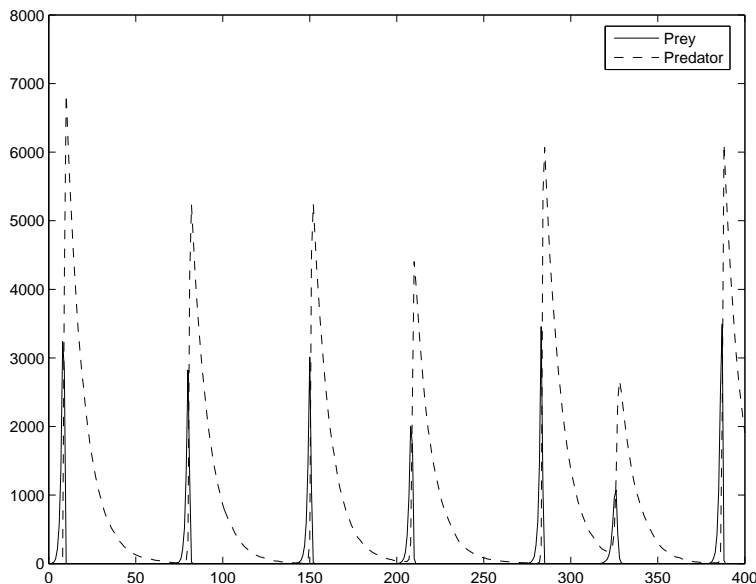


Figure 2.4: Stochastic solution obtained from simulating the prey-predator mesoscopic model (2.2). The solution is very remindful of that obtained from the microscopic model.

It is possible to proceed one step further and formulate problems for which the mesoscopic description in terms of a master equation is inaccurate. In the world of preys and predators, for example, one can easily devise rules for which the population strongly depends on the position by lowering the speed with which the animals move around. The simulation in Figure 2.5 uses the same parameters as before but now the average speed of each animal is about 8 times smaller. This makes the world inhomogenously populated and the impact of randomness more pronounced and difficult to characterize.

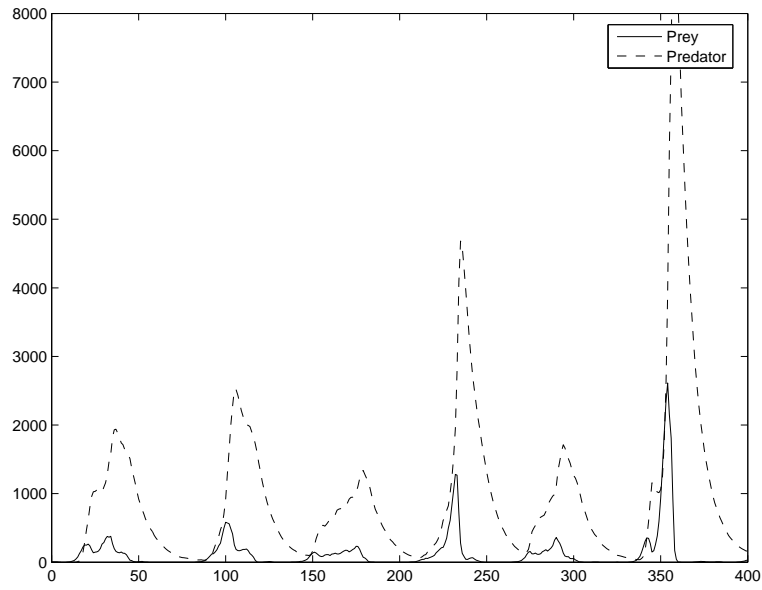


Figure 2.5: Same as in Figure 2.2 but now the animals move more slowly. This introduces a non-trivial spatial dependence on the solution and the population can no longer be regarded as well-stirred.

2.2 The Markov property and the master equation

A *stochastic process* [4, 7] is, loosely speaking, a time-dependent random variable $X(t)$. A certain *realization* of the process x_1, x_2, \dots and so on can be measured at times $t = t_1 \leq t_2 \leq \dots$ and we assume that the system can be described by a set of joint probabilities:

$$\Pr [x_n, t_n; x_{n-1}, t_{n-1}; \dots; x_1, t_1]. \quad (2.4)$$

In this thesis we exclusively treat *discrete* processes for which $X(t)$ takes values in the D -dimensional lattice space $\mathbf{Z}_+^D = \{0, 1, 2, \dots\}^D$, but generalizations to other types of processes are common [4].

