

*Analysis of*  
**M**arkovian  
**P**opulation  
**M**odels

MICHAEL BACKENKÖHLER



# ANALYSIS OF MARKOVIAN POPULATION MODELS

MICHAEL BACKENKÖHLER

DISSERTATION

zur Erlangung des Grades  
des Doktors der Naturwissenschaften  
der Fakultät für Mathematik und Informatik  
der Universität des Saarlandes

Saarbrücken, Juni 2022

Tag des Kolloquiums:

Dekan:

Prof. Dr. Thomas Schuster

Berichterstatler:

Prof. Dr. Verena Wolf,

Prof. Dr. Luca Bortolussi,

Prof. Dr. Tatjana Petrov

Vorsitzender:

Akademischer Mitarbeiter:

This dissertation was typeset in L<sup>A</sup>T<sub>E</sub>X using a design based on `classicthesis v4.6`. Hermann Zapf's *Palatino* (URW *Palladio*), *Euler* and *Optima* (URW *Classico*) type faces are used.

Final version as of May 5, 2022.

© Michael Backenköhler, June 2022

---

## ABSTRACT

---

Markovian population models are a powerful paradigm to describe processes of stochastically interacting agents. Their dynamics is given by a continuous-time Markov chains over the population sizes. Such large state-spaces make their analysis challenging.

In this thesis we develop methods for this problem class leveraging their structure. We derive linear moment constraints on the expected occupation measure and exit probabilities. In combination with semi-definite constraints on moment matrices, we obtain a convex program. This way, we are able to provide bounds on mean first-passage times and reaching probabilities. We further use these linear constraints as control variates to improve Monte Carlo estimation of different quantities. Two different algorithm for the construction of efficient variate sets are presented and evaluated.

Another set of contributions is based on a state-space lumping scheme that aggregates states in a grid structure. Based on the probabilities of these approximations we iteratively refine relevant and truncate irrelevant parts of the state-space. This way, the algorithm learns a well-justified finite-state projection for different scenarios.



---

## ZUSAMMENFASSUNG

---

Markowsche Populationsmodelle sind ein leistungsfähiges Paradigma zur Beschreibung von Prozessen stochastisch interagierender Akteure. Ihre Dynamik ist durch eine zeitkontinuierliche Markow-Kette über die Populationsgrößen gegeben. Solch große Zustandsräume machen ihre Analyse zu einer Herausforderung.

In dieser Arbeit entwickeln wir Methoden für diese Problemklasse, indem wir ihre Struktur nutzen. Wir leiten lineare Momentbeschränkungen für das erwartete Besetzungsmaß und die Austrittswahrscheinlichkeiten ab. In Kombination mit semidefiniten Nebenbedingungen für Momentmatrizen erhalten wir ein konvexes Programm. Auf diese Weise sind wir in der Lage, Schranken für mittlere Erstdurchlaufzeiten und Erreichbarkeitswahrscheinlichkeiten zu setzen. Außerdem verwenden wir diese linearen Nebenbedingungen als Kontrollvariablen, um die Monte-Carlo-Schätzung verschiedener Größen zu verbessern. Es werden zwei verschiedene Algorithmen für die Konstruktion effizienter Variablensätze vorgestellt und bewertet.

Eine weitere Gruppe von Beiträgen basiert auf einem Aggregationsschema, das Zustände in einer Gitterstruktur zusammenfasst. Auf der Grundlage der Wahrscheinlichkeiten dieser Näherungen verfeinern wir iterativ relevante und schneiden irrelevante Teile des Zustandsraums ab. Auf diese Weise erlernt der Algorithmus eine gut begründete endliche Zustandsprojektion für verschiedene Szenarien.





---

## CONTENTS

---

### I PRELIMINARIES

1	Introduction	3
1.1	Contributions	5
1.2	Organization	6
1.3	Previous Publications	7
2	Background	11
2.1	Continuous-time Markov Chains	11
2.1.1	Computing Transient Distributions	13
2.2	Markovian Population Models	15
2.3	State-Space Truncation	19
2.4	Stochastic Simulations	21
2.5	Moment Dynamics	22
2.5.1	Hybrid Representations	24
2.6	Stationary Distribution	25
2.7	Foster-Lyapunov Bounds	27
2.8	Augmented Foster-Lyapunov Bounds	28
2.8.1	The Drift and its Properties	30
2.8.2	Local Substitution	30
2.8.3	Illustrating Example	31
2.8.4	Concluding Remarks	34

### II MOMENT-BASED METHODS

3	Bounding Mean First-Passage Times	39
3.1	Related Work	41
3.2	Preliminaries	43
3.3	Martingale Formulation	44
3.4	Linear Model Constraints	48
3.5	Objective	51
3.6	Semi-Definite Constraints	52
3.6.1	Multi-Dimensional Generalization	53
3.6.2	A Semi-Definite Program	54
3.7	Implementation & Evaluation	55

3.7.1	Moment Scaling	56
3.7.2	Case Studies	57
3.7.3	Hybrid Models & Multi-Modality	60
3.8	Linear Hausdorff Constraints	63
3.8.1	A Linear Program	66
3.8.2	Case Studies	66
3.9	Conclusion	68
4	Linear Control Variates for Monte Carlo Estimation	71
4.1	Related Work	73
4.2	Moment Constraints	75
4.3	Control Variates	77
4.4	Moment-Based Variance Reduction	80
4.5	Case Studies	84
4.6	Resampling Algorithm	92
4.7	Conclusion	100
 <b>III AGGREGATION &amp; REFINEMENT</b>		
5	State-Space Aggregation	105
5.1	Related Work	105
5.2	Macro-States	106
5.3	Construction	107
5.4	Approximation Features	111
6	Truncations for Stationary Distributions	113
6.1	Related Work	114
6.2	Truncation-Based Approximation	116
6.2.1	Initial Aggregation	118
6.2.2	Iterative Refinement Algorithm	119
6.3	Results	121
6.3.1	Parallel Birth-Death Process	122
6.3.2	Exclusive Switch	123
6.3.3	p53 Oscillator	126
6.4	Conclusion	130
7	Analysis under Terminal Constraints	131
7.1	Related Work	132
7.2	Backwards Probabilities	134
7.3	Bridging Distribution	135
7.4	Bridge Truncation via Lumping Approximation	136

7.4.1	Finite State Projection	137
7.4.2	Iterative Refinement Algorithm	137
7.5	Results	139
7.5.1	Bounding Rare Event Probabilities	140
7.5.2	Mode Switching	142
7.5.3	Recursive Bayesian Estimation	146
7.6	Conclusion	149
8	Rare Event Probabilities	151
8.1	Related Work	152
8.2	Importance Sampling	153
8.3	Near-Optimal Biasing	154
8.4	Non-homogeneous Stochastic Simulation	156
8.5	Case Studies	160
8.5.1	Two Simple Examples	160
8.5.2	Toggle Switch	161
8.6	Conclusion	165
 <b>IV CONCLUDING REMARKS</b>		
9	Conclusion	169
9.1	Future Work	170
 <b>V APPENDIX</b>		
A	Additional Data	175
A.1	Control Variate Results	175
A.2	Aggregation for Stationary Distributions	175
 Bibliography		
List of Figures		201
List of Tables		203
List of Models		204
Acronyms		205



## Part I

### PRELIMINARIES



# 1

---

## INTRODUCTION

---

Any model is an abstraction of reality. Often this abstraction aims to capture a few aspects of interest while ignoring others. Modeling often is a delicate balance of adequate representation and abstraction. The former is the goal to capture all the relevant behaviors and effects in a model. Useful abstractions benefit both explainability and analysis. This balance between faithfully capturing reality and enabling an analysis is at the core of all modeling efforts.

In the natural sciences, we often deal with concentrations of different “things”. This could be the amount of some substance or the number of agents waiting for some service. Traditionally, ordinary differential equations are the most popular paradigm for dynamical models. Ordinary differential equations have two inherent simplifications: Firstly, they impose a continuous state space and, secondly, they are deterministic. The first assumption is appropriate in many circumstances. For example, chemical concentrations can usually be treated as continuous if the number of molecules in a given volume is large. With large populations, the influence of discrete effects decreases. In such cases, mean-field analyses are an adequate and efficient tool (Bortolussi et al., 2013). However, this simplification fails if discrete effects are central to the model. Consider, for example, the die-out of a species in the classical Lotka-Volterra predator-prey model (Lotka, 1925). In this case, the distinction between a low concentration and a zero concentration becomes relevant again: If the predator, e.g. the

*Think of mixing  
chemical substances,  
for example.*

foxes, die out the model's dynamic changes dramatically. Such an event can be missed by the deterministic model which instead gives a population size larger than zero but much smaller than a single individual. Mollison calls this problem humorously the *atto-fox* problem (Mollison, 1991). The second issue, i.e. the issue of determinism, is often closely connected to the first. While the issue of a deterministic vs. a stochastic world is more an issue of philosophy and physics, stochasticity undoubtedly provides an excellent abstraction for many phenomena. Staying in with the previous example, it makes sense to consider the dying out of the predatory species to be stochastic. That means, there is a non-zero probability for both, survival and die-out. And answering questions about such probabilities would be one of the central functions of such a model.

The previous example falls in the realm of systems biology. There are numerous applications in systems biology that benefit from a discrete and stochastic description (Ullah and Wolkenhauer, 2009; Wilkinson, 2018). Many biological processes such as cell functions are driven by stochastic effects. Consider for example the oscillatory processes such as the circadian cycle (Asgari-Targhi and Klerman, 2019) or even classical models such as the famous predator-prey model (Lotka, 1925). Another phenomenon central to many biological functions is a switch-like behavior.

Markovian population models (MPMs) are a versatile framework capturing stochastic interactions between groups of identical agents. States are discrete counts of agents and reaction events stochastically change these counts. Interactions between agents commonly referred to as reactions, happen at exponentially distributed random times. Their rate depends only on the current system state, i.e. the population sizes. This kind of population model introduces the major assumption that all agents behave similarly. Similarity entails both similar behavior of each agent and a homogenous mixture of all agents. This feature is a *Markov property* and therefore MPMs form a special class of structured continuous-time Markov chains (CTMCs) (Anderson, 2012). Accordingly, the Kolmogorov equations describes the time-

*Often the term well-stirredness is used to express this.*



evolution of the corresponding probability distribution. For small finite chains it is easy to compute accurate transient probabilities using the Kolmogorov equation. The CTMCs underlying most population models, however, can have large and often unbound populations. That makes their numerical solution extremely challenging. Many problems occurring in a wide range of areas such as chemistry (Gillespie, 1977), epidemiology (Mode and Sleeman, 2000), queuing systems (Breuer, 2003), finance (Pardoux, 2008), and performance analysis (Bortolussi et al., 2013; Gast, Bortolussi, and Tribastone, 2019) can be described using this formalism.

Typical research questions when using such models target the forward behavior. That is the problem of looking into the future given a starting distribution. For example, consider determining the probability that a system stays in some subset of the state space for a fixed amount of time. Mainly, we target these kinds of questions, but in Chapter 7 we look at a closely related class of problems – the bridging problem. In this problem, both the initial and the terminal distribution are fixed. Such questions are of particular interest in multimodal models to determine the switching behavior from one to another attracting region. A third area is the question of long-term behavior. In ergodic models, the forward probabilities converge to a stable equilibrium distribution which is independent of the initial distribution.

*In the context of (bio-)chemistry this is often called the chemical master equation.*

*This scenario is akin to the smoothing problem from stochastic filtering.*

↪ page 25

## 1.1 CONTRIBUTIONS

The goal of this thesis is to develop methods to analyze MPMs. A particular focus of these methods is scalability. All methods developed in this thesis aim to provide means to more efficiently analyze such systems – or at least uncover paths that can lead to scalable methodologies. Importantly, these methods are *reliable* approximations: There are no ad-hoc approximations influencing estimates. The methodological theme is to leverage (approximate) distributional information such as moment equations and state-space aggregation to obtain good approximations.

The first contribution proposes an augmentation scheme for simple Lyapunov functions in Section 2.8. Therein we suggest

locally altering valid proposal Lyapunov functions. Such local alterations are subject to very few constraints and Thus methods such as neural networks are applicable. The resulting sets can be much smaller.

In [Part ii](#) we focus on moment-based techniques. Methods to analyze such models without ignoring their inherent stochasticity received much attention in recent years. Facing state spaces too large to handle using the Kolmogorov equation directly, stochastic simulations (Gillespie, 1977) provide an alternative path. Such simulations have the advantage of providing an accurate approximation of the process. That is, they converge in the limit of simulation runs. Many extensions of stochastic simulations and Monte Carlo estimation have been proposed. In [Chapter 4](#) we develop such an extension for variance reduction of stochastic estimates. This variance reduction exploits moment dynamics to derive constraints on population averages. These constraints are a correction to the estimate. The corrected estimator is unbiased and has a variance lower or equal to the uncorrected estimator's variance.

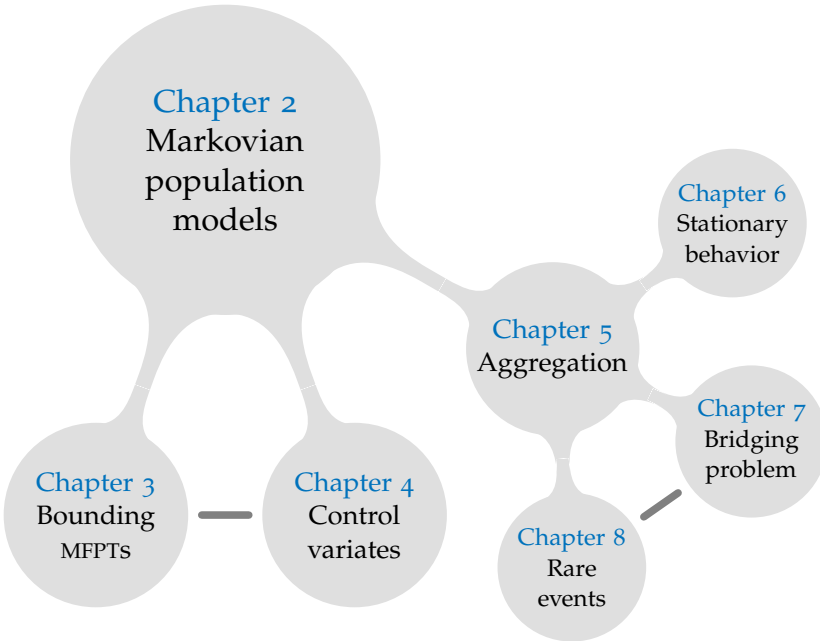
We use similar moment constraints – in combination with general moment constraints – to formulate convex optimization problems. This way, we compute rigorous bounds on reaching probabilities and mean first-passage times. This approach is developed in [Chapter 3](#). Fortunately, both of these moment-based approaches can do without any moment closures and thereby avoid any ad-hoc approximations.

In [Part iii](#) we use a hyper cube state-space aggregation scheme to gain a rough understanding of the dynamics. The aggregation is then further refined ([Chapter 6](#) and [Chapter 7](#)) or used to guide stochastic simulations ([Chapter 8](#)).

## 1.2 ORGANIZATION

[Figure 1.1](#) provides an overview of the dependencies between chapters. All contributions presented in this thesis are related to Markovian population models. Therefore the background chapter, i.e. [Chapter 2](#) is relevant to all later parts of the thesis. In [Part ii](#),

we present methods based on moment constraints. There is a connection between [Chapter 3](#) and [Chapter 4](#) in that they share the linear constraints on temporal moments. The prerequisite for [Chapter 6](#) and [Chapter 7](#) is the aggregation technique presented in [Chapter 5](#). The conceptual connection between the bridging problem ([Chapter 7](#)) and the rare event sampling ([Chapter 8](#)) lies in both, its setting and the utilization of backwards probabilities.



**Figure 1.1:** Chapter dependencies.

### 1.3 PREVIOUS PUBLICATIONS

The ideas and much of the presented results have appeared previously in the following publications. As such, the content of most chapters has undergone peer-review and been published in various conference proceedings. The publications and their respective sections are as indicated below.

- [Chapter 3](#) has with minor differences been published as

Michael Backenköhler, Luca Bortolussi, and Verena Wolf (2020). “Bounding Mean First Passage Times in Population Continuous-Time Markov Chains.” In: *17th International Conference on Quantitative Evaluation of SysTems*. Vol. 12289. Lecture Notes in Computer Science. Springer, pp. 155–174.

The approach was conceived by M. B. Author M. B. performed the implementation and evaluation with feedback from the other authors. All authors contributed to the text. The Hausdorff moments, presented in [Section 3.8](#), provide alternative moment constraints. This method was not subject of the above publication.

- [Chapter 4](#) has with minor differences been published as Michael Backenköhler, Luca Bortolussi, and Verena Wolf (2019). “Control Variates for Stochastic Simulation of Chemical Reaction Networks.” In: *17th International Conference on Computational Methods in Systems Biology*. Vol. 11773. Lecture Notes in Computer Science. Springer, pp. 42–59.

The control variate approach was conceived by M. B. and V. W. The refinement algorithm was developed during discussions of all authors. Author M. B. performed the implementation and evaluation with feedback from the other authors. All authors contributed to the text.

The contents of [Section 4.6](#) have been published in the article

Michael Backenköhler, Luca Bortolussi, and Verena Wolf (2021). “Variance Reduction in Stochastic Reaction Networks using Control Variates.” In: *Festschrift anlässlich Thomas Henzingers 60. Geburtstags*. Vol. to appear. Lecture Notes in Computer Science. Springer.

The control variate approach was conceived by M. B. and V. W. The resampling algorithm was developed during

discussions of all authors. Author M. B. performed the implementation and evaluation with feedback from the other authors. All authors contributed to the text.

- [Chapter 6](#) has with minor differences been published as Michael Backenköhler, Luca Bortolussi, Gerrit Großmann, and Verena Wolf (2021). “Abstraction-Guided Truncations for Stationary Distributions of Markov Population Models.” In: *18th International Conference on Quantitative Evaluation of SysTems*. Vol. 12846. Lecture Notes in Computer Science. Springer, pp. 351–371.

The lumping approach was conceived by M. B. All authors contributed to the text.

- [Chapter 7](#) has with minor differences been published as Michael Backenköhler, Luca Bortolussi, Gerrit Großmann, and Verena Wolf (2021). “Analysis of Markov Jump Processes under Terminal Constraints.” In: *27th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Vol. 12651. Lecture Notes in Computer Science. Springer, pp. 210–229.

The lumping approach was conceived by M. B. All authors contributed to the text. [Chapter 5](#) is in large part based on this and the latter two publication above. [Chapter 2](#) contains introductory material and examples from the two publications above.



