Monte Carlo simulation for estimating rare event probabilities and parameters in Markov process models

Author(s): Amrein, Michael

Publication Date: 2011

Permanent Link: https://doi.org/10.3929/ethz-a-006383172

Rights / License: In Copyright - Non-Commercial Use Permitted
Monte Carlo simulation for estimating rare event probabilities and parameters in Markov process models

A dissertation submitted to ETH ZURICH for the degree of Doctor of Sciences

presented by MICHAEL AMREIN
Dipl. Math. ETH
born November 27, 1980
citizen of Neudorf LU

accepted on the recommendation of
Prof. Dr. Hans Rudolf Künsch, examiner
Prof. Dr. Leonhard Held, co-examiner

2011
Acknowledgements

First and foremost, I would like to thank my advisor, Prof. Hans Rudolf Künsch, for supervising this thesis. With his calm nature and wisdom he was able to balance my sometimes impatient temper, letting me explore my own strategies and develop my own ideas at the same time. I also have a veneration for his immense knowledge in all fields related to statistics and his clear and precise way in explaining parts of it to others.

I would like to thank Prof. Leonhard Held for acting as co-examiner and having taken time to read and evaluate this thesis.

During my two years as organizer of the assistant group, Prof. Peter Bühlmann, head of the institute at that time, had always an open door to discuss issues arising in this context.

Further thanks go to Dr. Peter Gallin who aroused my interest for the fascinating world of formulas, theorems and proofs during high school, motivating me for studies in Mathematics.

Moreover, I want to thank the whole staff of the “Seminar für Statistik” (“SfS”), including some former members I had the pleasure to work with, for an excellent working atmosphere that could not have been better. In particular, the secretaries with their support in organisational – and sometimes also other – matters and Dr. Martin Mächler with his assistance in programming problems eased the live at the “SfS”. A special thank goes to Philipp Rütimann, with whom I shared the office most of the time at the “SfS”.

Last but not least, I want to thank my parents Susanne and Hansruedi and my girlfriend Olivia for their unconditional support, especially in times when I had to cope with setbacks in my work or during the double burden when I made the education in the Swiss Army to a company commander.
Everything should be made as simple as possible, but not simpler –
Albert Einstein
Contents

Abstract xi

Zusammenfassung xiii

1 Introduction 1

1.1 Markov processes . . . . . . . . . . . . . . . . . . . . . . . 1

1.1.1 Discrete-time Markov processes . . . . . . . . . . . . . 2

1.1.2 Continuous-time Markov processes . . . . . . . . . . 4

1.1.3 Continuous-time Markov chains . . . . . . . . . . . . . 5

1.1.4 Diffusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

1.1.5 General queueing models . . . . . . . . . . . . . . . . . 7

1.2 Problems considered in this thesis . . . . . . . . . . . . . . . 8

1.2.1 Rare entrance probabilities of Markov chains . . . . . . 8

1.2.2 Parameter estimation in Markov jump processes 10
2 Fixed numbers of successes in splitting simulation 13

2.1 Introduction ........................................... 13

2.2 Setting and definitions ............................... 15

2.2.1 Rare event simulation .............................. 15

2.2.2 Entrance probabilities of Markov chains ......... 17

2.2.3 Importance splitting ............................... 18

2.3 Unbiasedness of the importance splitting estimators . 21

2.3.1 Balanced selection for splitting with fixed num-
    ber of successes ........................................ 27

2.4 The optimal importance function .................... 29

2.4.1 Approximation for homogeneous Markov processes

    on countable state spaces ............................... 31

2.4.2 Approximation when estimating entrance prob-

    abilities of homogeneous Markov jump processes

    with a finite time horizon ............................. 32

2.5 Theoretical analysis in a simplified setting ........ 34

2.5.1 Optimal choice of levels ........................... 36

2.5.2 Comparison of efficiency .......................... 41

2.5.3 Asymptotic distribution ............................ 43

2.5.4 Two-stage fixed numbers of successes estimator . 45

2.6 Examples ................................................ 46
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6.1 High levels during a busy period of a G/G/1 queueing system</td>
<td>47</td>
</tr>
<tr>
<td>2.6.2 Buffer overflow in a tandem Jackson network during a busy period</td>
<td>51</td>
</tr>
<tr>
<td>2.6.3 Buffer overflows in finite time</td>
<td>56</td>
</tr>
<tr>
<td>2.7 Conclusions</td>
<td>60</td>
</tr>
<tr>
<td>2.A Counterexample to the claim (2.4)</td>
<td>61</td>
</tr>
<tr>
<td>2.B Proofs for optimality in the independence framework</td>
<td>64</td>
</tr>
<tr>
<td>2.B.1 Proof of Proposition 2.4</td>
<td>64</td>
</tr>
<tr>
<td>2.B.2 Proof of Proposition 2.5</td>
<td>66</td>
</tr>
<tr>
<td>2.C Proofs for asymptotic log-normality</td>
<td>67</td>
</tr>
<tr>
<td>2.C.1 Proof of Proposition 2.6</td>
<td>67</td>
</tr>
<tr>
<td>2.C.2 Proof of Proposition 2.7</td>
<td>69</td>
</tr>
<tr>
<td>2.D Numerical calculation of the overflow probabilities</td>
<td>69</td>
</tr>
<tr>
<td>2.E Simulation or numerical solution</td>
<td>71</td>
</tr>
<tr>
<td>3 Rate estimation in partially observed Markov jump processes with measurement errors</td>
<td>73</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>73</td>
</tr>
<tr>
<td>3.2 Setting and definitions</td>
<td>75</td>
</tr>
<tr>
<td>3.2.1 Model</td>
<td>75</td>
</tr>
</tbody>
</table>
3.2.2 Additional notation and formulae for densities  . 77
3.3 Bayesian estimation and Monte Carlo methods . . . . 78
3.4 Simulating a path given parameters and observations . 81
3.4.1 Generating new reaction totals . . . . . . . . . . 81
3.4.2 Generating a new path given the reaction totals 82
3.4.3 Acceptance probability of a new path . . . . . . 84
3.4.4 Choice of the subintervals \([a, b]\) . . . . . . . . . 85
3.4.5 Updating latent components . . . . . . . . . . . . 86
3.4.6 Updating the path at a border . . . . . . . . . . 87
3.5 Initialisation of \(\eta, \theta\) and \(y[t_0, t_n]\) . . . . . . . . 87
3.6 Examples . . . . . . . . . . . . . . . . . . . . . . . . . 89
3.6.1 Stochastic Oregonator . . . . . . . . . . . . . . . . 89
3.6.2 Prokaryotic auto-regulation . . . . . . . . . . . . 103
3.7 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . 113
3.8 Integer solutions of homogeneous linear equations . . 114
3.9 Inconsistency when using diffusions . . . . . . . . . . 115

Bibliography 120

Curriculum Vitae 127
Abstract

Markov processes models appear in many different fields, for example in queueing theory, epidemic models, stochastic kinetics of chemical reactions or financial engineering. Often, simple algorithms exist to generate sample paths of this processes. This makes analysis based on simulation, so called Monte Carlo methods, very attractive. We study two specific estimation problems in this context.

First we consider a simulation technique, called importance splitting, to estimate very small entrance probabilities for Markov processes by splitting sample paths at various level surfaces of the so called importance function before reaching the set of interest. This can be done in many ways, yielding different variants of the method.

In this context, we propose a new one, called fixed number of successes. We prove unbiasedness for the new and some known variants, because in many papers, the proof is based on an incorrect argument. Further, we analyze its behavior in a simplified setting, which is appropriate if the importance function is well chosen, in terms of efficiency and asymptotics in comparison to the standard variant. The main difference is that the new variant controls the precision of the estimator rather than the computational effort. Our analysis and simulation examples show that it is rather robust in terms of parameter choice and we present a two-stage procedure which also yields confidence intervals.

The choice of the importance function which governs the placement
of the splitting surfaces is a major issue in splitting estimation. Thus we present and motivate importance functions when estimating the entrance probability of a continuous-time Markov chain on a discrete state space into some set, for example a buffer overflow in a tandem Jackson network, before hitting another set or during some finite time interval. We show that they yield good results in practice, especially combined with our two-stage procedure based on fixed number of successes.

Second, we present a simulation methodology for Bayesian estimation of rate parameters in Markov jump processes arising for example in stochastic kinetic models. To handle the problem of missing components and measurement errors in observed data, we embed the Markov jump process into the framework of a general state space model. Markov chain Monte Carlo and particle filter type algorithms are introduced, which allow sampling from the posterior distribution of the rate parameters and the Markov jump process, also in data-poor scenarios. The algorithms are illustrated by applying them to rate estimation in a model for prokaryotic auto-regulation and in the stochastic Oregonator, respectively.
Zusammenfassung


In diesem Zusammenhang präsentieren wir eine neue Variante namens “fixed number of successes”. Wir beweisen Erwartungstreue für diese, aber auch für einige schon bekannte Varianten, da der Beweis in einigen Publikationen ein falsches Argument enthält. Weiter analysieren wir sie unter einer vereinfachenden Annahme – welche angemessen ist, wenn die “importance function” gut gewählt ist – in Bezug auf Effizienz und Asymptotik im Vergleich zur Standard-Variante. Der Hauptunterschied ist, dass unsere neue Variante die Präzision (im Sin-
ne der Varianz relativ zur quadrierten Schätzgröße) kontrolliert und nicht den Berechnungsaufwand. Weiter zeigen unsere Analyse und unsere Simulationsbeispiele, dass sie robust in Bezug auf die Wahl ihrer Simulations-Parameter ist. Basierend auf unseren Ergebnissen haben wir eine zweistufige Schätzprozedur entwickelt, die auch Vertrauensintervalle liefert.


Chapter 1

Introduction

1.1 Markov processes

Markov processes play a very important role throughout this thesis. On the one hand, we consider estimation problems arising in Markov process models used for example in queueing theory, epidemic models, stochastic kinetics of chemical reactions or financial engineering. On the other hand, they also build the basis for solving some of these problems via Markov chain Monte Carlo methods which allow simulation from very complicated probability distributions.

Let us consider a general Markov process with time index $t$ on a state space $\mathcal{E}$, that is $\mathcal{Y} = \{Y_t : t \geq t_0\}$ where $Y_t$ are elements of $\mathcal{E}$. Its defining feature is the so-called Markov property which states that given the present state of a process, the future is independent from the past, i.e., we have for times $p_2 < p_1 < s < t$, states $\ldots, z_2, z_1, y_s$ and $A \subset \mathcal{E}$ that

$$P[Y_t \in A|Y_s = y_s, Y_{p_1} = z_1, Y_{p_2} = z_2, \ldots] = P[Y_t \in A|Y_s = y_s]. \quad (1.1)$$
If in addition we have that
\[ P[Y_t \in A | Y_s = y_s] = P[Y_{t+h} \in A | Y_{s+h} = y_s] \]
for all \( h > 0, \ s < t, \ y_s \) and \( A \), we speak of a (time-) homogeneous Markov process.

There is a more restrictive form of the Markov property, called strong Markov property: Consider a stopping time \( \tau \) for the Markov process \( Y \), i.e., a random variable where the events \( \{ \tau \leq s \} \) are elements of the sigma algebra generated by \( \{ Y_t : t \leq s \} \). We call \( Y \) a strong Markov process when, conditioned on the event \( \tau < \infty \) and \( Y_\tau = y \), the process \( \{ Y_t : t \geq \tau \} \) is Markov process that behaves like \( Y \) started at \( y \), independently of \( \{ Y_t : t \leq \tau \} \).

There are several types of Markov processes often considered, distinguished mainly by the questions whether the time index \( t \) is discrete or continuous and whether the state space \( \mathcal{E} \) is countable (discrete) or not.

In the following, we will give a very brief overview on Markov processes, covering only the topics relevant for the core of this thesis. For further reading, consider e.g. Stirzaker (2005), Lefebvre (2007) or Stewart (2009).

### 1.1.1 Discrete-time Markov processes

When the time \( t \) is discrete, so that we can assume \( t \in \{ 0, 1, 2, \ldots \} = \mathbb{N}_0 \), we speak of a discrete-time Markov process (or sometimes discrete-time Markov chain, especially in the framework of Markov chain Monte Carlo methods). It can be defined iteratively by specifying an initial state \( y_0 \) and the so-called one-step transition kernels

\[ M_t(y, A) = P[Y_{t+1} \in A | Y_t = y] \quad (1.2) \]
at each time-step $t$, which defines the conditional distribution of the process at the consecutive time point $t + 1$ if we know the state at time $t$. Transition kernels of higher order $M_{s:t}(y, A) = P[Y_t \in A | Y_s = y]$ ($s < t$) can be calculated by (iterative) integration of the one-step kernels, that is,

$$M_{s:t}(y, A) = \int_{E} \cdots \int_{E} M_s(y, dy_{s+1})M_{s+1}(y_{s+1}, dy_{s+2}) \cdots M_{t-1}(y_{t-1}, A).$$

If $E$ is discrete, too (we assume without loss of generality that $E \subset \mathbb{N}_0$), we speak of a discrete-time Markov chain. Then the kernel (1.2) is usually given via the so-called (one-step) transition probabilities at step $t$

$$p_t(i, j) = P[Y_{t+1} = j | Y_t = i]$$

or the (one-step) transition (probability) matrix $P(t)$, formed by placing $p_t(i, j)$ in row $i$ and column $j$.

Thus if an algorithm to simulate according to the one-step transition kernels $M_t(y, A)$ is available (this is usually the case and also clearly true for the one-step transition probabilities $p_t(i, j)$), we can easily generate sample trajectories or sample paths of the process, i.e., sequences of states $Y_0, Y_1, Y_2, \ldots$ visited according to the law of the process.

If the chain is homogeneous we have that the transition probabilities or the one-step transition kernels, respectively, do not depend on the time index $t$, that is $p(i, j) \equiv p(i, j)$ and $P(.) \equiv P$ or $M_.(y, A) \equiv M(y, A)$, respectively. In this setting, we call $\pi$ an stationary or invariant distribution on $E$, if for all $A \subset E$

$$\pi(A) = \int_{E} M(y, A)d\pi(y),$$

in other words, if the chain state at time $t$, $Y_t$, is distributed according to $\pi$, then also the state at the consecutive time point $Y_{t+1}$. If the chain is positive recurrent \(^1\), the stationary distribution exists, if the

\(^1\)For the definitions of recurrence, irreducibility, aperiodicity and further details, see e.g. Stirzaker (2005) or Lefebvre (2007).
chain is in addition irreducible \(^1\), the stationary distribution is unique. Under the extra condition that the chain is aperiodic \(^1\), the stationary distribution corresponds to the limiting distribution of the process (the limit for \(t \to \infty\) of the distribution of \(Y_t\)), i.e.,

\[
\pi(A) = \lim_{t \to \infty} P[Y_t \in A].
\]

### 1.1.2 Continuous-time Markov processes

The difficulty in continuous time is that we need the transition kernels

\[
M_{s,t}(y, A) = P[Y_t \in A | Y_s = y]
\]

for all \(s < t\), since we do not have a smallest time-step as in the discrete-time framework. These transition kernels are related through the Chapman–Kolmogorov equations: For \(s < t < u\) we have

\[
M_{s,u}(y, A) = \int \mathcal{E} M_{s,t}(y, dy_t) M_{t,u}(y_t, A).
\]

One approach is to determine such transition kernels is to look at the infinitesimal transitions

\[
Q_t(y, A) = \lim_{h \to 0} \frac{1}{h} (M_{t+h,t}(y, A) - \mathbb{I}_A(y)),
\]

provided that

\[
M_{t+h,t}(y, A) \to \mathbb{I}_A(y)
\]

for \(h \to 0\), where \(\mathbb{I}_A(y)\) is the indicator function of the set \(A\). We mention that there are considerable technical difficulties for constructing the kernels \(M_{s,t}\) from a family of infinitesimal transitions \(Q_t\).

Alternatively a continuous-time Markov process can be given explicitly by a recipe to simulate sample paths. In the following, we look at some specific examples often used in practice for the two approaches.
1.1.3 Continuous-time Markov chains

Let $t \in [t_0, \infty)$ and $\mathcal{E}$ be countable. Analogously to the time-discrete case, we need the conditional probabilities

$$p_{s,t}(i,j) = P[Y_t = j | Y_s = i]$$

for $t_0 \leq s < t$ and states $i$ and $j$. Further, to avoid subtleties, we assume that the sample paths are right-continuous with a finite number of jumps on a finite time interval. Then we can define this type of Markov process by transition rates or transition intensities

$$q_t(i,j) = \lim_{\delta \to 0} \frac{p_{t,t+\delta}(i,j)}{\delta} (i \neq j) \quad \text{and} \quad q_t(i,i) = -\sum_{j \neq i} q_t(i,j)$$

or the generator (matrix) $Q(t)$ (sometimes $q(t)$), formed by placing $q_t(i,j)$ in row $i$ and column $j$, according to the first approach mentioned above.

If the chain is homogeneous, all $p_{s,t}(i,j)$ depend only on the difference $t - s$, i.e., $p_{s,t}(i,j) = p_{0,t-s}(i,j)$ for all $t_0 \leq s < t$ and $q(i,j) \equiv q(i,j)$ and $Q(.) \equiv Q$. In this case, one can show that law of this chain is equivalent using the transition kernel defined by the following rule. Denote with $\tau_0 = t_0$ and $\tau_k = \inf\{t > \tau_{k-1} : Y_t \neq Y_{\tau_{k-1}}\}$ the arrival times and with $\delta_k = \tau_k - \tau_{k-1}$ the inter arrival times. Then

$$\delta_k|Y_{\tau_{k-1}} = i \sim \text{Exp}(-q(i,i))$$

and

$$P[Y_{\tau_k} = j|Y_{\tau_{k-1}} = i] = \frac{q(i,j)}{-q(i,i)} (i \neq j).$$

On the one hand, this provides a simple method to simulate trajectories (via the so-called Gillespie algorithm, see Gillespie (1977)) and on the other, it establishes a connection to the discrete-time case: The embedded chain (also called jump chain) \{Y_k = Y_{\tau_k} : k \in \mathbb{N}_0\} is a homogeneous discrete-time Markov chain with $p(i,j) = -q(i,j)/q(i,i)$
Chapter 1. Introduction

(i \neq j) and \( p(i, i) = 0 \). Thus the concepts of stationary and limiting distributions can easily be transferred to the continuous case via the embedded chain.

Non-homogeneous continuous-time Markov chains can be simulated with the so-called Lewis’ method (see Ogata (1981)).

1.1.4 Diffusions

(\text{Itô}) diffusions are Markov processes where both time and state space – and also the sample paths – are continuous. They are defined by means of stochastic differential equations. We can not provide an introduction to this topic, because this would take us too far afield. Thus we recommend Øksendal (1995), Gardiner (2004) (with focus on applications in physics, chemistry and natural sciences) or Glasserman (2004) (with focus on applications in financial engineering) for readers unfamiliar with it.

Consider the diffusion

\[
dY_t = m(t, Y_t)dt + \Sigma^{1/2}(t, Y_t)dW_t, \ Y_0 = y_0,
\]

with drift \( m(t, y) \in \mathbb{R}^p \) and symmetric, positive-semidefinite diffusion term

\[
\Sigma(t, y) = \left( \Sigma^{1/2} \right)^T \Sigma^{1/2} \in \mathbb{R}^{p \times p},
\]

where \( W_t \) denotes the \( p \)-dimensional standard Wiener process. Under some smoothness conditions, existence and uniqueness of solutions is guaranteed.

Nevertheless, since sample trajectories are usually almost surely nowhere differentiable, we can not simulate diffusion processes. Thus we consider the so-called Euler–Maruyama discretization, which enables approximative simulation and also gives insight how the stochastic differential equation (1.3) is to be understood.

We discretize the time by steps of length \( \delta \). An increment \( \Delta W_i = W_i\delta - \)
W_{(i-1)\delta} of the Wiener process is normally distributed with mean 0 and covariance matrix \(\delta \cdot \mathbb{I}\). Similarly to the Euler method to solve ordinary differential equations, we define the Euler–Maruyama approximation \(\tilde{Y}_{i\delta}\) to \(Y_{i\delta}\) iteratively by setting \(\tilde{Y}_{0} = Y_{0}\) and

\[
\tilde{Y}_{i\delta} = m((i - 1)\delta, \tilde{Y}_{(i-1)\delta})\delta + \Sigma^{1/2}((i - 1)\delta, \tilde{Y}_{(i-1)\delta})\Delta W_{i}
\]

or equivalently

\[
\tilde{Y}_{i\delta} | \tilde{Y}_{(i-1)\delta} \sim \mathcal{N}(\tilde{Y}_{(i-1)\delta} + m((i - 1)\delta, \tilde{Y}_{(i-1)\delta})\delta, \Sigma((i - 1)\delta, \tilde{Y}_{(i-1)\delta})\delta).
\]

Under regularity conditions, we have for a fixed time \(T\) and \(n = \lfloor T/\delta \rfloor\) that \(\tilde{Y}_{n\delta} \to Y_{T}\) for \(\delta \to 0\) in mean, i.e., in \(L_{1}\)-norm, and thus also in probability. For more details and approximation schemes with somewhat better convergence properties, see Glasserman (2004), Chapter 6.

### 1.1.5 General queueing models

The basic element of queueing models are the servers, where a customer remains during the so-called service time before he leaves the system or goes to another server. Customers enter the system at certain servers driven by arrival processes, where the arrival time, i.e., the time between two consecutive arrivals, is given. The service and arrival times are specified by positive random variables. If a customer arrives at a server where another is being served at the moment, he has to wait until all customers arrived before him are served and thus a queue builds up at the server. A simple example is the G/G/1 queue, where customers arrive at the single server and are served according to general arrival and service distributions before they leave the system.

If all arrival and service times are exponentially distributed, the process \(\{Y_t : t \geq 0\}\), where the vector \(Y_t\) contains the lengths of the queues at each server at time \(t\), is a homogenous continuous-time Markov process with intensities equal the parameter of the corresponding exponential random time. Exemplarily for the G/G/1 queue with an \(\text{Exp}(\lambda)\) arrival
time, \( \text{Exp}(\mu) \) service and number of customers \( Y_t \) being at the single server, called M/M/1 queue in this framework, we find transition rates \( q(i, i + 1) = \lambda \) for \( i \in \mathbb{N}_0 \), \( q(i, i - 1) = \mu \) for \( i \in \mathbb{N} \) and 0 else.

For general arrival and service time distributions, i.e., non-exponential such as Weibull or Pareto, we need to include information on the remaining arrival and service times into the states to ensure that we get a Markov process, which is also easy to simulate. Exemplarily for the \( G/G/1 \) system, we write \( X^1_t \) for the number of customers in the system, \( X^2_t \) for the remaining time until the next arrival and \( X^3_t \) for the remaining service time of the customer being served at time \( t \). The Markov process \( \mathcal{X} = \{ X_t = (X^1_t, X^2_t, X^3_t)^T : t \geq 0 \} \) has the state space \( \mathcal{E} = \mathbb{N}_0 \times [0, \infty)^2 \), which is a mixture of discrete and continuous components.

For further reading on queueing theory, we recommend e.g. Kobayashi and Mark (2009) or Gross et al. (2008).

## 1.2 Problems considered in this thesis

As we have seen, for many many Markov process models, not least because of the Markov property (1.1), there exist rather simple algorithms to generate sample trajectories (or at least reasonable approximations in case of diffusions). This makes analysis based on simulation techniques, so-called Monte Carlo methods, very attractive.

In this thesis, two different problems in this context are studied. The following two sections in this chapter give a brief introduction to these problems.

### 1.2.1 Rare entrance probabilities of Markov chains

A first problem we consider in this thesis is to find the probability \( \gamma \) that a Markov process enters a set \( B \) before some fixed time \( T \) or before
hitting another set \( A \). This can be written as

\[ P[\tau_B < T] \text{ or } P[\tau_B < \tau_A], \]

respectively, if we denote by \( \tau_C \) the first entrance time of the Markov process in some set \( C \). Here, we assume \( \gamma \) to be small.

There are different methods available to solve such a problem including explicit calculation, which is possible in a few cases, and numerical approximation in certain situations, e.g. when we work with homogeneous continuous-time Markov chains (see Appendix 2.D or the books Stewart (2009) and Bolch et al. (2006)).

We focus on a third possibility to solve this kind of problems, namely Monte Carlo methods. Here, in principle one generates a certain number of trajectories of the Markov process, let us say \( n \), and estimates the probability \( \gamma \) by the fraction where the event of interest \( \{\tau_B < \zeta\} \) (\( \zeta = T \) or \( \zeta = \tau_A \)) occurred, that is,

\[ \hat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{\tau_B < \zeta\}}, \]

often called the naive Monte Carlo estimator. The relative variance or imprecision of this unbiased estimator is given by

\[ q = \frac{\text{Var}(\hat{\gamma})}{\gamma^2} = \frac{1 - \gamma}{\gamma n}. \]

To get a naive Monte Carlo estimate with imprecision \( q \), we need \( n = 1/(q\gamma) \) sample trajectories, which is often infeasible when \( \gamma \) is small. Thus acceleration methods are needed. The usual approaches to address these rare event problems are importance sampling and importance splitting. The main idea of the first technique is to change the probability law of the system in order to make the desired event more frequent and correcting the estimator by an appropriate likelihood ratio, whereas in the second case the probability law remains unchanged, but a drift toward the rare event is created by means of a so-called importance function, which allows to select trajectories which
are more promising than others. In Chapter 2 of this thesis, we present and analyze a new variant of the latter. The text is a revised version with additional material of the work published in Amrein and Künsch (2011). Further introductory remarks on importance splitting can be found in sections 2.1 and 2.2.

1.2.2 Parameter estimation in Markov jump processes

Markov process models \( \mathcal{Y} = \{Y_t, t \in [t_0, t_n]\} \) often include unknown parameters, for example the reaction rates in stochastic kinetic models. Therefore one wants to estimate them, written here in the form of a parameter vector \( \theta \), from time series data \( \mathcal{X} = \{X_t : t \in \{t_1, \ldots, t_n\}\} \). To include measurement errors and missing values in the time series data \( \mathcal{X} \), we regard the value \( X_{t_k} \) as a realisation of a conditional distribution \( g(\cdot | Y_{t_k}) \) based on the value \( Y_{t_k} \). Then we are in the framework of a general state space model: \( X_{t_0}, X_{t_1}, \ldots, X_{t_n} \) is an observed time series which is derived from the unobservable Markov chain \( Y_{t_0}, Y_{t_1}, \ldots, Y_{t_n} \).

For the Markov process models and the conditional distributions \( g \) considered here, the combined density of the full process and the data given the parameter \( \theta \) is explicitly known, that is, we can evaluate \( p(\mathcal{Y}, \mathcal{X} | \theta) \). It is usually too complicated to calculate the maximum likelihood estimator

\[
\hat{\theta}_{MLE} = \arg \min_{\theta} \left\{ p(\mathcal{X} | \theta) = \int_{\mathcal{Y}} p(\mathcal{Y}, \mathcal{X} | \theta) d\mathcal{Y} \right\}
\]

in such a framework, because the marginalization over \( \mathcal{Y} \) can not be done explicitly.

To circumvent this problem, one combines a Bayesian approach, which also allows to include prior knowledge about the parameter \( \theta \), with Markov chain Monte Carlo (MCMC) techniques. First, one chooses a suitable prior density for \( \theta \), say \( p(\theta) \). Then we have for the density \( f \)
of \( Z = (Y, \theta) \) conditioned on the data \( \mathcal{X} \) that

\[
f(Z) = \beta \cdot p(Y, \mathcal{X}|\theta)p(\theta),
\]

where

\[
\beta = \left( \int_Y \int_\theta p(Y, \mathcal{X}|\theta)p(\theta)d\theta dY \right)^{-1}
\]

is a normalization constant. Even if \( \beta \) is incalculable, a MCMC algorithm can be used to simulate from the distribution of \( Z \) which in particular yields samples from the posterior distribution of \( \theta \) given the data \( \mathcal{X} \) using a marginalization over \( Y \). These are used to construct (posterior) estimates of \( \theta \).

The main idea behind the MCMC algorithms is the following. One constructs a discrete-time Markov chain with states \( Z \) and stationary distribution corresponding to the the target density \( f(.) \). Because as mentioned, a Markov chain converges to its stationary distribution (under some additional conditions), one can produce samples according to the density \( f(.) \) simulating the chain long enough.