The celebrated *Markov property* of stochastic processes plays a crucial role in many fields of physics and mathematics. It states the rather drastic simplification that the conditional probability for the event (x_n, t_n) given the systems full history satisfies

$$\Pr [x_n, t_n | x_{n-1}, t_{n-1}; \dots; x_1, t_1] = \Pr [x_n, t_n | x_{n-1}, t_{n-1}], \quad (2.5)$$

i.e. that the dependence on past events can be captured by the dependence on the previous state (x_{n-1}, t_{n-1}) only — the process has no memory.

The Markov property (2.5) cannot always be taken for granted but frequently remains a very accurate approximation. The reason for this is that the discrete time-steps used for actual measurements of the process are usually order of magnitudes larger than the often very short auto-correlation time of the system.

The assumption (2.5) is also a very powerful one since Markovian systems can be described using only the initial probability $\Pr[x_1, t_1]$ and the *transition probability function* $\Pr[x_s, s | x_t, t]$ [4, 7]:

$$\begin{aligned} \Pr [x_n, t_n; x_{n-1}, t_{n-1}; \dots; x_1, t_1] &= \\ \Pr [x_n, t_n | x_{n-1}, t_{n-1}] \cdots \Pr [x_2, t_2 | x_1, t_1] \cdot \Pr [x_1, t_1]. \end{aligned} \quad (2.6)$$

Another important consequence of the Markov assumption can be derived as follows: for an arbitrary discrete stochastic process the conditional probability always satisfies

$$\begin{aligned} \Pr [x, t_3 | z, t_1] &= \sum_y \Pr [x, t_3; y, t_2 | z, t_1] \\ &= \sum_y \Pr [x, t_3 | y, t_2; z, t_1] \Pr [y, t_2 | z, t_1]. \end{aligned} \quad (2.7)$$

The Markov assumption applied to this expression immediately yields

$$\Pr[x, t_3|z, t_1] = \sum_y \Pr[x, t_3|y, t_2] \Pr[y, t_2|z, t_1]. \quad (2.8)$$

This is the important *Chapman-Kolmogorov equation* which we will now use to derive the master equation under the assumption of a jump process, that is,

$$w(x, y; t) \equiv \lim_{\Delta t \rightarrow 0} \frac{\Pr[x, t + \Delta t|y, t]}{\Delta t} \quad (2.9)$$

exists and is non-vanishing.

Fix an initial observation (y, s) and let $t \geq s$. Consider the time derivative of the conditional expectation of some arbitrary function f ,

$$\begin{aligned} \frac{\partial}{\partial t} E[f(X(t))|X(s) = y] = \\ \lim_{\Delta t \rightarrow 0} \left[\underbrace{\Delta t^{-1} \sum_x f(x) \Pr[x, t + \Delta t|y, s]}_{=:A} - \underbrace{\Delta t^{-1} \sum_x f(x) \Pr[x, t|y, s]}_{=:B} \right]. \end{aligned} \quad (2.10)$$

Introduce the dummy variable z by a creative summation using the Chapman-Kolmogorov equation (2.8),

$$A = \sum_{x,z} f(x) \Pr[x, t + \Delta t|z, t] \Pr[z, t|y, s], \quad (2.11)$$

$$B = \sum_{x,z} f(x) \Pr[z, t + \Delta t|x, t] \Pr[x, t|y, s]. \quad (2.12)$$

On taking limits using (2.9) we obtain

$$\begin{aligned} \sum_x \frac{\partial}{\partial t} f(x) \Pr[x, t|y, s] = \\ \sum_{x,z} f(x) w(x, z; t) \Pr[z, t|y, s] - \sum_{x,z} f(x) w(z, x; t) \Pr[x, t|y, s], \end{aligned} \quad (2.13)$$

or by the generality of f ,

$$\frac{\partial}{\partial t} \Pr[x, t|y, s] = \sum_z w(x, z; t) \Pr[z, t|y, s] - \sum_z w(z, x; t) \Pr[x, t|y, s]. \quad (2.14)$$

This is the *master equation* and is a formulation of the Markov assumption for discrete variables in continuous time. It can also be viewed as a differential form of the Chapman-Kolmogorov equation (2.8) — as such it has generalizations to other types of stochastic processes [4].