# 2

---

## BACKGROUND

---

### 2.1 CONTINUOUS-TIME MARKOV CHAINS

A *stochastic process* is a parameterized collection of random variables  $\{X_t\}_{t \in T}$  defined on some probability space  $(\Omega, \mathcal{F}, P)$  and takes values in complete metric space  $(S, r)$ . In most contexts the index set  $T$  models *time*. In a discrete setting  $T = \mathbb{N}$ , while  $T = [0, \infty)$  in the continuous setting, which we consider in this thesis. We observe the values  $X(t, \omega)$  for some fixed, but unknown  $\omega \in \Omega$ . The information of the process up to time  $t$  is given by the  $\sigma$ -algebra  $\mathcal{F}_t \subset \mathcal{F}$ . The increasing family of sigma algebras, i.e.  $\mathcal{F}_s \subseteq \mathcal{F}_t$  for  $s \leq t$  is called a *filtration*.

See Feller (1971) for stoch. processes in general.

More specifically, we study with models having continuous-time Markov chain (CTMC) semantics — a type of *Markov process*. Such processes satisfy the *Markov property*: For all Borel-measurable functions  $f$

$$E(f(X_{t+s}) \mid \mathcal{F}_t) = E(f(X_{t+s}) \mid X(t)) . \quad (2.1)$$

Intuitively, this property expresses that the future of the process depends only on the latest condition, i.e.  $X_t$  and not earlier conditions ( $\mathcal{F}_t$ ). A CTMC is a Markov process, that takes discrete values  $S = \{s_0, s_1, \dots\}$  over continuous time  $T = [0, \infty)$ . If further

The time-discrete analogue is the DTMC.

$$\Pr(X_t = s_1 \mid X_0 = s_0) = \Pr(X_{t+h} = s_1 \mid X_h = s_0) \quad (2.2)$$

for all  $t, h \geq 0$  and  $s_0, s_1 \in S$  the chain is *time-homogenous*: The absolute time point is irrelevant, and the dynamics do not change

if we shift in time. In this thesis we are only interested in tim-homogenous CTMCs, but most techniques developed should carry over. We define *transition probabilities*

$$p_{ij}(h) = \Pr(X_h = s_j \mid X_0 = s_i) . \quad (2.3)$$

Accordingly, the *transition matrix* is given by  $P(h)_{ij} = p_{ij}(h)$  for all indices  $i$  and  $j$ . The Chapman-Kolmogorov equation

$$P(s + t) = P(s)P(t) \quad (2.4)$$

follows directly from the law of total probability and the Markov property (2.1). (2.4) directly tells us that the transition probabilities form a semigroup. Studying Markov processes from this direction is a popular approach (Ethier and Kurtz, 2009).

The standard method of specifying a Markov process, and CTMCs in particular, is the *generator*. In general, this is an operator  $A$  on some class of functions and

$$\mathbb{E} (f(X_{t+h}) - f(X_t) \mid \mathcal{F}_t) = Af(X_t)h + o(h) , \quad (2.5)$$

where  $\mathcal{F}_t$  is the filtration up to time  $t$ . This equation can be interpreted as the requirement, that

$$f(X_t) - f(X_0) - \int_0^t Af(x_s) ds$$

is a martingale (see Kurtz, 1981, p. 5). We will use this in Chapter 3 to derive bounds on mean first-passage times.

In CTMCs the generator is defined by giving the *intensities* or *rates* of transitions. Such a rate  $q_{ij} > 0$  between state  $s_i$  and  $s_j$  implies

$$\Pr(X_h = s_j \mid X_0 = s_i) = q_{ij}h + o(h) . \quad (2.6)$$

Due to the discrete nature, the generator is a matrix, usually called the  $Q$ -matrix,

$$Q = \lim_{h \downarrow 0} \frac{1}{h} (P(h) - I) . \quad (2.7)$$

This is congruent with (2.5) under application the Markov property.



As such the change of the state probability distribution over time is fully characterized by the *Kolmogorov forward equation*

$$\frac{d}{dt}P(t) = P(t)Q. \quad (2.8)$$

Analogously, the *Kolmogorov backward equation* is

$$\frac{d}{dt}P(t) = QP(t)^T. \quad (2.9)$$

Distributions in the context of Markov chains are typically in row-vector form

$$\pi(t) := (\pi(x_1, t), \pi(x_2, t), \dots),$$

where we define

$$\pi(x_i, t) := \Pr(X_t = x_i), \quad \forall x_i \in \mathcal{S}, \forall t \geq 0.$$

Often, (2.8) and (2.9) are used in the context of an distributions. In this case it makes more sense to right-multiply the unit vector to these equations such that

$$\frac{d}{dt}\pi(t) = \pi(t)Q \quad (2.10)$$

and

$$\frac{d}{dt}\pi(t) = Q\pi(t)^T. \quad (2.11)$$

Given some initial distribution

$$\pi_0 := \pi(0),$$

the distribution  $\pi(t)$  is given by give simple initial value problems (IVPs) of (2.10).

### 2.1.1 Computing Transient Distributions

Stewart (1994) provides a comprehensive overview of solution methods. Here, we only provide a basic overview and intuition relevant for the rest of this thesis.

**MATRIX EXPONENTIAL** The transition probability matrix  $P(t)$  is the solution of (2.8)

$$P(t) = \exp(Qt),$$

where the matrix exponential for a square matrix  $M$

$$\exp(M) := \sum_{k=0}^{\infty} \frac{1}{k!} M^k.$$

Due to the factor  $h^k/k!$ , the sum of the matrix exponential can be truncated to get an estimate of high quality. In practice, this method is unsuited to many problems, because the factor  $Q^k$  becomes incurs a prohibitive cost, especially if the state space is large.

**NUMERICAL INTEGRATION** The Kolmogorov equations (2.8), (2.9) provide us with an **IVP** that can be solved numerically. This method scales much better than the matrix exponential method, especially if the whole time-series is of interest. If an initial distribution  $\pi_0$  is fixed, the ordinary differential equation (ODE) simplify further, such that we only have one equation per state. The drawback to this method is the error inherent to numerical integration schemes.

*The matrix exponential is the analytical solution.*

**UNIFORMIZATION** Uniformization is an elegant algorithm to compute transient solutions. Here, the **CTMC** is transformed into a discrete-time Markov chain (**DTMC**). Using a uniformization rate

$$\lambda_0 \geq \max_i |q_{ii}|$$

the transition probabilities of the **DTMC** become

$$P_{ij} = \begin{cases} Q_{ij}/\lambda_0, & \text{if } i \neq j \\ 1 - \sum_k Q_{ik}/\lambda_0, & \text{otherwise} \end{cases}.$$

The transient distribution at  $t$  can be obtained, by weighting the  $k$ -step probabilities of the **DTMC** by a Poisson distribution with rate  $\lambda_0 t$ :

$$\pi(t) = \sum_{k=0}^{\infty} \pi_0 P^k \frac{(\lambda_0 t)^k}{k!} \exp(-\lambda_0 t).$$

Truncation of this series clearly gives an underapproximation.

**MONTE CARLO SIMULATION** A simple way is estimation using Monte Carlo methods. This entails stochastic simulation of many trajectories of the [CTMC](#). Generating a trajectory is straightforward: Given that the process is in a particular state  $s_i$  the a transition has to be sampled along with the residence time in state  $s_i$ . The naive approach is to sample an exponential random variable for each  $q_{ij}$  and choose the one firing first. This algorithm can be improved by sampling a reaction directly and sampling the residence time separately. This algorithm will be shown later in the context of [MPMs](#).

↪ [page 22](#)

## 2.2 MARKOVIAN POPULATION MODELS

An Markovian population model ([MPM](#)) among agents of  $n_S$  distinct types in a well-stirred system. Other names for this model class are population [CTMC](#) ([pCTMC](#)), chemical reaction network ([CRN](#)), and stochastic reaction network ([SRN](#)). The system is given by a continuous-time stochastic process  $\{X_t\}_{t \geq 0}$ . It models only the number of agents according to their type. Therefore the process takes  $n_S$ -dimensional vectors of natural numbers as values, i.e. the state-space is  $\mathcal{S} \subseteq \mathbb{N}^{n_S}$ . By only considering the number of agents, we are neglecting factors such as spatial variations in agent density or other factors influencing interactions. The assumption of all agents being equally distributed in space is called the *well-stirredness* assumptions.

*“The secret to modeling is not being perfect.”*

— Karl Lagerfeld

These assumptions bring with them, the convenience of considering populations as a whole. The single agent is of no importance to the dynamics of the process. Just the overall size of each population determines the stochastic evolution of the process. Another consequence of this assumption is the limiting behavior. If all populations are proportionally scaled to infinity, their concentrations can be accurately described by deterministic [ODEs](#). In fact, this is the most widely used paradigm to analyze this kind of reaction networks. However this methodology may

*Other assumptions, such as exponentially distributed firing times, are discussed below.*

by widely inaccurate. Consider, for example, an epidemic process. Typically individuals are not well-stirred in a societal context and the influence of specific contact structures is crucial to the process' dynamics (Großmann, Backenköhler, and Wolf, 2020; Großmann, Backenköhler, and Wolf, 2021). Furthermore such a process typically exhibits discrete stochastic effects, such as the epidemic dying out. While the latter effect is retained in an *MPM*, the former is already lost. Therefore, a great deal of care has to be taken by the modeller which kinds of abstractions are appropriate for the chosen abstraction.

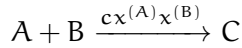
Interactions between agents are expressed as *reactions*. These reactions have associated gains and losses of agents, given by non-negative integer vectors  $v_j^-$  and  $v_j^+$  for reaction  $j$ , respectively. The overall change by a reaction is given by the vector  $v_j = v_j^+ - v_j^-$ . A reaction between agents of types  $S_1, \dots, S_{n_s}$  is specified in the following form:

$$\sum_{\ell=1}^{n_s} v_{j\ell}^- S_\ell \xrightarrow{\alpha_j(x)} \sum_{\ell=1}^{n_s} v_{j\ell}^+ S_\ell. \quad (2.12)$$

The propensity function  $\alpha_j$  gives the rate of the exponentially distributed firing time of the reaction as a function of the current system state  $x \in \mathcal{S}$ . Thus, for reaction  $j$  we have the intensity

$$\Pr(X_{t+h} = x + v_j \mid X_t = x) = \alpha_j(x) + o(ht). \quad (2.13)$$

In most physical models, *mass-action* propensities are most common. These model combinatorial nature of well-mixed molecules moving randomly through space: In a reaction



two molecules hit eachother with a probability, proportional to the product of their counts

$$c X_t^{(A)} X_t^{(B)}.$$

In general such rates are given by the product of the number of reactant combinations in  $x$  and a *rate constant*  $c_j$ , i.e.

$$\alpha_j(x) := c_j \prod_{\ell=1}^{n_s} \binom{x^{(S_\ell)}}{v_{j\ell}^-}. \quad (2.14)$$

In this case, we give the rate constant in (2.12) instead of the function  $\alpha_j$ . We use the superscript notation  $x^{(A)}$  to denote the index corresponding to species A in some vector of length  $n_S$ .

*This notation avoids conflicts, when we use the subscript for time or as an index.*

According to (2.13) the stochastic process  $\{X_t\}_{t \geq 0}$  describing the evolution of the population sizes over time  $t$  is a continuous-time Markov chain (CTMC). The infinitesimal generator matrix  $Q$  has the entries

$$Q_{x,y} = \begin{cases} \sum_{j: x+v_j=y} \alpha_j(x), & \text{if } x \neq y, \\ -\sum_{j=1}^{n_R} \alpha_j(x), & \text{otherwise.} \end{cases} \quad (2.15)$$

Note that in addition mild regularity assumptions are necessary for the existence of a unique CTMC  $X$ , such as non-explosiveness (Anderson, 2012). These assumptions are typically valid for realistic reaction networks. The probability distribution over time is given by an initial value problem. Given an initial state  $x_0$ , the probabilities

$$\pi(x_i, t) := \Pr(X_t = x_i \mid X_0 = x_0), \quad t \geq 0, \quad x \in S \quad (2.16)$$

evolve according to the Kolmogorov forward equation

$$\frac{d}{dt}\pi(t) = \pi(t)Q, \quad (2.17)$$

where  $\pi(t)$  is an arbitrary vectorization

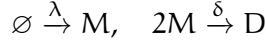
$$(\pi(x_1, t), \pi(x_2, t), \dots, \pi(x_{|S|}, t))$$

of the state probabilities. Often (2.17) is given for a single state. In this form – due to its usage in quantitative biology – it is commonly referred to as the *chemical master equation* (CME)

$$\frac{d\pi}{dt}(x, t) = \sum_{j=1}^{n_R} (\alpha_j(x - v_j)\pi(x - v_j, t) - \alpha_j(x)\pi(x, t)). \quad (2.18)$$

**EXAMPLE** Consider the following simple MPM with non-linear propensities as an example.

**Model 1** (Dimerization). We first examine a simple dimerization model on an unbounded state-space with reactions



and initial condition  $X_0^{(M)} = X_0^{(D)} = 0$ .

The semantics is given by a CTMC  $X_t = (X_t^{(M)}, X_t^{(D)})^T$ , where  $(S_1, S_2) = (M, D)$ . The reaction propensities according to (2.14) are

$$\alpha_1(x) = \lambda \quad \text{and} \quad \alpha_2(x) = \delta x^{(M)}(x^{(M)} - 1)/2.$$

The change vectors for the first reaction are

$$v_1^- = (0, 0)^T \quad \text{and} \quad v_1^+ = (1, 0)^T.$$

For the second reaction the change vectors are

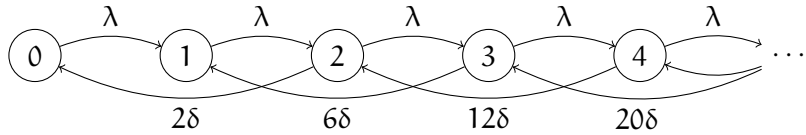
$$v_2^- = (2, 0)^T \quad \text{and} \quad v_2^+ = (0, 1)^T.$$

Consequently,  $v_1 = (1, 0)^T$  and  $v_2 = (-2, 1)^T$ .

The generator matrix

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & & & \dots \\ 0 & -\lambda & \lambda & 0 & & \dots \\ 2\delta & 0 & -(\lambda + 2\delta) & \lambda & 0 & \dots \\ 0 & 6\delta & 0 & -(\lambda + 6\delta) & \lambda & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Often it is a good idea to visualize the state-space and transitions as a graph. This graph is constructed by interpreting the Q-matrix as an adjacency matrix for some subset of states.



For a state  $(x^{(M)}, x^{(D)}) \in \mathbb{N}^2$ , where  $x^{(M)} \geq 2$ , the CME (2.18) becomes

$$\begin{aligned} & \frac{d}{dt} \pi((x^{(M)}, x^{(D)}), t) \\ = & \frac{\delta}{2} (x^{(M)} + 2)(x^{(M)} + 1) \pi((x^{(M)} + 2, x^{(D)} - 1), t) \\ & - (\lambda + \frac{\delta}{2} x^{(M)}(x^{(M)} - 1)) \pi((x^{(M)}, x^{(D)}), t) \\ & + \lambda \pi((x^{(M)} - 1, x^{(D)}), t) \end{aligned}$$

◇

The explicit representation of all state probabilities is often not possible, because there are infinitely many states. Usually the state-space is truncated to contain all relevant states (Andreychenko et al., 2011) or one switches to an approximation such as the mean-field (Bortolussi et al., 2013)

### 2.3 STATE-SPACE TRUNCATION

A complete solution of (2.18) is usually not possible. If the state-space with non-negligible probability is suitably small, a state space truncation can be performed. That is, (2.18) is integrated on a possibly time-dependent subset  $\hat{\mathcal{S}}_t \subseteq \mathcal{S}$  (Henzinger, Mateescu, and Wolf, 2009; Munsky and Khammash, 2006; Spieler, 2014). Transitions to states, that are not part of this subset are typically re-directed to a introduced sink-state. This state captures the mass “lost” by the approximation and gives the error up to the numerical integration scheme. Munsky and Khammash (2006) coined the term of finite state projection (FSP) for such a method.

*Uniformization can give a lower bound on the error.*

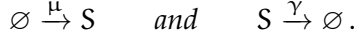
To analyze the stationary distribution (Section 2.6) the redirection scheme needs to be altered (Kuntz et al., 2021b): Instead of a re-redirection into a sink-state, transitions are redirected in *some fashion* back into the truncation set.

**EXAMPLE** Consider a birth-death process as a simple example. This model is used to describe a wide variety of phenomena and often constitutes a sub-module of larger models. For example, it represents an M/M/1 queue with service rates being linearly

This process occurs in many different varieties. For example, the state-space may be finite and propensities are non-polynomial (see Backenköhler and Großmann, 2020).

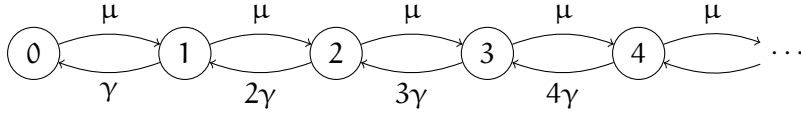
dependent on the queue length. Note that even for this simple model, the state-space is countably infinite.

**Model 2 (Birth-Death Process).** The model consists of exponentially distributed arrivals and service times proportional to queue length. It can be expressed using two mass-action reactions:



The initial condition  $X_0 = 0$  holds with probability one.

The underlying CTMC has the following infinite structure.



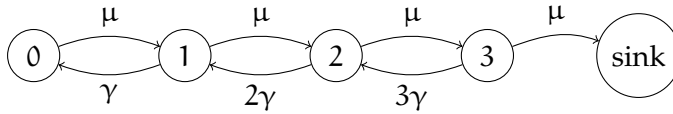
The change of probability mass in a single state  $x > 0$  is described by expanding (2.18) and

$$\begin{aligned} \frac{d}{dt}\pi_t(x) = & \mu\pi(x-1, t) + \gamma(x+1)\pi(x+1, t) \\ & - (\mu + \gamma x)\pi(x, t). \end{aligned} \quad (2.19)$$

A typical truncation scheme to  $[0, N]$  would use (2.19) for  $x \in [0, \dots, N-1]$  and for  $x = N$  the transition to  $N+1$  via the first reaction is re-directed into a sink state. We can drop the sink state because of the invariant  $\sum_x \pi_t(x) = 1, \forall t \geq 0$ . The ODE for the boundary state would consequently read

$$\frac{d}{dt}\pi_t(N) = \mu\pi(N-1, t) - (\gamma N + \mu)\pi(N, t).$$

Visualizing the resulting CTMC and letting  $N = 3$ .



In general, we can drop one equation in the CME system because of this invariant.