To ensure that the chain has the desired stationary distribution, one first chooses a conditional density \( q(.|Z) \) with respect to the dominating measure of the model, called proposal density. Then one defines the transition kernel by the Metropolis–Hastings recipe: Given the chain state \( Z^{(m)} \),

- generate \( Z^{new} \) according to the proposal density \( q(.|Z^{(m)}) \),

- take

\[
Z^{(m+1)} = \begin{cases} 
Z^{new} & \text{with probability } \alpha(Z^{new}, Z^{(m)}) \\
Z^{(m)} & \text{with probability } 1 - \alpha(Z^{new}, Z^{(m)}) 
\end{cases}
\]

where

\[
\alpha(Z^{new}, Z^{(m)}) = \min \left\{ \frac{f(Z^{new})q(Z^{(m)}|Z^{new})}{f(Z^{(m)})q(Z^{new}|Z^{(m)})}, 1 \right\}.
\]
The probability $\alpha$ is called the Metropolis–Hastings acceptance probability. Note also that it is sufficient to know the density $f(.)$ up to some unknown constant, since this constant cancels out in the formula for $\alpha$. The idea can also be extended to cases where we do not update the whole random variable $Z$ in one step of the chain, but only some of its components, for example the Markov process $\mathcal{Y}$ on a subinterval of $[t_0, t_n]$ or the parameter $\theta$. For further details about MCMC methods, see e.g. Robert and Casella (2004), Chapter 7, or Gilks et al. (1996).

Because here the random variable $Z$ has discrete parts and a rather complicated structure, it is not clear how to specify a valid starting value of the Metropolis–Hastings chain. To solve this problem and to speed up the convergence of the MCMC algorithm, we can exploit the general state space model framework using particle filter techniques, also called sequential Monte Carlo methods (see e.g. Künsch (2000) or Robert and Casella (2004), Chapter 14). They can be used to construct iteratively a starting value for the the process $\mathcal{Y}$ on the interval $(t_{k-1}, t_k]$ based on the data $X_{t_0}, X_{t_1}, \ldots, X_{t_k}$ and a rough estimate of $\theta$ given $X_{t_0}, X_{t_1}, \ldots, X_{t_{k-1}}$.

In Chapter 3 of this thesis, we present an algorithm in the above described fashion for a certain class of Markov processes used in context of epidemic models or stochastic reaction kinetics. The text is based on a working paper submitted to “Statistics & Computing”. Further introductory remarks on this subject can be found in sections 3.1 and 3.2.
Chapter 2

Fixed numbers of successes in splitting simulation

2.1 Introduction

Importance splitting or simply splitting is a simulation technique to estimate very small entrance probabilities for Markov processes (for general references see Garvels and Kroese (1998), Glasserman et al. (1999), Gillespie (1977), L’Ecuyer et al. (2006) and L’Ecuyer et al. (2009) and for a related method called RESTART see Villén-Altamirano and Villén-Altamirano (1994) and Villén-Altamirano and Villén-Altamirano (2006)). The basic idea is to consider the event of interest as the intersection of a nested sequence of events which are usually sublevel sets of the so called importance function. The rare event probability is thus the product of conditional probabilities, which can usually be estimated much more precisely.
There are different unbiased variants of the technique, e.g. fixed splitting or fixed effort. In this context, we propose a new one, called fixed number of successes, which is also shown to be unbiased. In contrast to the variant with fixed effort, it controls the imprecision of the estimator rather than the computational effort. Further, it never estimates the probability to be zero, and our analysis and simulation examples show that it is less sensitive to tuning issues like the choice of level sets and number of replicates per level.

A major issue in applications is the choice of the importance function, see e.g. the statement in L’Ecuyer et al. (2009) on page 48: “First, how should the importance function $h$ be defined? This is definitely the most important and most difficult question to address.” In principle, the optimal importance function is known (see Garvels (2000) or L’Ecuyer et al. (2009)), but its evaluation requires the quantity of interest and thus can not be used in applications. We give a computable approximations for this optimal importance function in two different cases:

- The Markov chain is time-homogeneous with a countable state space and the event of interest has the form $\{\tau_B < \tau_A\}$ for entrance times $\tau_B$ and $\tau_A$ in some set $A$ and $B$, respectively.

- Entrance probabilities of a homogeneous continuous-time Markov jump process into some set $B$ before a fixed time horizon $T$, that is, the event of interest has the form $\{\tau_B < T\}$. This is usually the harder problem, but important in many applications (see e.g. Asmussen and Jobmann (2002) or De Boer et al. (2007))

These approximations yield good results in applications, especially in combination with our two-stage estimation strategy based on the variant fixed number of successes as our simulation examples show.

The rest of the chapter is organized as follows. In Section 2.2, we describe the kinds of rare events we are going to consider, give a brief review of the importance splitting method and introduce the fixed number of successes variant. Section 2.3 deals with the unbiasedness of the
importance splitting estimators. We also cover some known variants, because the original proof on pages 16 and 17 in Garvels (2000) is based on an incorrect claim about conditional unbiasedness. Later papers either refer to Garvels (2000) without providing details on how to obtain a correct proof, or they derive the result in a in very general theoretical context which can be hard to follow or do not include some variants. Also, a second version of the fixed number of success estimator is introduced which is no longer unbiased, but is expected to be more concentrated around the true value. In Section 2.4, we review the derivation the optimal importance function and present our approximations in the two situations above. In Section 2.5, we analyze and compare our variant to the fixed effort estimator in a simplified setting, which is appropriate if the importance function is chosen described as in Section 2.4. We derive optimal parameters for both variants simultaneously, yielding slightly different results from Garvels (2000), and analyze the asymptotic behavior for a fixed number of sublevel sets. In addition, we compare the efficiency in a setting relevant for applications and look at the asymptotic distribution. From this, we derive some practical implications, in particular a two-stage estimation procedure, which yields confidence intervals. In Section 2.6, we present some simulation experiments for buffer overflows in a G/G/1 queue and in the well known tandem Jackson network. Conclusions are given in Section 2.7.

2.2 Setting and definitions

2.2.1 Rare event simulation

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $E \in \mathcal{F}$ a rare event, that is, $\gamma = P[E]$ is small. The naive Monte Carlo estimator $\hat{\gamma}_n$ of $\gamma$ generates $n$ samples $(\omega_1, \ldots, \omega_n)$ and calculates the proportion of samples $\omega_i \in E$. 


The mean squared relative error is
\[
\frac{\text{Var}(\hat{\gamma}_n)}{\gamma^2} = \frac{\gamma(1 - \gamma)/n}{\gamma^2} = \frac{1 - \gamma}{\gamma n}
\]
which diverges to infinity as $O(1/\gamma)$ when $\gamma$ goes to 0. Therefore it is unreliable when $\gamma$ is small. The best known alternative is to use importance sampling where the samples are generated from a distribution different from $P$ and weights equal to the likelihood ratio are used, see e.g. Asmussen and Glynn (2007), Glynn and Iglehart (1989), Bucklew (2004) or Robert and Casella (2004).

For a mathematical analysis of simulation estimators, one usually imbeds the given problem into a sequence where the probability $\gamma$ tends to zero by varying either the event $E$ or the probability $P$ or both. An unbiased Monte Carlo estimator $\hat{\gamma}$ which is reliable for small $\gamma$ should have bounded relative error (Bucklew (2004)):
\[
\limsup_{\gamma \to 0} \frac{\text{Var}(\hat{\gamma})}{\gamma^2} < \infty.
\]
Unbiased estimators usually satisfy
\[
\liminf_{\gamma \to 0} \frac{\text{Var}(\hat{\gamma})}{\gamma^2} > 0,
\]
because otherwise $\hat{\gamma}/\gamma$ converges to 1 in probability. These two properties imply the weaker (than bounded relative error) property
\[
\lim_{\gamma \to 0} \frac{\log(\text{Var}(\hat{\gamma}))}{\log(\gamma)} = 2.
\]
Simulation estimators have parameters like the number of replicates which influence the expected workload (computational complexity) $W(\hat{\gamma})$. Then the efficiency is defined as
\[
\text{Eff}(\hat{\gamma}) := (\text{Var}(\hat{\gamma})W(\hat{\gamma}))^{-1}.
\]
2.2. Setting and definitions

As before, one would like to have (see L’Ecuyer et al. (2007), page 3)

\[ \liminf_{\gamma \to 0} \gamma^2 \text{Eff}(\hat{\gamma}) > 0 \]  
\[ (2.1) \]

or

\[ \lim_{\gamma \to 0^+} \frac{\log(\text{Eff}(\hat{\gamma}))}{\log(\gamma)} = -2. \]  
\[ (2.2) \]

Unbiased estimators with this property are sometimes called work-normalized asymptotically efficient.

2.2.2 Entrance probabilities of Markov chains

We consider the following general framework. Let \( \mathcal{X} = \{X_t, t \geq 0\} \) be a strong Markov process on a state space \( \mathcal{E} \), where the time \( t \) can be continuous or discrete. In the continuous case, we assume that all the trajectories are right-continuous with left-hand limits (cadlag). Let \( B \) be a subset of \( \mathcal{E} \), \( x_0 \not\in B \) a starting point, \( \tau_B = \inf\{t > 0 : X_t \in B\} \) the first entrance time into \( B \) and \( \zeta \) another almost-surely finite stopping time. The rare event is then \( \tau_B < \zeta \), that is, we want to estimate

\[ \gamma = P[\tau_B < \zeta | X_0 = x_0]. \]

This problem can be reduced to the case where \( \zeta = \tau_A = \inf\{t > 0 : X_t \in A\} \) is the first entrance time in some set \( A \) by the process \( \mathcal{X} \), if we put enough information in the state, so that we are able to decide if \( \tau_B < \zeta \) has happened or not at any given state \( X_t \), see e.g. L’Ecuyer et al. (2009). For example if we want to calculate the first entrance time of a strong Markov process \( \mathcal{Y} = \{Y_t, t \geq 0\} \) with starting point \( y_0 \) into a set \( C \) before some deterministic time \( T \), we set

\[ X_t = (t, Y_t), \ A = [T, \infty) \times \mathcal{E}, \ B = [0, \infty) \times C \]  
\[ (2.3) \]

and \( x_0 = (0, y_0) \). So we will assume in the following that \( \zeta = \tau_A \) has the form of a entrance time and the Markov process contains the time \( t \) when necessary.
Sometimes, one is interested additionally in the conditional distribution of \(\{X_t, 0 \leq t \leq \tau_B\}\) given the rare event \(\{\tau_B < \tau_A\}\) has occurred, see e.g. Weinan and Vanden-Eijnden (2006) and Metzner et al. (2006). Although this can be estimated by the splitting method (see e.g. Del Moral and Garnier (2005)), it needs usually bigger sample sizes and we do not consider it here.

### 2.2.3 Importance splitting

Although importance sampling can be used to estimate the probability \(\gamma\), it is usually difficult to make a good choice for the proposal distribution, that is, how to modify the transition probabilities and/or jump rates of the chain such that the rare event occurs more easily, but the variance of the likelihood ratio does not become huge. Importance splitting (see e.g. L’Ecuyer et al. (2006), L’Ecuyer et al. (2007) or Garvels (2000)) is an appealing alternative.

The key ingredient for importance splitting is the importance function

\[
\Phi : \mathcal{E} \rightarrow \mathbb{R},
\]

which assigns importance values to the chain states. We assume \(B = \{x \in \mathcal{E} : \Phi(x) \geq L\}\) and \(\Phi(x_0) < L\). If we split the interval \([\Phi(x_0), L]\) into \(m\) subintervals with boundaries \(\Phi(x_0) = l_0 < l_1 < \cdots < l_m = L\), set \(T_k = \inf\{t \geq 0 : \Phi(X_t) \geq l_k\}\) and \(D_k = \{T_k < \tau_A\}\), we can decompose \(\gamma = P[D_m]\) into the conditional probabilities \(p_k = P[D_k]/P[D_{k-1}]\), i.e.,

\[
\gamma = \prod_{k=1}^{m} P[D_k|D_{k-1}] = \prod_{k=1}^{m} p_k,
\]

because \(D_m \subset D_{m-1} \subset \cdots \subset D_1 \subset D_0\) and \(P[D_0] = 1\) (assuming \(x_0 \notin A\)).

The estimation of the conditional probabilities \(p_k\) can now be done in principle by starting independent chains from states generated according to the distribution of \(X_{T_{k-1}}\) conditional on \(D_{k-1}\), called the
entrance distribution at threshold $l_{k-1}$ and denoted by $G_{k-1}$. Those chains where $D_k$ occurs provide also entrance states $X^i_{T_k}$ on level $k$, and their empirical distribution $\hat{G}_k$ is an estimate of the entrance distribution $G_k$. Thus we can proceed recursively, replacing $G_{k-1}$ with $\hat{G}_{k-1}$.

In actual implementations, the importance splitting algorithm works as follows.

- At the beginning, we choose $N_0 \in \mathbb{N} \setminus \{0\}$ and set $k = 0$, $R_0 = N_0$ and $T_0^i = 0$, $X_0^i = x_0$ for $i = 1, 2, \ldots, R_0$.

- At stage $k$ ($0 \leq k < m$), we have a sample $\{X^1_k, \ldots, X^{R_k}_k\}$ (writing $X^i_k$ for $X^i_{T_k}$) from $G_k$. From this, we resample $N_k$ entrance values (three resampling schemes normally used are described below) and simulate independent chains from there up to time $\min\{T_{k+1}, \tau_A\}$. The number of chains where $D_{k+1}$ occurs is denoted by $R_{k+1}$. If $R_{k+1} = 0$, we set $\hat{\gamma} = 0$ and stop the algorithm. Otherwise we set $\hat{p}_{k+1} = R_{k+1}/N_k$, take the $R_{k+1}$ values $X^i_{k+1}$ of the chains where $D_{k+1}$ occurred as the sample from $G_{k+1}$ and increase $k$ by one.

- If all $R_k > 0$, we estimate $\gamma$ by the product of the $\hat{p}_k$'s:

$$\hat{\gamma} := \prod_{k=1}^m \hat{p}_k = \prod_{k=1}^m \frac{R_k}{N_{k-1}}.$$  

Depending on how $N_k$ is chosen and how the resampling of entrance values is done, we obtain different variants of the algorithm (see L’Ecuyer et al. (2007), Garvels (2000), Dean and Dupuis (2009) and Del Moral and Garnier (2005)):

**Fixed splitting.** We resample each of the $R_k$ entrance states exactly $c_k \geq 1$ times where $c_k$ is fixed. Then $N_k = c_k R_k$ and both $N_k$ and $R_{k+1}$ are random. Note that with fixed splitting we can simulate each chain forward up to $\min\{\tau_B, \tau_A\}$, independently of
all other chains. We simply have to store the entrance value each time we cross a level, and later we have to start additional $c_k - 1$ independent chains at these entrance values. Because of this, fixed splitting is sometimes also called the global step approach (see Garvels (2000), pages 43 ff.).

**Fixed effort with random selection.** Here $N_k$ is fixed and resampling is done by random sampling with replacement, that is, we draw independently $N_k$ states at random from $\{X_k^1, \ldots, X_k^{R_k}\}$. Then $R_{k+1}$ and the number of times an entrance state is resampled are random.

**Fixed effort with balanced selection.** Here $N_k$ is again fixed, but resampling is done in a stratified fashion so that the variance is minimized: Let $c_k = \lfloor N_k / R_k \rfloor$ and $d_k = N_k \mod R_k$. Select $d_k$ states by sampling without replacement and resample these $c_k + 1$ times, and those not selected $c_k$ times.

We propose here a new variant, called **fixed number of successes**, where $R_{k+1} \geq 2$ is fixed. In this variant, we repeat independently the process of selecting an entrance state at random and simulating the chain starting from the selected state up to $\min\{T_k+1, \tau_A\}$ until we have observed $R_{k+1}$ chains where $D_{k+1}$ has occurred. We denote the number of failures, that is, the number of chains where $D_{k+1}$ did not occur, by $Z_{k+1}$. Hence the total number $N_k$ of chains generated at stage $k$ is equal to $R_{k+1} + Z_{k+1}$. In this variant we estimate $\gamma$ by

$$\tilde{\gamma} := \prod_{k=1}^{m} \tilde{p}_k \quad \text{where} \quad \tilde{p}_k = \frac{R_k - 1}{N_{k-1} - 1} = \frac{R_k - 1}{R_k + Z_k - 1}.$$ 

The reason for this choice of $\tilde{p}_k$ is that $(r-1)/(r+Z-1)$ is an unbiased estimator for $p$ if $Z \sim \text{NegBin}(r, p)$ and $r \geq 2$ (see Lemma 2.2). Moreover, it is easy to check that it is the only unbiased estimator.

The advantages of this new variant will be discussed in detail in the following sections. Intuitively it is clear that by choosing $R_{k+1}$ we control the imprecision of the estimator and adjust the computational
2.3 Unbiasedness of the importance splitting estimators

If we choose $N_k$, we control the effort and accept whatever imprecision we obtain. In particular, if the fixed effort produces a value $R_{k+1} = 0$ we have a useless estimator $\hat{\gamma} = 0$. In this situation it is tempting to increase $N_k$ which then comes very close to the variant with a fixed number of successes.

### 2.3 Unbiasedness of the importance splitting estimators

The original proof on pages 16 and 17 of Garvels (2000) for the unbiasedness of the importance splitting estimator is based on the claim that

$$E[\hat{p}_k|N_0, R_1, N_1, R_2, \ldots, R_{k-2}, N_{k-1}, R_{k-1}] = p_k.$$  

(2.4)

L’Ecuyer et al. (2006), page 138 or L’Ecuyer et al. (2007), page 7, refer to Garvels (2000) and give a short argument which is not clear unless one assumes (2.4).

However, (2.4) is not correct in general. Note that $\hat{p}_k = 0$ if any of the $R_i = 0$ for $i < k$. Hence it can hold at most in cases where $R_i > 0$ for all $i < k$. We give a counterexample in the Appendix 2.A, showing that $E[\hat{p}_k|N_i, R_i(i < k)]$ depends in general on the conditioning variables and that $E[\hat{p}_k] \neq p_k$, that is, $\hat{p}_k$ is not an unbiased estimator for $p_k$.

Nevertheless, unbiasedness of $\hat{\gamma}$, that is $E[\prod_{k=1}^m \hat{p}_k] = \prod_{k=1}^m p_k$, holds for all the variants described above, in particular also for the new variant with fixed number of successes. We give proofs circumventing claim (2.4) below. For fixed splitting, such a proof can also be found in Dean and Dupuis (2009) or Asmussen and Glynn (2007), and for fixed effort with random selection in Del Moral and Garnier (2005) or Asmussen and Glynn (2007).

**Proposition 2.1.** The importance splitting estimator is unbiased for each of the four following variants: Fixed splitting, fixed effort with random selection, fixed effort with balanced selection and fixed number
of successes.

Proof. First we introduce some notation. For \( k = 0, 1, \ldots, m \), we consider the stochastic process

\[
U_k = X_{T_k} (T_k < \tau_A)
\]

\[
= \Delta (T_k \geq \tau_A)
\]

with state space \( \bar{\mathcal{E}} = \mathcal{E} \cup \{\Delta\} \) where \( \Delta \) is an additional absorbing state. By the strong Markov property, \( \mathcal{U} = \{U_k : k \in \mathbb{N}_0\} \) is a Markov chain. We denote by \( M_k \) the transition kernel of this chain at step \( k \):

\[
M_k(u, A) = \mathbb{P}[U_k \in A | U_{k-1} = u]
\]

for any (measurable) subset \( A \) of \( \bar{\mathcal{E}} \) and any \( u \in \bar{\mathcal{E}} \) that can occur as possible value of \( U_{k-1} \). Note that because \( \Delta \) is absorbing, \( M_k(\Delta, \cdot) \) is a point mass at the point \( \Delta \).

Moreover, we introduce the transition kernels for several steps

\[
M_{k:l}(u, A) = \mathbb{P}[U_l \in A | U_k = u] \quad (0 \leq k < l \leq m).
\]

In particular, we then have

\[
\mathbb{P}[D_k] = 1 - M_{0:k}(x_0, \{\Delta\})
\]

By definition \( M_{k-1:k} = M_k \), and by the Markov property of \( \mathcal{U} \) we obtain for \( l > k \)

\[
M_{k-1:l}(u, A) = \int_{\mathcal{E}} M_k(u, du')M_{k:l}(u', A)du' + M_k(u, \{\Delta\})1_A(\Delta).
\]

(2.5)

For \( \tilde{\gamma} \), we denote by \( \mathcal{F}_k \) the \( \sigma \)-algebra generated by \( (Z_k, X^1_k, \ldots, X^{r_k}_k) \) for \( 0 \leq k \leq m \) where \( X^i_k = X^i_{T^l_k} \). By definition, \( r_0 = 1, Z_0 = 0, X^1_0 = x_0 \) and \( T^1_0 = 0 \). Then the claim follows from Proposition 2.2
2.3. Unbiasedness of the importance splitting estimators

below if we put $k = 1$.

For $\hat{\gamma}$, the argument can be given simultaneously for fixed effort with random or balanced selection or fixed effort with balanced selection and for fixed splitting. We define in this case $\mathcal{F}_k$ to be the $\sigma$-algebra generated by $(R_k, X^1_k, \ldots, X^R_k)$ for $0 \leq k \leq m$ where $R_0 \equiv 1$. For fixed effort $N_k \equiv n_k$ and for fixed splitting $N_k = c_k R_k$ with $c_0 = 1$. Then the claim follows from Proposition 2.3 below, again by putting $k = 1$.

We state and prove next two lemmas.

**Lemma 2.1.** Given $\mathcal{F}_{k-1}$, $Z_k$ has a Negative Binomial distribution with success parameter

$$\pi_{k-1} = 1 - \frac{1}{r_{k-1}} \sum_{i=1}^{r_{k-1}} M_k(X^i_{k-1}, \{\Delta\})$$

and for any $1 \leq i \leq r_k$ and any $A \subset \bar{E}$

$$P[X^i_k \in A | \mathcal{F}_{k-1}] = \frac{1}{r_{k-1} \pi_{k-1}} \sum_{i=1}^{r_{k-1}} M_k(X^i_{k-1}, A).$$

Moreover, all these variables are conditionally independent given $\mathcal{F}_{k-1}$.

**Proof.** Note that by the definition of the algorithm, at step $k$ we generate independent replicates from

$$\frac{1}{r_{k-1}} \sum_{i=1}^{r_{k-1}} M_k(X^i_{k-1}, .)$$

until we have obtained $r_k$ values different from $\Delta$ which then become the $X^i_k$. From this, the statement is obvious. □

**Lemma 2.2.**

$$\sum_{z=0}^{\infty} \frac{r-1}{r+z-1} \binom{r+z-1}{r-1} (1-p)^z = p^{1-r}.$$

Proof. The result follows directly from a Taylor series expansion of $p^{1-r}$ at $p = 1$. 

The next proposition is used for unbiasedness of $\tilde{\gamma}$.

Proposition 2.2.

\[
E \left[ \prod_{l=k}^{m} \tilde{p}_l \bigg| \mathcal{F}_{k-1} \right] = \frac{1}{r_k - 1} \sum_{j=1}^{r_k - 1} (1 - M_{k-1;m}(X^j_{k-1}, \{\Delta\}))
\]

for $1 \leq k \leq m$.

Proof (by induction). We start with $k = m$. Above the equality signs we indicate which lemma (number without brackets) or which formula (number with brackets) is used. We have

\[
E [\tilde{p}_m | \mathcal{F}_{m-1}] = E \left[ \frac{r_m - 1}{r_m + Z_{m-1} - 1} \bigg| \mathcal{F}_{m-1} \right]
\]

\[
\overset{2.1}{=} \sum_{z=0}^{2.1} \frac{r_m - 1}{r_m + z - 1} (1 - \pi_{m-1})^z \pi_m^{r_m} \overset{2.2}{=} \pi_m - 1.
\]
Next, we go from $k + 1$ to $k$:

$$E \left[ \prod_{l=k}^{m} \tilde{p}_l \left| \mathcal{F}_{k-1} \right. \right]$$

$$= E \left[ \tilde{p}_k E \left[ \prod_{l=k+1}^{m} \tilde{p}_l \left| \mathcal{F}_{k} \right. \right] \left| \mathcal{F}_{k-1} \right. \right]$$

$$= E \left[ \frac{r_k - 1}{r_k + Z_k - 1} \frac{1}{r_k} \sum_{i=1}^{r_k} (1 - M_{k:m}(X^i_k, \{\Delta\})) \left| \mathcal{F}_{k-1} \right. \right]$$

$$= \frac{r_k - 1}{r_k + Z_k - 1} E \left[ 1 - M_{k:m}(X^1_k, \{\Delta\}) \left| \mathcal{F}_{k-1} \right. \right]$$

$$= \pi_{k-1} \left( 1 - \frac{1}{r_k - 1} \sum_{j=1}^{r_k-1} \int_{\mathcal{E}} M_k(X^j_{k-1}, du) M_{k:m}(u, \{\Delta\}) du \right)$$

$$= \pi_{k-1} - \frac{1}{r_k - 1} \sum_{j=1}^{r_k-1} (M_{k-1:m}(X^j_{k-1}, \{\Delta\}) - M_k(X^j_{k-1}, \{\Delta\}))$$

$$= \pi_{k-1} - \frac{1}{r_k - 1} \sum_{j=1}^{r_k-1} M_{k-1:m}(X^j_{k-1}, \{\Delta\}) + 1 - \pi_{k-1}.$$

Finally, the next lemma and proposition are used for unbiasedness of $\hat{\gamma}$.

**Lemma 2.3.** For any $h : \mathcal{E} \to \mathbb{R}_+$,

$$E \left[ \sum_{i=1}^{R_k} h(X^i_k) \left| \mathcal{F}_{k-1} \right. \right] = \frac{N_{k-1}}{R_{k-1}} \sum_{i=1}^{R_{k-1}} \int_{\mathcal{E}} h(u) M_k(X^i_{k-1}, du).$$

**Proof.** In step $k - 1$ of the algorithm, we select first $N_{k-1}$ starting values $U^i_{k-1}$ with probabilities $\omega_{ij} = P[U^i_{k-1} = X^j_{k-1}]$ that depend on
the version we are using. For all versions however we have

\[ \sum_{i=1}^{N_{k-1}} \omega_{ij} = \frac{N_{k-1}}{R_{k-1}} \]

because the left-hand-side is the expected number of times the \( j \)-th entrance state is chosen. Given \( U^i_{k-1} \), we generate \( U^i_k \) according to \( M_k(U^i_{k-1}, \cdot) \) independently for different \( i \) and retain those \( U^i_k \) that are not equal to \( \Delta \). Hence if we define \( h(\Delta) = 0 \), we have

\[ \sum_{i=1}^{R_k} h(X^i_k) = \sum_{i=1}^{N_{k-1}} h(U^i_k). \]

Taking expectations, we obtain

\[
E \left[ \sum_{i=1}^{R_k} h(X^i_k) \bigg| \mathcal{F}_{k-1} \right] = \sum_{i=1}^{N_{k-1}} \sum_{j=1}^{R_{k-1}} \omega_{ij} \int_{\mathcal{E}} h(u)M_k(X^j_{k-1}, du) \\
= \frac{N_{k-1}}{R_{k-1}} \sum_{j=1}^{R_{k-1}} \int_{\mathcal{E}} h(u)M_k(X^j_{k-1}, du).
\]

\[ \square \]

**Proposition 2.3.** For \( 1 \leq k \leq m \), if \( R_{k-1} \neq 0 \)

\[
E \left[ \prod_{l=k}^{m} \hat{p}_l \bigg| \mathcal{F}_{k-1} \right] = \frac{1}{R_{k-1}} \sum_{i=1}^{R_{k-1}} (1 - M_{k-1;m}((T^i_{k-1}, X^i_{k-1}), \Delta))
\]

If \( R_{k-1} = 0 \), \( \prod_{l=k}^{m} \hat{p}_l = 0 \) by definition.