This thesis is concerned with solving the master equation for chemical systems. If we can describe a system of D reacting species by counting the number of molecules of each kind, then the master equation will govern the dynamics of the probability distribution for the system as follows.

Let $p(x, t)$ be the probability distribution of the states $x \in \mathbf{Z}_+^D = \{0, 1, 2, \dots\}^D$ at time t . That is, p simply describes the probability that a certain number of molecules is present at each time. Define R reactions as "moves" over the states x according to the *reaction propensities* $w_r : \mathbf{Z}_+^D \rightarrow \mathbf{R}_+$. These functions measure the transition probability per unit of time for moving from the state x_r to x ;

$$x_r = x + n_r \xrightarrow{w_r(x_r)} x, \quad (2.15)$$

where $n_r \in \mathbf{Z}^D$ is the transition step and is the r th column in the *stoichiometric matrix* n .

The master equation in this setting is then given by (compare (2.14))

$$\begin{aligned} \frac{\partial p(x, t)}{\partial t} &= \sum_{\substack{r=1 \\ x+n_r^- \geq 0}}^R w_r(x+n_r) p(x+n_r, t) - \sum_{\substack{r=1 \\ x-n_r^+ \geq 0}}^R w_r(x) p(x, t) \\ &=: \mathcal{M}p, \end{aligned} \quad (2.16)$$

where the transition steps are decomposed into positive and negative parts as $n_r = n_r^+ + n_r^-$ and where the summation is performed over feasible reactions only.

The description of a general chemical system of D species is now captured by a difference-differential equation in D spatial dimensions. This description suffers from the *curse of dimensionality* — with most existing methods for solving it, the memory and time complexity increases exponentially with D . The overall aim of the thesis is to investigate the

numerical properties of methods that reduce the computational complexity of the master equation so that relevant high dimensional models can be solved. The practical impact of such methods lies in the possibility to better understand and capture chemical processes which require the mesoscopic description. Many such processes are found inside living cells but relevant examples of systems obeying the master equation also exist in other fields of physics, statistics, epidemiology and socio-economics.

2.3 Mathematical properties of the master operator

When viewed purely as a mathematical problem, the master equation has several interesting properties of which we collect a few in the following section.

We use the usual Euclidean inner product (\cdot, \cdot) ,

$$(p, q) \equiv \sum_{x \geq 0} p(x)q(x). \quad (2.17)$$

It will be shown shortly that the most natural norm in the context of the master equation is the L^1 -norm,

$$\|p\|_{L^1} \equiv \sum_{x \geq 0} |p(x)|. \quad (2.18)$$

Recall that the *adjoint operator* \mathcal{M}^* of the master operator \mathcal{M} satisfies the relation $(\mathcal{M}p, q) = (p, \mathcal{M}^*q)$. One can show that there is a nice representation of the adjoint master operator:

$$\mathcal{M}^*q = \sum_{r=1}^R w_r(x)[q(x - n_r) - q(x)]. \quad (2.19)$$

This fact has an interesting application as follows: let $X = [X_1, \dots, X_D]$ be a description in terms of a stochastic variable in D dimensions. Suppose that this system obeys a master equation and consider the dynamics of the expected value of some unknown function T (independent of time),

$$\begin{aligned} \frac{d}{dt} E[T(X)] &= \sum_{x \geq 0} \frac{\partial p}{\partial t} T(x) = (\mathcal{M}p, T) = \\ &= (p, \mathcal{M}^*T) = \sum_{r=1}^R E[w_r(X) (T(X - n_r) - T(X))]. \end{aligned} \quad (2.20)$$

As a first example we may take $T(x) = 1$ and verify in this way the natural property that the master equation does not leak probability. As a second example we take $T(x) = x_i$ and obtain

$$\frac{d}{dt}E[X_i] = - \sum_{r=1}^R n_{ri} E[w_r(X)], \quad (2.21)$$

that is, this ODE gives the dynamics of the expectation value of X in each dimension. This is essentially the initial step for the approach investigated in paper A.

We now consider some spectral properties of the master operator. Let (λ, q) be an eigenpair of \mathcal{M}^* normalized so that the largest value of q is positive and real. Then we see from (2.19) that the real part of λ must be ≤ 0 so that all eigenvalues of \mathcal{M} share this property. In the cases when \mathcal{M} admits a full set of orthogonal eigenvectors this observation directly proves decay as measured in the L^2 -norm. However, this assumption is only rarely fulfilled in the problems considered in this thesis.