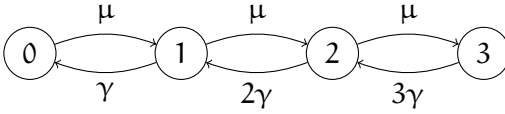


In a forward analysis, the sink state accumulates mass “lost” by the approximation. This provides a good practical error bound for the chosen truncation.

If one is interested in the stationary probability distribution (Section 2.6), a sink state is not practical, since – in the long run – the process will usually enter such a state. Instead, the cut transitions are re-directed back into the truncation. A plethora of re-directions is possible (Gupta, Mikelson, and Khammash, 2017; Kuntz et al., 2021b; Spieler, 2014), but in general it is reasonable to re-direct to states with cut incoming transitions. In this example, this is equivalent to cutting the transition, because the truncation has only this border state.

*The accuracy of the bound is dependent on the specific method of the forward analysis. Numerical integration, for example, it is less strict.*

*Such re-directions are used in Chapter 6.*



Thus, the ODE for state N here reads

$$\frac{d}{dt}\pi_t(N) = \mu\pi(N-1, t) - \gamma N\pi(N, t).$$

◇

## 2.4 STOCHASTIC SIMULATIONS

We can generate trajectories of this model using the stochastic simulation algorithm (SSA) (Algorithm 1) (Gillespie, 1977). The simulation algorithm consists of repeatedly evaluating the race condition and jump times induced by (2.15) until some terminal criterion such as a maximum simulation time  $T$  is reached (line 2). In particular, the algorithm iteratively chooses a reaction, with a probability that is proportional to its rate given the current state  $s$  (line 4). The jump time  $t_j - t_{j+1}$  is determined by sampling from an exponential distribution with rate  $\sum_i \alpha_i(s)$  (line 5).

**Algorithm 1:** Sample a trajectory

---

**input** :  $\pi_0, A$   
**output**: trajectory  $\tau$

```

1  $\tau \leftarrow$  empty list,  $s \leftarrow$  sample from  $\pi_0$ ,  $t \leftarrow 0$ ;
2 while  $t < T$  do
3    $\tau \leftarrow \text{append}(\tau, (s, t))$ ;
4    $k \leftarrow$  sample reaction  $i$  with probability  $\alpha_i(s) / \sum_i \alpha_i(s)$ ;
5    $\delta \sim \text{Exp}(\sum_i \alpha_i(s))$ ;
6    $s \leftarrow s + v_k$ ;
7    $t \leftarrow t + \delta$ ;
8 return  $\tau$ ;
```

---

The output of [Algorithm 1](#) is an alternating sequence of states and jump times

$$\tau = s_0 t_0 s_1 t_1 \dots t_n s_n, \quad t \in [0, T]$$

called a *trajectory*.

Monte Carlo estimation requires a sufficiently large number of such trajectories to be generated. This collection of realizations facilitates a statistical estimate of a wide range of quantities such as expected values and probabilities. The main benefit of this approach is its flexibility. It can solve essentially all relevant tasks. The main drawback is the cost associated with the generation of a sufficiently large trajectory ensemble. This problem becomes very pronounced in case of rare event probability estimation and stiff systems. Furthermore, the results only provide – by their very nature – statistical guarantees. Other methods such as explicit state-space representations can give stronger guarantees.

Bespoke techniques for these scenarios are available (Cao, Gillespie, and Petzold, 2005; Daigle Jr et al., 2011).

## 2.5 MOMENT DYNAMICS

Often times, the moments of a stochastic process provide sufficient information for its analysis. ODEs for the expected values of an MPM can be derived using its generator. We can apply  $Q$  to a

This is called the drift (see [Section 2.7](#)).

polynomial  $f$  such that

$$Qf(x) = \sum_{j=1}^{n_R} (f(x + v_j) - f(x)) \alpha_j(x). \quad (2.20)$$

By definition of the generator (2.5)

$$\frac{d}{dt} E(f(X_t) | X_t = x) = Qf(x).$$

Left-multiplying the probability of a given state and summing up over all states, we obtain the time derivative of the expected value  $E(f(X_t))$ . Written as a matrix vector product

$$\frac{d}{dt}(\pi f) = \pi Qf.$$

For **MPMs**, in particular, we arrive at the functional form

$$\frac{d}{dt} E(f(X_t)) = \sum_{j=1}^{n_R} E((f(X_t + v_j) - f(X_t)) \alpha_j(X_t)). \quad (2.21)$$

This equation is used to analyse (raw) *moments* of the process. A raw moment is

$$E(X^m) = E\left(\prod_{i=1}^{n_S} X_i^{m_i}\right), \quad m \in \mathbb{N}^{n_S}$$

Centered moments (e.g. variance) are equivalent via the binomial transform.

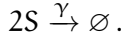
with respect to some probability measure. An **ODE** is given by (2.21), setting  $f(x) = x^m$ . The *order* of a moment  $E(X^m)$  is given by the sum of its exponents, i.e.  $\sum_i m_i$ . Note that the notion of expected value can be generalized to any measure  $\mu$  on a Borel-measurable space  $(E, \mathcal{B}(E))$ , where the  $m$ -th raw moment is  $\int_E x^m d\mu(x)$ . Throughout we assume that moments of arbitrary order remain finite over time, i.e.  $E(|X_t^m|) < \infty$ ,  $t \geq 0$ . In Gupta, Briat, and Khammash (2014) the authors propose a framework to verify this property for a given model.

**EXAMPLE** Let us express the dynamics of the first two uncentered moments for **Model 2** using (2.21).

$$\begin{aligned} \frac{d}{dt} E(X_t) &= \mu - \gamma E(X_t) \\ \frac{d}{dt} E(X_t^2) &= \mu(2E(X_t) + 1) - \gamma(2E(X_t^2) - E(X_t)) \end{aligned} \quad (2.22)$$

Setting initial moments these equations give as an [IVP](#), we can solve (see [Figure 2.1](#)). This, however, is more an exception than the norm: Unless all reactions have linear or constant rate functions  $\alpha_i(\cdot)$ ,  $\forall i$ , we would not end up with a closed system of [ODEs](#) as in [\(2.22\)](#). To illustrate, let us pretend the reaction  $(S \xrightarrow{\gamma} \emptyset)$  would become this non-linear reaction

*The model is then the same as [Model 1](#).*



Accordingly, due to mass-action [\(2.14\)](#)

$$\alpha_2(x) = \gamma(x^2 - x).$$

Therefore the first moment's derivative becomes

$$\frac{d}{dt} E(X_t) = \mu - \gamma(E(X_t^2) - E(X_t)).$$

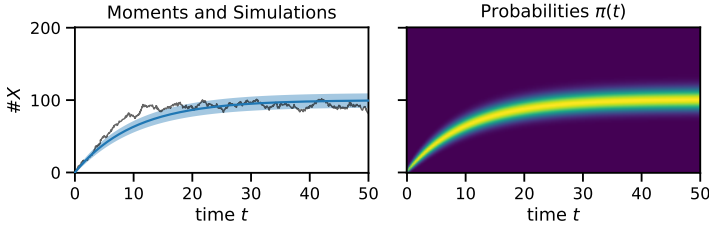
Note, that now the right-hand side of the derivative in the example depends on the value of the second moment  $E(X_t^2)$ .  $\diamond$

If we consider the general expression [\(2.21\)](#) for the moment of order  $k$  clearly a term of order  $k + 1$  occurs, that does (usually) not cancel out if a propensity function is at least quadratic. Therefore, researches commonly rely on ad-hoc approximations to truncate this infinite system of [ODEs](#) ([Hespanha, 2008](#); [Schnoerr, Sanguinetti, and Grima, 2014, 2015](#)). Unfortunately such schemes have typically no guarantees to converge – or even improve – with increasing truncation order ([Schnoerr, Sanguinetti, and Grima, 2014](#)) or increasing system size. Furthermore, fairly involved numerical schemes have to be employed to recover distributional approximations ([Andreychenko et al., 2017](#)). The only scheme with a convergence guarantee in the system size limit is the *mean-field* approximation ([Bortolussi et al., 2013](#)). Therein zero-covariances are assumed, i.e. the system is truncated at the first order equations using the approximation  $E(X_t^2) = E(X_t)^2$ .

*This finite moment problem could also be posed as a generalized moment problem. The resulting semi-definite program ([SDP](#)) can be solved numerically ([Lasserre, 2010](#)).*

### 2.5.1 Hybrid Representations

Especially in biological applications the abundancies of individual species vary by orders of magnitude within one model. A



**Figure 2.1:** The expected value  $\pm$  a standard deviation along with a sampled trajectory (left) and the probability distribution over time (right) of [Model 2](#) with  $\mu = 10$  and  $\gamma = 0.1$ .

canonical example of this fact are classic gene expression models, where proteins are expressed conditional on a binary gene state. To approximate such variables using a moment approximation, or even mean-field, does not make too much sense since such approximations either hinge on the high abundance of the species or try to alleviate state-space explosion stemming from it. Hybrid methods leverage the best of both worlds by approximating larger populations using continuous approximations such as moment closures, while small populations are modeled discretely.

An example of such an approach is the method of conditional moments (MCM) presented in Hasenauer et al. (2014) and Kazeroonian, Theis, and Hasenauer (2014). In this approach species are split into a high-count component  $\tilde{X}_t$  and a low-count component  $\hat{X}_t$ . Then the partial expectations

$$\mathbb{E}(\tilde{X}_t \mid \hat{X}_t = \hat{x}) \Pr(\hat{X}_t = \hat{x}) \quad (2.23)$$

are approximated for all or some valuations  $\hat{x}$  using a standard moment closure (Andreychenko, Mikeev, and Wolf, 2015). Equivalent results for (2.23) can be achieved by using that  $\mathbb{E}(\hat{X}) = \mathbb{E}(\hat{X}^m)$ ,  $m \geq 1$  for a Bernoulli random variable.

## 2.6 STATIONARY DISTRIBUTION

Assuming ergodicity of the underlying chain, a stationary distribution  $\pi_\infty$  is an invariant distribution, namely a fixed point of

the Kolmogorov forward equation (2.17). Let  $\pi_\infty$  be the vector description of a stationary distribution. It then satisfies

$$0 = \pi_\infty Q \quad (2.24)$$

as a fixed point of the Kolmogorov equation (2.17). Furthermore, the solution is constrained, to form a probability distribution, i.e. a measure with unit mass. Thus,

$$1 = \sum_{x \in \mathcal{S}} \pi_\infty(x). \quad (2.25)$$

Stationary distributions are connected to the *long-run* behavior of an MPM (Dayar et al., 2011), as the system's distribution will converge to the (unique) stationary distribution. The connection of the stationary distribution to the long-run behavior becomes clear when considering the ergodic theorem. For some  $A \subseteq \mathcal{S}$ ,

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T 1_A(X_t) dt = \sum_{x \in A} \pi_\infty(x). \quad (2.26)$$

Thus, the mean occupation time for set  $A$  over infinite trajectories is the stationary measure for  $A$ . Eq. (2.26) shows that we can assess long-run behavior using the stationary distribution and vice-versa.

**EXAMPLE** Returning to the example of Model 2 it is obvious that the state-space is irreducible. Further, we can easily show, that the stationary distribution is Poissonian with rate  $\mu/\gamma$ :

$$\pi_\infty(x) = \frac{(\mu/\gamma)^x \exp(-\mu/\gamma)}{x!}.$$

◇

For simplicity, we assume throughout that the state-space is composed of a single communicating class. Checking ergodicity given a countably infinite number of states is achieved by providing a suitable Foster-Lyapunov function (Meyn and Tweedie, 2012). Some automated techniques have been proposed for this task (Dayar et al., 2011; Gupta, Briat, and Khammash, 2014; Miliadis-Argeitis and Khammash, 2014).

## 2.7 FOSTER-LYAPUNOV BOUNDS

It is well-known that for a CTMC  $X$ , ergodicity can be proven by a Foster-Lyapunov function  $g : \mathcal{S} \rightarrow \mathbb{R}$  (Dayar et al., 2011; Meyn and Tweedie, 1993). This is the stochastic analogue of the Lyapunov functions, used to prove convergence of ODEs. The function  $g$  is required to have finite level sets:

$$|\{x \in \mathcal{S} \mid g(x) < l\}| < \infty, \quad \forall l > 0.$$

Typical choices for  $g$  are linear (Gupta, Briat, and Khammash, 2014; Miliadis-Argeitis and Khammash, 2014) or quadratic (Spieler, 2014). Given the  $g$ , we define its *drift*  $d$  as its average infinitesimal change, which is obtained applying the generator  $Q$  to  $g$ .

$$d(x) = Qg(x) = \sum_{j=1}^{n_R} \alpha_j(x)(g(x + v_j) - g(x)) \quad (2.27)$$

*Intuitively,  $g$  is interpreted as a vector with the values  $f(x_i)_i = f_i$  with the same ordering as in  $Q$ .*

As such the drift can be interpreted as the expected local tendency of change of a scalar valued function  $g$ , i.e.

$$d(x) = \frac{d}{dt} E(g(X_t) \mid X_t = x).$$

*Note that we end up with (2.21) taking the expectation.*

A Lyapunov function can be used to prove ergodicity of a CTMC: If there is a finite subset  $C \subset \mathcal{S}$  such that

$$Qg(x) \leq -1, \quad \forall x \in \mathcal{S} \setminus C, \quad (2.28)$$

$$Qg(x) < \infty, \quad \forall x \in C, \text{ and} \quad (2.29)$$

$$\|x\| \rightarrow \infty \Rightarrow g(x) \rightarrow \infty, \quad \text{where} \quad \|x\| = \sum_i x_i, \quad (2.30)$$

then the chain is non-explosive and ergodic (Miliadis-Argeitis and Khammash, 2014; Tweedie, 1975). Intuitively,  $g(X_t)$  should be a supermartingale outside  $C$  and a submartingale inside. Since the  $g$  is norm-like due to (2.30) this entails that the process tends towards  $C$  when outside, and out of from  $C$  when inside.

Given the above requirements

$$\mathcal{C}_{\epsilon_\ell} = \left\{ x \in \mathcal{S} \mid \frac{\epsilon_\ell}{c} d(x) > \epsilon_\ell - 1 \right\} \quad (2.31)$$

is finite, where

$$\infty > c \geq \sup_{x \in S} d(x).$$

In this case,  $\mathcal{C}_{\epsilon_\ell}$  contains at least  $1 - \epsilon_\ell$  of stationary probability mass for any  $\epsilon_\ell \in (0, 1)$  (Spieler, 2014, Thm. 8). Given that  $\mathcal{C}_{\epsilon_\ell}$  is finite, the chain is ergodic and

$$\sum_{x \in \mathcal{C}_{\epsilon_\ell}} \pi(x) > 1 - \epsilon_\ell \quad (2.32)$$

bounding the stationary probability mass contained within  $\mathcal{C}_{\epsilon_\ell}$ .

**EXAMPLE** We return to [Model 2](#) and choose  $g(x) = x$ . Then

$$d(x) = \mu - \gamma x.$$

The requirements (2.28), (2.29), and (2.30) are fulfilled for

$$C = \left\{ 0, \dots, \frac{1 + \mu}{\gamma} - 1 \right\}$$

and the underlying chain is ergodic. We can further bound the stationary distribution. Letting  $c = \mu$ ,

$$\frac{\epsilon_\ell}{\mu} d(x) = \epsilon_\ell - \frac{\epsilon_\ell \gamma}{\mu} x$$

and following (2.31) the states

$$0 \leq x < \frac{\mu}{\epsilon_\ell \gamma}$$

have at least  $1 - \epsilon_\ell$  stationary mass. ◇

## 2.8 AUGMENTED FOSTER-LYAPUNOV BOUNDS

The use of Foster-Lyapunov approaches, is often two-fold: They are used to prove ergodicity and to provide sets with a lower bound  $1 - \epsilon$  on their stationary probability mass. Often practical methods focus on proving *global* properties of their associated drift (Gupta, Briat, and Khammash, 2014; Spieler, 2014). In many



models, simple affine functions (Gupta, Mikelson, and Khammash, 2017) and even the identity (Spieler, 2014) is sufficient. While such simple forms provide some of the necessary ease of a global analysis, the task of optimizing them with performance in mind, is much more difficult (Miliadis-Argeitis and Khammash, 2014).

Ideally, one would have the best of both worlds: The ease of working with simple forms for the global guarantees of the Foster-Lyapunov criteria and the freedom of choice to fit efficient functions and get sets that are as small as possible. The approach presented here achieves this by *locally* altering a proposal Foster-Lyapunov function. The presence of such a valid proposal function is necessary for this method. Such a function can be identified by convex analysis, for example (Gupta, Briat, and Khammash, 2014) or using the approach described by Spieler (2014). Using a set of probability  $> 1 - \epsilon$  is identified using such a functional. On this set, the original function is then replaced by any other lower bounded function. At the set boundary this supplementary function is *phased out* and we switch back to the original function using some smooth step function. The choice of such a supplementary function offers much room for experimentation since all the necessary global criteria. For this function we can, for example, use any polynomial or even more variable models such as neural networks.

The procedure to identify an efficient function is thereby reduced to a simple machine learning problem. The objective exponentially rewards negative drift yielding tight sets. Its computation needs, in principle, all points in a sufficiently large set. For optimization however, we switch to an approach of sampling uniformly from the augmentation set.

This method yields sets that are smaller on the order of  $10^3$  for a fixed probability bound over a simple linear functional. We study neural networks and polynomial templates as candidates for local augmentation.

*Performance in this context means least states to cover most stationary probability mass.*

### 2.8.1 The Drift and its Properties

We explain the interpretation and give some basic use of the drift in [Section 2.7](#) on [page 27](#).

The drift (2.27) plays a central role in this chapter.

$$d(x; g) = Qg(x) = \sum_{j=1}^{n_R} \alpha_j(x)(g(x + v_j) - g(x)) \quad (2.33)$$

An interesting fact about the drift is, that it is invariant to linear transforms to  $g$ . That is, for a constant  $b \in \mathbb{R}$ , we can define  $h(x) = g(x) + b$  and  $Qh(x) = Qg(x)$ . Clearly, a positive linear factor  $m$  to  $g$  factors out, i.e.  $Q(f \circ g)(x) = mQg(x)$  for  $f(x) = mx$ ,  $m > 0$ . Consequently, if the drift is scaled by its maximum value, the scaled version is invariant to linear transforms of  $g$ :

$$\frac{Q(f \circ g)(x)}{\max_{x \in \mathcal{S}} Q(f \circ g)(x)} = \frac{Qg(x)}{\max_{x \in \mathcal{S}} Qg(x)} \quad (2.34)$$

Since the probability bounded sets  $C_{\epsilon_\ell}$  depend on the scaled drift. Therefore invariance under linear transformation implies that we cannot change – especially improve – the tightness of the sets by such a transform.