*Proof (by induction).* For \( k = m \), the proposition follows from Lemma 2.3 with \( h \equiv 1 \) on \( \mathcal{E} \). We go next from \( k + 1 \) to \( k \). Again, the number above the equality signs indicate which lemma (number without brackets) or formula (number with brackets) is used.
2.3. Unbiasedness of the importance splitting estimators

\[
E \left[ \prod_{l=k}^{m} \hat{p}_l \mid \mathcal{F}_{k-1} \right] \\
= E \left[ \hat{p}_k E \left[ \prod_{l=k+1}^{m} \hat{p}_l \mid \mathcal{F}_k \right] \mid \mathcal{F}_{k-1} \right] \\
= E \left[ \frac{R_k}{N_{k-1}} \frac{1}{R_k} \sum_{i=1}^{R_k} \left( 1 - M_k: m(X^i_k, \Delta) \right) \mid \mathcal{F}_{k-1} \right] \\
= \frac{1}{N_{k-1}} E \left[ \sum_{i=1}^{R_k} \left( 1 - M_k: m(X^i_k, \Delta) \right) \mid \mathcal{F}_{k-1} \right] \\
\overset{(2.3)}{=} \frac{1}{R_{k-1}} \sum_{j=1}^{R_{k-1}} \left( 1 - M_k(X^j_{k-1}, \Delta) - \int_{\mathcal{E}} M_k: m(u, \Delta)M_k(X^j_{k-1}, du) \right) \\
\overset{(2.5)}{=} \frac{1}{R_{k-1}} \sum_{j=1}^{R_{k-1}} \left( 1 - M_{k-1} : m(X^j_{k-1}, \Delta) \right).
\]

\[ \square \]

2.3.1 Balanced selection for splitting with fixed number of successes

For the version with fixed number of successes, there is also an analogue of balanced selection: Instead of choosing an entrance state at random independently at each repetition, one can use a random permutation of the available $R_k$ entrance states which is repeated until $R_{k+1}$ successes have been observed. In order to analyze this version, we look at the following situation: At some level, we have $r_0$ different entrance states with different probabilities to reach the next level. The average of these probabilities is denoted by $p$. We start an infinite number of independent chains, selecting the initial state either at random or by repeating a random permutation of the $r_0$ states and let $S_n$ and $S_n^*$, respectively, be the number of chains that reach the next level among
the first $n$ chains. Because $S_n^s$ has the same expectation, but a smaller variance than $S_n$ and both random variables are asymptotically normal, we expect that $P[S_n > k] > P[S_n^s > k]$ for $k > E[S_n] = np$ and $P[S_n < k] > P[S_n^s < k]$ for $k < np$. Hence if $Z = \min\{n; S_n = r\} - r$ and $Z^s = \min\{n; S_n^s = r\} - r$, then

$$P\left[\frac{r-1}{r+Z-1} < \frac{r-1}{r-1+j}\right] = P[Z > j] = P[S_{r+j} < r].$$

Moreover $r < (r+j)p$ and $r-1 < (r-1+j)p$ are almost equivalent, and thus we expect $P[\frac{r-1}{r+Z-1} < c] > P[\frac{r-1}{r+Z^s-1} < c]$ for $c < p$. For $c > p$, an analogous argument applies. Therefore, we expect that the distribution of the balanced estimator is more concentrated around the correct value.

Explicit computations are possible for $r_0 = 2$ and $r \geq 2$. We show that in this case $\frac{r-1}{r+Z^s-1}$ is no longer unbiased, but has a smaller mean square error than $\frac{r-1}{r+Z-1}$. Let $\alpha$ and $\beta$ be the probabilities to reach the next level from the two entrance states, so $p = 1/2 \cdot (\alpha + \beta)$. Then $Z \sim \text{NegBin}(r, p)$ and the distribution of $Z^s$ is given by

$$f_{Z^s}(z; r, \alpha, \beta) = P[Z^s = z] = \frac{1}{2}(f(z; r, \alpha, \beta) + f(z; r, \beta, \alpha)),$$

where $f(z; r, \alpha, \beta)$ is equal to

$$g(z; r, \alpha, \beta) = \sum_{s=\max\{0, r-1-t\}}^{\min\{r-1, u\}} \left\{ \binom{t}{u} \binom{t}{r-1-s} \times \alpha^s \beta^{r-1-s} (1-\alpha)^{u-s} (1-\beta)^{t-(r-1-s)} \right\}$$

with $t = \lfloor (r-1+z)/2 \rfloor$, $u = \lceil (r-1+z)/2 \rceil$ and

$$g(z; r, \alpha, \beta) = \begin{cases} \alpha & (r+z) \mod 2 = 0 \\ \beta & (r+z) \mod 2 = 1 \end{cases}.$$
For $\alpha \to 0$ and $\beta \to 1$, $P[Z^s = r]$ and $P[Z^s = r - 1]$ converge both to $\frac{1}{2}$ and thus

$$E[(r - 1)/(r - 1 + Z^s)] \to \frac{1}{2} \left( \frac{r - 1}{2r - 1} + \frac{r - 1}{2r - 2} \right) < \frac{1}{2}. $$

Hence there is a bias, but it is small for all practical situations.

The expected squared error for given $\alpha$, $\beta$ and $r$ can be computed numerically in both cases, summing over a large enough range of $z$-values.

In Figure 2.1, we show a perspective and a contour plot of the relative MSE

$$\frac{E[((r - 1)/(r - 1 + Z^s) - p)^2] - E[((r - 1)/(r - 1 + Z) - p)^2]}{E[((r - 1)/(r - 1 + Z) - p)^2]}$$

in dependence of $\alpha$ and $\beta$ for $r = 5$ (for other $r$, one yields qualitatively the same result). All the values are non-negative, so in terms of the expected squared error, it is better in this particular setting to use the biased estimator $\hat{p}^s = (r - 1)/(r - 1 + Z^s)$, especially when the difference between $\alpha$ and $\beta$ is big.

### 2.4 The optimal importance function

It is known that the efficiency of importance splitting depends strongly on the choice of the importance function (see e.g. Garvels (2000), L’Ecuyer et al. (2007), Glasserman et al. (1999) and L’Ecuyer et al. (2009)). Dean and Dupuis (2009) derive importance functions assuming that the probability of interest $\gamma$ satisfies a large deviation scaling. Garvels et al. (2002) derive that to minimize the residual variance of the splitting estimator from the current stage onward, the importance function should be chosen so that for all $k$

$$P[D_k|D_{k-1}, X_{k-1}] \overset{a.s.}{=} p_k. \quad (2.6)$$
In Section 2.5, we derive under this assumption and some conditions on the expected workload that the fixed number of successes estimator with the optimal choice of levels and optimal fixed numbers of successes is asymptotically efficient in the sense of (2.2). This also supports this choice of the importance function. It can be achieved with

\[ \Phi(x) = h(P[\tau_B < \tau_A|X_0 = x]) \]  \hspace{1cm} (2.7)

for some monotone increasing function \( h : [0, 1] \rightarrow \mathbb{R} \), assuming that the process \( X \) can not jump over levels, that is, \( \Phi(X_k) = l_k \). If we choose the above importance function (2.7) and, without loss of generality, \( h \) equals the identity, we have

\[ P[D_m|D_k, X_k] = \Phi(X_k) \geq l_k \]

and almost surely

\[ l_k = P[D_m|D_k, X_k] = E[P[D_m|D_{k+1}, X_{k+1}]|D_k, X_k] = l_{k+1}P[D_{k+1}|D_k, X_k], \]
2.4. The optimal importance function

which implies (2.6). We remark that the choice of $g$ does not matter if we allow the levels to be arbitrary. Of course, we can not use (2.7) in applications, because it requires the knowledge of the quantity of interest. Nevertheless, this suggests looking for an importance function which is a good approximation to (2.7). This is very similar to importance sampling, where the optimal proposal distribution also requires the knowledge of the quantity of interest and one tries to find a suitable approximation (see e.g. Robert and Casella (2004), pages 90 ff.).

In the remaining part of this section, we provide approximations to the optimal importance function for two settings which occur very often in applications. In the examples in Section 2.6, we illustrate that they yield good results, especially in combination with the two-stage fixed number of successes estimator introduced in Section 2.5.4.

2.4.1 Approximation for homogeneous Markov processes on countable state spaces

In the case where the state space $E$ is discrete and the Markov process $\mathcal{X}$ is homogeneous (in discrete or continuous time) with transition probabilities $p(x, \bar{x})$ (of the embedded chain in the continuous-time case) from the state $x$ to $\bar{x}$, we propose the following approximation of $P[\tau_B < \tau_A | X_0 = x]$, which has also been used by L’Ecuyer and Tuffin (2007) in the context of importance sampling. We introduce the modified state space $\mathcal{E}' = (A \cup B)^c \cup \beta$ where $\beta$ is an additional state which replaces all states in $B$ and for $x \in (A \cup B)^c$ we define $p(x, \beta) = \sum_{z \in B} p(x, z)$. Then we set

$$\Phi(x) = \begin{cases} 
0 & (x \in A) \\
1 & (x \in B) \\
\max \prod_{i=1}^{n} p(x_{i-1}, x_i) & (x \in (A \cup B)^c).
\end{cases}$$

(2.8)

Here, the maximum is taken over all $n$ and all paths in $\mathcal{E}'$ with $x_0 = x$, $x_i \neq \beta$ for $1 \leq i < n$ and $x_n = \beta$. This simply means that we take as importance function the probability of the most likely path from $x$
Chapter 2. Splitting Simulation

to $B$ without entering $A$. This importance function can be computed efficiently by Dijkstra’s algorithm (see Dijkstra (1959)): $-\log \Phi(x)$ for $x \in (A \cup B)^c$ is the length of the shortest path from $\beta$ to $x$ in the graph with set of vertices $\mathcal{E}'$ where the length of the directed edge from $\bar{x}$ to $x$ is set to $-\log(p(x, \bar{x}))$. If $\mathcal{E}'$ is infinite, we have in addition to reduce the state space or modify the transitions suitably.

2.4.2 Approximation when estimating entrance probabilities of homogeneous Markov jump processes with a finite time horizon

We now consider a homogeneous continuous-time Markov chain $Y$ on a discrete state space with generator $q$. We let $\lambda(y) = -q(y, y)$ be the jump rate and $p(y, \bar{y}) = -q(y, \bar{y})/q(y, y)$ the transition probabilities of the embedded chain. The optimal importance function for estimating the probability that $Y$ enters $C$ before the deterministic time $T$ is based on the general discussion above with $X$, $A$ and $B$ as in (2.3) –

$$\Phi(t, y) = P[X_{\tau_B < \tau_A} = (t, y)] = P[Y_{\tau_C < T - t} = y].$$

The first idea is to discretize time by considering the following time-homogeneous chain $X^d$ with

$$P[X^d_\delta = ((k + 1)\delta, y)|X^d_0 = (k, \delta, y)] = \exp(-\delta \lambda(y)),$$
$$P[X^d_\delta = ((k + 1)\delta, \bar{y})|X^d_0 = (k, \delta, y)] = p(y, \bar{y})(1 - \exp(-\delta \lambda(y))) \quad (\bar{y} \neq y)$$

and to use the same approximation as above, i.e., to compute the probability of the most likely path of this chain that goes from $(k\delta, y)$ to $B$ without hitting $A$. This however has two disadvantages: The number of states we have to consider is large, and the value depends strongly on the chosen time step $\delta$. To see why, note first that the most likely path does not include transitions where the second component of a state $x$ stays fixed since performing such steps only decreases the probability.
2.4. The optimal importance function

Hence this approach leads to the maximum of

\[
\prod_{i=1}^{m} p(y_{i-1}, y_i)(1 - \exp(-\delta \lambda(y_{i-1})))
\]

over all paths from \( y_0 = y \) to \( C \) subject to the restriction \((k + m)\delta < T\). If \( \delta \) is small, then this restriction is not active except when \( k\delta \) is very close to \( T \), and thus the importance function is almost time-independent. On the other hand, when \( \delta \) is large, the restriction \((k + m)\delta < T\) cannot be satisfied and \( \Phi \) becomes zero except for very small \( k\delta \). In our experience, finding a good value for \( \delta \) is a hard problem.

We suggest instead the following approximation

\[
\Phi(t, y) = \Psi(y)G(y, T - t)
\]

where \( \Psi(y) \) measures how easily the embedded chain can go from \( y \) to \( C \) and \( G(y, s) \) approximates the probability that the time of the fastest transition from \( y \) to \( C \) is less than \( s \). More precisely, \( \Psi \) is essentially the approximation (2.8) from above:

\[
\Psi(y) = \max \left\{ \prod_{i=1}^{m} p(y_{i-1}, y_i); y_0 = y, m \in \mathbb{N}, y_m \in C \right\}.
\]

In order to define \( G \), we introduce the smallest expected time of a path leading from \( y \) to \( C \),

\[
\mu(y) = \min \left\{ \sum_{i=1}^{m} \frac{1}{\lambda(y_{i-1})}; y_0 = y, m \in \mathbb{N}, y_m \in C \right\},
\]

and let \( m(y) \) be the number of time steps in the optimal path. Then \( G(y, \cdot) \) is the cumulative distribution function of the Gamma distribution with shape parameter \( m(y) \) and mean \( \mu(y) \). This means that we approximate the sum of \( m(y) \) exponential(\( \lambda_i \))-variables by a Gamma variable.

The multiplicative form of \( \Phi \) is justified by the fact that the waiting
time until the next jump and the position after this jump are independent. The approximation will not be very accurate in cases where the optimal paths behind $\Psi$ and $G$ are very different. However, this is compensated by computational advantages: All of $-\log(\Psi(.))$, $\mu(.)$ and $m(.)$ can be computed recursively with Dijkstra’s algorithm for the shortest path lengths from a given vertex in a graph to all other vertices as above.

De Boer et al. (2007) use a similar approximation in the context of importance sampling for the case where the jump rates converge to zero, that is the system slows down more and more and thus the entrance probability before some fixed time becomes smaller. This leads to a different behaviour in the time component of the importance function: In contrast to $G(y, T-t)$, their approximation is convex in the remaining time $T-t$.

2.5 Theoretical analysis in a simplified setting

Even for a fixed importance function, it is difficult to compare different versions of importance splitting or different strategies for choosing the levels. Some progress can be made under the simplifying assumption (2.6) which is appropriate if we choose an importance function that is a good approximation of (2.7) and the process does not make large jumps. This will be confirmed by the examples in Section 2.6.

So in this section, we work under the assumption (2.6). Then with the same arguments as in the proof of Proposition 2.1, one can easily see that for fixed efforts $n_1, \ldots, n_m$, the corresponding numbers of successes $R_1, \ldots, R_m$ are independent and $R_k \sim \text{Bin}(n_k, p_k)$. This holds both under random and balanced selection. Similarly, for fixed numbers of successes $r_1, \ldots, r_m$, the numbers of failures $Z_1, \ldots, Z_m$ are independent and $Z_k \sim \text{NegBin}(r_k, p_k)$. This independence is transferred to $\tilde{p}_1, \ldots, \tilde{p}_m$ and $\hat{p}_1, \ldots, \hat{p}_m$, respectively. In the literature, the
latter assumption is typically made at the outset. We believe that in most cases condition (2.6) is not only sufficient, but also necessary for the independence of the $\tilde{p}_i$ and $\hat{p}_i$.

Using the independence of $\hat{p}_1, \ldots, \hat{p}_m$, we obtain

$$\frac{\text{Var}(\hat{\gamma})}{\gamma^2} = \frac{\text{Var}(\prod_{k=1}^m \hat{p}_k)}{\prod_{k=1}^m p_k^2} = \prod_{k=1}^m \left( \frac{\text{Var}(\hat{p}_k)}{p_k^2} + 1 \right) - 1$$

$$= \prod_{k=1}^m \left( \frac{1}{n_k p_k} + 1 \right) - 1.$$

For $\text{Var}(\tilde{\gamma})$ the situation is slightly more complicated because there are no simple formulae for $\text{Var}(\tilde{p}_k)$. For our purposes, the following upper and lower bounds given in Mikulski and Smith (1976) for $r_k \geq 3$ are sufficient:

$$p_k^2(1 - p_k)/r_k \leq \text{Var}(\tilde{p}_k) \leq p_k^2(1 - p_k)/(r_k - 2).$$

The lower bound is the Cramer-Rao lower bound. So we get for the squared relative error

$$\prod_{k=1}^m \left( \frac{1}{r_k} + 1 \right) - 1 \leq \frac{\text{Var}(\tilde{\gamma})}{\gamma^2} \leq \prod_{k=1}^m \left( \frac{1}{r_k - 2} + 1 \right) - 1$$

$$\leq \prod_{k=1}^m \left( \frac{1}{r_k - 2} + 1 \right) - 1. \quad (2.11)$$

The ratio of the upper and the lower bound converges to 1 if all $r_k$’s go to infinity.

Our analysis also requires an expression for the average total workload $W$, the computational complexity of the estimators. We assume that the effort to choose an $X_{T_{i-1}}^i$, to simulate the chain until we can decide whether $D_k$ occurred or not and to store the state if needed depends
on $k$ and $m$ only. We denote it by $e_{k,m}$. Then

$$W(\hat{\gamma}) = \sum_{k=1}^{m} e_{k,m} n_k$$

and

$$W(\tilde{\gamma}) = \sum_{k=1}^{m} e_{k,m} E(r_k + Z_k) = \sum_{k=1}^{m} e_{k,m} \frac{r_k}{p_k}.$$

In the following, we will give various comparisons of the two estimators, based on the above expressions. A basic distinction can however be made immediately: With $\tilde{\gamma}$ we can control the relative mean square error, regardless of the values of $\gamma$ and $p_k$, and the estimator adjusts the workload automatically. With $\hat{\gamma}$ we can control the workload, regardless of the values of $\gamma$ and $p_k$, and the estimator adjusts the imprecision automatically.

### 2.5.1 Optimal choice of levels

We address here the question of the optimal choice of the number $m$ of levels, the values of the probabilities $p_k$ (which determine the values $l_k$ of intermediate levels) and the values for $r_k$ and $n_k$, respectively. For this we consider the minimization of the total workload to achieve a given bound $q > 0$ for the relative mean square error. We will make the simplifying assumption that the lower bound in (2.11) is exact. Further, for Proposition 2.4 below we assume that $e_{k,m} = e_m$; afterwards we set $e_{k,m} = 1$ for the rest of the Section 2.5.1. Using the parameter $w_k = n_k$ for fixed effort and $w_k = r_k/p_k$ for fixed number of successes, the problem becomes in both cases to minimize

$$W = e_m \sum_{k=1}^{m} w_k$$
with respect to \( m \) and \((p_1, w_1, \ldots, p_m, w_m)\) subject to

\[
\prod_{k=1}^{m} \frac{(1 - p_k)/(w_k p_k) + 1}{p_k} - 1 \leq q, \quad \prod_{k=1}^{m} p_k = \gamma.
\]

There are additional constraints: In general not all values of \( m \) and \( p_k \) are possible and \( n_k \geq 1 \) and \( r_k \geq 3 \) must be integers. We will ignore these constraints except for the obvious \( 0 \leq p_k \leq 1 \) and \( w_k \geq 1 \).

First, we analyze the problem for a fixed number of levels \( m \).

**Proposition 2.4.** The solution to the optimization problem

\[
\begin{align*}
\text{minimize} & \quad W = e_m \sum_{k=1}^{m} w_k \\
\text{with respect to} & \quad p_k, w_k \quad (k = 1, 2, \ldots, m) \\
\text{subject to} & \quad \prod_{k=1}^{m} \frac{(1 - p_k)/(w_k p_k) + 1}{p_k} - 1 \leq q \\
& \quad \prod_{k=1}^{m} p_k = \gamma \\
& \quad 0 \leq p_k \leq 1, \quad (k = 1, 2, \ldots, m) \\
& \quad w_k \geq 1, \quad (k = 1, 2, \ldots, m)
\end{align*}
\]

is of the following form: If \( q \geq 1/\gamma - 1 \), then all \( w_k = 1 \) and the \( p_k \) are arbitrary except for the constraint \( \prod_{k=1}^{m} p_k = \gamma \). Otherwise there is a subset \( \mathcal{I} \) of \( \{1, 2, \ldots, m\} \) such that

\[
p_k = \gamma^{1/(m-|\mathcal{I}|)}, \quad w_k = \frac{1 - \gamma^{1/(m-|\mathcal{I}|)}}{\gamma^{1/(m-|\mathcal{I}|)}((1 + q)^{1/(m-|\mathcal{I}|)} - 1)} \quad (k \notin \mathcal{I}).
\]

The objective value is

\[
W = e_m \left( |\mathcal{I}| + (m - |\mathcal{I}|) \cdot \frac{1 - \gamma^{1/(m-|\mathcal{I}|)}}{\gamma^{1/(m-|\mathcal{I}|)}((1 + q)^{1/(m-|\mathcal{I}|)} - 1)} \right).
\]

**Proof.** See Appendix 2.B.

Intuitively, a value of \( p_k = 1 \) means that two levels coincide so that \( m \) can be reduced by one. In our formulation of the problem, this is only true if \( e_m \) does not depend on \( m \), i.e., we set \( e_m = 1 \). Then we can
indeed restrict ourselves to values of \( m \) for which \( \mathcal{I} = \emptyset \). Thus we can find the optimal \( m \) simply by considering the symmetric solutions with 
\[ p_k = \gamma^{1/m}, \quad w_k = n_k = \frac{1-\gamma^{1/m}}{\gamma^{1/m}(1+q)^{1/m} - 1} \]
and 
\[ r_k = \frac{1-\gamma^{1/m}}{(1+q)^{1/m} - 1}. \]

\[ m_{\text{opt}}(\gamma, q) = \text{argmin}_m \{ W(\gamma, q, m) \} \tag{2.12} \]

where
\[ W(\gamma, q, m) = m \cdot \frac{1 - \gamma^{1/m}}{\gamma^{1/m}(1 + q)^{1/m} - 1}. \]

For given \( \gamma \) and \( q \), \( m_{\text{opt}} \) can be calculated. For more insight, we consider the case where the bound for the imprecision \( q \) tends to zero.

**Proposition 2.5.** If \( e_{k,m} \equiv 1 \), the optimal choice of levels \( m_{\text{opt}} \) is given by (2.12). If \( q \) is small enough, \( m_{\text{opt}} \) is equal to one of the two integers closest to
\[ m_{\text{opt}}^a = c|\log(\gamma)| \]
where \( c \approx 0.6275 \) is the solution of the equation
\[ \exp(1/c) = \frac{2c}{2c - 1}. \tag{2.13} \]

The corresponding optimal probabilities are
\[ p_{\text{opt}}^a = \frac{(2c - 1)}{(2c)} \approx 0.2032, \]
and the number of replicates and successes respectively on each level is
\[ n_k \sim \frac{c}{2c - 1} \frac{|\log(\gamma)|}{q} = 2.46 \frac{|\log(\gamma)|}{q}, \quad r_k \sim \frac{1}{2} \frac{|\log(\gamma)|}{q}. \]

Finally, the optimal total workload is
\[ W_{\text{opt}}(\gamma, q) \sim \frac{c^2}{2c - 1} \frac{\log(\gamma)^2}{q} = 1.544 \frac{\log(\gamma)^2}{q}. \]

**Proof.** See Appendix 2.B. \( \square \)

Another way to approach the problem would be to give a bound for the total average workload, i.e., \( W \leq w_{\text{tot}} \geq 1 \) and minimize the variance \( \text{Var}(\hat{\gamma}) \). For \( e_{k,m} = 1 \) this also leads to the solution \( p_k = \gamma^{1/m} \) for a
2.5. Theoretical analysis in a simplified setting

fixed $m$ (and $w_k = w_{tot}/m$). The optimal number of levels is then the minimizer of

$$
\left( \frac{1 - \gamma^{1/m}}{\gamma^{1/m}(w_{tot}/m)} + 1 \right)^m - 1 \approx \frac{m^2}{w_{tot}} \frac{1 - \gamma^{1/m}}{\gamma^{1/m}}, \tag{2.14}
$$

where the approximation is justified when $(1 - \gamma^{1/m})/\gamma^{1/m} \ll w_{tot}/m$, which leads to the same optimal $m$ as before. In L’Ecuyer et al. (2006), a less accurate approximation is used, that is, one omits the term $-\gamma^{1/m}$ in the numerator in the approximation in (2.14), which leads to $m = -1/2 \log(\gamma)$ and $p_k = 1/e^2 \approx 0.1353$, a rather significant difference to our value 0.2032.

The optimal workload above can be compared with the workload of the crude Monte Carlo estimator which corresponds to $m = 1$. In the case $\gamma = 10^{-6}$ for instance, we obtain a reduction by a factor of roughly 3400. The efficiency of importance splitting with the optimal choice of $m$ and $p_k$ is

$$
\text{Eff}(\hat{\gamma}) = \frac{(2c - 1)/c^2}{(\gamma \log(\gamma))^2} = \frac{0.648}{(\gamma \log(\gamma))^2}.
$$

Hence the method is asymptotically efficient in the sense of (2.2), but (2.1) does not hold. Note that this is also an indirect justification of choosing the importance function as in (2.7). One can easily see that this is true in any case where $m$ grows as $a|\log(\gamma)|$ for some $a > 0$ and $p_k = \gamma^{1/m} = \exp(-1/a)$, because then $\text{Eff}(\hat{\gamma})$ remains proportional to $(\gamma \log(\gamma))^{-2}$, but with factor $(a^2(\exp(1/a) - 1))^{-1}$. In particular, this is also true if $a = 1/2$, the value found with the approximation as in L’Ecuyer et al. (2006).

However, in some applications, there is a maximal number $M$ of levels because of the form of the state space $\mathcal{E}$ or the importance function $\Phi$ (see examples in Section 2.6) and it is impossible to let $m$ grow to $\infty$. So if $\gamma$ is small enough, we can only choose $m = M$. In addition, we rarely can choose the levels such that all $p_k$ are equal. So we discuss
next how to choose \( w_k \) for given \( q, m, p_k \) (and \( e_{k,m} \)). We approximate

\[
\prod_{k=1}^{m} \left( \frac{(1 - p_k)}{(w_k p_k)} + 1 \right) - 1 \approx \sum_{k=1}^{m} \frac{(1 - p_k)}{(w_k p_k)}
\]

which is reasonable when \( \frac{(1 - p_k)}{(w_k p_k)} \ll 1 \), i.e., we have a precise estimator \( q \ll 1 \). So

\[
\text{minimize} \quad W = \sum_{k=1}^{m} w_k e_{k,m} \\
\text{subject to} \quad \sum_{k=1}^{m} \frac{(1 - p_k)}{(w_k p_k)} = q \\
\quad w_k \geq 1, \quad k \in \{1, 2, \ldots, m\}
\]

with respect to \( w_k \).

Using Lagrange multipliers, we find

\[
w_k = n_k = \frac{1}{q} \left( \frac{1 - p_k}{p_k e_{k,m}} \right)^{1/2} \cdot \sum_{i=1}^{m} \left( \frac{(1 - p_i) e_{i,m}}{p_i} \right)^{1/2},
\]

\[
r_k = \frac{1}{q} \left( \frac{1 - p_k}{p_k e_{k,m}} \right)^{1/2} \cdot \sum_{i=1}^{m} \left( \frac{(1 - p_i) e_{i,m}}{p_i} \right)^{1/2} \cdot p_k \quad (2.15)
\]

and the corresponding efficiency is given by

\[
\left( \gamma^2 \cdot \left( \sum_{i=1}^{m} \left( \frac{(1 - p_i) e_{i,m}}{p_i} \right)^{1/2} \right)^2 \right)^{-1}.
\]

An interesting question is how the asymptotical behavior of the splitting estimator looks in this case. So we let \( p_k = p_k(\gamma) \to 0^+ \) as \( \gamma \to 0^+ \) and assume that \( \inf_\gamma \{e_{k,m}(\gamma)\} > 0 \) and \( \sup_\gamma \{e_{k,m}(\gamma)\} < \infty \). Then

\[
\lim_{\gamma \to 0^+} - \frac{\log(\text{Eff}(\gamma))}{\log(\gamma)} = 2 - \max_k \left\{ \lim_{\gamma \to 0^+} \frac{\log(p_k(\gamma))}{\log(\gamma)} \right\}. \quad (2.16)
\]

So asymptotically, the efficiency depends on the \( p_k \) that tend to zero most rapidly. So optimally, because \( \prod_{k=1}^{m} p_k(\gamma) = \gamma \), we would have that every \( p_k(\gamma) \) tends to 0 equally fast, that is, \( p_k(\gamma) = O(\gamma^{1/m}) \). In
this case \( \max_k \left\{ \lim_{\gamma \to 0^+} \frac{\log(p_k(\gamma))}{\log(\gamma)} \right\} = \frac{1}{m} \). Nevertheless, the estimator is no longer efficient. Interestingly, we can achieve the asymptotic performance given in (2.16) using the fixed number of successes estimator with constant \( r_k \geq 3 \) although this is not the optimal choice of \( r_k \) given in (2.15): Because of (2.11) we have that \( 0 < \lim_{\gamma \to 0^+} \frac{\text{Var}(\hat{\gamma})}{\gamma^2} < \infty \) and for the total average workload we get

\[
\lim_{\gamma \to 0^+} \frac{\log(W(\hat{\gamma}))}{\log(\gamma)} = -\max_k \left\{ \lim_{\gamma \to 0^+} \frac{\log(p_k(\gamma))}{\log(\gamma)} \right\}.
\]

Therefore, we expect that the fixed number of successes estimator yields acceptable results when we just choose the number of levels big enough together with constant \( r_k = r \) of moderate size (see also Section 2.6).