In paper C the strongest general result in this direction is proved: *Any solution to the master equation is non-increasing in L^1 .* Note that this holds true for a not necessarily normalized or positive solution p as long as it is L^1 -measurable. This is of course important to a numerical analyst since, by linearity, the error which usually not is a probability distribution, is advected under the master equation itself.

An even stronger result and for the physicist a more important result is the following one: *Let $p(x, 0)$ be a given discrete function. Then under certain restrictions on the structure of the master operator, the master equation (2.16) admits a unique steady-state solution as $t \rightarrow \infty$.* For a proof and a penetrating discussion we refer to [7, V.3].

2.4 Model problems

Let us first consider the very simplest *birth-death process* [1] which is mentioned in both paper A and C:



We recognize these reactions as one part of the prey-predator model (2.2). A rare feature of this problem is that it can be solved completely if initial

data is given in the form of a Poisson distribution of expectation a_0 ,

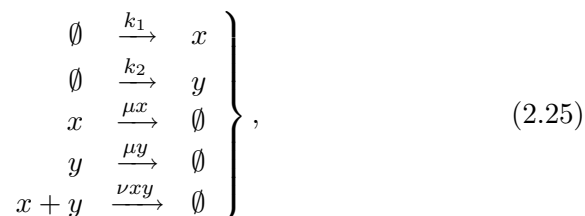
$$p(x, 0) = \frac{a_0^x}{x!} e^{-a_0}, \quad (2.23)$$

for which

$$p(x, t) = \frac{a(t)^x}{x!} e^{-a(t)}, \quad (2.24)$$

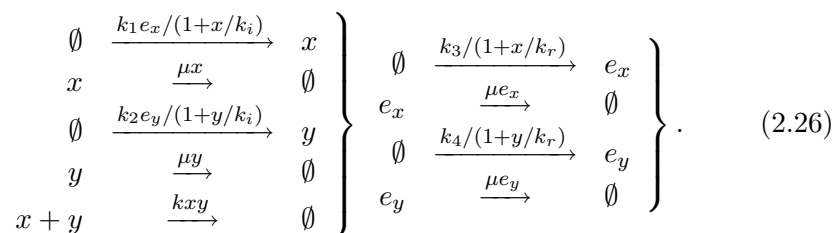
where $a(t) = a_0 \exp(-\mu t) + k/\mu \cdot (1 - \exp(-\mu t))$. Independently of the initial data, p approaches a Poisson distribution of expectation k/μ . Note that the speed with which the steady-state is reached essentially only depends on the death-rate constant μ .

As a very simple model containing *interaction*, consider the reactions



where birth-death equations control each species and where a single binary reaction couples the two dimensions of the problem. Despite the model's simplicity, no analytical solutions are available. A sample solution is shown in Figure 2.6.

A much more complicated example is treated in both paper A and paper C. The following example is found in [3] and is a model of the synthesis of two metabolites x and y by two enzymes e_x and e_y :



These reactions are not as before *elementary* but are the result of an *adiabatic* [4] simplification of a more complete model involving intermediate products. Plots of the steady-state solution of (2.26) are included in paper C.

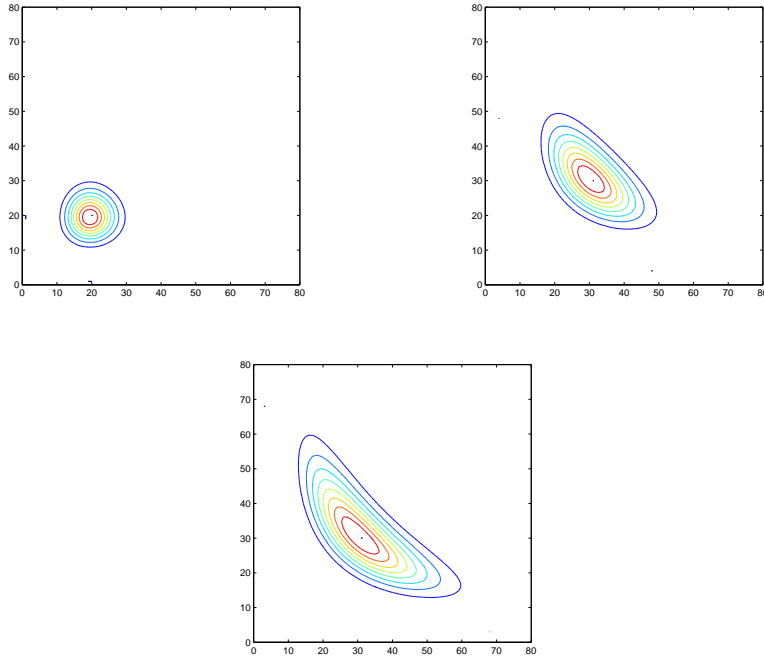
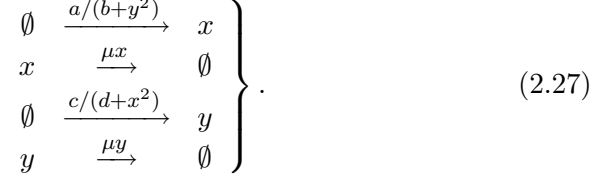