### 2.8.2 Local Substitution

We use a proved Foster-Lyapunov function as a starting point. For many relevant reaction networks, simple choices such as  $L_1$  or  $L_2$  norms are sufficient choices (Spieler, 2014). The resulting sets, however, are typically very large. Tasks such as computing approximate stationary distributions on truncations set up according to these sets can be very costly. This cost is exacerbated when a system has to be solved for a lot of different reentry matrices, which is necessary when state-wise bounds on the probability conditioned on a truncation are desired (Dayar et al., 2011).

We propose to augment the proposal function by a function, that is limited to local influence guided by the initial set. This supplementary function is phased out asymptotically using a simple sigmoid threshold function

$$\gamma_{k,z}(x) = \frac{1}{1 + k \exp(-x - z)} \cdot \quad (2.35)$$

Thus, in a one-dimensional model the augmented Lyapunov  $g'$  function becomes

$$g'(x) = \gamma_{k,z}(x)g(x) + (1 - \gamma_{k,z}(x))g^*(x). \quad (2.36)$$

The threshold function  $\gamma$  guarantees that  $g^*$  vanishes asymptotically. The drifts  $d'$  and  $d^*$  are defined accordingly.

### 2.8.3 Illustrating Example

Consider the example of Model 2. The question is: What makes the perfect Foster-Lyapunov function? Here, we are interested in the smallest set of states containing the largest amount of stationary probability mass. Since we know the stationary distribution to be Poissonian, its standard confidence intervals can be used as a reference. The confidence interval for level  $1 - \alpha$  of a Poisson with rate  $\mu$  is

*The beauty standard may vary for other applications.*

$$\frac{1}{2}P^{-1}(\alpha/2; 2k) \leq \mu \leq \frac{1}{2}P^{-1}(1 - \alpha/2; 2k + 2).$$

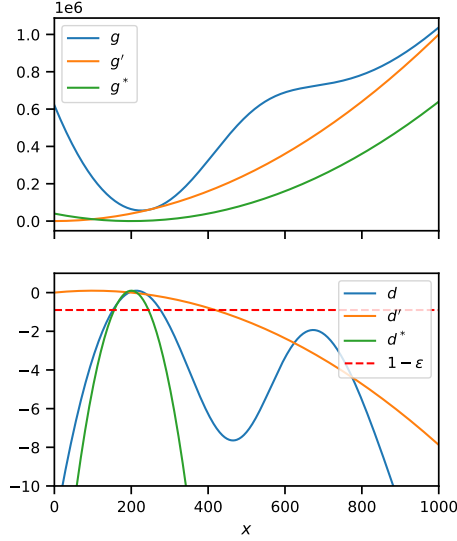
where  $P^{-1}(\cdot; l)$  is the inverse cumulative density function (CDF) of a  $\chi_l^2$  distribution. Thus, its the inverse of the regularized gamma function

$$\frac{1}{\Gamma(k/2)} \int_0^{x/2} t^{k/2-1} e^{-t} dt$$

w.r.t.  $x$ . This gives us the “ideal” intervals  $[l_\epsilon, h_\epsilon]$ ,  $(l_\epsilon, h_\epsilon) \in \mathbb{N}^2$  such that  $\pi_\infty([l_\epsilon, h_\epsilon]) = 1 - \epsilon$ . These sets are given as the dark area in Figure 2.3. Therefore, the perfect Lyapunov function would coincide with those sets. That is, a function  $g$  such that

$$C_\epsilon = [l_\epsilon, h_\epsilon], \quad \forall \epsilon > 0.$$

Using a simple L2 norm as an initially, i.e.  $g(x) = x^2$ , we obtain a fairly large set. In Figure 2.2, we contrast this result to the solution given by the choice of  $g^*(x) = (x - 250)^2$ , which gives a much tighter subset with the same guarantees. In this case, the guarantee is that the sets contain at least 0.9 stationary



**Figure 2.2:** Different example Lyapunov functions for Model 2. The drifts  $d$ ,  $d^*$ , and  $d'$  are scaled and the appropriate threshold for  $\epsilon = 0.1$  is given.

probability mass. We further demonstrate how the incorporation  $g^*$  into  $g$  significantly tightens the set proposed initially.

The benefit of the threshold-based construction (2.36) is that we only require  $g^*$  to be non-negative. All other properties are inherited from the proposal function. This freedom enables the use of flexible machine learning models to search an efficient  $g^*$ .

**NEURAL AUGMENTATION** The characteristics of the augmentation function  $g^*$  are typically not known beforehand. The formulation of augmented Foster-Lyapunov functions only places basic constraints on the function used: The function needs to be non-negative and an upper bound of the drift has to be known. Neural networks lend themselves naturally as an extremely flexible functional family.

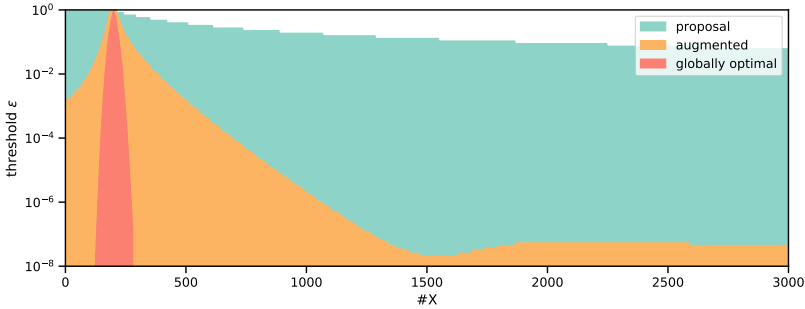
The central piece of fitting  $g^*$  is an objective function. Since the actual sets, bounded in probability, are defined in terms of their

drift, this objective needs to be a function of this drift. A desirable augmented drift has tight level sets with more emphasis placed on its peak regions. A natural way to express this prioritization is the objective

$$\sum_{x \in \mathcal{S}} \int_{-\infty}^{d^*(x)} \exp(y) dy = \sum_{x \in \mathcal{S}} \exp(d'(x))$$

based on the scaled augmented drift.

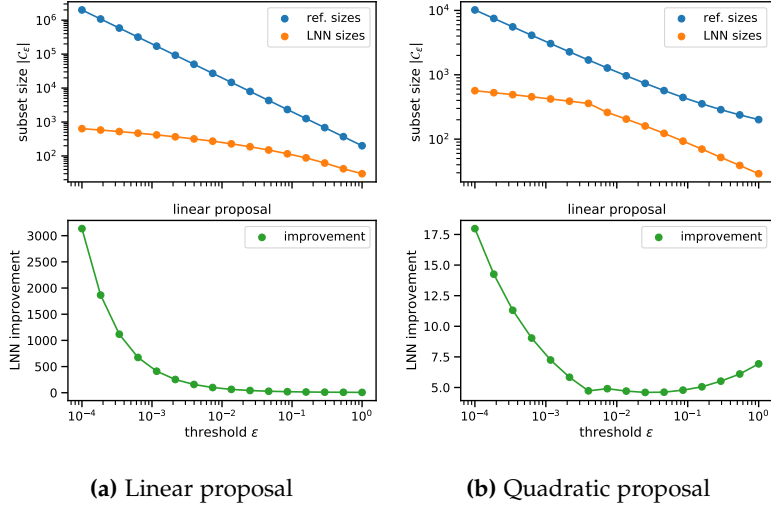
In [Figure 2.3](#) we give an example of different subsets  $|\mathcal{C}_\epsilon|$  across different thresholds  $\epsilon$ . For the augmentation we use a minimal neural network consisting of four radial basis functions and a threshold function with  $z = 1500$  and  $k = 0.01$ . While such a simple model performs well in this example, larger and higher-dimensional augmentation areas may require a more complex model. Considering the results, we observe a marked improvement over the proposal functions performance.



**Figure 2.3:** Lyapunov sets for the birth-death process for different probability thresholds  $\epsilon$  for the augmented function (orange) and the proposal (green). The “perfect” sets are computed using the confidence interval (red).

In [Figure 2.4](#), we compare the sets of Lyapunov subsets for varying thresholds  $\epsilon$  for the proposal and the augmented function. We observe a consistent improvement that is the strongest for smaller  $\epsilon$ . The overall improvement is significantly better using the linear proposal function. This is mainly due to the better performance of the quadratic proposal compared to the linear.

The subset sizes using the augmented function is qualitatively similar for both choices of proposal function.



**Figure 2.4:** The sizes of the reference Lyapunov sets compared to the sizes of the augmented Lyapunov sets along with the improvement given by the augmentation. Results are shown for both, a linear and a quadratic proposal function.

#### 2.8.4 Concluding Remarks

The local augmentation of simpler proposal Lyapunov functions is promising. To achieve more scalability and flexibility the method has some points that need refinement. The most obvious aspect is the optimization of the local augmentation. The evaluation of the objective necessitates the computation of the maximum drift. In previous works this has also been directly used as an objective (Miliadis-Argeitis and Khammash, 2014). This entails the evaluation of the drift across the entire augmentation region (and some global reasoning). Here, this issue has largely been ignored, but especially in higher dimensional models this may cause issues. There it might be better, to estimate the maxi-

mal drift using a – possibly random – grid of states instead of a dense state-space region.

Furthermore the design-space of augmentations is huge. More conventional models may prove to be easier to handle, especially with respect to the objective computation. Symbolic regression (Cranmer, 2020), for example, may be a useful alternative to derive more explainable Lyapunov functions.





## Part II

### MOMENT-BASED METHODS

We use moment properties to provide bounds on mean first-passage times and to improve statistical estimation of different quantities.



# 3

---

## BOUNDING MEAN FIRST-PASSAGE TIMES

---

For the quantitative analysis of CTMCs, many approaches have been developed, where properties of interest are often expressed in terms of temporal logics such as CSL (Aziz et al., 1996; Baier et al., 2000, 2003; Spieler, Hahn, and Zhang, 2014), MTL (Chen et al., 2011), and specifications for timed-automata (Chen et al., 2009; Mikeev et al., 2013). In addition, there exist efficient software tools (Dehnert et al., 2017; Hinton et al., 2006; Kwiatkowska, Norman, and Parker, 2011) that can be used to analyze and verify system properties. The computation of reachability probabilities is a central problem in this context.

Popular exact methods for CTMCs rely on numerical approaches that explicitly consider each system state individually. A major problem is that these methods cannot scale in the context of population models with large copy numbers of agents. A popular alternative to tackle this problem is statistical model checking, which is based on stochastic simulation (David et al., 2015). For MPMs arising in the context of chemical reaction networks, trajectories of the process are usually generated using the SSA (Gillespie, 1977). However, since the number of possible interactions grows with the number of agents, stochastic simulations of MPMs are time-consuming. Moreover, they are subject to inherent statistical uncertainty and give only statistically estimated bounds.

As an alternative, recent work concentrates on numerical methods that approximate the statistical moments of the system without the need to compute the probability of each state. For groups

↪ page 19

of identically behaving agents, it is possible to derive systems of differential equations for the evolution of the statistical population moments (Bogomolov et al., 2015; Bortolussi and Lanciani, 2013; Engblom, 2006; Gast, Bortolussi, and Tribastone, 2019; Schnoerr, Sanguinetti, and Grima, 2015, 2017). However, as the system of exact moment equations is infinite-dimensional, approximation schemes typically rely on certain assumptions about the underlying probability distribution to truncate it. For example, one might employ a “low dispersion closure” which assumes that higher-order moments are the same as those of a normal distribution (Hespanha, 2008). Such approximations are, by nature, ad-hoc and do not come with any guarantees.

Moment-based methods often scale well in terms of population sizes. However, it is not possible to control the effects of the introduced approximations, which in some cases can lead to large errors (Schnoerr, Sanguinetti, and Grima, 2015). This issue reverberates on the application of these methods to compute reachability probabilities and mean first-passage times (Bortolussi and Lanciani, 2013, 2014; Hayden, Stefanek, and Bradley, 2012). Moreover, they can suffer from numerical instabilities, in particular, when the maximum order of the considered moments has to be increased to more appropriately describe the underlying distribution.

Here, we put forward a method based solely on moments that gives *exact bounds* for mean first-passage times (MFPTs) and reachability probabilities in MPMs. For a set of states  $B$  and a time-horizon  $T$ , the first-passage time (FPT)

$$\tau = \inf\{t \geq 0 \mid X_t \in B\} \wedge T.$$

This mean of this stopping time  $E(\tau)$ , i.e. the mean first-passage time, directly characterizes the probability of reaching set  $B$  within  $T$  time units. Thus, safe upper and lower bounds on MFPTs can constitute a core component for the verification of properties in MPMs. Our approach extends recent work on moment bounds (Dowdy and Barton, 2018b; Sakurai and Hori, 2017) and it is based on a martingale formulation of the stopped process that we derive from the exact moment equations. From this formalization, we deduce a set of linear moment constraints from

which we derive upper and lower moment bounds using semi-definite program (SDP). Monotone sequences of both upper and lower bounds can be obtained by increasing the order of the relaxation. Crucially, no closure approximations are introduced. Therefore the bounds are exact up to the numerical accuracy of the SDP solver.

To experimentally validate our method in terms of accuracy and feasibility, we run some tests on examples from biology, leveraging an existing SDP solver and obtaining encouraging results. Comparing with other moment-based methods, our approach is not based on approximations due to closure schemes, thus providing guarantees on the bounds up to the numerical accuracy of the computations. However, similarly to other moment-based methods, we also found the insurgence of numerical instabilities because moments of higher order tend to span over many orders of magnitude. We ameliorate this problem by considering scaling strategies that reduce such variability. We also extend our approach to deal with MPMs exhibiting strong multimodal behavior, due to the presence of populations having low copy numbers. This extension exploits some ideas from hybrid moment closures (Kazeroonian, Theis, and Hasenauer, 2014). In summary, this chapter presents the following novel contributions:

- the derivation of moment constraints, based on a martingale formulation, for bounding mean first-passage times and reachability probabilities using a convex programming scheme;
- the extension of this scheme using hybrid moment conditions to systems exhibiting multimodal behavior;

### 3.1 RELATED WORK

**TRUNCATIONS AND ANALYTIC SOLUTIONS** Considerable effort has been directed at the analysis of first-passage time distributions in MPMs. Most works can either focus on an explicit state-space analysis (Barzel and Biham, 2008; Kuntz et al., 2019,

2021a; Munsky, Nemenman, and Bel, 2009) or employ approximation techniques for which, in general, no error bounds can be given (Bortolussi and Lanciani, 2014; Hayden, Stefanek, and Bradley, 2012; Schnoerr et al., 2017). For some model classes such as kinetic proofreading, analytic solutions are possible (Bel, Munsky, and Nemenman, 2009; Iyer-Biswas and Zilman, 2016; Munsky, Nemenman, and Bel, 2009).

Barzel and Biham (2008) propose a recursive scheme that consists of one equation for each state, expressing the average time the system needs to transition from that state to the target state. Kuntz et al. (2021a) propose to employ moment bounds in a linear programming approach to compute exit time distribution using state-space truncation schemes. In Kuntz et al. (2019) the authors propose a finite state-space projection scheme to bound first-passage time distributions

**MOMENTS APPROXIMATIONS** In Hayden, Stefanek, and Bradley (2012), the authors use moment closure approximations and Chebychev’s inequality to gain an understanding of first-passage time dynamics. Schnoerr et al. (2017) also employ a moment closure approximation and further approximate threshold functions to derive an approximate first-passage time distribution. Bortolussi and Lanciani (2014) use a mean-field approximation which is required to reach the target region.

**MOMENT BOUNDS** Recently, several groups independently suggested the use of semi-definite optimization for the computation of moment bounds for the limiting distribution (Dowdy and Barton, 2018a; Ghusinga et al., 2017; Kuntz, Juan And Thomas, Philipp And stan and Barahona, 2019; Sakurai and Hori, 2017). In this approach, the differential equations describing the moment dynamics are set to zero and form linear constraints (Backenköhler, Bortolussi, and Wolf, 2018). Alongside, semi-definite constraints can be placed on the *moment matrices*. These give a semi-definite program (SDP) that can be solved efficiently. Previously, SDPs have been used in the deterministic setting (Hasenauer et al., 2009).

This approach has been extended to the transient case (Dowdy and Barton, 2018b; Holtorf and Barton, 2021; Sakurai and Hori, 2019). The approach is similar in both works and is a cornerstone of the MFPT analysis presented here. They differ mainly by the fact that Sakurai and Hori (2019) apply a polynomial time-weighting, while Dowdy and Barton (2018b) use an exponential one. We adopt the former approach because it can be naturally adapted to the description of densities over time. The resulting forms can also be adapted to statistical estimation problems (Backenköhler, Bortolussi, and Wolf, 2019).

Semi-definite programming has been applied to a wide range of problems, including stochastic processes in the context of financial mathematics (Kashima and Kawai, 2009; Lasserre, Prieto-Rumeau, and Zervos, 2006). For a comprehensive introduction of SDP and its application areas, we refer the reader to Parrilo (2003) and, more recently, Lasserre (2010).

For the mean-field ODE model tropical analysis has been used to derive bounds on species using dominant terms (Beica, Feret, and Petrov, 2020).

**BOUNDING MFPTS USING OPTIMIZATION** Particularly relevant for this work is the application of convex optimization to FPT. Helmes, Röhl, and Stockbridge (2001) formulated a linear program (LP) using the Hausdorff moment conditions (Hausdorff, 1921) to bound moments of the FPT distribution in Markovian processes. Semi-definite optimization has been successfully applied in financial mathematics by Kashima and Kawai (2009), as well as Lasserre, Prieto-Rumeau, and Zervos (2006) to bound prices of exotic options.

### 3.2 PRELIMINARIES

In this work, we are interested in first-passage times (FPTs) of such processes. That is the time, the process first enters a set of target states  $B \subseteq S$ . Naturally, the analysis of FPTs is equivalent to the analysis of times at which the process exits the complement

$S \setminus B$ . More formally, the first-passage time  $\tau$  for some target set  $B$  is defined as the random variable

$$\tau = \inf\{t \geq 0 \mid X_t \in B\}. \quad (3.1)$$

In this example, we are interested in the time at which the number of type  $M$  agents exceed some threshold  $H$ . With the framework presented in the sequel, one can bound the expected value of this time using semi-definite programming. Further, it is possible to impose a time-horizon  $T$ , and find bounds on the probability of  $X_t^{(M)} \geq H$  for some  $0 \leq t \leq T$ . The employed framework is centered around semi-definite relaxations of the generalized moment problem (GMP) (Lasserre, 2010). These require linear constraints on the moments of measures. In the following section, we derive such constraints.