To get the same with \( \hat{\gamma} \), we would have to ensure that \( n_k = O(p_k^{-1}(\gamma)) \). For constant \( n_k \), a simple calculation yields

\[
\lim_{\gamma \to 0^+} -\log(\text{Eff}(\hat{\gamma})) / \log(\gamma) = 1,
\]

i.e., the estimator is asymptotically not better than the naive Monte Carlo estimator in this case.

### 2.5.2 Comparison of efficiency

In the situation \( n_k = r_k/p_k \), the workload of \( \hat{\gamma} \) and \( \tilde{\gamma} \) are the same, no matter what values \( e_{k,m} \) are, whereas

\[
\text{Var}(\hat{\gamma})/\gamma^2 = \prod_{k=1}^{m} \frac{((1 - p_k)/(n_k p_k) + 1) - 1}{((1 - p_k)/r_k + 1) - 1} \leq \text{Var}(\tilde{\gamma})/\gamma^2.
\]

Hence in this case, the fixed effort variant is preferable, but the gain is small when the \( r_k \) are big because of (2.11). However, in general we do not know \( p_k \) and thus can not choose \( n_k = r_k/p_k \). It is much
Chapter 2. Splitting Simulation

more relevant to compare the two versions for \( n_k \equiv n \) and \( r_k \equiv r \) and arbitrary values of \( p_k \). In order to simplify expressions, we assume that both \( n \) and \( r \) are large, \( m \) is fixed and \( e_{k,m} \equiv e_m = 1 \). Then

\[
\frac{\text{Var}(\hat{\gamma})}{\gamma^2} \sim \frac{m}{n} \left( \frac{1}{m} \sum_{k=1}^{m} \frac{1}{p_k} - 1 \right), \quad W(\hat{\gamma}) = mn,
\]

and

\[
\frac{\text{Var}(\tilde{\gamma})}{\gamma^2} \sim \frac{m}{r} \left( 1 - \frac{1}{m} \sum_{k=1}^{m} p_k \right), \quad W(\tilde{\gamma}) = r \sum_{k=1}^{m} \frac{1}{p_k}. \quad (2.17)
\]

Hence

\[
\frac{\text{Eff}(\hat{\gamma})}{\text{Eff}(\tilde{\gamma})} = \frac{1 - \frac{1}{m} \sum p_k}{1 - (\frac{1}{m} \sum 1/p_k)^{-1}} \leq 1
\]

because the harmonic mean is less or equal than the arithmetic mean. Moreover, the inequality is strict unless all \( p_k \) are equal.

Let us analyze this a little bit closer. We set \( p_{2l-1} = p^{1+\delta} \) and \( p_{2l} = p^{1-\delta} \) for \( l \in 1, \ldots, L \), \( \delta \in [-1, 1] \) and \( p \in [0, 1] \), so \( m = 2L \sim \log(\gamma)/\log(p) \) and \( p_{2k-1}p_{2k} = p^2 \). The deviation of \( \delta \) from 0 is now a measure for the unbalancedness of the probabilities \( p_k \). In Figure 2.2, we give the contour plots of the asymptotical efficiencies of \( \hat{\gamma} \) and \( \tilde{\gamma} \), respectively, in dependence of \( \delta \) and \( p \) relative to the optimal efficiency when \( \delta = 0 \) and \( p = p_{\text{opt}}^a = 0.2032 \). The number of levels \( m \) is 20, i.e., \( \gamma = 0.2032^{20} = 1.44 \cdot 10^{-14} \). One can clearly see that the variant with fixed number of successes is more robust in terms of deviations from the optimal parameter choice \( \delta = 0 \) and \( p = p_{\text{opt}}^a \). Hence if we do not know the \( p_k \)'s (or if we only have estimates of them), it is better to use the variant with fixed number of successes. In addition, if we assume that \( p \in [0.15, 0.67] \), we loose at most 50% compared to the maximal efficiency even in unbalanced cases. Checked against the ratio of the optimal efficiency with the one of the crude estimator with only one level \( (m = 1) \), that is, \( 0.648(1 - \gamma)/(\gamma \log^2(\gamma)) = 4.4 \cdot 10^{10} \), this loss seems not very relevant. Therefore, we expect to get good results with ad-hoc choices of the levels where we only ensure that the \( p_k \) are not
2.5. Theoretical analysis in a simplified setting

Figure 2.2: Contour plot of the asymptotical efficiencies relative to the optimal efficiency of $\hat{\gamma}$ (thin line) and $\tilde{\gamma}$ (thick line) when $p_{2l−1} = p^{1+δ}$, $p_{2l} = p^{1−δ}$ for $l \in 1, \ldots, L$ and $m = 2L \sim \log(γ)/\log(p)$ in dependence of $δ$ (measure for the inhomogeneity of two succeeding probabilities $p_k$) and $p$ (geometric average of the probabilities $p_k$).

too close to 0 or 1, respectively (see also Section 2.6).

2.5.3 Asymptotic distribution

Because of the product form of $\tilde{\gamma}$ and $\hat{\gamma}$, we expect these estimators to have asymptotically a log-normal distribution. However, for this we need to condition on $\hat{\gamma} > 0$ since $\hat{\gamma} = 0$ occurs with positive probability. We indicate this conditioning by $\hat{\gamma}^+$ and $\hat{p}^+_k$.

First we consider the case where the $p_k$ and $r_k$ or $n_k$, respectively, are fixed and the number of levels $m$ tends to infinity. Then $γ = γ^m \to 0$.

**Proposition 2.6.** Let $0 < p_{\text{min}} \leq p_k \leq p_{\text{max}} < 1$. For $2 \leq r_k \leq R$,

\[
\tilde{\mu}_m := \sum_{k=1}^{m} E[\log(\tilde{p}_k)] \text{ and } \tilde{\sigma}_m^2 := \sum_{k=1}^{m} \text{Var}(\log(\tilde{p}_k))
\]
we have
\[
\frac{\log(\tilde{\gamma}_m^m) - \hat{\mu}_m}{\hat{\sigma}_m} \to \mathcal{N}(0, 1) \quad (m \to \infty).
\]
and for \(2 \leq n_k \leq N\),
\[
\hat{\mu}_m := \sum_{k=1}^{m} E[\log(\hat{\rho}_{+, k})] \quad \text{and} \quad \hat{\sigma}_m^2 := \sum_{k=1}^{m} \text{Var}(\log(\hat{\rho}_{+, k}))
\]
we have
\[
\frac{\log(\tilde{\gamma}_+^m) - \hat{\mu}_m}{\hat{\sigma}_m} \to \mathcal{N}(0, 1) \quad (m \to \infty),
\]
i.e., \(\log(\tilde{\gamma}_m^m)\) and \(\log(\tilde{\gamma}_+^m)\) are asymptotically normally distributed.

Proof. See Appendix 2.C.

Note that by Jensen's inequality \(\tilde{\mu}_m < \log \gamma^m\) and \(\hat{\gamma}_+^m < \log \gamma^m\), that is, on the log scale the estimators are biased.

Alternatively, we can fix \(m\) and let \(r_k\) and \(n_k\), respectively, go to infinity. In this case, the bias on the log scale disappears asymptotically.

**Proposition 2.7.** Let \(m\) and \(0 < p_k < 1\) be fixed. For \(r_k \sim c_k r\) with \(c_k > 0\) we have
\[
r^{1/2}(\log(\tilde{\gamma}) - \log(\gamma)) \to \mathcal{N} \left( 0, \sum_{k=1}^{m} \frac{1 - p_k}{c_k} \right) \quad (r \to \infty) \tag{2.18}
\]
and for \(n_k \sim d_k n\) with \(d_k > 0\) we have
\[
n^{1/2}(\log(\hat{\gamma}) - \log(\gamma)) \to \mathcal{N} \left( 0, \sum_{k=1}^{m} \frac{1 - p_k}{p_k d_k} \right) \quad (n \to \infty). \tag{2.19}
\]

Proof. See Appendix 2.C.

Therefore, we can expect that both \(\tilde{\gamma}\) and \(\hat{\gamma}\) show approximately a log-normal behavior in typical situations.
2.5.4 Two-stage fixed numbers of successes estimator

Based on the discussion in the previous sections, the following estimation strategy seems reasonable for applications. Choose an importance function using (2.7) as guideline, see e.g. the approximations in sections 2.4.1 and 2.4.2. Then in a first stage, use a fixed number of successes estimator $\tilde{\gamma}_p$ with a large number of levels $m$ and $r_k \equiv r$ of moderate size, e.g. $r = 50$. If it takes too long to produce a result, the problem we are facing is probably too hard and we would need more computing resources or a better importance function (remember that from Section 2.5.1 the fixed number of successes estimator should have a good efficiency in such situations). Otherwise, we have at least an idea of the order of magnitude of $\gamma$ and rough estimates of the $p_k$’s. Then we choose new levels with $m \approx -0.6275 \log(\tilde{\gamma}_p)$, $p_k \approx 0.2032$ and $r_k \approx -1/(2q) \log(\tilde{\gamma}_p)$, to get a rough control on the imprecision of $\tilde{\gamma}$ in the second stage. Alternatively, if most of the $p_k$ are not too close to 0 or 1, a very simple and effective strategy for the second stage is to take the same levels as in the first stage and to set $r_k = r \sim (1 - \tilde{\gamma}_p^{1/m})/(1 + q)^{1/m - 1}$, because as we have seen in Section 2.5.2, the estimator is quite robust in terms of parameter choice.

Clearly, with this strategy $\text{Var}(\tilde{\gamma})/\gamma^2 \approx q$ holds only when the assumption (2.6) exactly holds. So we want to compute confidence intervals for $\gamma$ which do not rely on this assumption. Because we expect to have approximately normally distributed estimators on the log scale (see Section 2.5.3), we use intervals of the form $I_{a,b} = [a \tilde{\gamma}, b \tilde{\gamma}]$ ($a, b > 1$). Values of $a$ and $b$ can be obtained at the cost of a little loss of efficiency: Choose $r_k = \lceil r^{1/2} \rceil$, repeat the second stage $\lceil r^{1/2} \rceil$ times yielding $\tilde{\gamma}^1, \tilde{\gamma}^2, \ldots, \tilde{\gamma}^{\lceil r^{1/2} \rceil}$ and calculate

$$\tilde{\gamma} = \frac{1}{\lceil r^{1/2} \rceil} \sum_{l=1}^{\lceil r^{1/2} \rceil} \tilde{\gamma}^l. \quad (2.20)$$

This has approximately the same expected total workload as before. The leading term of the asymptotical variance for $r \to \infty$ remains also
the same as in (2.17), but the remainder can shown to be $O(r^{-3/2})$
instead of $O(r^{-2})$. In order to compute the bounds $a$ and $b$ of our
confidence interval, we now can use the bootstrap for $\log(\tilde{\gamma}/\gamma)$ (see e.g.
Efron and Tibshirani (1993)).

The log estimators $\log(\tilde{\gamma}^1), \ldots, \log(\tilde{\gamma}^{r^1/2})$ provide also a possibility to
detect statistically if we have violated the assumption (2.6), which indicates that probably the importance function was chosen poorly. Under
the null hypothesis (2.6) with $m = \lfloor -0.6275 \log(\gamma) \rfloor$, $p^k = \gamma^{1/m}$, and
$r_k = \lfloor r^1/2 \rfloor$ where $r = -1/(2q) \log(\gamma)$, $\log(\tilde{\gamma}^l)$ has the same distribution as

$$\sum_{k=1}^{m} \log \left( \frac{\lfloor r^1/2 \rfloor - 1}{\lfloor r^1/2 \rfloor + Z_k - 1} \right) \text{ with } Z_k \sim i.i.d \\text{NegBin}(\lfloor r^1/2 \rfloor, \gamma^{1/m}).$$

We found by simulations of this distribution that even for relatively big
values of $q$ and $\gamma$, yielding rather small values of $m$ and $r$, it is almost a
normal distribution. Thus normal Q-Q plots of $\log(\tilde{\gamma}^1), \ldots, \log(\tilde{\gamma}^{r^1/2})$
can be a helpful tool to validate the choice of the importance function in
retrospect. Typically, asymmetry of the confidence interval is already
a good indicator of a violation of (2.6) (see also the examples in Section
2.6).

### 2.6 Examples

Here, we we mainly want to illustrate the behavior of the strategy
introduced in Section 2.5.4 and its effectivity when applied together
with well chosen approximations to the optimal importance function
such as the ones introduced in sections 2.4.1 and 2.4.2.
2.6. Examples

2.6.1 High levels during a busy period of a G/G/1 queueing system

We first consider a G/G/1 queueing system (see e.g. Gross et al. (2008), Kobayashi and Mark (2009) or Bolch et al. (2006)) where both inter arrival times and service times are Weibull distributed with shape parameters $k_a$ and $k_s$ and scale parameters $1/\lambda$ and $1/\mu$, respectively. If we write $X^1_t$ for the number of customers in the system, $X^2_t$ for the remaining time until the next arrival and $X^3_t$ for the remaining service time of the customer being served at time $t$, then $X = \{X_t = (X^1_t, X^2_t, X^3_t)^T : t \geq 0\}$ is a Markov chain with state space $\mathcal{E} = \mathbb{N}_0 \times [0, \infty)^2$, which is easy to simulate. We set $A = \{x : x^1 = 0\}$, $B = \{x : x^1 \geq L\}$ and $x_0 = (1,0,0)^T$ and want to estimate $\gamma = P[\tau_B < \tau_A | X_0 = x_0]$, the probability to reach $B$ before entering into $A$ (see Section 2.2.2).

For $k_a = k_s = 1$, we get the special case of a M/M/1 queueing system. Then the sub process $\mathcal{X}^1 = \{X^1_t : t \geq 0\}$ is already Markov with transition intensities $\lambda$ for $x^1 \to x^1 + 1$ if $x^1 \geq 0$, $\mu$ for $x^1 \to x^1 - 1$ if $x^1 \geq 1$ and 0 otherwise. It can be simulated without any information on remaining times, using the fact that it is a homogeneous Markov jump process, with the Gillespie algorithm (see Section 3.2, in particular (3.1) and (3.2)). In this case, we have for $\rho = \mu/\lambda \neq 1$ and $\tau_l = \inf\{t > 0 : X^1_t \geq l\}$ that

$$P[\tau_l < \tau_A | X^1_0 = x^1] = \frac{1 - \rho^{x^1}}{1 - \rho^L} \text{ for } 1 \leq x^1 \leq l$$

and so $\gamma = (1 - \rho)/(1 - \rho^L)$. Moreover, if we just work with the sub chain $\mathcal{X}^1$, we know from Section 2.4 that the optimal importance function is given by $h(P[\tau_0 < \tau_A | X^1_0 = x^1])$ for some monotone increasing function $h : [0,1] \to \mathbb{R}$. Thus for simplicity we choose the optimal importance function $\Phi(x) = x^1$ for which the assumption (2.6) is obviously fulfilled. On the other hand, if we simulate the M/M/1 queue including the information on the remaining arrival and service times, the optimal importance function now also depends on these times: If for example
the remaining service time of an entrance state at some level is relatively big, it is more likely to reach a high level, because jumps downwards are blocked during this time. Thus in this case, $\Phi(x) = x^1$ is no longer optimal. This also applies for the cases where $k_a$ or $k_s$ are not equal to 1, because the simulation method we then apply requires knowledge of the remaining times. Nevertheless, we use this importance function throughout this example, since it approximates the optimal importance function good enough.

Further, it is sufficient to choose the levels only from $\{1, \ldots, L\}$, in particular the maximal number of thresholds is $L$. If $l_{k-1} = a < l_k = b$ for $a$ and $b$ in $\mathbb{N}$, then $p_k = P[D_k|D_{k-1}] = (1 - \rho^a)/(1 - \rho^b)$ for the M/M/1 queue and thus $\lim_{\rho \to \infty} p_k(\rho)/\gamma(\rho) = (b - a)/(L - 1)$. Because of (2.16), this means that if we choose the maximal number of levels we can achieve a best possible asymptotic performance of

$$\lim_{\gamma \to 0^+} \frac{-\log(\text{Eff}(\gamma))}{\log(\gamma)} = 2 - \frac{1}{L - 1}.$$ 

Therefore the simple M/M/1 queueing system shows that one is not able to handle arbitrary hard problems with importance splitting.

**Simulation experiments**

First, we look at the M/M/1 case. We choose $\lambda = 1$, $\mu = \rho = 0.2032^{-1/8} = 1.2204$ and $L = 104$, i.e. $\gamma = 2.220 \cdot 10^{-10}$. Here, as already mentioned, we have two variants, namely simulating only the jump chain of $X^1$, where we have only one possible entrance state on each level, and simulating $X$, where we have infinitely many entrance states on each level with different probabilities to reach the next level. In addition, we look at the following three situations:

- Near optimal choice of levels, that is, $l_1 = 3$, $l_k = 8(k - 1)$ for $k \in \{2, \ldots, 13\}$, and optimal choice of $r_k$ with $q = 0.01$, i.e., $r_k \approx 1120$. 
Table 2.1: Overview of results for the $M/M/1$ queueing system

<table>
<thead>
<tr>
<th>Situation</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\gamma$</td>
<td>$\gamma$</td>
<td>$\gamma$</td>
</tr>
<tr>
<td>$\langle \gamma \rangle$</td>
<td>$2.22 \cdot 10^{-10}$</td>
<td>$2.24 \cdot 10^{-10}$</td>
<td>$2.24 \cdot 10^{-10}$</td>
</tr>
<tr>
<td>$\text{Var}(\gamma)/\gamma^2$</td>
<td>0.0112</td>
<td>0.0097</td>
<td>0.0116</td>
</tr>
<tr>
<td>$\langle W(\gamma) \rangle$</td>
<td>$2.677 \cdot 10^7$</td>
<td>$2.664 \cdot 10^7$</td>
<td>$2.655 \cdot 10^7$</td>
</tr>
<tr>
<td>all 3 components</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle \gamma \rangle$</td>
<td>$2.22 \cdot 10^{-10}$</td>
<td>$2.21 \cdot 10^{-10}$</td>
<td>$2.26 \cdot 10^{-10}$</td>
</tr>
<tr>
<td>$\text{Var}(\gamma)/\gamma^2$</td>
<td>0.0124</td>
<td>0.0124</td>
<td>0.0182</td>
</tr>
<tr>
<td>$\langle W(\gamma) \rangle$</td>
<td>$2.676 \cdot 10^7$</td>
<td>$2.632 \cdot 10^7$</td>
<td>$2.646 \cdot 10^7$</td>
</tr>
</tbody>
</table>

- Many levels such that the probabilities $p_k$ are unbalanced but not too close to 0 or 1, that is,

$$(l_0, \ldots, l_{51}) = (1, 4, 5, 8, 9, 12, 13, 16, \ldots, 104),$$

with constant $r_k$ such that the total workload is comparable to the one above, i.e., $r_k \approx 2060$.

- The same levels as in the second situation, but with $r_k = 20$. To get a comparable workload to the situations above, we took an average over 103 repetitions.

The workload is measured in simulated random numbers, which is on average around $2.65 \cdot 10^7$ in all the above $2 \cdot 3 = 6$ settings (see Table 2.1).

Table 2.1 reveals that the difference in terms of the squared relative error between the three situations is small, confirming the expectation that we have good efficiency properties even when we just work with $p_k$'s not too close to 0 or 1 and a moderate fixed number of successes. Further, the desired imprecision is achieved.

Next we look at the Weibull cases with $k_a = k_b = k = 0.75$ and $k_a = k_b = k = 1.25$, respectively, and the same values for $\lambda$, $\mu$ and $L$ as before. Note that the expectation of a Weibull($k; 1/\lambda$) distributed
Table 2.2: Overview of results for the G/G/1 queueing system

<table>
<thead>
<tr>
<th>k</th>
<th>0.75</th>
<th>1.00</th>
<th>1.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r \approx )</td>
<td>1290</td>
<td>2060</td>
<td>2910</td>
</tr>
</tbody>
</table>

\[ \langle \gamma \rangle \]

\[ \text{Var}(\gamma)/\langle \gamma \rangle^2 \]

\[ \langle W(\gamma) \rangle \]

\[ 0.0134 \quad 0.0102 \quad 0.0116 \]

\[ 7.067 \cdot 10^6 \quad 2.254 \cdot 10^7 \quad 5.304 \cdot 10^7 \]

random variable is \( 1/\lambda \cdot \Gamma(1 + 1/k) \). Thus if we were to replace the Weibull arrival and service times in our examples with exponential random variables with the same expectation, then we would find for the resulting M/M/1 queue that

\[ \rho = \frac{(1/\lambda \cdot \Gamma(1 + 1/k))}{(1/\mu \cdot \Gamma(1 + 1/k))} = 1.2204 \]

and therefore obtain the same \( \gamma \) as above. So we also look at the case \( k = 1 \). In each situation, we apply the two-stage procedures outlined in Section 2.5.4 exactly 100 times. That is, we first use a prerun estimator \( \tilde{\gamma}_p \) with \( r = 20 \) and the ad hoc choice of levels \( l_k = k \) for \( k \in \{0, 1, 2, \ldots, L\} \) to get a first estimate \( \tilde{\gamma} \). Next we compute the estimator with the same levels and \( r_k = r \sim (1-\tilde{\gamma}_p^{1/m})/(1+q)^{1/m} - 1 \) with desired imprecision \( q = 0.01 \), and the averaged estimator (2.20). The total workload turns out to be about \( 7.1 \cdot 10^6, 2.25 \cdot 10^7 \) and \( 5.3 \cdot 10^7 \) simulated random numbers for \( k = 0.75, k = 1 \) and \( k = 1.25 \), respectively.

In Table 2.2, we see that the different shapes of the distributions for inter arrival and service times lead to substantially different values of \( \gamma \). Both estimators achieve the desired imprecision nearly, but we have a slightly larger variance for the averaged estimator. This is the price we pay for being able to compute confidence intervals.

In Figure 2.3, plots of the bootstrapped 95% confidence intervals for the estimator (2.20) on a log scale are given, where the solid line indicates the average estimates \( \langle \tilde{\gamma} \rangle \) from Table 2.2 (\( r_k = r^{1/2} \) with averaging) or the true value of \( \gamma \), respectively, and the dotted lines indicate the
2.6. Examples

![Figure 2.3: Bootstrapped 95% confidence intervals (log scale) of the estimator (2.20) for the G/G/1 queueing system. The solid lines indicate the true overflow probability $\gamma$ or the average $\langle \tilde{\gamma} \rangle$ over the 100 repetitions, respectively, and the dashed lines the expected width of 95% confidence intervals with imprecision $\text{Var}(\tilde{\gamma})/\gamma^2 = q$.](image)

expected width of confidence intervals assuming log-normality for the estimators with imprecision $\text{Var}(\tilde{\gamma})/\gamma^2 = q$. They yield satisfactory coverages, but seem to be slightly wider than expected. Further, normal Q-Q plots (not shown) of $\log(\tilde{\gamma}^1), \ldots, \log(\tilde{\gamma}^{r^{1/2}})$ reveal that the log estimators seem normally distributed, as expected regarding the results in Section 2.5.4.

### 2.6.2 Buffer overflow in a tandem Jackson network during a busy period

Here we consider the open tandem Jackson queueing network with two queues that has had a great deal of attention in rare event simula-
Chapter 2. Splitting Simulation

In this case, we are able to calculate the importance function (2.8) explicitly. We find for $1 \leq x_2 < L$

$$
\Phi(x_1, x_2) = \left( \frac{\mu_1}{\lambda + \mu_1 + \mu_2} \right)^{L-x_2} \left( \frac{\lambda}{\lambda + \mu_2} \right)^{\max(L-x^2-x^1,0)}
$$

(2.21)

This can be seen as follows: To reach the set $B$ from some state $(x^1, x^2)$ with $x_2 < L$, we need at least $L - x_2$ transitions from the first to the second queue and thus at least $\max(L - x^2 - x^1, 0)$ arrivals at the first queue. In the most likely sequence, there are clearly no exits from the second queue. A transition from the first to the second queue has always probability $\mu_1/(\lambda + \mu_1 + \mu_2)$ if $x_1 > 0$. An arrival at the first queue has probability equal to $\lambda/(\lambda + \mu_1 + \mu_2)$ if $x_1 > 0, x_2 > 0$ and equal to $\lambda/(\lambda + \mu_2)$ if $x_1 = 0, x_2 > 0$. The second variant has a higher probability, so the most likely sequence first empties the first buffer and then alternates between arrivals at the first queue and transitions to the second one until $x_2$ reaches the level $L$. Note that this sequence does not have to correspond to a typical sequence of the process which reaches the set $B$ from a state $(x^1, x^2)$ without entering $(0, 0)$. For $x_2 = 0$ the value of $\Phi(x)$ is slightly larger because an exit from the second queue is impossible. In this case, it is also not difficult to obtain an explicit formula for the importance function.

We will see that this choice yields good results, in particular in the case when the naive choice $\Phi(x^1, x^2) = x^2$ leads to underestimation, that is, when $\rho_1 > \rho_2$ or equivalently $\mu_1 < \mu_2$ (see Glasserman et al. (1998) or Garvels et al. (2002)).
2.6. Examples

Table 2.3: Overview of results for the tandem Jackson network

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>1/2</th>
<th>1/3</th>
<th>1/5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>$1.86 \cdot 10^{-9}$</td>
<td>$1.94 \cdot 10^{-14}$</td>
<td>$8.59 \cdot 10^{-21}$</td>
</tr>
<tr>
<td>$\langle \tilde{\gamma} \rangle$</td>
<td>$1.84 \cdot 10^{-9}$</td>
<td>$1.93 \cdot 10^{-14}$</td>
<td>$8.52 \cdot 10^{-21}$</td>
</tr>
<tr>
<td>$\text{Var}(\tilde{\gamma})/\gamma^2$</td>
<td>0.00763</td>
<td>0.00726</td>
<td>0.00583</td>
</tr>
<tr>
<td>$\langle W(\tilde{\gamma}) \rangle$</td>
<td>$1.456 \cdot 10^7$</td>
<td>$2.978 \cdot 10^7$</td>
<td>$8.571 \cdot 10^7$</td>
</tr>
</tbody>
</table>

Working with a large enough bound on the first buffer, we get an accurate numerical approximation of $\gamma$, which we use to check our simulations (see Appendix 2.D for more details).

A first series of simulation experiments

We choose $L = 30$, $\rho_1 = 1/2$ and $\rho_2$ equal 1/2, 1/3 and 1/5. We worked with $q = 0.0025$ and applied the two-stage procedure with the averaged estimator (2.20) 100 times. The results are summarized in Table 2.3. The actual imprecision is slightly larger than the desired imprecision, because there is presumably a slight dependence between the probabilities to reach the next level from entrance states on a certain level.

Figure 2.4 shows the estimates with corresponding confidence intervals. The solid and the dotted lines have the same meaning as in Figure 2.3. The coverage is again satisfactory. Although the normal Q-Q plots of $\log(\tilde{\gamma}_1), \ldots, \log(\tilde{\gamma}_{r^{1/2}})$ show no deviation from log-normality, the confidence intervals are slightly wider than expected under the log-normal distribution.

Non-exponential arrival and service times

As in the example in Section 2.6.1, let us investigate the case of Weibull arrival and service times with scale parameters $1/\lambda'$, $1/\mu'_1$ and $1/\mu'_2$ and and shape parameters $k_1$, $k_2$ and $k_3$, respectively. Similarly, we include
Chapter 2. Splitting Simulation

Figure 2.4: Bootstrapped 95% confidence intervals (log scale) of the estimator (2.20) for the tandem Jackson network. The solid lines indicate the true overflow probability $\gamma$ and the dashed lines the expected width of 95% confidence intervals with imprecision $\text{Var}(\hat{\gamma})/\gamma^2 = q$.

the information on the remaining times into the state – it has now five dimensions – to get a Markov process, which we can simulate.

We use the importance function (2.21) with $\lambda = \lambda'/\Gamma(1 + 1/k_1)$, $\mu_1 = \mu'_1/\Gamma(1 + 1/k_2)$ and $\mu_2 = \mu'_2/\Gamma(1 + 1/k_3)$. The idea behind is simple: We replace the Weibull times with exponential times with the same expectation and only consider the numbers of customers which form now a Markov process as before.