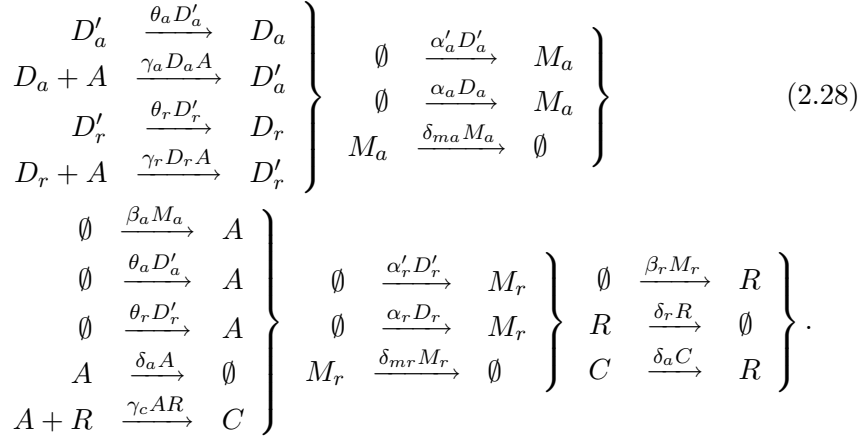
Figure 2.6: Solution contours for (2.25) at times $t = 0, 100$ and 400 . Notice how the bulk of probability rapidly arrives near the steady-state solution while the tail of probability forms much more slowly. The stiffness of the problem is thus clearly visible.

A set of reactions for which the method proposed in paper C is suitable is the *toggle switch*. Such switches can be formed by two mutually cooperatively repressing products x and y [5]. The equations are



Again, these equations are found by adiabatic simplifications of a more complicated model. The behavior of (2.27) can easily be understood at the intuitive level of preys and predators. Suppose that the population of x -molecules dominates over the number of y -molecules. Then we see that the birth-rate of y -molecules is kept in check so that the population can find a stable state. However, by a certain small probability the natural noise in the population can make the number of y -molecules eventually grow. This switches the population by reversing the roles played by x and y since then the birth-rate of x -molecules will instead be controlled.

As a final and quite complicated example we quote from paper A the *circadian clock* [2]:



This fully elementary set of reactions produces solutions that oscillate in time and was originally proposed as an explanation of biological systems that are able to keep track of time. The products R and C can be viewed as the "output" of the clock and its precise behavior is determined by the various parameters. It is shown in paper A how non-physical solutions to this model can be produced by a deterministic approach where the effects of stochasticity are not properly included.

3 Summary of papers

In this section the three papers upon which the thesis is based on are summarized. The main contributions of each paper is highlighted without going into the details. Paper A and C contains the suggested methods for solving the master equation while the short paper B contains a numerical technique needed in paper C.

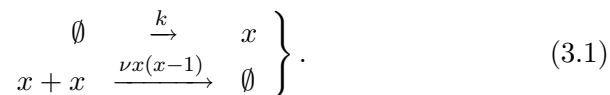
3.1 Paper A

This paper investigates a generalization of the deterministic reaction-rate approach. The reaction-rate equation is essentially formed as in (2.20) above or, in the language of preys and predators, is equivalent to the macroscopic model (2.1). The *method of moments* is an attempt to formulate equations for the central moments of order n for any given master equation. The main advantage of such an approach is efficiency: in general, the equations of order n can be solved in D^n time, where D is the number of dimensions. In this way problems of rather high dimensionality can be solved at the cost of obtaining a reduced, but often very useful, form of the solution.