### 3.3 MARTINGALE FORMULATION

Next, we will discuss the ordinary differential equations for the evolution of the statistical moments of the process. The moments over the state-space are then used to derive temporal moments, i.e. moments of measures over both the state-space and the time. This extended description results in a process with the martingale property. This property can be used to formulate linear constraints on the temporal moments and, as a special case, the mean first-passage time. In combination with semi-definite properties of moment matrices, we can formulate mathematical programs that yield upper and lower bounds on mean first-passage times.

↪ page 17

Let  $f$  be a polynomial function,  $t \geq 0$ . Using the CME (2.18), we can derive ODEs describing the dynamics of  $E(f(X_t))$  (Engblom, 2006). Specifically,

More details on the derivation of the moment ODEs is given in Section 2.5.

$$\frac{d}{dt} E(f(X_t)) = \sum_{j=1}^{n_R} E((f(X_t + v_j) - f(X_t)) \alpha_j(X_t)). \quad (3.2)$$

↪ page 18

**EXAMPLE** Let us consider Model 1 and agent type  $M$  as an example. Further, let  $X_t = X_t^{(M)}$  for ease of exposition. When



choosing  $f(X_t) = X_t^m$ ,  $m = 1$  and  $m = 2$  we obtain two differential equations describing the change of the first two moments of species  $M$ ,  $E(X_t)$  and  $E(X_t^2)$ , respectively.

$$\begin{aligned} \frac{d}{dt} E(X_t) &= \lambda E(X_t^0) - 2\delta (E(X_t^2) - E(X_t)) \\ \frac{d}{dt} E(X_t^2) &= \lambda(2E(X_t) + 1) \\ &\quad - 4\delta (E(X_t^3) - 2E(X_t^2) + E(X_t)) . \end{aligned} \quad (3.3)$$

Fixing initial moments, the ODE system describes the moments over time exactly. However, these ODEs cannot be integrated because the system is not closed. The right-hand side for moment  $E(X_t^m)$  always contains  $E(X_t^{m+1})$ .  $\diamond$

To solve the IVP, one typically resorts to ad-hoc approximations of the highest order moments to close the system. Here we do *not* need such approximations because we do not numerically integrate the moment equations. Instead we adopt an approach (Dowdy and Barton, 2018b; Sakurai and Hori, 2019) that extends the description of state-space moments to a temporal one.

This is achieved by the introduction of a time-dependent polynomial  $w(t)$  that is multiplied to (3.2). An integration by parts on  $[0, T]$  yields (Dowdy and Barton, 2018b; Sakurai and Hori, 2019)

$$\begin{aligned} w(T)E(f(X_T)) - w(t_0)E(f(X_{t_0})) - \int_{t_0}^T \frac{dw(t)}{dt} E(f(X_t)) dt \\ = \sum_{j=1}^{n_R} \int_{t_0}^T w(t) E((f(X_t + v_j) - f(X_t)) \alpha_j(X_t)) dt . \end{aligned} \quad (3.4)$$

Starting from this equation, it is possible to derive a martingale process, i.e. a process that has an expected value equal to 0, regardless of time.

We now want to interchange the order of integration and the summation due to the expected value. To this end, we have to assume the absolute convergence of the integrals. On finite time intervals  $[0, T]$  this holds because  $w$  is polynomial and we assumed finite moments for all  $t \geq 0$ . Interchanging the summa-

tion and integral of a monomial  $x^m$ , i.e. pulling all expectation operators outside

$$\begin{aligned}
& \int_0^T g(t) E(X_t^m) dt \\
&= \int_0^T \sum_{x \in S} g(t) \Pr(X_s = x) x^m dt \\
&= \int_0^T \int_{\Omega} g(t) X_s(\omega)^m dP(\omega) dt \\
&= \int_{\Omega} \int_0^T g(t) X_s(\omega)^m dt dP(\omega) \\
&= E \left( \int_0^T g(t) X_t^m dt \right).
\end{aligned}$$

Hence, we are able to pull out the expectation operator in (3.4).

$$\begin{aligned}
0 &= w(T) E(f(X_T)) - w(0) E(f(X_0)) - E \left( \int_0^T \frac{dw(t)}{dt} f(X_t) dt \right) \\
&\quad - \sum_{j=1}^{n_R} E \left( \int_0^T w(t) (f(X_t + v_j) - f(X_t)) \alpha_j(X_t) dt \right),
\end{aligned} \tag{3.5}$$

This gives us the expected value of a time-dependent function of the original process. The function can be viewed as a stochastic process of its own where the time-horizon  $T$  is the index variable. A key property of this process is also illustrated by (3.5): The process' expected value remains zero, regardless of the choice of  $T$ . This martingale property is particularly useful because it can be used to formulate linear constraints on stopping times of the process. Explicitly, we can define this process  $\{Z_T\}_{T \geq 0}$  parameterized by the time-weighting  $w$  and polynomial  $f$ .

$$\begin{aligned}
Z_T &:= w(T) f(X_T) - w(0) f(X_0) - \int_0^T \frac{dw(t)}{dt} f(X_t) dt \\
&\quad - \sum_{j=1}^{n_R} \int_0^T w(t) (f(X_t + v_j) - f(X_t)) \alpha_j(X_t) dt.
\end{aligned} \tag{3.6}$$

A useful choice for  $f$  and  $w$  are monomials. When choosing  $w(t) = t^k$  with  $k \in \mathbb{N}$  and  $f(X) = X^m$  the process takes the form

$$Z_T^{(m,k)} = T^k X_T^m - 0^k X_0^m + \sum_i c_i \int_0^T t^{k_i} X_t^{m_i} dt \quad (3.7)$$

In [Chapter 4](#) we use an exponential weighting.

where  $(m_i)_i$ ,  $(k_i)_i$ , and  $(c_i)_i$  are finite sequences resulting from the substitution of  $f$  and  $w$  and expansion of (3.6).

When choosing  $w(t) = t^k$  with  $k \in \mathbb{N}$  and  $f(X) = X^m$  this process takes the form

$$Z_T^{(m,k)} = T^k X_T^m - 0^k X_0^m + \sum_i c_i \int_0^T t^{k_i} X_t^{m_i} dt \quad (3.8)$$

where  $(m_i)_i$ ,  $(k_i)_i$ , and  $(c_i)_i$  are finite sequences resulting from the substitution of  $f$  and  $w$ . This choice allows to naturally characterize the behavior in time and state-space as moments, because the expected value of (3.8) then becomes a linear form of moments. We will use these as constraints in the semi-definite program used to bound MFPTs.

**EXAMPLE** If we apply this to our previous example (cf. (3.3)), letting  $m = 1$  and  $k = 1$  we obtain the following process for [Model 1](#).

$$\begin{aligned} Z_T^{(1,1)} = & TX_T - \int_0^T X_t dt - \lambda \int_0^T t dt \\ & - 2\delta \int_0^T t X_t dt + 2\delta \int_0^T t X_t^2 dt, \end{aligned}$$

where the sequences according to (3.8) are  $(m_i)_i = (1, 0, 1, 2)$ ,  $(k_i)_i = (0, 1, 1, 1)$ , and  $(c_i)_i = (-1, -\lambda, -2\delta, 2\delta)$ .  $\diamond$

We now turn to the analysis of first passage times within some time-bound  $T > 0$ . Given some subset of the state-space  $B \subseteq \mathcal{S}$  the first passage time is given by the continuous random variable

$$\tau = \inf\{t \geq 0 \mid X_t \in B\} \wedge T$$

where  $a \wedge b := \min\{a, b\}$ . For this chapter, we only look at threshold hitting times, i.e. we set a threshold  $H$  for species  $S$  and thus

$$B = \{x \mid x^{(S)} \geq H\}.$$

Note, that this framework allows for a more general class of target sets, which are discussed in [Section 3.6.1](#). In the sequel, we will use  $\tau$  as a stopping time in our martingale formulation and consider  $Z_\tau^{(m,k)}$  instead of  $Z_T^{(m,k)}$ . Since (3.8) defines a martingale,  $Z_\tau^{(m,k)}$  remains a martingale by Doob's optional sampling theorem (Gihman and Skorohod, 1975). In particular, this implies that

$$E(Z_\tau^{(m,k)}) = 0 \quad (3.9)$$

for all moment orders  $m$  and degrees  $k$  in the weighting function  $w(t)$ .

### 3.4 LINEAR MODEL CONSTRAINTS

To simplify our presentation, we fix an initial state  $x_0$ , i.e.  $P(X_0 = x_0) = 1$ . Expanding (3.9) using (3.8) for  $Z_\tau^{(m,k)}$  yields the following linear constraint on expected values.

$$0 = E(\tau^k X_\tau^m) - 0^k x_0^m + \sum_i c_i E\left(\int_0^\tau t^{k_i} X_t^{m_i} dt\right), \quad (3.10)$$

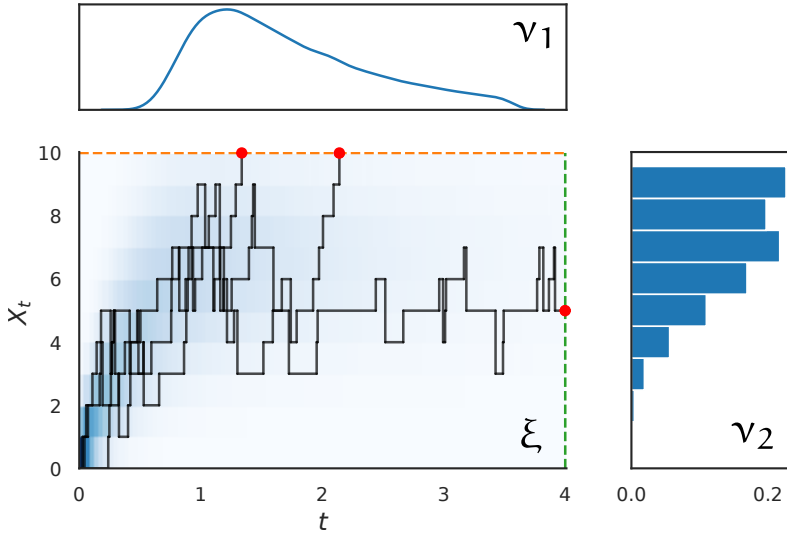
where  $0^0 = 1$ . Hence, we have established a relationship between the process dynamics up to the hitting time via expected values of the time-integrals and the final process state at the hitting time via  $E(\tau^k X_\tau^m)$ .

For the ease of exposition, we now turn to the analysis of first passage times in one-dimensional processes w.r.t. an upper threshold  $H$ . In particular, we will consider moments  $E(X^m)$ ,  $m = 0, 1, 2, \dots$ , of a one-dimensional process. The approach proposed in the sequel, however, can be extended to multi-dimensional processes and more complex target sets  $B$ .

**EXAMPLE** Consider again [Model 1](#) and assume that we are interested in the time at which species  $M$  exceeds threshold  $H$  while fixing the considered time-horizon to  $T = 4$ . That is, we are interested in the stopping time

$$\tau = \inf\{t \geq 0 \mid X_t \geq 10\} \wedge 4.$$

Since the abundance of  $D$  does not influence  $M$ , we can ignore species  $D$  and treat the process as one-dimensional. Figure 3.1 shows three example trajectories: Two reach an upper threshold  $H = 10$ , while one reaches the final time-horizon  $T = 4$ . The figure also illustrates another aspect present in (3.10). It gives a connection between the terminal distribution, i.e. the distribution of  $X_\tau$ , and the dynamic behavior up to  $\tau$ . The statistics at  $\tau$  are described by a distribution whose moments are represented by the  $E(\tau^k X_\tau^m)$  term in (3.10). This distribution corresponding two moments encompasses both cases of how  $\tau$  can be reached. In the first case threshold  $H$  is reached and the second case the process reaches the time-horizon  $T$ . In the following we will define the interplay between these measures more formally.  $\diamond$



**Figure 3.1:** The relationship between the occupation measure  $\xi$ , and the exit location probability measures  $\nu_1$  and  $\nu_2$ . The shaded area indicates the structure of the occupation measure. Three example trajectories are additionally plotted with their exit location highlighted. The plots are based on 10,000 sample trajectories.

Equation (3.10) describes a relationship between two measures (Lasserre, 2010, Chapter 9.2):

EXPECTED OCCUPATION MEASURE  $\xi$  describes the expected residence time inside a subset of the state-space and time. As such it is supported on  $[0, H] \times [0, T]$ :

$$\xi(A \times C) := E \left( \int_{[0, \tau] \cap C} 1_{\in A}(X_t) dt \right), \quad (3.11)$$

EXIT LOCATION PROBABILITY  $\nu$  gives the state probability associated with the stopping time  $\tau$ . Therefore it is supported on  $(\{H\} \times [0, T]) \cup ([0, H] \times \{T\})$ :

$$\nu(A \times C) := \Pr((X_\tau, \tau) \in A \times C), \quad (3.12)$$

where  $A \times C$  is a measurable set, i.e.  $A$  and  $C$  are elements of the Borel  $\sigma$ -algebras on  $[0, H]$  and  $[0, T]$ , respectively.

Using [Figure 3.1](#), one can gain an intuition for these two measures. The expected occupation measure is shaded in blue. As the name implies  $\xi(A \times C)$  tells us how much time the process spends in  $A$  up to  $\tau$  restricting to the time instants belonging to  $C$ . In particular,  $\xi([0, H] \times [0, T]) = E(\tau)$ . The exit location probability  $\nu$ , while being a two-dimensional distribution, can be viewed as a composition of a density describing the time at which the process reaches  $H$  (if it does) and a probability mass function on the states of the process if the time-horizon is reached without exceeding  $H$ . We partition the measure  $\nu$  into  $\nu_1$  and  $\nu_2$  by conditioning on  $\tau = T$ . Thus,

$$\nu_1(C) := \Pr(\tau \in C, \tau < T)$$

and

$$\nu_2(A) := \Pr(X_T \in A, \tau = T)$$

and hence  $\nu(A \times C) = \nu_1(C) + \nu_2(A)$ . To refer to the moments of these measures, we define *partial moments*

$$E(g(X); f(Y) = y) := E(g(X) \mid f(Y) = y) \Pr(f(Y) = y),$$

for some polynomial  $g$  and some indicator function  $f$ . Then

$$E(\tau^k X_\tau^m) = T^k E(X_\tau^m; \tau = T) + H^m E(\tau^k; \tau < T, X_\tau = H).$$

Therefore the linear moment constraints are

$$0 = T^k E(X_\tau^m; \tau = T) + H^m E(\tau^k; \tau < T, X_\tau = H) - 0^k x_0^m + \sum_i c_i E\left(\int_0^\tau t^{k_i} X_t^{m_i} dt\right). \quad (3.13)$$

Next, we consider infinite sequences of partial moments  $y_1 = (y_{1k})_{k \in \mathbb{N}}$ ,  $y_2 = (y_{2m})_{m \in \mathbb{N}}$ , and  $z = (z_{mk})_{(m,k) \in \mathbb{N}^2}$  of  $\nu_1$ ,  $\nu_2$ , and  $\xi$ , respectively. In particular,

$$y_{1k} := E(\tau^k; \tau < T),$$

$$y_{2m} := E(X_\tau^m; \tau = T),$$

and

$$z_{km} := E\left(\int_0^\tau t^k X_t^m dt\right).$$

### 3.5 OBJECTIVE

Given the above measures and their corresponding moments, we can now identify the moments we are particularly interested in. We formulate an optimization problem with variables corresponding to the moments defined above. The [MFPT](#) is exactly the zeroth moment of  $\xi$ ,

$$z_{00} = E\left(\int_0^\tau 1_{\leq H}(X_t) dt\right) = E(\tau).$$

Therefore  $z_{00}$  corresponds to the objective of the optimization problem that gives bounds for the [MFPT](#). Furthermore, we can easily change the objective to the zeroth moment of  $\nu_1$ ,

$$y_{10} = E(\tau^0; \tau < T) = \Pr(\tau < T).$$

This moment is the probability of reaching threshold  $H$  before reaching time-horizon  $T$ . Since the target set can be more complex, this formulation can be used to perform model checking on a wide variety of properties.

Moreover, it is possible to formulate objectives not directly corresponding to a raw moment such as the variance (Dowdy and Barton, [2018a](#); Sakurai and Hori, [2019](#)).

## 3.6 SEMI-DEFINITE CONSTRAINTS

The linear constraints alone are not sufficient to identify moment bounds. We further leverage the fact that a necessary condition for a positive measure that the *moment matrices* are positive semi-definite. A matrix  $M \in \mathbb{R}^{n \times n}$  is positive semi-definite, denoted by  $M \succeq 0$  if and only if

$$v^T M v \geq 0 \quad \forall v \in \mathbb{R}^n.$$

**EXAMPLE** As an example, let us consider a one-dimensional random variable  $Z$  with moment sequence  $z$ . For moment order  $r$ , the entries of the  $(r+1) \times (r+1)$  moment matrix  $M_r(x)$  are given by the raw moments. In particular,

$$(M_r(z))_{ij} = z_{i+j-2} = E(Z^{i+j-2})$$

for  $i, j \in \mathbb{N}_r$  where  $\mathbb{N}_r = \{0, 1, \dots, r\}$  and the maximum order in the matrix is  $2r$ . For instance,

$$M_1(x) = \begin{bmatrix} x_0 & x_1 \\ x_1 & x_2 \end{bmatrix} \tag{3.14}$$

needs to be positive semi-definite. By Sylvester's criterion this means  $\det M_1 \geq 0$  and  $x_0 \geq 0$ . We can easily see that in this case this entails

$$\det M_1 = x_0 x_2 - x_1^2 = E(X^2) - E(X)^2 = \text{Var}(X) \geq 0.$$

This restriction is natural since the variance cannot be negative.  $\diamond$

The restriction of the non-negative variance we saw in the example generalizes to moment matrices in form of a positive semi-definite constraint (Parrilo, 2003). This gives us the following restrictions on the moment matrices.

$$M_r(z) \succeq 0, \quad M_r(y_1) \succeq 0, \quad \text{and} \quad M_r(y_2) \succeq 0 \tag{3.15}$$

for arbitrary orders  $r$ , providing a first tranche of moment constraints.

Furthermore, we need to enforce the restriction of the measures  $\xi$ ,  $\nu_1$ , and  $\nu_2$  to their supports. This can be done, by defining non-negative polynomials on the intended support of the measure.