As in the example in Section 2.6.1, if we use simulation including the remaining times, the optimal importance function would take into consideration the remaining times, even in the case $k_1 = k_2 = k_3 = 1$ which corresponds to exponential arrival and service times. Our simulations will show that in deed, the estimates based on this choice of the
2.6. Examples

Figure 2.5: Bootstrapped 95% confidence intervals (log scale) of the estimator (2.20) for the Weibull tandem Jackson network. The solid lines indicate the true overflow probability $\gamma (k = 1)$ or the average $\langle \tilde{\gamma} \rangle$ over the 100 repetitions ($k = 0.75, 1.25$), respectively, and the dashed lines the expected width of 95% confidence intervals with imprecision $\text{Var}(\tilde{\gamma})/\gamma^2 = q$.

importance function do not achieve the desired accuracy, in particular also in the case $k_1 = k_2 = k_3 = 1$. The results are acceptable, but only to a limited extent.

A second series of simulation experiments

We choose $L = 30$, $\mu_1 = 2$, $\mu_3 = 3$, $q = 0.0025$ and once again run the two-stage procedure with the averaged estimator (2.20) 100 times. The case $k_1 = k_2 = k_3 = k = 1$ corresponds to the situation with $\rho = 1/3$ from before and thus $\gamma = 1.94 \cdot 10^{-14}$. Further, we present simulations for $k_1 = k_2 = k_3 = k = 0.75$ and $k_1 = k_2 = k_3 = k = 1.25$, where we do not know any method for numerical approximation of the true overflow
probabilities.

In Figure 2.5 we see that the confidence intervals are clearly wider than expected and we do not get the desired coverage of 95%. Note also that the computed coverage rate is too optimistic in the cases \( k_1 = k_2 = k_3 = k = 0.75 \) and \( k_1 = k_2 = k_3 = k = 1.25 \), respectively, since we use the average over the 100 repetitions as true value. Further, the normal Q-Q plots of \( \log(\tilde{\gamma}_1), \ldots, \log(\tilde{\gamma}_{r_1/2}) \) (not shown) indicate a moderate deviation from log normality and thus a relevant violation of (2.6).

### 2.6.3 Buffer overflows in finite time

We consider the M/M/1 queue and the tandem Jackson network from Sections 2.6.1 and 2.6.2 once again, but now the entrance into the set of interest should occur before a fixed time \( T \). In both cases, we are able to approximate the rare event probabilities numerically using the techniques described in Appendix 2.D for transient analysis of homogenous continuous-time Markov chains.

**M/M/1 queue**

Here, we write \( Y_t \) for the number of customers in the system at time \( t \) and we want to find the overflow probability that it reaches some high level \( L \) before time \( T \), starting with an empty queue, that is, \( \gamma = P^{Y}[\tau_C < T|Y_0 = 0] \), where \( C = \{y : y \geq L\} \).

We want to compare the performance of the importance function proposed in (2.9) with the optimal time-independent importance function for estimating \( P^{Y}[\tau_C < \tau_0|Y_0 = 1] \) where \( \tau_0 \) is the first time the queue reaches to empty state. This is \( \Phi(y) = h(y) \) for some monotone increasing function \( h \), we choose \( \Phi(y) = y \). In order to do this comparison, we calculate 100 times the two-stage fixed number of successes estimator (2.20) and the corresponding bootstrapped 95% confidence intervals.
for a desired imprecision of $q = \frac{\text{Var}(\tilde{\gamma})}{\gamma^2} = 0.01$. Parameter values are $\lambda = 1$, $\mu = 1.1$, $L = 41$ and $T$ equals 10, 20 and 40, respectively, yielding true overflow probabilities $\gamma$ equal to $5.84 \cdot 10^{-17}$, $1.74 \cdot 10^{-10}$ and $2.06 \cdot 10^{-06}$, respectively. The results are shown in Figure 2.6.

For all three time horizons $T$, the time-dependent importance functions yield good results: The bootstrapped 95% confidence intervals seem symmetric on the log scale and contain the true values nicely, although they are slightly wider than expected by our choice of $q$. On
the other hand, the time-independent importance function completely fails for the smallest value of $T$. Mostly, the true value is underestimated severely, the variance of the estimator is much too big and the confidence intervals indicate a right-skewed distribution of $\log(\tilde{\gamma})$. The problem diminishes when the time horizon becomes bigger and seems nearly inapparent for $T = 40$. This is because for big values of $T$, the change of $P^Y[\tau_C < T - t | Y_0 = y]$ after a jump of the Markov chain depends for most general states mainly on the state space component, so that the time-independent importance function already creates enough drift towards the rare event.

**Tandem Jackson network**

Also for the tandem Jackson network, we consider the finite time horizon problem. The arrival rate at the first queue is $\lambda = 1$ while $\mu_1$ and $\mu_2$ are the service rates at the first and the second queue, respectively. We write $Y^i_t$ for the number of customers in buffer $i$.

We want to estimate $\gamma = P^Y[\tau_C < T | Y_0 = y_0]$ for a fixed time $T$. We compare the time-dependent importance function (2.9) and the time-independent counterpart defined in (2.10). For our simulation, we use $\lambda = 1$, $\mu_1 = 2$, $\mu_2 = 3$ and $L = 20$ for $T$ values 4, 8 and 50 yielding true overflow probabilities $\gamma$ equal to $1.54 \cdot 10^{-17}$, $1.45 \cdot 10^{-12}$ and $1.16 \cdot 10^{-8}$, respectively. As in the examples before, we calculate 100 times the two-stage fixed number of successes estimator (2.20) and the corresponding bootstrapped 95% confidence intervals for a desired imprecision of $q = \text{Var}(\tilde{\gamma})/\gamma^2 = 0.01$, see Figure 2.7.

Normal Q-Q plots (not shown) of $\log(\tilde{\gamma}^1), \ldots, \log(\tilde{\gamma}^{\lfloor r^{1/2} \rfloor})$ exhibit a normal distribution when the time-dependent importance function is used (indicated also by the symmetric confidence intervals on the log scale), which is expected under the assumption (2.6) (see Section 2.5.4), and the confidence intervals are only slightly too wide. This is also true for time-independent importance function when $T = 50$ is big, where one can not recognize any difference to the time-dependent choice. For
the smallest $T$-value, we observe as in the M/M/1 example asymmetric confidence intervals, i.e., $\log(\hat{\gamma}^1), \ldots, \log(\hat{\gamma}^{|r/2|})$ are not normally distributed, and a huge variance, therefore the time-independent importance function is inappropriate in this situation.
2.7 Conclusions

In this chapter, we have introduced a new splitting variant and a corresponding rare-event probability estimator. Among other benefits, our procedure avoids the problem of producing zero-valued estimates. We have shown that our estimator is unbiased, and have also provided a new unbiasedness proof for previous splitting variants, because the proof usually given is based on an incorrect claim.

We have analyzed the behavior of our new estimator, as well as that of the fixed effort estimator, in a simplified setting that is appropriate when the importance function is a good approximation to the optimal one (whose evaluation requires the quantity of interest). Our key observations are as follows:

- The fixed effort method controls the workload, whereas the fixed number of successes method controls the imprecision.

- The optimal choice of parameters has been derived and with these both variants are asymptotically efficient.

- Working with moderate constant fixed number of successes, we are able to get rough probability estimates in an efficient way.

- If the probabilities $p_k$ are unknown or we have only rough estimates of them, we have seen that it is preferable to use the fixed number of successes estimator, because it is more robust in terms of deviations from the optimal parameter values.

- The estimators are asymptotically log-normally distributed as the number of levels converges to infinity. This is also true if the number of successes or the number of trials, respectively, converges to infinity.

Further, we have provided approximations for the optimal importance function, also for problems with a finite time horizon, which lead to
excellent results in practice, especially when combined with our two-stage fixed numbers of successes estimator (which needs no tuning and yields confidence intervals), as our examples illustrate. In particular, the results in the case of the tandem Jackson network when the service rate at the second queue is bigger, which is known to be a hard problem, are convincing.

2.A Counterexample to the claim (2.4)

Let \( \{X_t, t \in \mathbb{N}_0\} \) be a discrete-time Markov chain with state space \( \mathcal{E} = \{1, 2, 3, 4, 5, 6, 7\} \), initial state \( x_0 = 1 \) and transition matrix

\[
Q = \begin{pmatrix}
0 & 1/3 & 2/3 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/10 & 2/10 & 0 & 7/10 \\
0 & 0 & 0 & 8/10 & 1/10 & 0 & 1/10 \\
0 & 0 & 0 & 0 & 0 & 9/10 & 1/10 \\
0 & 0 & 0 & 0 & 1/10 & 9/10 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1/10 & 1/10 \\
\end{pmatrix}.
\]

We set \( A = \{7\} \) and \( B = \{6\} \) and are interested in the probability \( \gamma = P[\tau_B < \tau_A] \). The importance function is chosen to be

\[
\Phi(x) = \begin{cases}
-1 & x = 7 \\
0 & x = 1 \\
1 & x \in \{2, 3\} \\
2 & x \in \{4, 5\} \\
3 & x = 6
\end{cases},
\]

so that \( A = \{x \in \mathcal{E} : \Phi(x) \leq -1\} \) and \( B = \{x \in \mathcal{E} : \Phi(x) \geq 3\} \) and we set \( l_i = i \) for \( i \in \{0, 1, 2, 3\} \). The situation is summarized in Figure 2.8.

First, we compute \( p_i \) for \( i = 1, 2, 3 \). Obviously, \( P[D_1] = 1 \). The event \( D_2 \) is equivalent to one of the jump sequences \( (1, 2, 4), (1, 2, 5), (1, 3, 4) \)
Figure 2.8: Counterexample.

or \((1, 3, 5)\) for \((X_0, X_1, X_2)\). Similarly for \(D_3\) we need in addition that \(X_3 = 6\). From this, we easily obtain \(P[D_2] = 7/10, P[D_3] = \gamma = 157/300\) and

\[
p_3 = P[D_3|D_2] = P[D_3]/P[D_2] = 157/210 \approx 0.7476.
\]

Next, we consider fixed effort importance splitting with \(N_0 = 1, N_1 = 2\) and \(N_2 = 1\). In this example, random and balanced selection coincide. Moreover, \(R_1 = 1\) always holds, \(R_2\) can take the values 0, 1 or 2 and
$R_3$ can take the values 0 or 1. For $R_2 = 1$ one of the events $1 \rightarrow 2 \rightarrow (1, 4)$ or $(4, 1)$, $1 \rightarrow 2 \rightarrow (1, 5)$ or $(5, 1)$, $1 \rightarrow 3 \rightarrow (1, 4)$ or $(4, 1)$ or $1 \rightarrow 3 \rightarrow (1, 5)$ or $(5, 1)$ must occur. For $R_2 = 1, R_3 = 1$ the next state in the above sequences has to be 6. From this, we obtain $P[R_2 = 1] = 13/50$ and $P[R_2 = 1, R_3 = 1] = 223/1500$. Similarly, we find $P[R_2 = 2] = 57/100$, $P[R_3 = 1, R_2 = 2] = 449/1000$. Therefore

\[
E[\hat{p}_3|R_2 = 1] = P[R_3 = 1|R_2 = 1] = \frac{223}{390} < p_3,
\]

\[
E[\hat{p}_3|R_2 = 2] = P[R_3 = 1|R_2 = 2] = \frac{449}{570} > p_3.
\]

Moreover $E[\hat{p}_3] = P[R_3 = 1] = 1793/3000 < p_3$, but unbiasedness of $\hat{\gamma}$ holds because

\[
E[\hat{p}_3\hat{p}_2\hat{p}_1] = P[R_2 = 2, R_3 = 1] + \frac{1}{2}P[R_2 = 1, R_3 = 1] = \gamma.
\]

Finally, we consider fixed splitting with $N_0 = 1$, $c_0 = 1$, $c_1 = 2$ and $c_2 = 1$. Again, $R_1 = 1$ always, but $R_2$ and $R_3$ can now both take the values 0, 1, or 2. The probabilities $P[R_2 = 1]$, $P[R_2 = 1, R_3 = 1]$ and $P[R_2 = 2]$ are the same as for fixed effort. Because $P[R_2 = 1, R_3 = 2] = 0$, $E[\hat{p}_3|R_2 = 1]$ is the same as for fixed effort. By similar computations as before, we obtain $P[R_2 = 2, R_3 = 1] = 897/5000$ and $P[R_2 = 2, R_3 = 2] = 3593/10000$ and therefore

\[
E[\hat{p}_3|R_2 = 2] = \frac{1}{2}P[R_3 = 1|R_2 = 2] + P[R_3 = 2|R_2 = 2] = \frac{449}{570} \neq p_3.
\]

Again, unbiasedness of $\hat{\gamma}$ can be verified and also

\[
E[\hat{p}_3] = E[\hat{p}_3|R_2 = 1]P[R_2 = 1] + E[\hat{p}_3|R_2 = 2]P[R_2 = 2] = \frac{1793}{3000} \neq p_3.
\]
2.B Proofs for optimality in the independence framework

2.B.1 Proof of Proposition 2.4

The case \( q \geq 1/\gamma - 1 \) is trivial. So we assume \( q < 1/\gamma - 1 \). Then the first constraint becomes an equality. Using the reparametrization \( v_k = (1 - p_k)/(w_k p_k) + 1 \), so \( w_k = (1 - p_k)/((v_k - 1)p_k) \) we get

\[
\begin{align*}
\text{minimize} & \quad W = e_m \sum_{k=1}^m (1 - p_k)/((v_k - 1)p_k) \\
\text{with respect to} & \quad p_k, v_k, k \in 1, 2, \ldots, m \\
\text{subject to} & \quad \prod_{k=1}^m v_k = q + 1 \leq \frac{1}{\gamma} \\
& \quad \prod_{k=1}^m p_k = \gamma \\
& \quad 0 \leq p_k \leq 1, k \in 1, 2, \ldots, m \\
& \quad 1 \leq v_k \leq 1/p_k, k \in 1, 2, \ldots, m. 
\end{align*}
\]

(2.22)

It remains to prove the following:

**Proposition 2.8.** The solution to the optimization problem (2.22) is of the following form: There is a subset \( I \) of \( \{1, 2, \ldots, m\} \) such that

\[
\begin{align*}
p_k = 1, v_k = 1 & \quad (k \in I) \\
p_k = \gamma^{1/(m-|I|)}, v_k = (1 + q)^{1/(m-|I|)} & \quad (k \notin I)
\end{align*}
\]

with objective value

\[
e_m \left(|I| + (m - |I|) \cdot \frac{1 - \gamma^{1/(m-|I|)}}{\gamma^{1/(m-|I|)}((1 + q)^{1/(m-|I|)} - 1)} \right).
\]

Let us first consider the simple case \( m = 2 \).

**Lemma 2.4.** For \( m = 2 \), the only solutions of (2.22) are:

If \( 0 < q < (1 - 4\gamma^{1/2} + 3\gamma)/\gamma \), then \( p_1 = p_2 = \gamma^{1/2}, v_1 = v_2 = (1 + q)^{1/2} \).

If \( (1 - 4\gamma^{1/2} + 3\gamma)/\gamma \leq q \leq 1/\gamma - 1 \), then \( p_1 = 1, p_2 = \gamma, v_1 = 1, v_2 = 1 + q \) or \( p_1 = \gamma, p_2 = 1, v_1 = 1 + q, v_2 = 1 \).

In the first case, \( W = 2e_2(1 - \gamma^{1/2})/(\gamma^{1/2}(1 + q)^{1/2} - 1)) \), and in the
second case \( W = e_2(1 + (1 - \gamma)/(\gamma q)). \)

Proof. We set \( q' = (1 + q)^{1/2}, v_1 = aq', v_2 = q'/a, p_1 = b\gamma^{1/2} \) and \( p_2 = (1/b)\gamma^{1/2}. \) Then for
\[
\gamma^{1/2} \leq b \leq 1/\gamma^{1/2}
\]
and
\[
\max\left\{1/q', q'\gamma^{1/2}/b\right\} \leq a \leq \min\left\{1/b\gamma^{1/2}q', q'\right\}
\]
all the constraints are fulfilled. The objective becomes (setting \( e_2 = 1 \) without loss of generality)
\[
f(a, b) = \left(\frac{1 - b\gamma^{1/2}}{b\gamma^{1/2}(aq' - 1)}\right) + \left(\frac{1 - (1/b)\gamma^{1/2}}{(1/b)\gamma^{1/2}(q'/a - 1)}\right)
\]
\[
= \frac{1}{b\gamma^{1/2}(aq' - 1)} + b\gamma^{1/2}(q'/a - 1) - \frac{1}{aq' - 1} - \frac{1}{q'/a - 1}.
\]
A long, but not difficult argument shows that the only candidate for a local minimum of \( f \) on \( \mathbb{R}^+ \times \mathbb{R}^+ \) is \( a = b = 1 \), which corresponds to \( p_1 = p_2 = \gamma^{1/2}, v_1 = v_2 = q' \) and \( W = 2(1 - \gamma^{1/2})/(\gamma^{1/2}(q' - 1)). \)

It remains to check the situation on the boundary of the region given by (2.23) and (2.24). It is enough to consider \( a = \min\left\{1/b\gamma^{1/2}q', q'\right\} \) for \( \gamma^{1/2} \leq b \leq 1/\gamma^{1/2} \) because of symmetry. One can show that there \( f \) is minimized for \( p_1 = 1, p_2 = \gamma, v_1 = 1 \) and \( v_2 = (1 + q) \) with objective value \( 1 + (1 - \gamma)/(\gamma q) \), which is smaller or equal than \( 2(1 - \gamma^{1/2})/(\gamma^{1/2}(q' - 1)) \) when \( (1 - 4\gamma^{1/2} + 3\gamma)/\gamma \leq q \leq 1/\gamma - 1. \)

Proof of Proposition 2.8. Now we turn to the case for an arbitrary \( m. \)

Assertion 2.1. An optimal solution \((p_1, \ldots, p_m, v_1, \ldots, v_m)\) to (2.22) is of the following form:

\[
\exists \mathcal{I} \subset \{1, 2, \ldots, m\}:
\]
\[
(\forall k \in \mathcal{I} : p_k = 1, v_k = 1 \text{ and } \forall k \in \mathcal{I}^C : p_k = p < 1, v_k = v > 1).
\]
From this, the result is deduced directly, because to fulfill the constraints, \( p \) must be equal \( \gamma \frac{1}{m-|I|} \) and \( v = (1 + q)^{1/(m-|I|)} \). The objective value is then
\[
e_m \left( |I| + (m - |I|) \cdot \frac{1 - \gamma^{1/(m-|I|)}}{\gamma^{1/(m-|I|)}((1 + q)^{1/(m-|I|)} - 1)} \right).
\]

Proof of Assertion 2.1. Let us assume we have an optimal solution with \( p_i < 1 \) (\( \Rightarrow v_i > 1 \)) and \( p_j < 1 \) (\( \Rightarrow v_j > 1 \)) with \( v_i \neq v_j \) or \( p_i \neq p_j \). Without loss of generality, \( i = 1 \) and \( j = 2 \). We set \( \gamma^* = p_1 p_2 \) and \( q^* + 1 = v_1 v_2 > 1 \).

If \( q^* < (1 - 4 \gamma_*^{1/2} + 3 \gamma_*)/\gamma_* \), then for \( p_1^* = p_2^* = \gamma_*^{1/2} \), \( p_k^* = p_k \), \( k \in \{3, \ldots, m\} \) and \( v_1^* = v_2^* = (q^* + 1)^{1/2} \), \( v_k^* = v_k \), \( k \in \{3, \ldots, m\} \) we have a strictly smaller total average effort \( W^* \) because of Lemma 2.4, which is a contradiction.

If \( (1 - 4 \gamma_*^{1/2} + 3 \gamma_*)/\gamma_* \leq q^* \leq \frac{1}{\gamma_*} - 1 \), we get a strictly better solution for \( p_1^* = 1 \), \( v_1^* = 1 \), \( p_2^* = \gamma^* \) and \( v_2^* = 1 + q^* \) using Lemma 2.4.

If \( q^* > \frac{1}{\gamma_*} - 1 \), then we simply choose \( v_1^* = v_2^* = p_1^* = p_2^* = w_1^* = w_2^* = 1 \) to reduce the total average effort. \( \square \)

2.B.2 Proof of Proposition 2.5

Define
\[ g_\gamma(x) = x^2(\gamma^{-1/x} - 1) \]
and
\[ \bar{g}(\gamma) = \min\{g_\gamma(m); m \in \mathbb{N}^+\}, \quad M_\gamma = \{m \in \mathbb{N}^+; g_\gamma(m) = \bar{g}(\gamma)\}. \]
It is easy to check that $g_\gamma(x)$ is decreasing for $0 < x \leq c|\log(\gamma)|$ and increasing for $x \geq c|\log(\gamma)|$ where $c$ is the solution of the equation (2.13). Hence $M_\gamma$ consists of one or two elements. Furthermore, we note that for any $m$

$$\frac{q/m}{(1 + q)^{1/m} - 1} \downarrow 1 \quad (q \to 0).$$

Hence if $q$ is chosen small enough such that for all $m \in M_\gamma$

$$\frac{q/m}{(1 + q)^{1/m} - 1} < \frac{\min\{g_\gamma(m); m \notin M_\gamma\}}{\bar{g}(\gamma)},$$

then for $m_1 \in M_\gamma$ and $m \notin M_\gamma$

$$W(\gamma, q, m) = \frac{1}{q} g_\gamma(m) \frac{q/m}{(1 + q)^{1/m} - 1} \quad \frac{1}{q} g_\gamma(m_1) \frac{q/m_1}{(1 + q)^{1/m_1} - 1} = W(\gamma, q, m_1).$$

The rest of the proof follows by simple calculations.

2.C Proofs for asymptotic log-normality

2.C.1 Proof of Proposition 2.6

First, we prove some lemmas:

**Lemma 2.5.** If $Z \sim \text{NegBin}(r, p)$ with $r \geq 2$, $0 < p < 1$ and $k \in \mathbb{N}_0$. Then

$$E\left[\log^k \left( \frac{r - 1}{r - 1 + Z} \right) \right] \leq \frac{E[Z^k]}{(r - 1)^k} < \infty.$$

**Proof.** We have

$$\left| \log^k \left( \frac{r - 1}{r - 1 + Z} \right) \right| = \log^k \left( 1 + \frac{Z}{r - 1} \right) \leq \frac{Z^k}{(r - 1)^k}.$$
Chapter 2. Splitting Simulation

The rest follows from the monotonicity of the expected value and the fact that the Negative Binomial distribution has bounded moments.

Lemma 2.6. For $2 \leq r \leq R$, $0 < p_{\text{min}} < p_{\text{max}} < 1$ and $Z \sim \text{NegBin}(r, p)$, we have

$$\min_{2 \leq r \leq R} \left\{ \min_{p \in [p_{\text{min}}, p_{\text{max}}]} \left\{ \text{Var} \left( \log \left( \frac{r-1}{r-1+Z} \right) \right) \right\} \right\} > 0.$$  

Proof. Set $f(r, p) := \text{Var} \left( \log \left( \frac{r-1}{r-1+Z} \right) \right) < \infty$. Because $f(r, p)$ is continuous in $p$ and $f(r, p) > 0$ unless $p = 0$ or $p = 1$, the assertion follows.

Lemma 2.7. For $2 \leq r \leq R$, $0 < p_{\text{min}} < p_{\text{max}} < 1$ and $Z \sim \text{NegBin}(r, p)$, the maximum over $2 \leq r \leq R$ and $p \in [p_{\text{min}}, p_{\text{max}}]$ of

$$E\left[ \left| \log \left( \frac{r-1}{r-1+Z} \right) - E \left[ \log \left( \frac{r-1}{r-1+Z} \right) \right] \right|^3 \right]$$

is finite.

Proof. We have with Lemma 2.5

$$E\left[ \left| \log \left( \frac{r-1}{r-1+Z} \right) - E \left[ \log \left( \frac{r-1}{r-1+Z} \right) \right] \right|^3 \right] \leq \frac{E[Z^3]}{(r-1)^3} + 3 \frac{E[Z^2]E[Z]}{(r-1)^3} + 4 \frac{E[Z]^3}{(r-1)^3} \leq \frac{1}{(r-1)^3} \left[ r(1-p)(2-p) + 3r^2(1-p)^2 + 4r^3(1-p)^3 \right].$$

Clearly, this expression remains bounded for $2 \leq r \leq R$ and $p_{\text{min}} < p < p_{\text{max}}$.

With the help of these lemmas, Proposition 2.6 can be proved by applying Lyapunov’s Central Limit Theorem.
2.C.2 Proof of Proposition 2.7

Clearly it is sufficient to prove the Proposition for \( m = 1 \) since the \( \hat{p}_k \) and \( \tilde{p}_k \) respectively are independent for different \( k \)'s. By the Central Limit Theorem for \( n \to \infty \)

\[
n^{1/2}(\hat{p}_k - p_k) \to \mathcal{N}(0, p_k(1-p_k)/d_k).
\]

Hence (2.19) follows by the delta technique for \( f(u) = \log(u) \). For (2.18), we combine the delta technique with the next lemma.

**Lemma 2.8.** Let \( 0 < p < 1 \). Set \( \hat{p} = (r - 1)/(r - 1 + Z) \) where \( Z \sim \text{NegBin}(r, p) \). Then we have

\[
r^{1/2}(\hat{p} - p) \to \mathcal{N}(0, p^2(1-p)) \quad (r \to \infty).
\]

**Proof.** First, observe that \( Z = \sum_{i=1}^{r} Z_i \), where \( Z_i \sim \text{i.i.d} \text{ NegBin}(1, p) \). Therefore, it follows from the Central Limit Theorem that

\[
r^{1/2}(Z/r - \mathbb{E}[Z/r]) = r^{1/2}(Z/r - (1-p)/p) \to \mathcal{N}(0, (1-p)/p^2) \quad (r \to \infty).
\]

Clearly, this remains true if we substitute \( Z/r \) with \( Z/(r - 1) \). Now we use the delta technique with \( f(u) = 1/(1 + u) \).

2.D Numerical calculation of the overflow probabilities

In some examples in Section 2.6, we are able to calculate the overflow probabilities numerically. Here we describe briefly how this can be done. For more details and further calculation methods we refer to e.g. Bolch et al. (2006) and Stewart (2009).

Consider a homogeneous continuous-time Markov chain on a finite state space \( \{0, 1, \ldots, E\} \) with generator matrix \( Q \) and transition matrix \( P \) of the embedded discrete-time Markov chain, which is assumed to be
aperiodic. Further, we write $\pi^{(0)} = \nu^{(0)}$ for the initial probability distribution vector, $\pi(t)$ for the probability distribution vector at time $t$ and $\nu^{(n)}$ for the probability distribution vector after $n$ time-steps of the embedded chain. In this situation, the limiting distribution $\pi$ exists and is given by

$$\pi = \lim_{n \to \infty} \nu^{(n)}.$$  

Because $\nu^{(n+1)} = \nu^{(n)} P$, one can numerically approximate the limiting distribution $\pi$ by just repeatedly multiplying the initial probability vector $\nu^{(0)}$ with $P$ until some convergence criterion is met (see Bolch et al. (2006) Section 3.5.2).

Note that we do not need irreducibility with this technique, which is important since we will consider chains with absorbing states in the following. In the reducible cases, the limiting distribution $\pi$ depends in general on the initial distribution.

If we are interested in the transient probabilities $\pi(t)$, one can use the technique of uniformization (see Bolch et al. (2006), Section 5.1.4, or Stewart (2009), Section 10.7.2). Setting $A = Q/q + I$ for some $q > \max_{i,j} |Q_{i,j}|$ one can show

$$\pi(t) = \sum_{n=0}^{\infty} \alpha^{(n)} \exp(-qt) \frac{(qt)^n}{n!}, \quad \alpha^{(n)} = \alpha^{(n-1)} A, \quad \alpha^{(0)} = \pi^{(0)}.$$

Using an appropriate truncation point for the infinite series, we have a numerical approximation of $\pi(t)$.