The paper explicitly gives equations for the first moment (the expected value), and for the second moment (the covariance matrix). While the first moment equations can be consistently derived by assuming that the full solution of the master equation is a discrete Dirac density, the second moment equations must be regarded as an approximation which is valid under restrictions on the true solution and on the master equation itself.

General equations for higher order moments are also constructed and their validity is discussed from a theoretical as well as from an experimental point of view.

The theoretical part of the paper is built around a certain model problem for which an exact steady-state solution is available:



The motivation for considering this set of reactions as a suitable model problem is that it captures *interaction* to a certain extent while still being simple enough to solve and analyze explicitly.

The assumption $k \gg \nu$ turns out to be critical for obtaining an accurate method. This assumption is reasonable from physical considerations

since ν plays the role of the combined probability that two molecules meet, have compatible inner states and finally decide to react with each other. In many encountered systems this rather small probability is completely dominated by the birth-rate constant k .

The experimental part of the paper treats two quite different numerical examples and suggests some additional analysis of the behavior of the method. One example is the circadian clock (2.28) and has the interesting property that the presence of stochastic noise is critical for the clock to continuously producing reliable oscillations.

In summary, the paper demonstrates that the method of moments is a very useful approach to solving and analyzing chemical reactions. The main advantages of the method are that it has low computational complexity and frequently produces the output that one is really interested in. On the other hand, the disadvantages of the method are that it is difficult to analyze *a priori* and that the produced system of ODEs can become very stiff for higher order moments.

3.2 Paper B

This short paper contains a numerical investigation of Gaussian quadratures for series on the form

$$\sum_{x \in \Omega} f(x)w(x) = \sum_{i=1}^n f(x_i)w_i + R_n, \quad (3.2)$$

where Ω is a real but possibly unbounded set of points. The purpose of the paper is supportive; the resulting formulas are used extensively in the numerical experiments in paper C.

The paper briefly reviews the classical theory of mechanical quadratures aiming specifically at discrete measures. Three Gaussian summation formulas are explicitly constructed and numerical experiments on quite general series indicate their performance. It is demonstrated that the formulas generally work well as a numerical tool for summation. Some difficulties that are not usually encountered for continuous quadratures are also discussed.

The suggested technique opens up for some interesting applications. Most notable are *discrete spectral methods* for difference equations in general, and for the master equation in particular. This is the main topic of paper C.

3.3 Paper C

This lengthy paper describes a discrete spectral method for the master equation. The motivation for trying this approach is that spectral methods are efficient and natural solution strategies for any linear equation when the computational domain and the boundary conditions are "simple". The master equation satisfies these requirements as it is defined over the set of non-negative integers and requires no boundary conditions at all.

The proposed scheme involves certain polynomials that are orthogonal with respect to a discrete measure — these are *Charlier's* polynomials [8]. The constructed basis avoids the need for a continuous approximation of discrete solutions and yet allows for an efficient representation of "smooth" solutions, where smoothness has to be defined in this new discrete context.

The theoretical part of the paper starts with an introductory section containing discussions of the master equation and some results covering the behavior of its solutions. The paper continues with an interesting theory for approximation of discrete functions defined over the semi-infinite discrete set of points $\{0, 1, \dots\}$ which is remindful of classical results for continuous approximation. Conditional stability of the proposed scheme is established in a non-standard way where certain crucial properties of the master operator are made use of.

Feasibility of the proposed method is shown by the numerical solution of two different models from molecular biology. The first model is the four-dimensional example (2.26) and was also encountered in paper A. The second model (2.27) takes place in two spatial dimensions and provides a setting for which the reaction-rate approach fails. In both models, *spectral convergence* is obtained. This means that the error decays as $\exp(-cN)$ where $c > 0$ is a constant and where N is the order of the scheme.

In summary, the numerical experiments suggest that the scheme is an effective, accurate and stable alternative to traditional solution methods when the dimensionality of the problem is sufficiently small.

4 Future work

Two points for intended future work deserves to be described here. The first is a coupling between the two methods presented in this thesis and the second is a theoretical study of the possibility of improving the quality

of certain inverse problems by solving the master equation.