*This constraint is valid for general positive measures — even if they do not model probabilities.*



**EXAMPLE** The exit location probability  $\nu_2$  has support  $[0, H]$ . We can now define

$$u_H(t, x) = Hx - x^2, \quad x \in \mathbb{R}$$

as a polynomial that is non-negative on  $[0, H]$ . Using such polynomials, we can construct *localizing matrices*, which have to be positive semi-definite (Lasserre, 2010). Applying  $u_H$  to the moment matrix of measure  $\nu_2$ , i.e.  $M_1(y_2)$

$$M_1(u_H, y_2) = \begin{bmatrix} Hy_{21} - y_{22} & Hy_{22} - y_{23} \\ Hy_{22} - y_{23} & Hy_{23} - y_{24} \end{bmatrix}$$

with the constraint  $M_1(u_H, y_2) \succeq 0$ , where the application of a polynomial such as  $u_H$  to a moment matrix is formally defined for the multidimensional case in Section 3.6.1. Similarly, let  $u_T(t, x) = Tt - t^2$  to restrict  $\nu_1$  to  $[0, T]$ . The expected occupation measure  $\xi$  is constrained similarly to its domain  $[0, H] \times [0, T]$ . This gives us the following restrictions on the moment matrices.

$$\begin{aligned} M_r(u_T, z) &\succeq 0, & M_r(u_T, y_1) &\succeq 0, \\ M_r(u_H, z) &\succeq 0, & M_r(u_H, y_2) &\succeq 0. \end{aligned} \tag{3.16}$$

◇

### 3.6.1 Multi-Dimensional Generalization

For a general multi-dimensional moment sequence

$$y = (E(X^m))_{m \in \mathbb{N}^{n_s}},$$

the moment matrix is (Lasserre, 2010)

$$M_r(y)(\alpha, \beta) = y_{\alpha+\beta}, \quad \forall \alpha, \beta \in \mathbb{N}_r^n \tag{3.17}$$

where row and column indices,  $\alpha$  and  $\beta$ , are ordered according to the canonical basis

$$\begin{aligned} v_r(x) = (1, x_1, x_2, \dots, x_n, x_1^2, x_1x_2, \dots, \\ \dots, x_1x_n, \dots, x_1^r, \dots, x_n^r)^T. \end{aligned} \tag{3.18}$$

Equivalently to (3.17),

$$M_r(y) = E \left( v_r(x) v_r(x)^T \right).$$

For a moment sequence the semi-definite restriction  $M_r(y) \succeq 0$  must hold.

Measures can be restricted to semi-algebraic sets

$$\{x \in \mathbb{R}^n \mid u_j(x) \geq 0, j = 1, \dots, m\},$$

where  $u_j, j = 1, \dots, m$  are polynomials (Lasserre, 2010). This is done by placing restrictions on the localizing matrices. For each polynomial  $u_i \in \mathbb{R}[x]$  with coefficient vector  $u = \{u_\gamma\}$ , i.e.

$$u(x) = \sum_{\gamma \in \mathbb{N}^n} u_\gamma x^\gamma,$$

the localizing matrix is

$$M_r(u, y)(\alpha, \beta) = \sum_{\gamma \in \mathbb{N}^n} u_\gamma y_{\gamma + \alpha + \beta}, \quad \forall \alpha, \beta \in \mathbb{N}_r^n.$$

Requiring that this matrix is positive semi-definite restricts the measure to  $\{x \mid u_i(x) \geq 0\}$ . This way we can, for example, restrict the moment sequence  $y$  to measures that are positive w.r.t. dimension  $j$ . Simply letting  $u(x) = x_j$  and requiring  $M_1(u, y) \succeq 0$  for  $i = 1, \dots, n_S$  gives us this restriction.

### 3.6.2 A Semi-Definite Program

With the linear constraints given in (3.10) and the semi-definite constraints (3.15) and (3.16) discussed in the previous sections, we can now formulate a semi-definite program (SDP) for any relaxation order  $0 < r < \infty$ . An SDP is a convex optimization problem over the set of positive semi-definite  $n \times n$ -matrices  $X$  under linear constraints:

$$\begin{aligned} \min_{X \in \mathcal{X}} \quad & \sum_{i,j} A_{ij}^{(0)} X_{ij} \\ \text{such that} \quad & X \succeq 0 \\ & \sum_{i,j} A_{ij}^{(k)} X_{ij} \leq b_k, \quad k = 1, \dots, m \end{aligned} \tag{3.19}$$

with constant matrices  $A^{(i)} \in \mathbb{R}^{n \times n}$ ,  $i = 0, \dots, m$  and constants  $b_k \in \mathbb{R}$ ,  $k = 1, \dots, m$  to define a set of  $m$  linear constraints. Such a problem is convex and can be solved efficiently using off-the-shelf solvers (Vandenberghe, 2010).

With each moment sequence  $x$  we associate a sequence proxy variables  $x'$  used in the optimization problem.

$$\begin{aligned}
 & \min/\max \quad z'_{00} \\
 & \text{such that} \quad M_r(z') \succeq 0, M_r(u_T, z') \succeq 0, M_r(u_H, z') \succeq 0, \\
 & \quad M_r(y'_1) \succeq 0, M_r(u_T, y'_1) \succeq 0, \\
 & \quad M_r(y'_2) \succeq 0, M_r(u_H, y'_2) \succeq 0, \\
 & \quad 0 = y'_{1k} H^m - y'_{2m} T^k - 0^k x_0^m \\
 & \quad \quad + \sum_i c_i z'_{k_i m_i}, \quad \forall m, k.
 \end{aligned} \tag{3.20}$$

This SDP can be compiled to the canonical form. To this end, the moment matrices can be arranged in a block-diagonal form and the localizing constraints (3.16) can be encoded by the introduction of new variables and appropriate equality constraints. This transformation can be done automatically using modeling frameworks such as CVXPY (Diamond and Boyd, 2016). We therefore only give the SDP in the more intuitive format. This problem can be solved using off-the-shelf SDP solvers such as MOSEK (MOSEK ApS, 2018), CVXOPT (Vandenberghe, 2010), or SCS (O'Donoghue et al., 2017).

In principle, we can choose an arbitrarily large order  $r$  for the moment matrices and their corresponding constraints, because there are infinitely many moments. In practice, however, the order is bounded by practical issues such as the program size (number of constraints and variables) and numerical issues. These issues are discussed in Section 3.7 in more detail. Choosing a finite  $r$  is a relaxation of the problem since it removes constraints regarding higher-order moments.

### 3.7 IMPLEMENTATION & EVALUATION

The implementation of the SDP (3.20) is straightforward using modeling frameworks and off-the-shelf solvers. However, as

noted in previous work (Dowdy and Barton, 2018a,b; Sakurai and Hori, 2017, 2019) on moment-based SDPs the direct implementation of the problem may lead to difficulties for the solver. A source of these is that moments of various orders by nature may differ by many orders of magnitude. A re-scaling of the moments (Dowdy and Barton, 2018a; Sakurai and Hori, 2019) such that moments only vary by few orders of magnitude may alleviate this problem. In other scenarios such as the bounding of general transient or steady-state moments, the scaling can be particularly difficult, because the magnitude of moments is generally not known a priori. In the context of MFPTs with a finite time-horizon moments are trivially bounded.

### 3.7.1 Moment Scaling

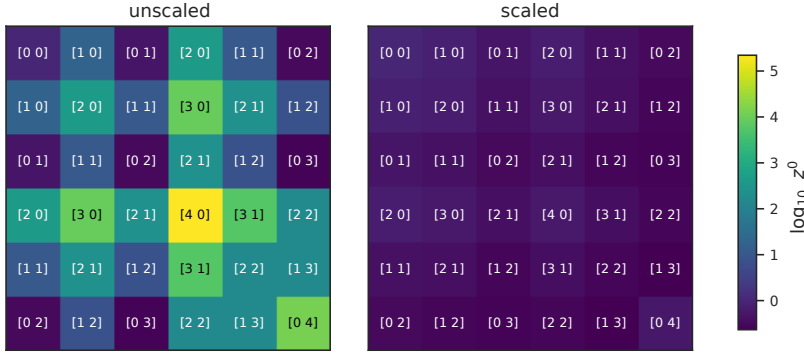
Using the fact that  $S \setminus B$  is often finite, it is possible to derive trivial bounds, which can be used to scale moments. If, for example, we have a one-dimensional process  $X_t$  with  $X_0 = 0$  a.s. and are interested in the hitting time of an upper threshold  $H > 0$  until time  $T > 0$  for  $i, k \in \mathbb{N}$

$$\begin{aligned} z_{ik} &= E \left( \int_0^T t^i X_t^k dt \right) \leq E \left( \int_0^T t^i X_t^k dt \right) \\ &\leq H^k \int_0^T t^i dt = \frac{T^{i+1} H^k}{i+1}. \end{aligned}$$

Thus, we fix a scaling vector  $d$  with entries  $d_{ik} = T^{i+1} H^k$  in the same order as the canonical base vector (3.18). Using this scaling vector, we can define a scaling matrix  $D = dd^T$ . Clearly,  $D \succeq 0$ . Now we can formulate the optimization (3.20) over a scaled version  $D^{-1}M(z')$  instead of  $M(z')$ . The moment matrices of the exit location probabilities are scaled in the same way. Alternatively, one could use approximations such as moment closures or bounds obtained by lower-order relaxations or solve a sequence of problems, incrementally increasing the time-horizon, and adjust the scaling accordingly (suggested in Dowdy and Barton (2018b)).

*These scaling strategies have not been evaluated so far.*

In [Figure 3.2](#) we illustrate the influence the scaling has on the optimization variables. While the unscaled version shows large differences between values, these differences become significantly smaller in the scaled version of the problem.



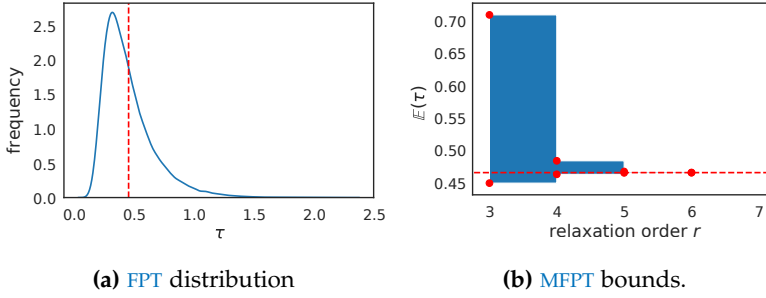
**Figure 3.2:** The unscaled and scaled value of the moment matrix proxy variable  $M(z')$  after optimization using MOSEK. The indices are given along the logarithmic (base 10) values. The unscaled version (left) shows large differences in magnitudes, while on the scaling suppresses these large variations (right). The case study used here is [Model 1](#), with a threshold  $H = 25$  for species  $M$  and a time-horizon  $T = 1$ . The relaxation order  $r = 2$ . Therefore moments of orders up to  $2r = 4$  appear.

### 3.7.2 Case Studies

The model constraints are derived using the symbolic math toolkit Sympy (Meurer et al., 2017). We implemented and solved the [SDP](#) programs described above using optimization suite MOSEK version 9.1.2 (MOSEK ApS, 2018). Both optimizers are used through the CVXPY interface (Diamond and Boyd, 2016).

#### Dimerization

As a first case study, we use [Model 1](#) with parameters  $\lambda = 100$  and  $\delta = 0.2$ . In this model, we are interested in the time at which



**Figure 3.3:** First-passage time characteristics for [Model 1](#) with  $\tau = \inf\{t \geq 0 \mid X_t \geq 10\} \wedge \infty$ . The dashed red line denotes the sampled MFPT based on 100,000 SSA samples. Bounds are based on the SDP (3.20) with different moment orders

the number of agents of type M surpasses a threshold of 25 before some time-horizon  $T$ , i.e.

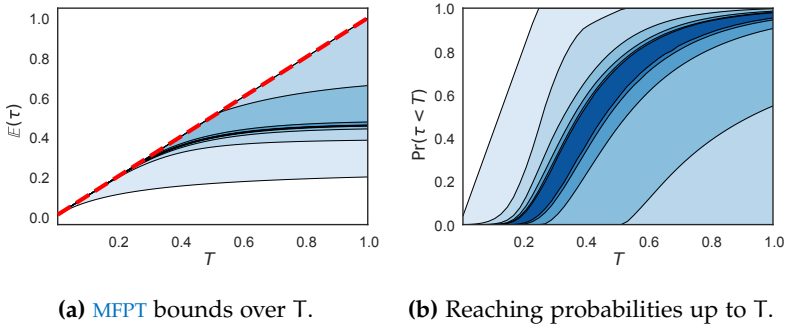
$$\tau = \inf\{t \geq 0 \mid X_t \geq 25\} \wedge T.$$

*We can let  $T \rightarrow \infty$  because this system is ergodic and therefore  $\tau < \infty$  a.s.*

First, we set no finite time-horizon  $T$ , i.e.  $T = \infty$ . This is achieved by dropping the moments  $y_2$  of measure  $\nu_2$  in the linear constraints (3.20). This can be done because the threshold on  $M$  makes the state-space finite and therefore the first passage time distribution is a phase-type distribution which possesses finite moments (Stewart, 2009, Chapter 7.6).

The empirical FPT distribution based on 100,000 SSA simulations is given in [Figure 3.3a](#) and the bounds, given different moment orders, are given in [Figure 3.3b](#). As we can see in [Figure 3.3b](#), the bounds capture the MFPT precisely for orders 5 and 6. The difference between upper and lower bound decreases roughly exponentially with increasing relaxation order  $r$ . We found that this trend was consistent among the case studies presented here (cf. [Figure 3.5](#)).

Next, we look at first passage times within a finite time-horizon  $T$ . In [Figure 3.4a](#) we summarize the bounds obtained for the MFPT over  $T$ . While low-order relaxations (light) give rather loose bounds, the bounds are already fairly tight when using  $r = 4$ . In many cases, hitting probabilities, that is, the probability of



**Figure 3.4:** MFPTs and reaching probability bounds for the dimerization model with  $\tau = \inf\{t \geq 0 \mid X_t \geq 25\} \wedge T$  and varying  $T$ . The results for SDP relaxations of orders 1 (light) to 6 (dark) are shown.

reaching the threshold before time  $T$ , are of particular interest. This is done by switching the optimization objective in (3.20) from the mass of the expected occupation measure  $\xi$ , to the mass of  $\nu_1$ . In terms of moments, the objective changes from  $z_{00}$  to  $y_{10}$ . The need for such a scenario often arises in the context of model checking, where one might be interested in the probability of a population exceeding a critical threshold. By varying the time-horizon, we are able to recover bounds on the cumulative density

$$F(t) = \Pr(X_s = H \mid s < t)$$

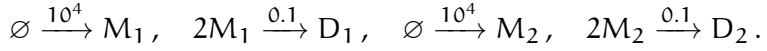
of the first passage time (Figure 3.4b).

Finally, we look at turn to the dimer species  $D$  that is synthesized by the combination of two monomers  $M$ . Here, we look at the time until the agents of type  $D$  exceed a threshold of five with a time-horizon  $T = 1$ . Note that we do not limit the number of  $M$  agents. Therefore the analyzed state-space is countably infinite. As in the previous two examples, we observe a roughly exponential decrease in interval size with increasing relaxation order  $r$  (cf. Figure 3.5 and Table 3.1).

### Parallel Dimerizations

As a second study, we consider a two-dimensional model by combining two independent dimerizations.

**Model 3** (Parallel Independent Dimerizations). *This model consists of two independent versions of [Model 1](#). The reactions are*



Initially,

$$0 = X_0^{(M_1)} = X_0^{(M_2)} = X_0^{(D_1)} = X_0^{(D_2)}.$$

As a [FPT](#) we consider the time at which either  $M_1$  or  $M_2$  surpasses a threshold of 200 or a time-horizon of  $T = 10$  is reached, i.e.

$$\tau = \inf\{t \geq 0 \mid X_t^{(M_1)} \geq 200\} \wedge \inf\{t \geq 0 \mid X_t^{(M_2)} \geq 200\} \wedge 10.$$

As before, we ignore the product species  $D_1$  and  $D_2$  since they do not influence  $\tau$ . The [SSA](#) (using  $n = 10,000$  runs) gives the estimate  $E(\tau) \approx 0.028378$  which is captured tightly by the [SDP](#) bounds (cf. [Table 3.1](#)). For higher relaxation orders  $r \geq 5$  numerical issues prevented the solution of the corresponding [SDPs](#).

### 3.7.3 Hybrid Models & Multi-Modality

The analysis of switching times is a particularly interesting case of [FPTs](#) that arises in many contexts. Often mode switching in such systems can be described a modulating Markov process whose switching rates may depend on the system state (e.g. the population sizes). In biological applications, mode switching often describes a change of the DNA state (Hasenauer et al., [2014](#); Stekel and Jenkins, [2008](#)) and the analysis of switching time distribution is of particular interest (Barzel and Biham, [2008](#); Spieler, Hahn, and Zhang, [2014](#)).

In the context of [MPMs](#), typically the state-space  $\mathcal{S} = \mathbb{N}^{\tilde{n}_s} \times \{0, 1\}^{\hat{n}_s}$ . This state is modeled by  $\hat{n}_s$  population variables with binary domains. Therefore, at each time point, the state of these



modulator variables is given by a set of Bernoulli random variables. When considering the moments of such a variable  $X$ , clearly

$$E(X^m) = E(X) = \Pr(X = 1), \quad \forall m \geq 1$$

We can use this fact two ways: We could use the same moment constraints as above and impose additional equality constraints on the moments matrices to ensure  $E(X^m) = E(X)$ ,  $m \geq 1$ . Alternatively, we can apply this simplification to the moment equation, which we choose here.

We apply a split of variables  $X_t$  into the high count part  $\tilde{X}_t$  and the binary part  $\hat{X}_t$  to the expectations in (3.2). Similarly, we split  $v_j$  and with a case distinction over the mode variable, we arrive at a similar result as in (Hasenauer et al., 2014):

$$\begin{aligned} & \frac{d}{dt} E(\tilde{X}_t^m 1_{=y}(\hat{X}_t)) \\ &= \sum_{j=1}^{n_R} E\left(\left((\tilde{X}_t + \tilde{v}_j)^m 1_{=y}(\hat{X}_t + \hat{v}_j) - \tilde{X}_t^m 1_{=y}(\hat{X}_t)\right) \alpha(X_t)\right) \\ &= \sum_{j=1}^{n_R} E\left(\left(\tilde{X}_t + \tilde{v}_j\right)^m \alpha_j(\tilde{X}_t, y - \hat{v}_j) 1_{=y-\hat{v}_j}(\hat{X}_t)\right) \\ & \quad - \sum_{j=1}^{n_R} E\left(\tilde{X}_t^m \alpha_j(\tilde{X}_t, y) 1_{=y}(\hat{X}_t)\right). \end{aligned} \tag{3.21}$$

Similarly to the general moment case, we can derive a constraint, by multiplying with a time-weighting factor and integrating.

For simplicity, here we assume  $\tilde{n}_S = \hat{n}_S = 1$ . Fixing appropriate sequences  $(c_i)_i$ ,  $(m_i)_i$ ,  $(k_i)_i$ , and  $(y_i)_i$  the constraint has the following form.