We first consider the case $\tau_B < \tau_A$, that is, the homogeneous continuous-time Markov chain $\mathcal{X}$ starting in $x_0$ on countable state space $\mathcal{E}$ enters the set $B$ before $A$ ($x_0 \notin (A \cup B)$. We label the states in $\mathcal{E} \setminus (A \cup B)$ with $\{1, \ldots, E - 1\}$ assuming $x_0$ has the label 1, using a suitable truncation if $\mathcal{E} \setminus (A \cup B)$ is infinite, replace all the states in $A$ with the absorbing state 0 and all the states in $B$ with the absorbing state $E$. This means all transitions going to some state in $A$ and $B$, respectively, go to 0 and $E$, respectively, in the resulting chain with finite state space. We then
have
\[ P^X[\tau_B < \tau_A | X_0 = x_0] = \pi_E, \]
i.e., the probability to enter \( B \) before \( A \) is equal to the limiting probability being in state \( E \) of the resulting chain with initial distribution \( \pi(0) = (0, 1, 0 \ldots, 0)^T \). The statement is only approximately true if we have truncated the state space.

For the case \( \tau_B < T \) for a fixed time \( T \) we proceed analogously (there is no need to replace \( A \)) and find
\[ P^X[\tau_B \leq T | X_0 = x_0] = \pi_E(T), \]
i.e., the probability to hit \( B \) before time \( T \) is (approximately) equal the transient probability being in state \( E \) at time \( T \) of the resulting chain with initial distribution \( \pi(0) = (0, 1, 0 \ldots, 0)^T \). As above, the statement only approximately holds when a truncation of the state space was done.

Now, we can use the methods described above to get numerical approximations of \( \pi_E \) and \( \pi_E(T) \), respectively.

There exist other numerical methods to calculate such probabilities for homogeneous continuous-time Markov chains, see Stewart (2009), Chapter 10, or Bolch et al. (2006), Chapter 4 and Chapter 5. For methods based on Laplace transform calculation, see also Asmussen and Jobmann (2002) for a transient buffer overflow in a M/M/1 queue.

### 2.E Simulation or numerical solution

The numerical methods from Section 2.D above are fast compared to simulation and work well for systems of moderate size. But for bigger systems, underflow can occur and rounding errors can be substantial, since the probability distribution vectors have components of very different orders of magnitude. Thus even for homogenous continuous-time
Markov chains, splitting simulation can be advantageous, since it does not suffer from this problem.

However, the main advantage of splitting simulation is that in principle we only need an algorithm to generate trajectories of the process (at least approximately), which can usually be done in a simple and effective way (see Section 1.1). Thus these methods are usually more easily and more generally applicable, for example also in queueing systems with non-exponential arrival and service times (where the system can not be described with a continuous-time Markov chain, since the state space is uncountable; see the examples with Weibull times in Section 2.6) or when complicated diffusion processes (such as in financial engineering) are involved.
Chapter 3

Rate estimation in partially observed Markov jump processes with measurement errors

3.1 Introduction

It is generally accepted that many important intracellular processes, e.g. gene transcription and translation, are intrinsically stochastic, because chemical reactions occur at discrete times as results from random molecular collisions (McAdams and Arkin (1997) and Arkin et al. (1998)). These stochastic kinetic models correspond to a Markov jump process and can thus be simulated using techniques such as the Gillespie algorithm (Gillespie (1977)) or – in the time-inhomogeneous case – Lewis' thinning method (Ogata (1981)). Many of the parameters in such models are uncertain or unknown, therefore one wants to es-
timate them from times series data. One possible approach is to approximate the model with a diffusion and then to perform Bayesian (static or sequential) inference based on the approximation (see Golightly and Wilkinson (2005), Golightly and Wilkinson (2006), Golightly and Wilkinson (2008) and Golightly and Wilkinson (2009)). This gives more flexibility to generate the proposals (see Durham and Gallant (2002)), but it is difficult to quantify the approximation error. In the Appendix 3.B, we even provide simple examples where we show that this approach can lead to inconsistent estimates. Therefore, depending on the application, it might be preferable to work with the original Markov jump process. This possibility is mentioned in Wilkinson (2006), Chapter 10, and Boys et al. (2008) demonstrate in the case of the simple Lotka-Volterra model that this approach is feasible in principle, but in more complex situations it is difficult to construct a Markov chain Monte Carlo (MCMC) sampler with good mixing properties. The key problems in our view are to construct good proposals for the latent process on an interval when the values at the two end points are fixed and the process is close to the boundary of the state space, and to construct reasonable starting values for the process and the parameters, in particular when some of the components are observed with small or zero noise. We propose here solutions for both of these problems that go beyond Wilkinson (2006), Chapter 10, and Boys et al. (2008) and thus substantially enlarge the class of models that are computationally tractable.

The rest of the chapter is organized as follows. In Section 3.2, we describe the model, establish the relation to stochastic kinetics and introduce useful notation and densities. In Section 3.3, we motivate the Bayesian approach and present the base frame of the MCMC algorithm. Section 3.4 describes in detail certain aspects of the algorithm, mainly the construction of proposals for the latent Markov jump process. In Section 3.5, the particle filter type algorithm to initialize values for the parameters and for the latent Markov jump process is presented. In Section 3.6, we look at two examples. First, the stochastic Oregonator (see Gillespie (1977)) is treated in various scenarios, including some data-poor ones, to show how the algorithm works. Then, we turn to
a model for prokaryotic auto-regulation introduced in Golightly and Wilkinson (2005) and reconsidered in Golightly and Wilkinson (2009). Finally, conclusions are given in Section 3.7.

3.2 Setting and definitions

3.2.1 Model

Consider a Markov jump process

$$\mathcal{Y} = \{y_t = (y_1^t, \ldots, y_p^t)^T : t \geq t_0\}$$

on a state space $\mathcal{E} \subset \mathbb{N}_0^p$ with jump vectors $A_i \in \mathbb{Z}^p$ for $i \in \{1, \ldots, r\}$ and possibly time dependent transition intensities $\mu_i(t, y) = \theta_i \cdot h_i(t, y)$:

$$P[y_{t+\delta} = y + A_i | y_t = y] = \mu_i(t, y)\delta + o(\delta) \quad (\delta > 0).$$

We denote the total transition intensity by

$$\mu_0(t, y) = \sum_{i=1}^r \mu_i(t, y).$$

We assume that the functions $h = \{h_i\}_{i \in \{1, 2, \ldots, r\}}$, called the standardized transition intensities, the jump matrix $A$ with columns $A_i$ and the initial distribution $f_0$ of $y_{t_0}$ are known. The goal is to estimate the hazard rates $\theta = (\theta_1, \ldots, \theta_r)$ from partial measurements $x_0, x_1, \ldots, x_n$ of the process at discrete time points $0 = t_0 < t_1 < \cdots < t_n$. Unobserved components are set to na and we assume

$$x_l | \mathcal{Y} = x_l | y_{t_l} \sim g_\eta(\cdot | y_{t_l}),$$

where $g_\eta(x_l | y_{t_l})$ is a density with respect to some $\sigma$-finite measure (with possibly unknown) nuisance parameter $\eta$. We specify this more precisely in the examples in Section 3.6.
If a row in the matrix \( A \) a is a linear combination of the others, say

\[
A_{lj} = \sum_{i \neq l}^{p} \lambda_i A_{ij} \quad \forall j \in \{1, \ldots, r\},
\]

then

\[
y_l^t - \sum_{i \neq l}^{p} \lambda_i y_i^t = \text{const} \quad \forall t.
\]

Throughout the chapter, we assume these conservation constants to be known. Therefore, we can remove \( y_l^t \) from the system and assume in the following that \( \text{rank}(A) = p \leq r \).

This framework can be regarded as a general state space model: \( x_0, x_1, \ldots \) is an observed times series which is derived from the unobservable Markov chain \( y_{t_0}, y_{t_1}, \ldots \) (see Künsch (2000) or Doucet et al. (2001)).

For computational reasons, we further assume that we can easily evaluate the time-integrated standardized transition intensities

\[
H_i(s, t, y) := \int_{s}^{t} h_i(u, y)du.
\]

Models of the above form arise for example in the context of stochastic kinetics. Consider a biochemical reaction network with \( r \) reactions \( R_1, \ldots, R_r \) and \( p \) species \( Y^1, \ldots, Y^p \), i.e.,

\[
R_i : v_{i1} Y^1 + \cdots + v_{ip} Y^p \longrightarrow u_{i1} Y^1 + \cdots + u_{ip} Y^p
\]

for \( i = 1, \ldots, r \). Let \( y_{i}^j \) denote the number of species \( Y^j \) at time \( t \), \( y_t = (y_1^1, \ldots, y_p^p)^T \), \( V = (v_{ij}) \) and \( U = (u_{ij}) \). Then, according to the mass action law, we can describe \( \{y_t : t \geq t_0\} \) as a Markov jump process with jump matrix \( A = (U - V)^T \) and standardized reaction intensities

\[
h_i(y) = \prod_{j, v_{ij} \geq 1} \left( \frac{y^j}{v_{ij}} \right).
\]
3.2. Setting and definitions

For further details, see e.g. Gillespie (1977), Golightly and Wilkinson (2005) or Golightly and Wilkinson (2009). We will use in the following terminology from this applications: We will call the jump times reaction times and classify a jump as one of the \( r \) possible reaction types.

3.2.2 Additional notation and formulae for densities

A possible path \( y_{[a,b]} \) on an interval \([a,b]\) in our model is uniquely characterized by the total number of reactions \( n_{\text{tot}} \), the initial state \( y_a \), the successive reaction times \( a < \tau_1 < \ldots < \tau_{n_{\text{tot}}} \leq b \) and the reaction types (or indices) \( r_1, r_2, \ldots, r_{n_{\text{tot}}} \in \{1, \ldots, r\} \). The states at the reaction times are then obtained as

\[
y_{\tau_k} = y_a + \sum_{i=1}^{k} A_{r_i}.
\]

We write for simplicity \( y_k \) instead of \( y_{\tau_k} \). Furthermore, \( r_{i_{\text{tot}}}^i \) is the total number of reactions of type \( i \) and \( r_{\text{tot}} \) is the vector with components \( r_i^{i_{\text{tot}}} \). All these quantities depend on the interval \([a,b]\). If this interval is not clear from the context, we write \( n_{\text{tot}}([a,b]), \tau_k([a,b]), \text{ etc.} \)

The density \( \psi_\theta \) of \( y_{[a,b]} \) given \( y_a \) is well known, see e.g. Wilkinson (2006), Chapter 10. Defining \( \tau_0 = a, \tau_{n_{\text{tot}}+1} = b \) and \( y_{\tau_0} = y_0 = y_a \), it is given by

\[
\psi_\theta(y_{[\tau_0,\tau_n]}|y_0) = \exp \left( - \sum_{i=1}^{r} \theta_i \int_a^b h_i(s, y_s) ds \right) \cdot \prod_{k=1}^{n_{\text{tot}}} \theta_{r_k} h_{r_k}(\tau_k, y_{k-1})
\]

\[
= \exp \left( - \sum_{k=1}^{n_{\text{tot}}+1} \sum_{i=1}^{r} \theta_i H_i(\tau_{k-1}, \tau_k, y_{k-1}) \right) \cdot \prod_{k=1}^{n_{\text{tot}}} \theta_{r_k} h_{r_k}(\tau_k, y_{k-1}).
\]
In the time-homogeneous case, i.e., \( h_i(t, y) = h_i(y) \), we have

\[
H_i(\tau_{k-1}, \tau_k, y_{k-1}) = h_i(y_{k-1}) \delta_k
\]

with \( \delta_k = \tau_k - \tau_{k-1} \). Therefore

\[
\delta_k | \tau_{k-1}, y_{k-1} \sim \text{Exp}(\mu_0(y_{k-1})) \tag{3.1}
\]

and

\[
P[r_k = i | \tau_{k-1}, y_{k-1}] = \frac{\mu_i(y_{k-1})}{\mu_0(y_{k-1})}, \tag{3.2}
\]

and we can exactly simulate the Markov jump process using the Gillespie algorithm (see Gillespie (1977)) or some faster versions thereof (see Gibson and Bruck (2000)). Replacing \( h_i(y_{k-1}) \) by \( h_i(\tau_{k-1}, y_{k-1}) \) in (3.1) and (3.2), this can be done “approximately” in the inhomogeneous case. An exact simulation algorithm based on a thinning method is described in Ogata (1981).

We write the density of all observations in \([a, b]\) as

\[
g_\eta(x[a,b]|y[a,b]) = \prod_{t,a \leq t \leq b} g_\eta(x_t|y_{t})
\]

where the empty product is interpreted as 1. The joint density of \( y_{[t_0,t_n]} \) and \( x_{[t_0,t_n]} \) (given the parameters \( \theta \) and \( \eta \)) is then

\[
p(y_{[t_0,t_n]}, x_{[t_0,t_n]}| \theta, \eta) = f_0(y_{t_0}) \cdot \psi_\theta(y_{[t_0,t_n]}|y_{t_0}) \cdot g_\eta(x_{[t_0,t_n]}|y_{[t_0,t_n]}).
\tag{3.3}
\]

### 3.3 Bayesian estimation and Monte Carlo methods

The maximum likelihood estimator is too complicated to compute because we are not able to calculate the marginalisation of the density in (3.3) over \( y_{[t_0,t_n]} \) explicitly. It seems easier to combine a Bayesian
approach with Monte Carlo methods, that is we will sample from the posterior distribution of the parameters and the underlying Markov jump process $y_{[t_0,t_n]}$ given the data (see Robert and Casella (2004)). This has also the additional advantage that prior knowledge about the reaction rates can be used. Assuming $\theta$ and $\eta$ to be independent a priori, the joint distribution of $y_{[t_0,t_n]}$, $x_{[t_0,t_n]}$, $\theta$ and $\eta$ has the form

$$p(y_{[t_0,t_n]}, x_{[t_0,t_n]}, \theta, \eta) = p(y_{[t_0,t_n]}, x_{[t_0,t_n]} | \theta, \eta) \cdot p(\theta) \cdot p(\eta).$$

We want to simulate from the conditioned density $p(y_{[t_0,t_n]}, \theta, \eta | x_{[t_0,t_n]})$, which yields also samples from $p(\theta, \eta | x_{[t_0,t_n]})$ using a marginalisation over $y_{[t_0,t_n]}$. The standard approach to do this is iterating between block-wise updates of the latent process $y_{[t_0,t_n]}$ on subintervals of $[t_0, t_n]$ with Metropolis-Hastings steps, updates of $\theta$ and updates of $\eta$ (see e.g. Gilks et al. (1996), Chapter 1, Boys et al. (2008) or Golightly and Wilkinson (2009)).

As in Boys et al. (2008), we choose independent Gamma distributions with parameters $\alpha_i$ and $\beta_i$ as priors for $\theta_i$:

$$p(\theta) \propto \prod_{i=1}^{r} \theta_i^{\alpha_i-1} \exp(-\beta_i \theta_i).$$

We write this distribution as $\Gamma_r(\alpha, \beta)$ where $\alpha$ and $\beta$ are vectors of dimension $r$. Conditionally on $y_{[t_0,t_n]}, x_{[t_0,t_n]}$ and $\eta$, the components $\theta_i$ have then again independent Gamma distributions, more precisely

$$\theta | y_{[t_0,t_n]}, x_{[t_0,t_n]}, \eta \sim \theta | y_{[t_0,t_n]}$$

$$\sim \Gamma_r \left( \tilde{\alpha}(y_{[t_0,t_n]}), \tilde{\beta}(y_{[t_0,t_n]}) \right),$$

with

$$\tilde{\alpha}_i(y_{[t_0,t_n]}, \alpha_i) = \alpha_i + r_i^{tot}$$

(3.4)
and

\[
\tilde{\beta}_i(y_{[t_0,t_n]}, \beta_i) = \beta_i + \int_{t_0}^{t_n} h_i(s; y_s) ds \nonumber
\]
\[
= \beta_i + \sum_{k=1}^{ntot+1} H_i(\tau_{k-1}, \tau_k, y_{k-1}).
\]

Choosing a suitable prior for \(\eta\) depends heavily on the error distribution, so we refer to the examples in Section 3.6.

We propose the following algorithm, which will be explained in more detail in the next sections. The generation of initial values \(y_{(0)}[t_0, t_n], \theta^{(0)}\) and \(\eta^{(0)}\) will be discussed in Section 3.5. The choice of the set \(\mathcal{I}_{[t_0, t_n]}\) of overlapping subintervals \([a, b] \subset [t_0, t_n]\) for updating \(y\) will be discussed in Section 3.4.4.

**Algorithm 3.1** (Simulation from \(y_{[t_0,t_n]}, \theta, \eta | x_{[t_0,t_n]}\)).

For \(m = 1, 2, \ldots, M\):

1. Set \(y_{[t_0,t_n]} = y_{(m-1)}^{(m-1)}, \theta = \theta^{(m-1)}, \eta = \eta^{(m-1)}\). Update \(y_{[a,b]}\) for all \([a, b] \in \mathcal{I}_{[t_0, t_n]}\) sequentially in a random order by proposing \(y_{[a,b]}^{new}\) as described in sections 3.4.1, 3.4.2, 3.4.5 and 3.4.6 and replacing \(y_{[a,b]}\) by \(y_{[a,b]}^{new}\) with probability \(\alpha(y_{[a,b]}^{new}|y_{[a,b]}, \theta, \eta)\) (see (3.12)). Set \(y_{[t_0,t_n]}^{(m)} = y_{[t_0,t_n]}\).

2. Simulate \(\theta^{(m)} \sim \Gamma_r(\tilde{\alpha}(y_{[t_0,t_n]}^{(m)}), \tilde{\beta}(y_{[t_0,t_n]}^{(m)})\).

3. Generate \(\eta^{(m)}\) given \(y_{[t_0,t_n]}^{(m)}\) in a suitable fashion.
3.4 Simulating a path given parameters and observations

We assume now that \( \theta \) and \( \eta \) are fixed and we want to modify \( y_{[a,b]} \) on subintervals \([a, b]\) of \([t_0, t_n]\). First we consider the case \( t_0 < a < b < t_n \) where the values \( y_a \) and \( y_b \) remain unchanged. The boundary cases will be discussed in Section 3.4.6. Exact methods to simulate from a continuous time Markov chain conditioned on both endpoints are reviewed and discussed in Hobolth and Stone (2009). The rejection method is too slow in our examples, and direct sampling and uniformization require finite state space and eigendecompositions of the generator matrix. This would require truncating the state space and is too time-consuming in our examples. Hence we use a Metropolis-Hastings procedure. Our proposal distribution \( q \) first generates a vector of new total reaction numbers \( r_{tot}^{new} \) on \([a, b]\) and then, conditioned on \( r_{tot}^{new} \), generates a value \( y_{[a,b]}^{new} \).

3.4.1 Generating new reaction totals

Because the values \( y_a \) and \( y_b \) are fixed, we must have that

\[
Ar_{tot}^{new} = y_b - y_a = Ar_{tot} \iff A(r_{tot}^{new} - r_{tot}) = 0. \quad (3.5)
\]

If \( p = r \iff \text{rank}(A) = r \), \( A \) is invertible and the reaction totals remain unchanged. Otherwise it is known that \( \text{kernel}(A) := \{x \in \mathbb{Z}^r : A \cdot x = 0\} \) forms a lattice and can be written as \( \{a_1 \cdot v_1 + \cdots + a_d \cdot v_d : a_1, \ldots, a_d \in \mathbb{Z}\} \) with \( d = \dim(\text{kernel}(A)) = r - p \) and basis vectors \( v_l \in \mathbb{Z}^r, l \in \{1, 2, \ldots, d\} \) (note that these vectors are not unique). Appendix 3.A describes how to compute a basis vector matrix

\[
V(A) = (v_1, \ldots, v_d).
\]

This enables us to generate a vector \( r_{tot}^{new} \) which respects (3.5) in a
simple way:
\[ r_{\text{tot}}^{\text{new}} = r_{\text{tot}} + V(A) \cdot Z, \quad Z \sim q^Z_i, \quad (3.6) \]
where \( q^Z_i \) is a symmetric proposal distribution \( q^Z_i \) on \( \mathbb{Z}^d \), i.e., \( q^Z_i(z) = q^Z_i(-z) \), with parameter \( \iota \). If \( r_{\text{tot}}^{\text{new}} \) has a negative component, we stop and set \( y_{[a,b]}^{\text{new}} = y_{[a,b]} \).

### 3.4.2 Generating a new path given the reaction totals

The new path \( y_{[a,b]}^{\text{new}} \) depends only on \( y_a \) and the new reaction totals \( r_{\text{tot}}^{\text{new}} \), and not on the old path \( y_{[a,b]} \). The constraint \( y_{b}^{\text{new}} = y_{b} \) is satisfied automatically by our construction of \( r_{\text{tot}}^{\text{new}} \). Therefore our algorithm simply generates a path on \([a,b]\) with given initial value and given reaction totals, and we can omit the superscripts \( \text{new} \).

A first possibility is to generate the path according to \( r \) independent inhomogeneous Poisson processes with intensities
\[
\lambda_i(t) = \mu_i(a, y_a) \frac{b - t}{b - a} + \mu_i(b, y_b) \frac{t - a}{b - a}, \quad (3.7)
\]
conditioned on the totals \( r_{\text{tot}}^{\text{new},i} \) as in Boys et al. (2008) for the simple Lotka-Volterra reaction system. In many cases this leads to proposals which approximate the true jump process nicely. But in situations where the standardized reaction intensities \( h_i \) depend strongly on \( y \), this proposal often generates paths that are impossible under the model. This is typically the case when the number of molecules of some species is small. In order to address this problem, we constructed the following proposal which first decides the order in which the reactions take place, that is we first generate \( r_k \) for \( k = 1, 2, \ldots, n_{\text{tot}} \). In a second step, we generate the reaction times \( \tau_k \). In the first step, we take into account both the possibility of a reaction of a given type to occur at the current state of the process and the remaining number of reactions \( S_k^i \) of type \( i \) after time \( \tau_k \) that still have to occur in order to reach the prescribed total. In the second step, we take the values of the intensities into
3.4. Simulating a path given parameters and observations

account. In order to make the description of the algorithm easier to read, we mention that \( t_k^* \) is a first guess for \( \tau_{k-1} \) (needed only if the intensities are time-inhomogeneous). Also remember that \( y_k = y_{\tau_k} \).

**Algorithm 3.2** (Generating \( y_{[a,b]} \) given \( r_{\text{tot}} \) and \( y_a \)).

1. Set \( S^i_0 = r^i_{\text{tot}} \) for \( i \in \{1, \ldots, r\} \) and \( y_0 = y_a \).

2. For \( k = 1, \ldots, n_{\text{tot}} \) do the following:
   Set \( t^*_k = a + (b - a)(k - 1)/n_{\text{tot}} \). If \( \mu_l(t^*_k, y_{k-1}) = 0 \) for all \( l \) with \( S^l_{k-1} > 0 \), stop. Otherwise, generate \( r_k \) with probabilities
   \[
   P[r_k = i] \propto \mathbb{I}_{(0, \infty)}(\mu_i(t^*_k, y_{k-1})) \cdot S^i_{k-1}. \tag{3.8}
   \]
   If \( r_k = i \), set \( S^i_k = S^i_{k-1} - 1 \), \( S^l_k = S^l_{k-1} \) for \( l \neq i \) and \( y_k = y_{k-1} + A_i \).

3. Generate \((\delta_k; k \in \{1, \ldots, n_{\text{tot}} + 1\})\) according to a Dirichlet distribution with parameter \( \alpha = (\alpha_k; k \in \{1, \ldots, n_{\text{tot}} + 1\}) \) where
   \[
   \alpha_k = \mu_0^{-1}(t^*_k, y_{k-1}) \sum_l \mu_0^{-1}(t^*_l, y_{l-1}) \sum_l \mu_0^{-2}(t^*_l, y_{l-1}), \tag{3.9}
   \]
   and set \( \tau_k = \tau_{k-1} + (b - a)\delta_k \) for \( k = 1, \ldots, n_{\text{tot}} \).

The algorithm stops in step 2 when we can no longer reach the state \( y_b \) on a possible reaction path using the available remaining reactions. This means that an impossible path is proposed which has acceptance probability 0.

The heuristics behind the steps in the above algorithm is the following. The probabilities (3.8) are an attempt to ensure that a reaction of type \( i \) at the current state is possible according to the law of the process and we nevertheless reach the prescribed reaction total. Of course, there are many other possibilities to define the probabilities in (3.8), e.g. the geometric mean

\[
\sqrt{S^i_{k-1}\mu_i(t^*_k, y_{k-1})}.
\]
Empirically, we found that the above variant leads to good acceptance rates in the examples in Section 3.6.

The Dirichlet distribution in (3.9) is used as an approximation of the distribution of independent exponential-$\mu_0(t^*_k, y_{k-1})$ waiting times $\delta_k$ conditioned on the event that their sum is equal to $b - a$ because of the following considerations. If all $\mu_0(t^*_k, y_{k-1})$ are equal, the conditional first two moments are

$$
E\left[\delta_k \mid \sum_l \delta_l = b - a\right] = (b - a) \frac{E[\delta_k]}{\sum_l E[\delta_l]} \quad (3.10)
$$

and

$$
\text{Var}\left[\delta_k \mid \sum_l \delta_l = b - a\right] = (b - a)^2 \times
\left(\frac{\text{Var}(\delta_k) + E[\delta_k]^2}{\sum_l \text{Var}(\delta_l) + (\sum_l E[\delta_l])^2} - \left(\frac{E[\delta_k]}{\sum_l E[\delta_l]}\right)^2\right), \quad (3.11)
$$

and moreover the conditional distribution is Dirichlet with parameters $\alpha_k = 1$, scaled by $b - a$, see e.g. Bickel and Doksum (1977), Section 1.2. In the general case, we use a Dirichlet distribution as approximation and determine the parameters such that the expectation matches the right-hand side of (3.10) for all $k$. This implies that

$$
\alpha_k \propto \mu_0^{-1}(t^*_k, y_{k-1}).
$$

Finally, the proportionality factor is determined such that the sum of the variances matches the sum of the right-hand sides of (3.11).

### 3.4.3 Acceptance probability of a new path

By construction, the proposal density $q(y_{[a,b]}^{\text{new}} | y_{[a,b]}, \theta)$ has the form

$$
q(y_{[a,b]}^{\text{new}} | y_a, r_{tot}^{\text{new}}, \theta)q(r_{tot}^{\text{new}} | r_{tot}).
$$
Because of the symmetry of $q^Z$, we have

$$q(r_{\text{tot}}^{\text{new}} | r_{\text{tot}}) = q(r_{\text{tot}} | r_{\text{tot}}^{\text{new}}).$$

So it will cancel out in the acceptance probability and we do not need to consider it.

Next, $q(y_{[a,b]}|y_a, r_{\text{tot}}, \theta)$ is equal to

$$\prod_{k=1}^{n_{\text{tot}}} \frac{\mathbb{I}_{(0,\infty)}(\mu_i(t^*_k, y_{k-1}))S^i_{k-1}}{\sum_{l=1}^{r} \mathbb{I}_{(0,\infty)}(\mu_l(t^*_k, y_{k-1}))S^l_{k-1}} \times f^{\text{Dir}}(\tau_k - \tau_{k-1}) / (b-a) : k \in \{1, \ldots, n_{\text{tot}} + 1\}$$

where $f^{\text{Dir}}$ is the density of the Dirichlet distribution with parameter $\alpha$ from (3.9).

Hence, according to the Metropolis-Hastings recipe, the acceptance probability $\alpha(y_{[a,b]}^{\text{new}}|y_{[a,b]}, \theta, \eta)$ is

$$\min \left\{ 1, \frac{\psi_\theta(y_{[a,b]}^{\text{new}}|y_a)g_\eta(x_{[a,b]}|y_{[a,b]}^{\text{new}})q(y_{[a,b]}|y_a, r_{\text{tot}}, \theta)}{\psi_\theta(y_{[a,b]}|y_a)g_\eta(x_{[a,b]}|y_{[a,b]})q(y_{[a,b]}^{\text{new}}|y_a, r_{\text{tot}}^{\text{new}}, \theta)} \right\}. \quad (3.12)$$

### 3.4.4 Choice of the subintervals $[a, b]$  

To ensure that the process can be updated on the whole interval $[t_0, t_n]$, we have to choose a suitable set of subintervals $I_{[t_0, t_n]}$ for which we apply the above updating algorithms. As a general rule, one can say that they should be overlapping. Also it is often useful to include subintervals which do not lead to a change of the process at the observation times $t_1 < t_2 < \cdots < t_n$. In such situations, the terms $g_\eta(x_{[a,b]}|y_{[a,b]}^{\text{new}})$ and $g_\eta(x_{[a,b]}|y_{[a,b]})$ are equal and therefore cancel out in the acceptance probability.

In cases where the observations are complete and noise-free, we need only the subintervals of the form $[t_{l-1}, t_l]$. However, because it is some-
times a non-trivial problem to find a realization of the Markov jump process which matches all observations, we found that it is sometimes useful to include a tiny noise in the model and to choose also subintervals with a $t_l$ as interior point. By this trick we can often obtain realizations that match all observation by the above updating algorithms.