4.1 Coupling of the macro-meso scales

In paper A it is mentioned that the deterministic (first order) equation can be derived by assuming the solution of the master equation to be a point-mass. This assumption is frequently a rather drastic approximation but can be useful in a few of the dimensions. We therefore divide the dimensions into two parts as follows: let S_1 be the set of "thin" dimensions and let S_2 be the set of "full" dimensions. If D is the total number of dimensions, then $D_1 + D_2 = D$ where D_1 and D_2 denote the number of elements in S_1 and S_2 . The intent is now to remove all thin dimensions by approximating p by a point-distribution along all dimensions in S_1 . A suitable representation for the solution using the basis functions C_γ (arbitrary for now) is then

$$p(x, t) = \sum_{\gamma} c_{\gamma}(t) C_{\gamma}(x_{(2)}) [x_{(1)} = m], \quad (4.1)$$

where $[f]$ is 1 if f is true and zero otherwise and where $x_{(1)}$ and $x_{(2)}$ is used to denote the thin and full dimensions respectively. Note that the outer sum is of reduced dimensionality since the basis only depends on $x_{(2)}$. This is possible thanks to the introduction of the special degrees of freedom m , a vector of length D_1 containing expected values for all thin dimensions.

Using this representation, it is an easy task to evolve the degrees of freedom $[c_{\gamma}, m]$ by a Galerkin formulation for c_{γ} as in paper C and by using the first order moment equations for m as in paper A. The total cost of obtaining this reduced form of the solution is then determined by the dimensionality of S_2 plus a much smaller cost associated with the thin dimensions in S_1 .

Similar ideas for the related *Fokker-Planck* equation are presented in [10].

4.2 Strong parameter estimation

Consider the macro-scale description of a chemical system of D reacting species:

$$\frac{dm}{dt} = f(m; k), \quad (4.2)$$

that is, the usual reaction-rate equation in a simplified notation. Suppose that this system has been observed at discrete times so that a set of observations (t_i, \bar{m}_i) is available. The *inverse* problem is now: given the empirical data, determine the reaction-rate constants $k = [k_1, \dots, k_n]$ as accurately as possible under the assumption of the model (4.2). The intuitive interpretation of finding the "most likely" coefficients for the data is misleading since there is just one set of coefficients, namely the correct set!

Perhaps the most well-known solution procedure for this problem is the *maximum likelihood estimation* [9]. Under this interpretation, the question posed is instead "What is the set of parameters that generates the observed data *at the highest possible probability?*" In contrast to the "likeliness" of coefficients, the probability of obtaining a certain data set *given* the coefficients of the model is a definite and well-defined number that in principle can be computed — or at least estimated.

The traditional way of estimating this probability is to assume that the observations \bar{m}_i are normally distributed, independently of each other, around the "true" model $m(t_i)$ with the same standard deviation σ . Then the probability of obtaining the given data is the product of the probability of each measurement,

$$\Pr \left[\bigcap_i |m(t_i) - \bar{m}_i| \leq \Delta m \right] \propto \prod_i \left\{ \exp \left[-\frac{1}{2} \left(\frac{m(t_i) - \bar{m}_i}{\sigma} \right)^2 \right] \Delta m \right\}. \quad (4.3)$$

Maximizing this expression is immediately seen to be equivalent to minimizing the more familiar expression

$$M(k) = \sum_i (m(t_i) - \bar{m}_i)^2, \quad (4.4)$$

that is, maximum likelihood estimation is in this setting the very same thing as the usual least-squares fit.

Consider now the mesoscopic description corresponding to (4.2),

$$\frac{\partial p}{\partial t} = \mathcal{M}_k p, \quad (4.5)$$

where the subscript indicates the dependence of the master operator \mathcal{M} on the coefficients k . It is straightforward to write down the maximum

likelihood setup using this formulation as follows: find the set of coefficients k that maximize

$$N(k) = \prod_i p_i(\bar{m}_{i+1}, t_{i+1}), \quad (4.6)$$

in terms of which the conditional probabilities p_i satisfy (4.5) together with the initial condition

$$p_i(x, t_i) = [x = \bar{m}_i]. \quad (4.7)$$

The point of using (4.6) instead of (4.4) is that *no assumption on the probability distribution is made*. The master equation produces the probability density itself and this stronger form of the solution makes (4.6) a "stronger" estimate than (4.4). The cost of strong parameter estimation is the need for solving the full master equation (4.5) rather than the much simpler macro-description (4.2). Although this cost is certainly prohibitive in all except for a few special cases, the macro-meso scale method sketched in Section 4.1 could well be an interesting alternative for more realistic situations.