We fix  $0^0 = 1$ .

$$\begin{aligned} & \sum_{y \in \{0,1\}} H^m E(\tau^k; \hat{X}_\tau = y, \tau < T) \\ & \quad + T^k E(\tilde{X}_T^m; \hat{X}_T = y, \tau = T) \\ &= 0^k \tilde{x}_0^m 1_{=y}(\hat{x}_0) + \sum_i c_i E\left(\int_0^\tau t^{k_i} \tilde{X}_t^{m_i} dt; \hat{X}_t = y_i\right) \end{aligned} \tag{3.22}$$

This way we can decompose the moment matrices such that for each mode  $y \in \{0, 1\}$ , we have moment matrices composed of the

**Table 3.1:** MFPT bounds on [Model 1](#) for  $X_t \geq 5$  and  $T = 1$ , [Model 3](#) for  $X_t^{(M_1)} \geq 200$  or  $X_t^{(M_2)} \geq 200$  and  $T = 10$ , and [Model 4](#) for  $X_t^{(P)} \geq 5$  and  $T = 10$ .

		relaxation order $r$				
		1	2	3	4	5
<a href="#">Model 1</a>	lower	0.0909	0.2661	0.2845	0.2867	0.2871
	upper	1.0000	0.3068	0.2932	0.2886	0.2875
<a href="#">Model 3</a>	lower	0.0010	0.0250	0.0275	0.0280	0.0280
	upper	10.0000	0.0575	0.0323	0.0299	0.0290
<a href="#">Model 4</a>	lower	4.0000	6.0028	6.2207	6.3377	6.3772
	upper	10.7179	6.4619	6.4079	6.4004	6.3835

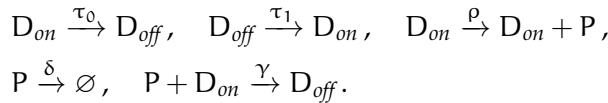
respective partial moments. To this end, let  $z_m^{(y)}$  be the partial moment w.r.t.  $\hat{X} = y$ . The moment constraint over the partial moments has a linear structure:

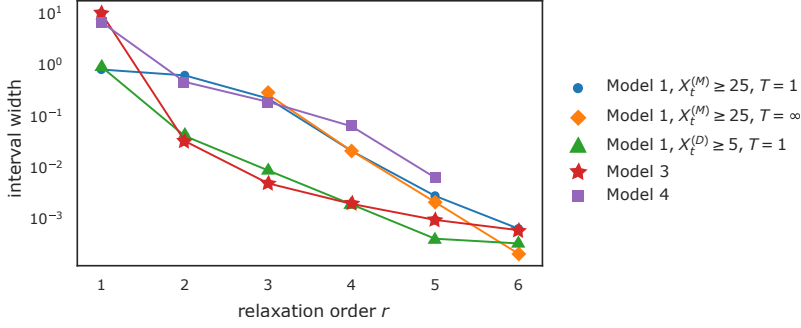
$$0 = y_{1k}H^m - y_{2m}T^k - 0^k x_0^m + \sum_i c_i z_{k_i m_i}^{(y_i)}. \quad (3.23)$$

#### Gene Expression with Negative Feedback

As an instance of a multi-modal system, we consider a simple gene expression with self-regulating negative feedback which is a common pattern in many genetic circuits (Stekel and Jenkins, 2008).

**Model 4** (Negative Self-Regulated Gene Expression). *This model consists of a gene state that is either on or off, i.e.  $X_t^{D_{on}} + X_t^{D_{off}} = 1$ ,  $\forall t \geq 0$ . Therefore the system has two modes.*





**Figure 3.5:** The interval width, i.e. the difference between upper and lower bound, for different case studies and targeted first passage times against the order  $r$  of the SDP relaxation.

The model parameters are  $(\tau_0, \tau_1, \rho, \delta, \gamma) = (10, 10, 2, 0.1, 0.1)$  and  $X_0^{(D_{off})} = 1, X_0^{(P)} = 0$  a.s.

As a first passage time we consider

$$\tau = \inf\{t \geq 0 \mid X_t^{(P)} \geq 5\} \wedge 20.$$

The results are summarized in Table 3.1. The estimated MFPT based on 100,000 SSA samples is  $E(\tau) \approx 6.37795 \pm 0.02847$  at 99% confidence level. Note that our SDP solution for  $r = 5$  yields tighter moment bounds than the statistical estimation.

In Figure 3.5 we summarize our results about the decrease of the interval widths for increasing relaxation order  $r$  by plotting them on a log-scale. We see an approximately exponential decrease with increasing  $r$ . The SDPs above were all solved within at most a few seconds.

### 3.8 LINEAR HAUSDORFF CONSTRAINTS

Before closing this chapter, we will briefly discuss the Hausdorff moment constraints. These constraints offer linear constraints on moments of bounded measures. As such they can – in principle – be used as both, a replacement and an addition to the semi-definite constraints discussed above.

The *Hausdorff moment problem* is the question whether an infinite sequence  $(m_0, m_1, \dots)$  is a sequence of moments

$$m_k = \int_0^1 x^k d\mu(x), \quad \forall k \in \mathbb{N}$$

for some Borel-measure supported on the unit interval  $[0, 1]$ .

See also Feller (1971) for additional details.

Hausdorff (1921) came up with a necessary condition that

$$\int_0^1 x^\ell (1-x)^k d\mu(x) \geq 0, \quad \forall \ell, k \in \mathbb{N}. \quad (3.24)$$

The validity of this condition is easy to see since

$$x^\ell (1-x)^k \geq 0, \quad \forall x \in [0, 1]. \quad (3.25)$$

Expanding this term yields a polynomial, which – by integration – provides a linear constraint on the moments. This moment condition has been used by Helmes, Röhl, and Stockbridge (2001) to bound MFPTs in a variety of stochastic processes. In Helmes and Röhl (2008) the authors provide an extension taking into account the geometry of the Hausdorff polytope. With this extension the method was competitive with the SDP approach on their selection of case studies. Since these conditions are linear in the moments they can solve LPs instead of the semi-definite programs shown below.

**EXAMPLE** Let  $\ell = 2, k = 2$ . Then the (3.24) becomes

$$\int_0^1 x(1-x)^2 d\mu(x) = m_2 - 2m_3 + m_4 \geq 0$$

constraining the moments on  $[0, 1]$ .  $\diamond$

Since the Hausdorff moment problem is defined on  $[0, 1]$ , we need to adjust accordingly. Let the interval be  $[0, H]$ ,  $H > 0$  then we can simply change the variables. That means considering  $y = x/H$  instead of  $x$ . Equivalently, we can apply this rescaling in (3.24) such that

$$\int_0^H x^\ell (H-x)^k d\mu(x) \geq 0, \quad \forall \ell, k \in \mathbb{N} \quad (3.26)$$

remains a valid constraint on  $[0, H]$ .

Generalizing (3.24) to multiple dimensions can be done by simply multiplying terms for each dimension in a similar manner. For  $n$  dimensions

$$\int_{[0,1]^n} \prod_{i=0}^n x_i^{\ell_i} (1 - x_i)^{k_i} d\mu(x) \geq 0$$

Using multi-index notation

$$\int_{[0,1]^n} x^\ell (1 - x)^k d\mu(x) \geq 0$$

for all  $\ell, k \in \mathbb{N}^n$ . We treat the time horizon  $T$  exactly the same as the other bounds  $H_i$  here. With arbitrary positive upper bounds  $H = (H_1, H_2, \dots, H_n)^T$  this becomes

$$\int_{\times_{i=1}^n [0, H_i]} x^\ell (H - x)^k d\mu(x) \geq 0 \quad (3.27)$$

This equation can simply be expanded using the multi-binomial theorem. This is just a vector version of the regular binomial theorem. That is, given  $n$ -dimensional vectors  $k$ ,  $x$ , and  $y$

$$(y + x)^k = \sum_{j \leq k} \binom{k}{j} y^j x^{k-j}.$$

Thus, (3.27) becomes the linear moment constraint

$$\sum_{j \leq k} \binom{k}{j} H^j (-1)^{k-j} \int x^{k-j+\ell} d\mu(x) \geq 0. \quad (3.28)$$

We can interchange the integrals since all measures have finite support and mass.

### 3.8.1 A Linear Program

Combining the model constraints (3.13) and the linear Hausdorff-type constraints (3.28) for both expected occupation and the exit location measure yields a LP.

$$\begin{aligned}
 & \min/\max \quad z'_{00} \\
 & \text{such that} \quad \text{for all measures and bounds} \\
 & \quad (3.28) \text{ holds } \forall m, k, \\
 & \quad 0 = y'_{1k} H^m - y'_{2m} T^k - 0^k x_0^m \quad (3.29) \\
 & \quad \quad + \sum_i c_i z'_{k_i m_i}, \quad \forall m, k, \\
 & \quad 1 = y'_{10} + y'_{20}.
 \end{aligned}$$

Using a framework such as CVXPY, this convex program can easily be coded and solved using various state-of-the-art solvers.

### 3.8.2 Case Studies

The LP in (3.29) is encoded using CVXPY, as well. The resulting LPs are solved using the Gurobi (Gurobi Optimization, LLC, 2021) solver. For a fixed moment order  $k$  the following constraints are used:

1. model constraints corresponding to moments  $E(X^\ell \tau^m)$  where  $\ell, \tau < k$  and
2. Hausdorff constraints (3.28) for all moments smaller and equal  $k$  in each dimension.

As a simple example we consider a Poisson process with rate 100, i.e.  $\emptyset \xrightarrow{100} X$ . We are interested in the threshold event of reaching 25 before a time-horizon of  $T = 10$ , i.e.

$$\tau = \inf\{t \geq 0 \mid X_t \geq 25\} \wedge 10.$$

The results are given in Table 3.2. We can see, that the interval quickly converges to the actual value of 0.25. Even for relatively small orders (2 or 3) the intervals are very tight. Furthermore

order	2	3	4	5
lower	0.2438	0.2500	0.2500	0.2500
upper	0.2500	0.2500	0.2500	0.2500
width	$6.1 \times 10^{-3}$	$4.8 \times 10^{-5}$	$8.0 \times 10^{-7}$	$8.9 \times 10^{-10}$
# constr.	140	299	792	1395

**Table 3.2:** MFPT bounds for a Poisson process and  $\tau = \inf\{t \geq 0 \mid X_t \geq 25\} \wedge 10$  using LP.

the interval width decreases by orders of magnitudes with each increase in relaxation order.

As a second example we consider a birth-death model (Model 2) with rates  $\mu = 100$  and  $\gamma = 0.1$ . We are interested in the threshold event of reaching 25 before a time-horizon of  $T = 1$ , i.e.

$$\tau = \inf\{t \geq 0 \mid X_t \geq 25\} \wedge 1.$$

The results are given in Table 3.3. Here, too, the interval converges quickly, albeit not as fast as in the first example.

order	2	3	4	5
lower	0.2000	0.2742	0.2836	0.2850
upper	0.3333	0.2895	0.2857	0.2852
width	$1.3 \times 10^{-1}$	$1.5 \times 10^{-2}$	$2.1 \times 10^{-3}$	$1.6 \times 10^{-4}$
# constraints	140	299	792	1395

**Table 3.3:** MFPT bounds for a birth-death process and  $\tau = \inf\{t \geq 0 \mid X_t \geq 25\} \wedge 1$  using LP.

As a third example we consider a dimerization model (Model 1) the threshold event

$$\tau = \inf\{t \geq 0 \mid X_t \geq 25\} \wedge 1.$$

The results are given in Table 3.4. Convergence for this case study was slightly slower than in the examples above. For order 6 and larger, the solver showed signs of numerical instability.

order	2	3	4	5
lower	0.2000	0.2736	0.3884	0.4935
upper	1.0000	1.0000	1.0000	0.5777
width	$8.00 \times 10^{-1}$	$7.3 \times 10^{-1}$	$6.1 \times 10^{-1}$	$8.4 \times 10^{-2}$
# constraints	140	299	792	1395

**Table 3.4:** MFPT bounds for [Model 1](#) and  $\tau = \inf\{t \geq 0 \mid X_t \geq 25\} \wedge 1$  using LP.

The inclusion of the linear constraints into the [SDP](#) did not yield any improvement of the bounds. This, of course, may be specific to this problem setting and the inclusion of linear inequalities may lead to improvements on other models and stopping times.

In the above experiment the moment order was not increased above 5. For some of these examples this is possible, but high moment orders tend to cause numerical issues. The reasons for these problems are analogue to the [SDP](#) formulation: Moments tend to get exponentially larger with order. Therefore the constraints couple values varying by multiple orders of magnitude. Similar stiffness originates from the Hausdorff constraints (3.28) where terms such as  $H^j$  may cause coupling of vastly different magnitudes of coefficients. It is possible, that some sort of scaling approach similar to the one used before might mitigate these problems at least partially.

*Note, that the above experiments were performed without any scaling.*

Overall the [LP](#) formulation has the advantage of being easier to formulate and implement than the [SDP](#) approach. However, as it is based on the Hausdorff moment problem, the domain needs to be finite. Thus, infinite state-spaces cannot directly be analyzed. In case of infinite state-spaces truncations such as the one presented in Kim et al. (2020) could be a viable option.

### 3.9 CONCLUSION

Numerical methods to compute reachability probabilities and first passage times for continuous-time Markov chains that are based on an exhaustive exploration of the state-space are ex-



act up to numerical precision. Such methods, however, do not scale and cannot be efficiently applied to models with large or infinite state-spaces, an issue exacerbated in population models. Moment-based methods offer an alternative analysis approach for [MPMs](#), which scales with the number of different populations in the system but are approximations with little or no control of the error. In this work, we bridge this gap by proposing a rigorous approach to derive bounds on first passage times and reachability probabilities, leveraging a semi-definite programming formulation based on appropriate moment constraints.

The method we propose is shown to be accurate in several examples. It does, however, suffer, like all moment-based methods, from numerical instabilities in the [SDP](#) solver, caused by the fact that moments typically span several orders of magnitude. We proposed a scaling of moments to mitigate this effect. However, the scaling only addresses the moment matrices but not the linear constraints which still contain values with varying orders of magnitudes. Therefore, we plan as future work to investigate an appropriate scaling for the linear constraints or to redefine the moment constraints (e.g. using an exponential time weighting (Dowdy and Barton, [2018b](#))). Based on this investigation, we expect to make this approach applicable to more problems including, for example, the computation of bounds of rare event probabilities. We also expect that the development of more sophisticated scaling techniques will improve approximate moment-based methods.

Furthermore, moment-based analysis approaches have shown to be successful in a wide range of applications such as optimal control problems or the estimation of densities (Lasserre, [2010](#)). We expect that our proposed ideas can be adapted to a wider range of stochastic models such as stochastic hybrid systems, exhibiting partly deterministic dynamics.



# 4

---

## LINEAR CONTROL VARIATES FOR MONTE CARLO ESTIMATION

---

Analysis approaches based on sampling, such as the stochastic simulation algorithm (SSA) (Gillespie, 1977), can be applied independent of the size of the model's state-space. However, statistical approaches are costly since a large number of simulation runs is necessary to reduce the statistical inaccuracy of estimators. This problem is particularly severe if reactions occur on multiple time scales or if the event of interest is rare. A particularly popular technique to speed up simulations is  $\tau$ -leaping which applies multiple reactions in one step of the simulation. However, such multi-step simulations rely on certain assumptions about the number of reactions in a certain time interval. These assumptions are typically only approximately fulfilled and therefore introduce approximation errors on top of the statistical uncertainty of the considered point estimators.  $\hookrightarrow$  page 21

Moment-based techniques offer a fast approximation of the statistical moments of the model. The exact moment dynamics can be expressed as an infinite-dimensional system of ODEs, which cannot be directly integrated for a transient analysis. Hence, ad-hoc approximations need to be introduced, expressing higher order moments as functions of lower-order ones (Ale, Kirk, and Stumpf, 2013; Engblom, 2006). However, moment-based approaches rely on assumptions about the dynamics that are often not even approximately fulfilled and may lead to high approximation errors. Recently, equations expressing the mom-

ent dynamics have also been used as constraints for parameter estimation (Backenköhler, Bortolussi, and Wolf, 2018) and for computing moment bounds using semi-definite programming (Dowdy and Barton, 2018b; Ghusinga et al., 2017).

In this work, we propose a combination of such moment constraints with the SSA approach. Specifically, we interpret these constraints as random variables that are correlated with the estimators of interest usually given as functions of chemical population variables. These constraints can be used as (linear) control variates (CVs) in order to improve the final estimate and reduce its variance (Lavenberg, Moeller, and Welch, 1982; Szechtman, 2003). The method is easy on an intuitive level: If a control variate is positively correlated with the function to be estimated then we can use the estimate of the variate to adjust the target estimate.

The incorporation of control variates into the SSA introduces additional simulation costs for the calculation of the constraint values. These values are integrals over time, which we accumulate based on the piece-wise constant trajectories. This introduces a trade-off between the variance reduction that is achieved by using control variates versus the increased simulation cost. This trade-off is expressed as the product of the variance reduction ratio and the cost increase ratio.

For a good trade-off, it is crucial to find an appropriate set of control variates. Here we propose a class of constraints which is parameterized by a moment vector and a weighting parameter, resulting in infinitely many choices. We present an algorithm that samples from the set of all constraints and proceeds to remove constraints that are either only weakly correlated with the target function or are redundant in combination with other constraints.

In a case study, we explore different variants of this algorithm both in terms of generating the initial constraint set and of removing weak or redundant constraints. We find that the algorithm's efficiency is superior to a standard estimation procedure using stochastic simulation alone in almost all cases.

Although in this work we focus on estimating first order moments at fixed time points, the proposed approach can in principle deal with any property that can be expressed in terms of

expected values such as probabilities of complex path properties. Another advantage of our technique is that an increased efficiency is achieved without the price of an additional approximation error as it is the case for methods based on moment approximations or multi-step simulations.

#### 4.1 RELATED WORK

**STATE-SPACE TRUNCATIONS** If the state-space is finite and small enough one can deal with the underlying Markov chain directly. But there are also cases where the transient distribution has an infinitely large support and one can still deal with explicit state probabilities. To this end, one can fix a finite state-space, that should contain most of the probability (Munsky and Khammash, 2006). Refinements of the method work dynamically and adjust the state-space according to the transient distributions (Andreychenko et al., 2011; Henzinger, Mateescu, and Wolf, 2009; Mateescu et al., 2010).