In general, good choices of the subintervals can be very dependent on the given situation. The standard one is to let $\mathcal{I}_{[t_0,t_n]}$ consist of all intervals of the form $[t_{l-1}, t_l]$ and $[(t_{l-1} + t_l)/2, (t_l + t_{l+1})/2]$.

### 3.4.5 Updating latent components

Let us assume that at the time $t_l$ the $j$-th component is not observed, i.e., $x^j_l = \text{na}$, but the others are. Especially when working with small or no noise, updating with the above proposal on $[a,b]$ where $a < t_l < b$ is problematic because of the following reason: Assume $y_l$ already matches $x_l$ nicely on the observed components. Then a new proposal can only be accepted when $y^\text{new}_l$ matches the observed components, too. Thus not only the acceptance rate is low, but more severely, $y^j_l$ remains usually unchanged although we do not have any information on $y^j_l$.

To circumvent this problem, we construct a proposal on the interval $[a,b]$, $a < t_l < b$, which generates simultaneously new reactions totals on the intervals $[a,t_l]$ and $[t_l,b]$, $r^\text{new}_{tot,a}$ and $r^\text{new}_{tot,b}$, respectively, so that the values $y_{ta}$, $y_{tb}$ and the observed values of $y_l$ remain fixed, but $y^j_l$ can change. Similarly to (3.5), we consider solutions of

$$A_{-j,.}(r^\text{new}_{tot,a} - r_{tot,a}) = 0$$

where $A_{-j,.}$ denotes the reaction matrix without the $j$-th row. Because of the assumption that $\text{rank}(A) = p \leq r$, $\text{kernel}(A_{-j,.}) \setminus \text{kernel}(A)$ is non-empty. Thus we can draw $v$ from $\text{kernel}(A_{-j,.}) \setminus \text{kernel}(A)$ in a
symmetric manner, so that the reaction totals
\[ r_{\text{tot},a}^{\text{new}} = r_{\text{tot},a} + v \quad \text{and} \quad r_{\text{tot},b}^{\text{new}} = r_{\text{tot},b} - v \] (3.13)
fulfill the above requirements. To propose the jump processes on the two intervals \([a, t_l] \) and \([t_l, b]\), we can use the techniques from Section 3.4.2. The calculation of the acceptance probability is similar to the one described in Section 3.4.3. The symmetric proposal distribution for the vector \(v\) from \(\text{kernel}(A_{-j, \cdot}) \setminus \text{kernel}(A)\) is specified for the examples considered in Section 3.6.

### 3.4.6 Updating the path at a border

In the cases \(b = t_n\) or \(a = t_0\) we also want to change the values of \(y_{t_n}\) and \(y_{t_0}\), respectively (unless \(f_0\) is a Dirac measure). We recommend to propose first a change in \(r_{\text{tot}}^{\text{new}}\), that is
\[ r_{\text{tot}}^{\text{new}} = r_{\text{tot}} + r', \quad r' \sim q_{r'}, \] (3.14)
where \(q_{r'}\) is a symmetric distribution on \(\mathbb{Z}^r\). Then either \(y_a\) or \(y_b\) remains unchanged and the other value follows from \(y_b - y_a = A r_{\text{tot}}^{\text{new}}\). The rest can be done again with Algorithm 3.2. If \(y_{t_0}^{\text{new}} \neq y_{t_0}\), the factor \(f_0(y_{t_0}^{\text{new}})/f_0(y_{t_0})\) is needed additionally in the acceptance probability (3.12).

If one wants to change only some components of \(y_{t_0}\) or \(y_{t_n}\), respectively, the same ideas as in Section 3.4.5 can be used. For more details, see the examples in Section 3.6.

### 3.5 Initialisation of \(\eta, \theta\) and \(y_{[t_0, t_n]}\)

The form of the trajectories of the underlying Markov jump process depends strongly on the parameter \(\theta\) and the value at \(t_0\). So just choosing \(\eta^{(0)}\) and \(\theta^{(0)}\) and then simulating \(y_{[t_0, t_n]}^{(0)}\) leads usually to processes which
match the observed data badly. It then takes very many iterations in
the algorithm until we obtain processes that are compatible with the
data.

In our experience, generating the starting values by algorithm 3.3 below
leads to substantial increases in computational efficiency. It is inspired
by the particle filter: We select the most likely particle, perform a
number of Metropolis-Hastings steps (similarly to Gilks and Berzuini
(2001)) and propagate with the Gillespie algorithm.

An additional trick can bring further improvement. Because the speed
of the techniques described depends heavily on the number of reactions
in the system, one wants to ensure that the initial value \( y_{t_0,t_n} \) for
Algorithm 3.1 has rather too few than too many reactions. We can
achieve this with a simple shrinkage factor \( \nu \) between 0 and 1 for \( \theta \)
during the initialisation, that is replacing \( \theta \) after simulation with \( \nu \cdot \theta \).
This acts like a penalisation on the reaction numbers: It does not
affect the probabilities in (3.2) (time-homogeneous case), but makes
the system slower, resulting in fewer reactions.

**Algorithm 3.3** (Generating starting values).

1. Choose \( \eta^{(0)} \).

2. Simulate \( S^{[1]} \) i.i.d starting values \( y^s_{t_0} \sim p(y_{t_0}|x_{t_0}) \) and gener-
ate \( y^s_{[t_0,t_1]} \) for \( s \in \{1,2,\ldots,S^{[1]}\} \) using the Gillespie algorithm
with the normalized standardized reaction intensities \( \mathbb{I}_{\{h_i>0\}} \) \( (i = \)
1,\ldots,r) and equal hazard rates \( 1/(t_1 - t_0) \). Set \( y_{[t_0,t_1]}^{[1]} = y^s_{[t_0,t_1]} \),
where \( s' = \text{argmax}_s \{g_{\eta^{(0)}}(x_{[t_0,t_1]}|y^s_{[t_0,t_1]})\} \). Simulate
\( \theta^{[1]} \sim \Gamma_r \left( \tilde{\alpha}(y_{[t_0,t_1]}^{[1]}), \tilde{\beta}(y_{[t_0,t_1]}^{[1]}) \right) \).

3. For \( l = 1,\ldots,n - 1 \):

   a) Use \( M^{[l]} \) steps of algorithm 3.1 on \( [t_0,t_l] \) with shrinkage fac-
tor \( \nu \) and starting values \( y_{[t_0,t_l]}^{[l]} \) and \( \theta^{[l]} \) to generate \( y_{[t_0,t_l]}^{[l+1]} \)
and \( \theta^{[l+1]} \).
3.6. Examples

b) Generate $S^{[l]}$ paths $y_{[t_0, t_{l+1}]}^s$ which are independent continuations of $y_{[t_0, t_{l}]}^{[l+1]}$ on $(t_l, t_{l+1}]$, based on the Gillespie algorithm with $\theta^{[l+1]}$ and set $y_{[t_0, t_{l+1}]}^{[l+1]} = y_{[t_0, t_{l+1}]}^{s'}$, where $s'$ is equal to $\arg\max_s \{g_{\eta^{(0)}}(x_{l+1} | y_{[t_0, t_l]}^{s})\}$.

4. Set $\theta^{(0)} = \theta^{(n)}$ and $y_{[t_0, t_n]}^{(0)} = y_{[t_0, t_n]}^{(n)}$.

So to propagate to the process on the interval $(t_l, t_{l+1}]$ (for $l = 1, \ldots, n-1$), we use $\theta^{[l+1]}$ which should roughly follow the distribution of $\theta$ given $x_{[t_0, t_l]}$ and $\eta^{(0)}$, because of step 3.a).

3.6 Examples

3.6.1 Stochastic Oregonator

First we consider the stochastic Oregonator to illustrate the algorithms. It is a highly idealized model of the Belousov-Zhabotinskii reactions, a non-linear chemical oscillator. It has 3 species and the following 5 reactions:

\[
\begin{align*}
R_1: & \quad Y^2 \quad \rightarrow \quad Y^1 \\
R_2: & \quad Y^1 + Y^2 \quad \rightarrow \quad \emptyset \\
R_3: & \quad Y^1 \quad \rightarrow \quad 2Y^1 + Y^3 \\
R_4: & \quad 2Y^1 \quad \rightarrow \quad \emptyset \\
R_5: & \quad Y^3 \quad \rightarrow \quad Y^2
\end{align*}
\] (3.15)

For further details, see Gillespie (1977). Following Section 3.2.1, the process $\{y_t : t \geq t_0\}$, where $y_t = (y^1_t, y^2_t, y^3_t)^T$ and $y^i_t$ is the number of species $Y^i$ at time $t$, is a Markov jump process with standardized reaction intensities

\[
h(y) = (y^2, y^1 y^2, y^1, y^1(y^1 - 1)/2, y^3)^T
\]
and the jump matrix

\[ A := \begin{pmatrix} 1 & -1 & 1 & -2 & 0 \\ -1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & -1 \end{pmatrix}. \]

As starting distribution \( f_0 \) for \( y_{t_0} \), we use the uniform distribution on \( \{0, \ldots, K\}^3 \) with \( K = 25 \). The measurement errors are normally distributed with precision \( \eta \), that is

\[ g_\eta(x, y) = \prod_{j:x^j \neq y^j} \frac{\sqrt{\eta}}{\sqrt{2\pi}} \exp \left( -\frac{\eta}{2} (x^j - y^j)^2 \right). \]  

(3.16)

In Figure 3.1, a sample trajectory for \( \theta = (0.1, 0.1, 2.5, 0.04, 1) \), simulated with the Gillespie algorithm, is shown, observed every \( 1/3 \) units of time during a time period of 20 units.

**Figure 3.1:** Sample trajectory of the Oregonator Markov jump process at observation times \( 0, \frac{1}{3}, \frac{2}{3}, \ldots, 20 \): \( y^1_t \) (squares), \( y^2_t \) (circles) and \( y^3_t \) (triangles). The thin lines indicate the process between the observation times.
If we choose a Gamma($\alpha, \beta$) prior for $\eta$, then the full conditional distribution of $\eta$ in the posterior is again a Gamma distribution with parameters $\tilde{\alpha}^\eta(x_{[t_0,t_n]}, \alpha)$ equal to

$$\alpha + \frac{1}{2} \#\{(l, j) \in \{1, \ldots, n\} \times \{1, \ldots, r\} : x^j_l \neq \text{na}\}$$

and

$$\tilde{\beta}^\eta(y_{[t_0,t_n]}, x_{[t_0,t_n]}, \beta) = \beta + \frac{1}{2} \sum_{(l,j):x^j_l \neq \text{na}} \left( x^j_l - y^j_{t_l} \right)^2 .$$

where $\text{na}$ denotes an unobserved value. This yields a simple way to perform step 3. in Algorithm 3.1.

We now want to estimate the parameters and the Markov jump process given in Figure 3.1 with total reaction numbers

$$r_{tot} = (76, 417, 518, 92, 508)^T$$

from observations at the times $T = \{0, \frac{1}{3}, \frac{2}{3}, \ldots, 20\}$. We analyze the following situations:

**Full observation (F):** We observe every species at the time points $T$

(a) exact ($\eta = \infty$);

(b) with error ($\eta = 1$ unknown).

**Species 1 latent (S1):** We observe only $Y^2$ and $Y^3$ at the time points $T$

(a) exact ($\eta = \infty$);

(b) with error ($\eta = 1$ known).

**Species 2 latent (S2):** We observe only $Y^1$ and $Y^3$ at the time points $T$

(a) exact ($\eta = \infty$);

(b) with error ($\eta = 1$ known).
Species 3 latent (S3): We observe only $Y^2$ and $Y^3$ at the time points $T$

(a) exact ($\eta = \infty$);
(b) with error ($\eta = 1$ known).

We use near uniform $\Gamma(1, 0.01) = \text{Exp}(0.01)$ priors on the parameters $\theta$, so that the mode of the posterior should be near to the maximum likelihood estimator. When $\eta$ is unknown, we use a $\Gamma(2, .2)$ prior. It is also rather flat, but ensures that the precision is greater than 0.

The scenarios S1, S2 and S3 are expected to be difficult for the MCMC algorithm because of the additional complication of having to mix over the uncertainty of the latent species. In particular the scenario S1, where we do not observe $Y_1$, because it is the most involved species, as we can see from the reaction system (3.15): It appears in 3 reactions as reactant and in 2 as product. The other species $Y_2$ and $Y_3$ occur at most twice as reactant and only once as product.

Specifications of the algorithm

We specify the proposal distributions and further details in our algorithm as follows. A basis vector matrix is given by

$$V(A) = \begin{pmatrix} 1 & -1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix}^T$$

and we simulate $Z$ in (3.6) as follows

$$Z = (2B_s - 1) \cdot (B_c \cdot B_1, (1 - B_c) \cdot B_2)^T$$

where $B_s \sim \text{Bin}(1, 1/2)$, $B_c \sim \text{Bin}(1, 1/3)$, $B_1 \sim \text{Bin}(1, 9/10)$ and $B_2 \sim \text{Bin}(6, 1/2)$. We need the larger variance in the second component to ensure good mixing properties since simulation shows that the total numbers of reaction 2,3 and 5 vary over a much bigger range than
3.6. Examples

reactions 1 and 4. Further, we use the standard set of subintervals described in Subsection 3.4.4.

For the initialisation (Algorithm 3.3), we use $M^{[l]}$ and $S^{[l]}$ around 100 to 200, slight shrinking and $\eta = 8$ when unknown. We use the standard set of subintervals described in Section 3.4.4.

In the scenarios with exact observation (a), we add a tiny normal noise to the model ($\sigma^2 = 1/\eta = 10^{-4}$). By this trick, we obtain underlying jump processes that match all observations after several iterations of Algorithm 3.1 as mentioned in Section 3.4.4.

In the scenarios with the latent components (S1, S2 and S3), we replace the standard update from above on intervals $[(t_{l-1} + t_l)/2, (t_l + t_{l+1})/2]$ containing an observation every second MCMC iteration with the special update described in Section 3.4.5 on $[t_{l-1}, t_{l+1}]$, so that the latent component mixes better. To find a suitable distribution to simulate the vector $v$ in (3.13), we look exemplarily at the scenario S1. First, we need the integer solutions to

$$A_{-1,.} x = 0.$$

With the techniques from Appendix 3.A, we find the basis vectors $v_1 = (1, -1, 0, 0, 0)^T$, $v_2 = (0, 0, 0, 1, 0)^T$ and $v_3 = (0, 1, 1, 0, 1)^T$. Because the last one is already in the kernel of $A$, we can restrict ourselves to $v_1$ and $v_2$ for the proposal of $v$, i.e., we choose $\pm v_1$ or $\pm v_2$ with equal probability $1/4$.

For the new reaction number at the beginning on the interval $[t_{n-1}, t_n]$ or at the end on the interval $[t_{n-1}, t_n]$, we want updates which change only one component of $y_{t_0}$ or $y_{t_n}$, respectively, to get better acceptance. In order to do this for the first component, can use the same proposal as above and add the resulting vector to the total reaction number to get the new one.

For the scenarios S2 and S3, basis vectors of $\text{kernel}(A_{-2,.})$ are

$$(1, 1, 0, 0, 0)^T, (0, -2, 0, 1, 0)^T, (0, 1, 1, 0, 1)^T$$
and for kernel($A_{-3,.}$)

$$(-1, 1, 2, 0, 0)^T, \ (0, 0, 2, 1, 0)^T, \ (0, 1, 1, 0, 1)^T,$$

where only $(0, 1, 1, 0, 1)$ is in kernel($A$).

**Results**

First, we analyze average acceptance rates for the different scenarios separately for updates which do not change the values of the process at the observation times, i.e., updates on intervals $[t_{l-1}, t_l]$ (A), updates on intervals $[(t_{l-1} + t_l)/2, (t_l + t_{l+1})/2]$ (B) with the standard technique and updates on intervals $[t_{l-1}, t_{l+1}]$ with the special update for the scenarios with latent components described in Section 3.4.5 (C). An overview is given in Table 3.1.

We see that it is a good idea to include the intervals with no changes of the process at observation times since the acceptance rate is much higher on these. When observation is exact (a) in the scenarios with latent components, the need of using the special update (C) is evident, for noisy observations (b), it is not that important.

Here, we proposed the underlying jump process according to Algorithm 3.2. The performance of the proposal based on independent Poisson processes (3.7) was comparable. But between times 3 to 6, where at least one component is very small, Algorithm 3.2 was clearly better (acceptance rates are up to 2.5 times that big), as expected from the heuristics given in Section 3.4.2.

In Figure 3.2 and 3.3, we show the trace plots for the parameters ($\theta, \eta$) exemplarily in scenario Fb and S1a. On the whole, mixing seems satisfactory, although not optimal for some parameters in the scenario S1a with the latent species. Nevertheless, we can produce good estimates as we will see below. In addition, the initialisation process yields starting values which are already very close to the true values. As expected, it reduces the burn in considerably and thus makes parallelization (if
Table 3.1: Average acceptance rates in % for the different Oreg-onator scenarios and process updates on intervals \([t_{l-1},t_l](A),\]
\([\frac{t_{l-1}+t_l}{2}, \frac{t_l+t_{l+1}}{2}](B)\) and \([t_{l-1},t_{l+1}](C)\) when latent components occur with the technique from Section 3.4.5.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>F a</td>
<td>17.1</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>17.0</td>
<td>2.6</td>
</tr>
<tr>
<td>S1 a</td>
<td>13.3</td>
<td>0.5</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>9.9</td>
<td>3.1</td>
</tr>
<tr>
<td>S2 a</td>
<td>16.9</td>
<td>0.5</td>
<td>7.8</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>17.1</td>
<td>4.7</td>
</tr>
<tr>
<td>S3 a</td>
<td>16.6</td>
<td>0.5</td>
<td>7.5</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>17.2</td>
<td>4.6</td>
</tr>
</tbody>
</table>

needed) more efficient.

To analyze if our chains are long enough and to compare the algorithm in the different scenarios, we calculate effective sample sizes. The effective sample size (ESS) gives the size of an i.i.d sample with the same variance as the current sample and thus indicates the loss of of the efficiency due the use of the Markov chain (see e.g. Robert and Casella (2004), Section 12.3.5 and 12.6). Note that the estimation of the ESS is a delicate issue, so the values should be interpreted only as an indication of the order. In Table 3.2, we estimated the ESS per 10^5 iterations of the MCMC scheme (Algorithm 3.1) with the function `effectiveSize` from the package `coda` in the language for statistical computing R (see R Development Core Team (2010)). This function fits an AR\(p\) process to the traces of each parameter. The value of the asymptotic variance is then given by a well-known formula.

We observe that the ESS for the parameters in the scenarios S1, S2 and S3 is much smaller than in the case of full observation. Especially for the reaction rates corresponding to standardized transition intensities which depend on the latent component. For example in scenario S1, \(Y^1\) is not observed, leading to a loss in terms of mixing for reactions
Figure 3.2: Traces for the parameters in scenario Fb for the Oreg
onator example. The origin on the abscissa marks the last iteration of
the initialisation (Algorithm 3.3). True values are indicated with a
horizontal line.
Figure 3.3: Traces for the parameters in scenario S1a for the Oregonator example. The origin on the abscissa marks the last iteration of the initialisation (Algorithm 3.3). True values are indicated with a horizontal line.
rates $\theta_2$, $\theta_3$ and $\theta_4$. Or in scenario S2, where the ESS of $\theta_1$ and $\theta_2$ is very low. The reason for this is that the Algorithm has to mix over the latent components (see Figure 3.6). Nevertheless, we are able to produce reasonable estimates, as shown below.

In Figure 3.5, we compare the posterior densities of the parameters in the scenarios Fa, Fb, S1a, S2a and S3a. Densities were calculated using a retransformed estimate of the logarithmic densities of the parameters with the function `density` in the language for statistical computing R (see R Development Core Team (2010)) with a Gaussian kernel and the default function internal rule to determine the bandwidth. The vertical dotted line indicates the true values of $\theta$ and $\eta$, respectively. In Table 3.3, we give an overview of estimated modes and 95% credible intervals of the parameters in the different scenarios. Figure 3.4 shows corresponding box plots of the simulated parameters $\theta$. All calculations are based on 200'000 iterations of Algorithm 3.1.

In the scenarios with full observation (F), the estimated mode or the median (Figure 3.4) is always near the true value. When a component is missing, this is not always the case, but nevertheless, the true value is mostly in regions where the posterior is high. In the scenario S1, the posterior of $\theta_2$, $\theta_3$ and $\theta_4$ seem rather spread out. But this are exactly the reaction rates depending directly on the unobserved component.

### Table 3.2: ESS per $10^5$ iterations (Oregonator).

<table>
<thead>
<tr>
<th></th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\theta_3$</th>
<th>$\theta_4$</th>
<th>$\theta_5$</th>
<th>$\eta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fa</td>
<td>a 13055</td>
<td>1050</td>
<td>1062</td>
<td>11338</td>
<td>1053</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b 9861</td>
<td>783</td>
<td>791</td>
<td>7887</td>
<td>792</td>
<td>150</td>
</tr>
<tr>
<td>S1</td>
<td>a 4389</td>
<td>192</td>
<td>138</td>
<td>73</td>
<td>498</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b 3442</td>
<td>84</td>
<td>62</td>
<td>45</td>
<td>268</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>a 62</td>
<td>57</td>
<td>598</td>
<td>1344</td>
<td>552</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b 56</td>
<td>53</td>
<td>507</td>
<td>992</td>
<td>482</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>a 1157</td>
<td>354</td>
<td>328</td>
<td>1041</td>
<td>326</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b 514</td>
<td>369</td>
<td>355</td>
<td>455</td>
<td>169</td>
<td></td>
</tr>
</tbody>
</table>
3.6. Examples

Table 3.3: Estimated 2.5% quantiles, modes and 97.5% quantiles (Oregonator).

<table>
<thead>
<tr>
<th></th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\theta_3$</th>
<th>$\theta_4$</th>
<th>$\theta_5$</th>
<th>$\eta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>0.100</td>
<td>0.100</td>
<td>2.500</td>
<td>0.040</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>a</td>
<td>0.083</td>
<td>0.087</td>
<td>2.204</td>
<td>0.027</td>
<td>0.866</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.106</td>
<td>0.104</td>
<td>2.549</td>
<td>0.034</td>
<td>1.007</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.137</td>
<td>0.125</td>
<td>2.977</td>
<td>0.043</td>
<td>1.178</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>0.087</td>
<td>0.089</td>
<td>2.252</td>
<td>0.028</td>
<td>0.892</td>
<td>0.737</td>
</tr>
<tr>
<td></td>
<td>0.111</td>
<td>0.107</td>
<td>2.627</td>
<td>0.035</td>
<td>1.042</td>
<td>1.163</td>
</tr>
<tr>
<td></td>
<td>0.143</td>
<td>0.132</td>
<td>3.113</td>
<td>0.044</td>
<td>1.236</td>
<td>2.115</td>
</tr>
<tr>
<td>S1</td>
<td>a</td>
<td>0.089</td>
<td>0.102</td>
<td>2.555</td>
<td>0.036</td>
<td>0.795</td>
</tr>
<tr>
<td></td>
<td>0.114</td>
<td>0.134</td>
<td>3.276</td>
<td>0.060</td>
<td>0.933</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.146</td>
<td>0.187</td>
<td>4.400</td>
<td>0.119</td>
<td>1.116</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>0.091</td>
<td>0.120</td>
<td>2.918</td>
<td>0.040</td>
<td>0.776</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.116</td>
<td>0.168</td>
<td>3.925</td>
<td>0.079</td>
<td>0.926</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.149</td>
<td>0.272</td>
<td>6.211</td>
<td>0.260</td>
<td>1.154</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>a</td>
<td>0.098</td>
<td>0.088</td>
<td>2.258</td>
<td>0.027</td>
<td>0.889</td>
</tr>
<tr>
<td></td>
<td>0.136</td>
<td>0.119</td>
<td>2.651</td>
<td>0.034</td>
<td>1.048</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.218</td>
<td>0.187</td>
<td>3.152</td>
<td>0.045</td>
<td>1.248</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>0.083</td>
<td>0.071</td>
<td>2.355</td>
<td>0.027</td>
<td>0.934</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.113</td>
<td>0.096</td>
<td>2.779</td>
<td>0.035</td>
<td>1.103</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.221</td>
<td>0.191</td>
<td>3.358</td>
<td>0.047</td>
<td>1.335</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>a</td>
<td>0.073</td>
<td>0.079</td>
<td>1.992</td>
<td>0.023</td>
<td>0.935</td>
</tr>
<tr>
<td></td>
<td>0.097</td>
<td>0.101</td>
<td>2.452</td>
<td>0.031</td>
<td>1.139</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.130</td>
<td>0.127</td>
<td>2.992</td>
<td>0.041</td>
<td>1.394</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>0.107</td>
<td>0.089</td>
<td>2.279</td>
<td>0.029</td>
<td>0.659</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.141</td>
<td>0.111</td>
<td>2.747</td>
<td>0.04</td>
<td>0.807</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.181</td>
<td>0.139</td>
<td>3.305</td>
<td>0.053</td>
<td>1.006</td>
<td></td>
</tr>
</tbody>
</table>

Finally, Figure 3.6 displays point-wise 95% credible bands of the latent components in the process for the scenarios S1a, S2a and S3a. For comparison, we also indicate the true values of the latent component with a thin line. We can see that they mostly lie within our credible bands which shows that our algorithm can reliably recover the unobserved process component. Note that the values for $Y^1$ in scenario S1a are rather underestimated, leading to lower values for the standardized
Figure 3.4: Box plots of the parameters $\theta$ generated with the MCMC Algorithm 3.1 for the different scenarios (Oregonator model). True values are indicated with a horizontal dotted line.
Figure 3.5: Posterior densities of the parameters $\theta$ and $\eta$ for the Oregonator model in the scenarios Fa (thick-solid), Fb (thin-solid), S1a (dashed), S2a (dotted) and S3a (dot-dashed). True values are indicated with a vertical dotted line.
Figure 3.6: Point-wise 95% credible bands (indicated by the thick lines) of the totally latent components for the Oregonator model in the scenarios S1a (top), S2a (middle) and S3a (bottom), respectively. The true values are shown as thin line.
reaction intensities depending on $y^1$. This explains the overestimation of the directly associated parameters $\theta_2$, $\theta_3$ and $\theta_4$ (see Figure 3.5 or 3.4).

### 3.6.2 Prokaryotic auto-regulation

We look at the simplified model for prokaryotic auto-regulation introduced in Golightly and Wilkinson (2005) and reconsidered in Golightly and Wilkinson (2009). It is described by the following set of 8 chemical reactions.

\[
\begin{align*}
R_1 &: \text{DNA} + P_2 \rightarrow \text{DNAP}_2 \\
R_2 &: \text{DNAP}_2 \rightarrow \text{DNA} + P_2 \\
R_3 &: \text{DNA} \rightarrow \text{DNA} + \text{RNA} \\
R_4 &: \text{RNA} \rightarrow \text{RNA} + P \\
R_5 &: 2P \rightarrow P_2 \\
R_6 &: P_2 \rightarrow 2P \\
R_7 &: \text{RNA} \rightarrow \emptyset \\
R_8 &: P \rightarrow \emptyset
\end{align*}
\]

In this system, the sum $\text{DNAP}_2 + \text{DNA}$ remains constant, and we assume that this constant $K$ is known and equal to 10 in our simulation. Therefore it is enough to consider the four species $y = (y^1, y^2, y^3, y^4)^T = (\text{RNA}, P, P_2, \text{DNA})^T$, where RNA, P, P_2 and DNA are now interpreted as numbers of the corresponding species. According to the mass action law, the standardized transition intensities are

\[
h(y) = \begin{pmatrix}
\text{DNA} \times P_2 \\
K - \text{DNA} \\
\text{DNA} \\
\text{RNA} \\
P \times (P - 1)/2 \\
P_2 \\
\text{RNA} \\
P
\end{pmatrix}
\]
and the jump matrix is given by

$$A := \begin{pmatrix}
0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 1 & -2 & 2 & 0 & -1 \\
-1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.$$  

As starting distribution, we assume that the number of DNA molecules is uniformly distributed on \(\{0, \ldots, K\}\) and the other species are initially 0. Following Golightly and Wilkinson (2009), we again use normally distributed measurement errors, see (3.16). The update for \(\eta\) (step 4. in Algorithm 3.1), can then be done using Gamma distributions. Also we consider different scenarios in a similar manner to the last example.