In summary, it seems reasonable to believe that strong parameter estimation is a much more accurate alternative to other methods whenever the effects of stochasticity must be correctly modeled.

5 Conclusions

The master equation is an accurate description of highly general physical systems described by discrete coordinates. The description is a direct consequence of the Markov assumption on the nature of the underlying stochastic process.

For chemical systems with many participating molecules a usually very accurate and attractive solution method is the reaction-rate equation which completely avoids the curse of dimensionality.

In situations where this approach fails, a usually better result can be obtained by adding higher order moments to the set of equations, still avoiding the high computational complexity while better capturing stochastic effects. The accuracy of this method depends on the ratio between reaction-rate constants and inflow parameters as well as on the solution itself and a rigorous analysis *a priori* is very difficult.

A viable method for the full discrete master equation is a discrete Galerkin spectral method. Here, efficiency is obtained by a compact representation of smooth solutions defined over discrete sets. In contrast to the method of moments, the solution thus obtained is the full probability density but the cost is prohibitive when many dimensions are considered.

References

- [1] W. J. Anderson. *Continuous-Time Markov Chains*. Springer Series in Statistics. Springer-Verlag, New York, 1991.
- [2] N. Barkai and S. Leibler. Circadian clocks limited by noise. *Nature*, 403:267–268, 2000.
- [3] J. Elf, P. Lötstedt, and P. Sjöberg. Problems of high dimension in molecular biology. In W. Hackbusch, editor, *Proceedings of the 19th GAMM-Seminar in Leipzig "High dimensional problems - Numerical Treatment and Applications"*, pages 21–30, 2003.
- [4] C. W. Gardiner. *Handbook of Stochastic Methods*. Springer Series in Synergetics. Springer-Verlag, Berlin, 3rd edition, 2004.
- [5] T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, 403:339–342, 2000.
- [6] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403–434, 1976.
- [7] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, Amsterdam, 5th edition, 2004.
- [8] R. Koekoek and R. F. Swarttouw. The Askey-scheme of hypergeometric orthogonal polynomials and its q -analogue. Technical Report 98-17, Delft University of Technology, Faculty of Information Technology and Systems, Department of Technical Mathematics and Informatics, 1998. Available at <http://aw.twi.tudelft.nl/~koekoek/askey.html>.
- [9] R. J. Larsen and M. L. Marx. *An Introduction to Mathematical Statistics and its Applications*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1986.

- [10] P. Lötstedt and L. Ferm. Dimensional reduction of the Fokker-Planck equation for stochastic chemical reactions. *Multiscale Meth. Simul.*, 5:593–614, 2006.

Recent licentiate theses from the Department of Information Technology

- 2006-007** Stefan Engblom: *Numerical Methods for the Chemical Master Equation*
- 2006-006** Anna Eckerdal: *Novice Students' Learning of Object-Oriented Programming*
- 2006-005** Arvid Kauppi: *A Human-Computer Interaction Approach to Train Traffic Control*
- 2006-004** Mikael Erlandsson: *Usability in Transportation – Improving the Analysis of Cognitive Work Tasks*
- 2006-003** Therese Berg: *Regular Inference for Reactive Systems*
- 2006-002** Anders Hessel: *Model-Based Test Case Selection and Generation for Real-Time Systems*
- 2006-001** Linda Brus: *Recursive Black-box Identification of Nonlinear State-space ODE Models*
- 2005-011** Björn Holmberg: *Towards Markerless Analysis of Human Motion*
- 2005-010** Paul Sjöberg: *Numerical Solution of the Fokker-Planck Approximation of the Chemical Master Equation*
- 2005-009** Magnus Evestedt: *Parameter and State Estimation using Audio and Video Signals*
- 2005-008** Niklas Johansson: *Usable IT Systems for Mobile Work*
- 2005-007** Mei Hong: *On Two Methods for Identifying Dynamic Errors-in-Variables Systems*
- 2005-006** Erik Bängtsson: *Robust Preconditioned Iterative Solution Methods for Large-Scale Nonsymmetric Problems*
- 2005-005** Peter Naucér: *Modeling and Control of Vibration in Mechanical Structures*
- 2005-004** Oskar Wibling: *Ad Hoc Routing Protocol Validation*



UPPSALA
UNIVERSITET