**Full observation (F):** Observation of every species

(a) exact \((\eta = \infty)\);
(b) with error \((\eta = 4\) unknown).

**DNA is latent (L):** Observation of species RNA, P and P\(_2\), i.e., the DNA is not observed,

(a) exact \((\eta = \infty)\);
(b) with error \((\eta = 4\) unknown).

The true values of the parameters are

$$\theta = (0.1, 0.7, 0.6, 0.085, 0.05, 0.2, 0.2, 0.015)$$

and we observe the process every 0.5 units of time on the interval \([0, 50]\). The total reaction numbers for the true Markov jump process are

$$r_{tot} = (192, 190, 122, 53, 116, 99, 117, 7)^T.$$  

As reported in Golightly and Wilkinson (2005) and Golightly and Wilkinson (2009), ratios of the parameters \(\theta_1/\theta_2\) and \(\theta_5/\theta_6\), connected to the
reversible reaction pairs $R_1, R_2$ and $R_5, R_6$, respectively, are more precise than the individual rates. We found a similar behavior also for $\theta_3/\theta_7$ and $\theta_4/\theta_8$. This is related to the fact that adding or subtracting an equal number of the corresponding reaction between two consecutive observation times does not change the values of the Markov jump chain at these time points, making it rather difficult to tell how many of these reaction events should be there from discrete observations only. This implies also that there is a strong positive dependence in the posterior between these pairs of parameters.

Therefore we analyse the MCMC algorithm when working with the following reparameterization:

$$\begin{align*}
\rho_1 &= \theta_1 + \theta_2, \\
\rho_3 &= \theta_3 + \theta_7, \\
\rho_5 &= \theta_4 + \theta_8, \\
\rho_7 &= \theta_5 + \theta_6, \\
\rho_2 &= \frac{\theta_1}{\theta_1 + \theta_2}, \\
\rho_4 &= \frac{\theta_3}{\theta_3 + \theta_7}, \\
\rho_6 &= \frac{\theta_4}{\theta_4 + \theta_8}, \\
\rho_8 &= \frac{\theta_5}{\theta_5 + \theta_6}.
\end{align*}$$

(3.17)

In this parameterization, the true value is

$$\rho = (0.8, 0.125, 0.8, 0.75, 0.1, 0.85, 0.25, 0.2).$$

We use $\Gamma(a, b)$ priors for $\rho_l$ ($l = 1, 3, 5, 7$) and $\text{Beta}(d, e)$ priors for $\rho_k$ ($k = 2, 4, 6, 8$). For updating e.g. $(\rho_1, \rho_2)$, we factor the joint density of $(\rho_1, \rho_2)$ given $y_{[t_0, t_n]}$ as $p(\rho_1 | \rho_2) p(\rho_2)$. Then $p(\rho_1 | \rho_2)$ is a $\Gamma(a + r_{tot}^1 + r_{tot}^2, b + \rho_2 I_1 + (1 - \rho_2) I_2)$ density, and

$$p(\rho_2) \propto (b + \rho_2 I_1 + (1 - \rho_2) I_2)^{-(a + N_1 + N_2)} \rho_2^{d + N_1 - 1} (1 - \rho_2)^{e + N_2 - 1},$$

with $I_i = \sum_{k=1}^{n_{tot}+1} h_i(y_{k-1}) \delta_k$ ($i = 1, 2$). The factor

$$(\beta + \rho_2 I_1 + (1 - \rho_2) I_2)^{-(a + N_1 + N_2)}$$

can be approximated with piecewise linear upper bounds, so we can simulate from $p(\rho_2)$ using an adaptive accept-reject-method with mixtures of truncated Beta distributions as proposals. After we have generated $\rho$ is this way, we can easily find the corresponding $\theta$ using the inverse transform of (3.17). Thus we have now two ways to update $\theta$. The
standard variant (S) and the one via the transformation (3.17) from above (T). We will see in Section 3.6.2, that T has somewhat better mixing properties.

In the scenario F, we use near uniform $\Gamma(1,0.1)$ priors for $\theta$ as before. When working with $\rho$, we also use $\Gamma(1,0.1)$ priors for $\rho_l$ ($l = 1, 3, 5, 7$) and $B(1,1)$ priors, i.e., uniform priors on $[0,1]$, for $\rho_k$ ($k = 2, 4, 6, 8$), so that the mode of the posterior should be near to the maximum likelihood estimator.

In the scenario L, when working with the above priors, we have the problem that sometimes reaction 1 or 2 are removed from the system, i.e., the corresponding rates are estimated to be 0 or at least very small. Since we assume that these occur, we use for $\theta_1, \theta_2$ and $\rho_1 \Gamma(2,5)$ and for $\rho_2 \text{Beta}(1.2,1.2)$ priors instead. This provides the prior information that corresponding rate parameters are unlikely to be near zero and are around 0.1 to 1.

**Specifications of the algorithm**

The basis vector matrix is given by

$$V(A) = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}^T$$

and for $q_i^Z$ we choose $q_i^Z(\pm \vec{e}_i) = 0.1$ for $i \in \{1, 2, 3, 4\}$ and $q_i^Z(\vec{0}) = 0.2$ (see (3.6)).

To get the new total reaction number for the update at the beginning, i.e., on the interval $[t_0, t_1]$, we have to respect that $y_{t_0}^1 = y_{t_0}^2 = y_{t_0}^3 = 0$. We therefore only want to change the fourth component of $y_{t_0}$. So

$$A_{-4,\cdot}(r_{new}^{tot}(y_{[t_0,t_1]}) - r_{tot}(y_{[t_0,t_1]})) = 0.$$
With the techniques from Appendix 3.A, we find the same basis vectors as in \( V(A) \) plus the vector \( v_5 = (0, -1, 0, 2, 1, 0, 0, 0)^T \). So we use (3.14) where \( q_t^R \) draws not only from \( \pm v_5 \), but also from some other specific vectors in \( \ker(A_{-4,}) \setminus \ker(A) \), e.g. \( v_5 + V(A)_1 \). For scenario L, where the DNA is latent, we construct a proposal for \( v \) in (3.13) in the same way, since we also want to change the fourth component only.

**Results**

We analyze each of the four different scenarios Fa, Fb, La and Lb with both methods to generate the rate parameter, i.e., method S (standard) and T (using the transformation). As before, we first look at acceptance rates, see Table 3.4. Acceptance rates for proposals which update process values at observation times are much lower. In the scenarios where DNA is latent (L), the update according to (3.13) is not as important as in the previous example, since acceptance rates on intervals \( [(t_{l-1} + t_l)/2, (t_l + t_{l+1})/2] \) are similar.

Compared to the Poisson process proposal, Algorithm 3.2 was slightly better on average, with clear advantage when DNA is on a low level, i.e., 0 or 1.

**Table 3.4:** Average acceptance rates in % for the different scenarios and process updates on intervals \( [t_{l-1}, t_l] \) (A), \( [(t_{l-1} + t_l)/2, (t_l + t_{l+1})/2] \) (B) and \( [t_{l-1}, t_{l+1}] \) when latent components occur with the technique from Section 3.4.5 (C).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>FaS</td>
<td>39.4</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>FaT</td>
<td>39.3</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>FbS</td>
<td>39.8</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>FbT</td>
<td>39.3</td>
<td>4.6</td>
<td></td>
</tr>
<tr>
<td>LaS</td>
<td>42.0</td>
<td>2.3</td>
<td>2.8</td>
</tr>
<tr>
<td>LaT</td>
<td>41.0</td>
<td>2.5</td>
<td>2.4</td>
</tr>
<tr>
<td>LbS</td>
<td>43.6</td>
<td>5.6</td>
<td>2.3</td>
</tr>
<tr>
<td>LbT</td>
<td>42.7</td>
<td>6.3</td>
<td>1.7</td>
</tr>
</tbody>
</table>
Figure 3.7: Traces of $(\theta, \eta)$ for the prokaryotic auto-regulation model in scenario FbS. The origin on the abscissa marks the last iteration of the initialisation (Algorithm 3.3). True values are indicated with a horizontal line.
Figure 3.8: Pairs plot of $\theta$ in scenario FbS.

Figure 3.9: Pairs plot of $\rho$ in scenario FbT.
In Figure 3.7, we exemplarily show the trace plots of the initialisation and 250’000 iterations of Algorithm 3.1 for the scenario FbS. Stationary behaviour seems to be achieved a few iterations after the initialisation. Once again, Algorithm 3.3 is of utter utility.

In figures 3.8 and 3.9, we see that the parameter pairs \((\theta_1, \theta_2), (\theta_3, \theta_7), (\theta_4, \theta_8)\) and \((\theta_5, \theta_6)\) are positively correlated as expected. In the reparameterized case, correlation is reduced on the whole, but the pairs \((\rho_5, \rho_6),\) corresponding to \((\theta_4, \theta_8),\) and \((\rho_7, \rho_8),\) corresponding to \((\theta_5, \theta_6),\) are now negatively correlated.

To analyze gain of the reparameterization, we once again compute the ESS. Results are given in Table 3.5. For most parameters, the variant with the transformation \(T\) yields an improvement in terms of ESS and thus has better mixing properties. The ESS usually decreases comparing a scenario with observation errors to the same one without.

| Table 3.5: ESS per \(10^5\) iterations (prokaryotic auto-regulation). |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| \(\theta_1\)                    | \(\theta_2\)    | \(\theta_3\)    | \(\theta_4\)    | \(\theta_5\)    | \(\theta_6\)    | \(\theta_7\)    | \(\theta_8\)    | \(\eta\)        |
| FaS                             | 162             | 154             | 1507            | 21093           | 406             | 384             | 1463            | 6130            |
| FaT                             | 189             | 197             | 1609            | 23804           | 459             | 428             | 1718            | 6671            |
| FbS                             | 176             | 160             | 700             | 533             | 303             | 300             | 668             | 251             | 316             |
| FbT                             | 188             | 180             | 731             | 645             | 355             | 343             | 757             | 261             | 319             |
| LaS                             | 118             | 143             | 836             | 351             | 306             | 292             | 1696            | 259             |
| LaT                             | 212             | 217             | 661             | 403             | 505             | 463             | 1414            | 268             |
| LbS                             | 100             | 192             | 271             | 133             | 280             | 267             | 394             | 116             | 158             |
| LbT                             | 205             | 269             | 330             | 122             | 337             | 343             | 421             | 121             | 139             |

Figure 3.10 shows box plots and Table 3.6 and displays an overview of estimated modes and 95% credible intervals of the parameters in the different scenarios using the transformation to generate the proposals for \(\theta\), based on 100’000 iterations of the MCMC algorithm, which should be enough considering Table 3.5. We see that in the scenarios where every species is observed \((F)\), posterior modes or medians (see the box plots in Figure 3.10) are near the true values. When the DNA is latent \((L)\), true values are almost always contained in the credible intervals.
3.6. Examples

Figure 3.10: Box plots of the parameters $\theta$ generated with the MCMC Algorithm 3.1 for the different scenarios (prokaryotic auto-regulation). True values are indicated with a horizontal dotted line.
Figure 3.11: 95% credible bands (indicated by the thin lines) of the latent DNA in the scenarios LaT (top) and LbT (below) for the prokaryotic auto-regulation model. True values are shown as thick line.
3.7. Conclusions

Table 3.6: Estimated 2.5% quantiles, modes and 97.5% quantiles for the prokaryotic auto-regulation.

<table>
<thead>
<tr>
<th></th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\theta_3$</th>
<th>$\theta_4$</th>
<th>$\theta_5$</th>
<th>$\theta_6$</th>
<th>$\theta_7$</th>
<th>$\theta_8$</th>
<th>$\eta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FaT</td>
<td>.075</td>
<td>.370</td>
<td>.063</td>
<td>.082</td>
<td>.122</td>
<td>.108</td>
<td>.015</td>
<td>4.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.104</td>
<td>.505</td>
<td>.082</td>
<td>.050</td>
<td>.177</td>
<td>.148</td>
<td>.013</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.163</td>
<td>.707</td>
<td>.110</td>
<td>.078</td>
<td>.303</td>
<td>.207</td>
<td>.034</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FbT</td>
<td>.071</td>
<td>.395</td>
<td>.375</td>
<td>.061</td>
<td>.037</td>
<td>.126</td>
<td>.110</td>
<td>.001</td>
<td>3.570</td>
</tr>
<tr>
<td></td>
<td>.100</td>
<td>.554</td>
<td>.083</td>
<td>.049</td>
<td>.182</td>
<td>.159</td>
<td>.006</td>
<td>4.699</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.157</td>
<td>.787</td>
<td>.122</td>
<td>.082</td>
<td>.322</td>
<td>.231</td>
<td>.057</td>
<td>6.075</td>
<td></td>
</tr>
<tr>
<td>LaT</td>
<td>.053</td>
<td>.400</td>
<td>.302</td>
<td>.077</td>
<td>.031</td>
<td>.099</td>
<td>.110</td>
<td>.013</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.074</td>
<td>.417</td>
<td>.103</td>
<td>.046</td>
<td>.158</td>
<td>.151</td>
<td>.034</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.115</td>
<td>.609</td>
<td>.152</td>
<td>.070</td>
<td>.263</td>
<td>.210</td>
<td>.093</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LbT</td>
<td>.030</td>
<td>.297</td>
<td>.334</td>
<td>.058</td>
<td>.035</td>
<td>.145</td>
<td>.146</td>
<td>.009</td>
<td>2.772</td>
</tr>
<tr>
<td></td>
<td>.050</td>
<td>.484</td>
<td>.087</td>
<td>.053</td>
<td>.232</td>
<td>.208</td>
<td>.030</td>
<td>4.479</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.098</td>
<td>.730</td>
<td>.184</td>
<td>.087</td>
<td>.429</td>
<td>.309</td>
<td>.169</td>
<td>7.418</td>
<td></td>
</tr>
</tbody>
</table>

Finally, we compare the point-wise 95% credible bands of the latent component, that is the number of DNA molecules, in the scenarios LaT and LbT based on 100’000 iterations in Figure 3.11. In the case with observation error (LbT), there seems to be a slight overestimation, whereas for exact observation (LaT), the true underlying process component lies nicely within our credible bands.

3.7 Conclusions

In this chapter, we have presented a technique to infer rate constants and latent process components of Markov jump processes from time series data using fully Bayesian inference and Markov chain Monte Carlo algorithms. We have used a new proposal for the Markov jump process and - exploiting the general state space framework - a filter type initialisation algorithm to render the problem computationally more tractable. Even in very data-poor scenarios in our examples, e.g.
one species is completely unobserved, we have been able to estimate parameter values and processes and the true values are contained in the posterior credible bands.

The techniques are generic to a certain extend, but as our examples have shown, they have to be adapted to the situation at hand, which makes their blind application rather difficult. Clearly, the speed of our algorithm scales with the number of jump events, so it is less suitable in situations with many jumps. In such a situation, using the diffusion approximation can be useful. However, we believe that the statement “It seems unlikely that fully Bayesian inferential techniques of practical value can be developed based on the original Markov jump process formulation of stochastic kinetic models, at least given currently available computing hardware” in the introduction of Golightly and Wilkinson (2009) is too pessimistic.

3.A Integer solutions of homogeneous linear equations

Let \( A \in M_{p\times r}(\mathbb{Z}) \) be an integer \( p \times r \) matrix. We want to determine the set

\[
\mathbb{L} = \{ x \in \mathbb{Z}^r : Ax = 0 \}.
\]  

(3.18)

Obviously, it is enough to consider only linear independent rows of \( A \), so we assume \( \text{rank}(A) = p \leq r \). The case \( p = r \) is then trivial, so \( p < r \). The main idea is to transform the matrix \( A \) into the so-called Hermite normal form. We denote with \( \lfloor x \rfloor \) the largest integer smaller or equal \( x \). For the following, see Newman (1972), pages 15 ff, or Cohen (1993), pages 66 ff.

**Definition 3.1** (Hermite normal form). A matrix \( H \in M_{p\times r}(\mathbb{Z}) \) with rank \( s \) is in Hermite normal form if

\begin{enumerate}
  \item \( \exists i_1, \ldots, i_s \) with \( 1 \leq i_1 < \cdots < i_s \leq p \) with \( H_{i_j,j} \in \mathbb{Z}\backslash\{0\} \) for \( 1 \leq j \leq s \).
\end{enumerate}
2. \( H_{i,j} = 0 \) for \( 1 \leq i \leq i_j - 1, \ 1 \leq j \leq s \).

3. The columns \( s+1 \) to \( r \) are 0.

4. \( \lfloor H_{i,j,l}/H_{i,j,j} \rfloor = 0 \) for \( 1 \leq l < j \leq s \).

**Proposition 3.1.** For every \( A \in \mathbb{M}_{p \times r}(\mathbb{Z}) \) exists a unique unimodular matrix \( U \) (\( U \in GL_r(\mathbb{Z}) := \{ B \in M_{r \times r}(\mathbb{Z}) : \det(B) = \pm 1 \} \)), such that \( H = AU \) is in Hermite normal form.

There exist many algorithms to calculate \( H \) and \( U \), see e.g. Storjohann and Labahn (1996) or Jäger (2001).

The Hermite normal form allows us to determine the set (3.18). Because \( A \) is assumed to have maximal rank, by definition \( H = (B,0) \), where \( B \) is an invertible, lower triangular \( p \times p \) matrix. For \( y = U^{-1}x \) we have the equation

\[
0 = Ax = AUy = (B,0)y,
\]

so \( y \) has zeroes in the first \( p \) positions and arbitrary integers in the remaining positions. Hence a basis vector matrix \( V \) for (3.18) is given by \( v_i = u_{r+i} \). If necessary, one can reduce the length of the \( v_i \) by the Algorithm 2.3 in Ripley (1987).

### 3.B  Inconsistency when using diffusions

In Golightly and Wilkinson (2005), Golightly and Wilkinson (2006) and Golightly and Wilkinson (2009) the Markov jump process with jump Matrix \( A \) and reaction intensities \( \mu \) is replaced with the following diffusion

\[
d\bar{Y}_t = m(\bar{Y}_t, \theta)dt + \Sigma^{1/2}(\bar{Y}_t, \theta)dW_t
\]

with drift

\[
m(y, \theta) = \lim_{\delta \to 0} \frac{1}{\delta} E_\theta [Y_{t+\delta} - Y_t \mid Y_t = y] = \mu(y, \theta) \cdot A^T
\]
and diffusion term

\[ \Sigma(y, \theta) = \lim_{\delta \to 0} \frac{1}{\delta} \text{Cov}_\theta [Y_{t+\delta} - Y_t \mid Y_t = y] = A \cdot \text{diag}(\mu(y, \theta)) \cdot A^T. \tag{3.21} \]

Based on this diffusion, one can define a pseudo maximum likelihood estimator for \( \theta \), i.e.,

\[ \hat{\theta}_d = \arg \min_\theta \sum_{l=1}^n \log(dP_\theta[\tilde{Y}_{t_l} = Y_l \mid \tilde{Y}_{t_{l-1}} = Y_{l-1}]), \]

where \( dP_\theta[\tilde{Y}_{t_l} = Y_l \mid \tilde{Y}_{t_{l-1}} = Y_{l-1}] \) denotes the transition density given by the stochastic differential equation (3.19). Further, the diffusion can be approximated using the Euler–Maruyama method, i.e., \( \tilde{Y}_{t_i} \mid \tilde{Y}_{t_{i-1}} \) has the distribution

\[ \mathcal{N}(\tilde{Y}_{t_{i-1}} + m(\tilde{Y}_{t_{i-1}}, \theta)(t_i - t_{i-1}), \Sigma(\tilde{Y}_{t_{i-1}}, \theta)(t_i - t_{i-1})). \tag{3.22} \]

This enables us to define a second pseudo maximum likelihood estimator

\[ \hat{\theta}_{dEM} = \arg \min_\theta \sum_{l=1}^n \log(dP_\theta[\tilde{Y}_{t_l} = Y_l \mid \tilde{Y}_{t_{l-1}} = Y_{l-1}]), \]

where \( dP_\theta[\tilde{Y}_{t_l} = Y_l \mid \tilde{Y}_{t_{l-1}} = Y_{l-1}] \) denotes the transition density induced by (3.22).

We designed the examples in a way so that we are able to calculate the estimators \( \hat{\theta}_d, \hat{\theta}_{dEM} \) and their expectations under the law of the Markov jump process explicitly.

For a first example, we set

\[ \mu(y, \theta) = \left( \frac{2y^2 \theta}{(1 + (y - 1)^2)^3}, \frac{2(y - 2)^2 \theta}{(1 + (y - 1)^2)^3} \right). \tag{3.23} \]

and \( A = (-2, 2) \). Assuming \( Y_{t_0} \in \{0, 2\} \), we can easily see that this corresponds to a Markov jump process which alternates between the states 0 and 2 with transition intensity \( \theta \). The observation times are \( t_l = \Delta \cdot l, l \in \{0, 1, \ldots, n\} \). In this case the number of jumps between
two consecutive observation times is $\text{Poi}(\Delta \theta)$-distributed and therefore

$$P[Y_i = Y_{i-1} \mid Y_{i-1}] = \exp(-\Delta \theta \sum_{k=0}^{\infty} \frac{(\Delta \theta)^{2k}}{(2k)!})$$

$$= \exp(-\Delta \theta) \cosh(\Delta \theta)$$

(3.24)

and

$$P[Y_i \neq Y_{i-1} \mid Y_{i-1}] = \exp(-\Delta \theta \sum_{k=0}^{\infty} \frac{(\Delta \theta)^{2k+1}}{(2k+1)!})$$

$$= \exp(-\Delta \theta) \sinh(\Delta \theta).$$

(3.25)

We now calculate $\hat{\theta}_d$. With (3.20) and (3.21), we find

$$d\bar{Y}_t = -\frac{16(\bar{Y}_t - 1)\theta}{(1 + (\bar{Y}_t - 1)^2)^3} dt + \frac{4\sqrt{\theta}}{(1 + (\bar{Y}_t - 1)^2)} dW_t.$$

This stochastic differential equation is reducible to a linear stochastic differential equation (see Kloeden and Platen (1999), pages 117 ff.). Using the transformation

$$h(y) = \frac{y(y^2 - 3y + 6)}{3}$$

we find with the Itô formula (see Øksendal (1995), page 41) that

$$d(h(\bar{Y}_t)) = 4\sqrt{\theta} dW_t$$

and

$$h(\bar{Y}_{\Delta l}) - h(\bar{Y}_{\Delta(l-1)}) \overset{i.i.d.}{\sim} N(0, 16\Delta \theta)$$

for $l \in \{1, \ldots, n\}$. Thus

$$\hat{\theta}_d = \frac{1}{16\Delta n} \frac{1}{n} \sum_{l=1}^{n} (h(Y_{\Delta l}) - h(Y_{\Delta(l-1)}))^2.$$

With $(h(0) - h(2))^2 = (h(2) - h(0))^2 = 64/9$, $(h(0) - h(0))^2 = (h(2) -
Chapter 3. Rate Estimation

\( h(2)^2 = 0 \), (3.24) and (3.25) we derive

\[
E[(h(Y_{\Delta l}) - h(Y_{\Delta (l-1)}))^2] = E[E[(h(Y_{\Delta l}) - h(Y_{\Delta (l-1)}))^2 \mid Y_{\Delta (l-1)}]]
\]

\[
= \frac{64}{9} \exp(-\Delta \theta) \sinh(\Delta \theta)
\]

and therefore for \( n \to \infty \) we have

\[
\hat{\theta}_d \overset{a.s.}{\to} \frac{4}{9\Delta} \exp(-\Delta \theta) \sinh(\Delta \theta) = \frac{4}{9} \theta - \frac{4}{9} \Delta \theta^2 + O(\Delta^2),
\]

i.e., \( \hat{\theta}_d \) is not consistent, not even in the case where \( \Delta \) becomes very small.

To find \( \hat{\theta}_{dEM} \), we have to maximize

\[
K - \frac{n}{2} \log(\theta) - \sum_{l=1}^{n} \frac{1}{2\Sigma^2(Y_{\Delta (l-1)}, \theta)} (Y_{\Delta l} - Y_{\Delta (l-1)} - m(Y_{\Delta (l-1)}, \theta) \Delta)^2,
\]

where \( K \) does not depend on \( \theta \). Using that summands in the summation are equal to \(-\Delta \theta/2 \) if \( Y_{\Delta l} = Y_{\Delta (l-1)} \) and \((\Delta \theta - 1)^2/(2\Delta \theta) \) if \( Y_{\Delta l} \neq Y_{\Delta (l-1)} \), differentiation of the above expression with respect to \( \theta \) yields

\[
-\frac{n}{2} \frac{1}{\theta} - \frac{\Delta}{2} \#\{Y_{\Delta l} = Y_{\Delta (l-1)}\} + \left(-\frac{\Delta}{2} + \frac{1}{2\Delta \theta^2}\right) \#\{Y_{\Delta l} \neq Y_{\Delta (l-1)}\}
\]

and thus

\[
\hat{\theta}_{dEM} = \frac{1}{2\Delta} \left(\sqrt{\frac{1}{n} \#\{Y_{\Delta l} \neq Y_{\Delta (l-1)}\}} + 1 - 1\right)
\]

and for \( n \to \infty \)

\[
\hat{\theta}_{dEM} \overset{a.s.}{\to} \frac{1}{2\Delta} \left(\sqrt{4 \exp(-\Delta \theta) \sinh(\Delta \theta) + 1} - 1\right) = \theta - 2\Delta \theta^2 + O(\Delta^2).
\]

This estimator is also inconsistent, but for small times between two observations, it is reasonable.
Replacing the transition intensities (3.23) by

$$\mu(y, \theta) = \left( \frac{1}{2} y(y - 1)\theta, \frac{1}{2} (2 - y)(1 - y)\theta \right)$$  \hspace{1cm} (3.26)

for the second example, we would still have the same Markov jump process (assuming \(Y_{t_0} \in \{0, 2\}\)) and therefore the same estimator \(\hat{\theta}_{dEM}\).

The diffusion is now given by

$$d\bar{Y}_t = -2\theta(\bar{Y}_t - 1)dt + \sqrt{4\theta} \mid \bar{Y}_t - 1 \mid dW_t.$$  

With the transformation \(h(y) = \log(\mid y - 1 \mid)\) we have \(d(h(\bar{Y}_t)) = -4\theta dt + \sqrt{4\theta} dW_t\). A short calculation yields

$$\hat{\theta}_d = \frac{1}{8\Delta} \left( \sqrt{\frac{4}{n} \sum_{l=1}^{n} \log^2 \left( \frac{\mid Y_{\Delta l} - 1 \mid}{\mid Y_{\Delta(l-1)} - 1 \mid} \right) + 1} - 1 \right).$$

Under the law of the jump process, the summand in the summation is always 0 and thus \(\hat{\theta}_d = 0\), so the estimator is useless in this situation.

One might criticize that the intensities (3.26) can assume negative values. But they are derived from the reaction system

$$2Y^1 \rightarrow 2Y^2$$
$$2Y^2 \rightarrow 2Y^1$$

with the law of mass action (see Golightly and Wilkinson (2005)), assuming equal rate constants \(\theta\) for both reactions and using the fact that \(Y^1 + Y^2\) must be constant, here \(Y^1 + Y^2 = 2\) or \(Y^2 = 2 - Y^1\).

Regarding this examples one can not expect estimates based on an approximation with a diffusion defined as in (3.19) to be consistent, because there are essentially no conditions on the behavior of the jump rates for non integer-valued vectors, so that one can specify them in many different ways yielding different diffusion approximations for the same Markov jump process.
Bibliography


Curriculum Vitae

I was born on November 27, 1980 in Zweisimmen (BE) and grew up in Pfäffikon (ZH), attending primary school and 2 years of “Sekundarschule” there. I subsequently spent five years at the “Kantonsschule Zürcher Oberland” in Wetzikon (ZH), where I completed a science (Type C) Matura in 2001.

Afterwards, I started my diploma studies in Mathematics at ETH Zurich. In the latter part, I focussed on probability theory and statistics and wrote my diploma thesis with title ”Parameterschätzung in zeitstetigen Markov-Prozessen” (parameter estimation in continuous-time Markov processes) under the guidance of Prof. Hans Rudolf Künsch.

Right after graduation in November 2006, I enrolled as PhD student, also under the supervision of Prof. Hans Rudolf Künsch, and worked since as a teaching assistant and group organizer at the ”Seminar für Statistik”.

Besides my studies, I also made career in the Swiss Military. In 2004, I was promoted to lieutenant. At the moment, I am commander in the rank of a captain of an artillery battery (“Fest Art Bttr 13/1”) with about 200 personnel.