**MOMENT APPROXIMATIONS** On the other end of the spectrum there are mean-field approximations, which model the mean densities faithfully in the system size limit (Bortolussi et al., 2013). In between there are techniques such as moment closure (Singh and Hespanha, 2006), that not only consider the mean, but also the variance and other higher order moments. These methods depend on ad-hoc approximations of higher order moments to close the ODE system given by the moment equations. Yet another class of methods approximate molecular counts continuously and approximate the dynamics in such a continuous space, e.g. the system size expansion (Kampen, 1992) and the chemical Langevin equation (Gillespie, 2000).

While the moment closure method uses ad-hoc approximations for high order moments to facilitate numerical integration, they can be avoided in some contexts. For the equilibrium distribution, for example, the time-derivative of all moments is equal to zero. This directly yields constraints that have been used for parameter estimation at steady-state (Backenköhler, Bortolussi, and Wolf,

2018) and bounding moments of the equilibrium distribution using semi-definite programming (Ghusinga, Lamperski, and Singh, 2018; Ghusinga et al., 2017; Kuntz, Juan And Thomas, Philipp And stan and Barahona, 2019). The latter technique of bounding moments has been successfully adapted in the context of transient analysis (Dowdy and Barton, 2018b; Sakurai and Hori, 2017, 2019). We adapt the constraints proposed in these works to improve statistical estimations via stochastic simulation (cf. Section 4.2).

**MONTE CARLO SIMULATION** While the above techniques give a deterministic output, stochastic simulation generates single executions of the stochastic process (Gillespie, 1977). This necessitates accumulating large numbers of simulation runs to estimate quantities. This adds a significant computational burden. Consequently, considerable effort has been directed at lowering this cost. A prominent technique is  $\tau$ -leaping (Gillespie, 2001), which in one step performs multiple instead of only a single reaction. Another approach is to find approximations that are specific to the problem at hand, such as approximations based on time-scale separations (Bortolussi, Milios, and Sanguinetti, 2015; Cao, Gillespie, and Petzold, 2005).

Recently, multilevel Monte Carlo methods have been applied in to time-inhomogenous MPM (Anderson and Yuan, 2018). In this techniques estimates are combined using estimates of different approximation levels.

The most prominent application of a variance reduction technique in the context of stochastic reaction networks is importance sampling (Kuwahara and Mura, 2008). This technique relies on an alteration of the process and then weighting samples using the likelihood-ratio between the original and the altered process.

We refer the reader to Beentjes (2021) for a recent survey of variance reduction methods applied to MPMs.

## 4.2 MOMENT CONSTRAINTS

The time-evolution of  $E(f(X_t))$  for some function  $f$  is given by (2.21). The integration of (2.21) with such functions  $f$  is well-known in the context of moment approximations of MPM. For most models the arising ODE system is infinitely large, because the time-derivative of low order moments usually depends on the values of higher order moments. To close this system, *moment closures*, i.e. ad-hoc approximations of higher order moments are applied (Schnoerr, Sanguinetti, and Grima, 2015). The main drawback of this kind of analysis is that it is not known whether the chosen closure gives an accurate approximation for the case at hand. Here, such approximations are not necessary, since we will apply the moment dynamics in the context of stochastic sampling instead of trying to integrate (2.21). ↪ page 23

Apart from integration strategies, setting (2.21) to zero has been used as a constraint for parameter estimation at steady-state (Backenköhler, Bortolussi, and Wolf, 2018) and bounding moments at steady-state (Dowdy and Barton, 2018a; Ghusinga et al., 2017; Kuntz, Juan And Thomas, Philipp And stan and Barahona, 2019). The extension of the latter has recently lead to the adaption of these constraints to a transient setting (Dowdy and Barton, 2018b; Sakurai and Hori, 2019). These two transient constraint variants are analogously derived by multiplying (2.21) by a time-dependent, differentiable weighting function  $w(t)$  and integrating

As in the previous chapter, multiplication with  $w(t)$  and integration on  $[t_0, T]$  yields (Dowdy and Barton, 2018b; Sakurai and Hori, 2019) ↪ page 45

$$\begin{aligned} & w(T)E(f(X_T)) - w(t_0)E(f(X_{t_0})) - \int_{t_0}^T \frac{dw(t)}{dt} E(f(X_t)) dt \\ &= \sum_{j=1}^{n_R} \int_{t_0}^T w(t) E((f(X_t + v_j) - f(X_t)) \alpha_j(X_t)) dt \quad (4.1) \end{aligned}$$

While many choices of  $f$  are possible, for this work we will restrict ourselves to monomial functions  $f(x) = x^m$ ,  $m \in \mathbb{N}^{n_s}$  i.e. the *non-central* or *raw moments* of the process. In the context

In [Chapter 3](#) we chose the monomial, such that [\(4.1\)](#) would admit the interpretation of temporal moments. Here the exponential is a good choice, because it can model both increasing and decreasing weighting.

Contrary to the claim in Beentjes ([2021](#), Ch. 2.4.2), the approach is not restricted to mass-action kinetics.

of computing moment bounds via semi-definite programming the polynomial  $w(t) = t^s$  (Sakurai and Hori, [2019](#)) and the exponential  $w(t) = e^{\lambda(T-t)}$  (Dowdy and Barton, [2018b](#)) have been proposed. While both choices proved to be effective in different case studies, relying solely on the latter choice, i.e.

$$w(t) = e^{\lambda(T-t)}$$

was sufficient.

For this chapter, we assume that propensity functions  $\alpha_i$  are polynomial. This restriction is only due to the simplification of only having to record one value for each tuple of moment vector  $\mathbf{m}$  and weighting coefficient  $\lambda$ . Given more complex propensity functions, we would need to accumulate other terms during simulation accordingly. By expanding the rate functions and  $f$  in [\(4.1\)](#) and substituting the exponential weight function we can re-write [\(4.1\)](#) as

$$0 = \mathbb{E}(f(X_T)) - e^{\lambda T} \mathbb{E}(f(X_{t_0})) + \sum_k c_k \int_{t_0}^T e^{\lambda(T-t)} \mathbb{E}(X_t^{m_k}) dt \quad (4.2)$$

with coefficients  $c_k$  and vectors  $m_k$  defined accordingly. Assuming the moments remain finite on  $[0, T]$ , we can define the random variable

$$Z = f(X_T) - e^{\lambda T} f(X_{t_0}) + \sum_k c_k \int_{t_0}^T e^{\lambda(T-t)} X_t^{m_k} dt \quad (4.3)$$

with  $\mathbb{E}(Z) = 0$ .

**EXAMPLE** For [Model 2](#) the moment equation for  $f(x) = x$  becomes

$$\frac{d}{dt} \mathbb{E}(X_t) = \gamma - \delta \mathbb{E}(X_t) .$$

The corresponding constraint [\(4.2\)](#) with  $\lambda = 0$  gives

$$0 = \mathbb{E}(X_T) - \mathbb{E}(X_0) - \gamma T + \delta \int_0^T \mathbb{E}(X_t) dt .$$



In this instance the constraint leads to an explicit function of the moment over time. If  $X_0 = 0$  w.p. 1, then (4.2) becomes

$$E(X_T) = \frac{\gamma}{\delta} (1 - e^{-\delta T}) \quad (4.4)$$

when choosing  $\lambda = -\delta$ .  $\diamond$

In general, a realization of  $Z$  depends on the whole trajectory  $\tau = x_0 t_1 x_1 t_2 \dots t_n x_n$  over  $[t_0, T]$ . Thus, for the integral terms in (4.3) we have to compute sums

$$\frac{1}{\lambda} \sum_{i=1}^n \left( e^{\lambda(T-t_{i+1})} - e^{\lambda(T-t_i)} \right) x_i^{m_k} \quad (4.5)$$

over a given trajectory. This accumulation is best done during the simulation to avoid storing the whole trajectory. Still, the cost of a simulation run increases. For the method to be efficient, the variance reduction (Section 4.3) needs to overcompensate for this increased cost of a simulation run.

### 4.3 CONTROL VARIATES

Now, we are interested in the estimation of some quantity  $E(V)$  by stochastic simulation. Let  $V_1, \dots, V_n$  be independent samples of  $V$ . Then the sample mean

$$\hat{V}_n = \frac{1}{n} \sum_{i=1}^n V_k$$

is an estimate of  $E(V)$ . By the central limit theorem

$$\sqrt{n} \hat{V}_n \xrightarrow{d} N(E(V), \sigma_V^2).$$

Now suppose, we know of a random variable  $Z$  with  $0 = E(Z)$ . The variable  $Z$  is called a *control variate* (CV). If a control variate  $Z$  is correlated with  $V$ , we can use it to reduce the variance of  $\hat{V}_n$  (Glasserman and Yu, 2005; Nelson, 1990; Szechtman, 2003; Wilson, 1984). For example, consider we are running a set of simulations and consider a single constraint. If the estimated value of this constraint is larger than zero and we estimate a

---

**Algorithm 2:** SSA<sub>CV</sub>: SSA with accumulator updates
 

---

**input** :  $\pi_0, T, P, n$   
**output**: trajectory  $\tau$

```

1 initialize accumulator map A for P;
2 for  $i = 1, \dots, n$  do
3    $\tau \leftarrow$  empty list,  $s \leftarrow$  sample from  $\pi_0$ ,  $t \leftarrow 0$ ;
4   while  $t < T$  do
5      $\tau \leftarrow \text{append}(\tau, (s, t))$ ;
6      $k \leftarrow$  sample reaction  $i$  with probability  $\propto \alpha_i(s)$ ;
7      $\delta \sim \text{Exp}(\sum_i \alpha_i(s))$ ;
8     for  $(m, \lambda) \in \text{keys}(A)$  do
9        $A[(m, \lambda)] \leftarrow A[(m, \lambda)]$ 
10         $+ \frac{1}{\lambda} (e^{\lambda(t+\delta)} - e^{\lambda t}) \chi^m$ ;
11      $s \leftarrow s + v_k$ ;
12      $t \leftarrow t + \delta$ ;
13   update means  $\hat{V}$ ,  $\hat{Z}$  and covariances  $\hat{\Sigma}$  using A;
14   for  $(m, \lambda) \in \text{keys}(A)$  do
15      $A[(m, \lambda)] \leftarrow 0$ 
16 return  $(\hat{\Sigma}, \hat{V}, \hat{Z})$ ;

```

---

positive correlation between the constraint  $Z$  and  $V$ , we would, intuitively, like to decrease our estimate  $\hat{V}_n$  accordingly. This results in an estimation of the mean of the random variable

$$Y_\beta = V - \beta Z$$

instead of  $V$ . The variance

$$\sigma_{Y_\beta}^2 = \sigma_V^2 - 2\beta \text{Cov}(V, Z) + \beta^2 \sigma_Z^2.$$

The optimal choice  $\beta$  can be computed by considering the minimum of  $\sigma_{Y_\beta}^2$ . Then

$$\beta^* = \text{Cov}(V, Z) / \sigma_Z^2.$$

Therefore

$$\sigma_{Y_{\beta^*}}^2 = \sigma_Z^2 (1 - \rho_{VZ}^2),$$

where  $\rho_{VZ}$  is the correlation of  $Z$  and  $V$ .

If multiple control variates are available, we can proceed in a similar fashion. Now, let  $Z$  denote a vector of  $d$  control variates and let

$$\Sigma = \begin{bmatrix} \Sigma_Z & \Sigma_{VZ} \\ \Sigma_{ZV} & \sigma_V^2 \end{bmatrix}$$

be the covariance matrix of  $(Z, V)$ . As above, we estimate the mean of

$$Y_\beta = V - \beta^\top Z.$$

The ideal choice of  $\beta$  is the result of an ordinary least squares regression between  $V$  and  $Z_i$ ,  $i = 1, \dots, n$ . Specifically,

$$\beta^* = \Sigma_Z^{-1} \Sigma_{ZV}.$$

Then, asymptotically the variance of the estimator

$$\hat{Y}_{\beta^*} = \hat{V} - \beta^{*\top} \hat{Z} \tag{4.6}$$

is (Szechtman, 2003),

$$\sigma_{\hat{Y}_{\beta^*}}^2 = (1 - R_{ZV}^2) \sigma_V^2, \quad (4.7)$$

where

$$R_{ZV}^2 = \Sigma_{ZV} \Sigma_Z^{-1} \Sigma_{ZV} / \sigma_V^2.$$

In practice, however,  $\beta^*$  is unknown and needs to be replaced by an estimate  $\hat{\beta}$ . Then the estimator

$$\hat{Y}_{\hat{\beta}} = \hat{V} - \hat{\beta}^\top \hat{Z}. \quad (4.8)$$

This leads to an increase in the estimator's variance. Under the assumption of  $Z$  and  $V$  having a multivariate normal distribution (Cheng, 1978; Lavenberg, Moeller, and Welch, 1982), the variance of the estimator is

$$\sigma_{\hat{Y}_{\hat{\beta}}}^2 = \frac{n-2}{n-2-d} (1 - R_{ZV}^2) \sigma_V^2. \quad (4.9)$$

Clearly, a control variate is “good” if it is highly correlated with  $V$ . The constraint in (4.4) is an example of the extreme case. When we use this constraint as a control variate for the estimation of the mean at some time point  $t$ , it has a correlation of  $\pm 1$  since it describes the mean at that time precisely. Therefore the variance is reduced to zero. We thus aim to pick control variates that are highly correlated with  $V$ .

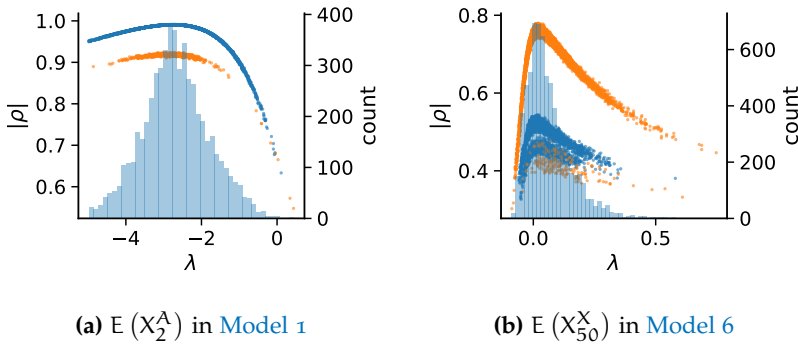
**EXAMPLE** Consider, for example, the above case of the birth-death process. If we choose (4.4) as a constraint, it would always yield the exact difference of the exact mean to the sample mean and therefore have a perfect correlation. Clearly,  $\hat{\beta}$  reduces to 1 and  $\hat{Y}_1 = E(X_t)$ .  $\diamond$

#### 4.4 MOMENT-BASED VARIANCE REDUCTION

We propose an adaptive estimation algorithm (Algorithm 3) that starts out with an initial set of control variates and periodically removes potentially inefficient variates. The “accumulator set”  $A$

represents the time-integral terms (4.5). The size of  $A$  has the most significant impact on the overall speed of the algorithm since it represents the only factor incurring a direct cost increase in the SSA itself (line 4).

The algorithm consists of a main loop which performs  $n$  simulation runs (line 3). Between each run the mean and covariance estimates of  $[Z, V]$  are updated (line 5). Every  $d < n$  iterations, the control variates are checked for *efficiency* and *redundancy* (lines 6–9).



**Figure 4.1:** The absolute correlation of different constraints to  $V$  arising from different choices of  $\lambda$ . The blue dots represent constraints based on first order moments, while the orange refers to control variates derived from second order moments. In both cases 10,000 samples were used with 30 initial samples for  $\lambda$  from  $N(0, 1)$  and  $k_{\min} = 2$ . A quadratic decision bound was used for the redundancy removal. Furthermore, a histogram of control variates selected by Algorithm 3 is given. In (a)  $E(X_2^A)$  in the dimerization model was estimated. In (b)  $E(X_{50}^X)$  in the distributive modification model was estimated.

Checking both conditions is based on the correlation  $\rho_{ij}$  between the  $i$ -th and  $j$ -th control variate and the correlation  $\rho_{iv}$

of a control variate  $i$  to  $V$ . These are elements of the correlation matrix

$$C = \begin{bmatrix} 1 & \dots & \rho_{1k} & \rho_{1v} \\ \vdots & \ddots & \vdots & \vdots \\ \rho_{k1} & \dots & 1 & \rho_{kv} \\ \rho_{v1} & \dots & \rho_{vk} & 1 \end{bmatrix}.$$

The first condition is a simple lower threshold  $\rho_{\min}$  for a correlation  $\rho_{iv}$ . This condition aims to remove those variates from the control variate set that are only weakly correlated to  $V$  (line 7). The rationale is that, if variate  $i$  has a low correlation with the variable of interest  $V$ , its computation may not be worth the costs. Here, we propose to set  $\rho_{\min}$  heuristically as

$$\rho_{\min} = \min \left( 0.1, \frac{\max_i \rho_{iv}}{k_{\min}} \right),$$

where  $k_{\min} > 1$  is an algorithm parameter.

The second condition aims to remove redundant conditions. This is not only beneficial for the efficiency of the estimator, but also necessary for the matrix inversion (4.7) because perfectly and highly correlated constraints will make the covariance matrix estimate  $\hat{\Sigma}_Z$  (quasi-) singular. For all considered criteria we iterate over all tuples  $(i, j) \in \{1, \dots, k\}^2$ ,  $i \neq j$ , removing the weaker of the two, i.e.  $\arg \min_{k \in \{i, j\}} \rho_{kv}$ , if the two control variates are considered redundant (line 8).

There are many ways to define such a redundancy criterion. Here, we focus on criteria that are defined in terms of the average correlation

$$\bar{\rho}_{ij} = (\rho_{iv} + \rho_{jv})/2.$$

For two variates  $i$  and  $j$  we then check if their mutual correlation  $\rho_{ij}$  exceeds a some function  $\phi$  of  $\bar{\rho}_{ij}$ , i.e. we check the inequality

$$\phi(\bar{\rho}_{ij}) \leq \rho_{ij}.$$

If this inequality holds, constraint  $\arg \min_{k \in \{i, j\}} \rho_{kv}$  is removed. Naturally, there are many possible choices for the above decision boundary  $\phi$  (cf. Figure 4.2).

The simplest choice is to ignore  $\bar{\rho}_{ij}$  and just fix a constant close to 1 as a threshold, e.g.

$$\phi_c(\bar{\rho}_{ij}) = 0.99.$$

While this often leads to the strongest variance reduction and avoids numerical issues in the control variate computation, it turns out that the computational overhead is not as well-compensated as by other choices of  $\phi$  (see [Section 4.5](#)).

Another option is to fix a simple linear function, i.e.

$$\phi_\ell(\bar{\rho}_{ij}) = \bar{\rho}_{ij}.$$

For this choice the intuition is, that one of two constraints is removed if their mutual correlation exceeds their average correlation with  $V$ .

Here, we also assess two quadratic choices for  $\phi$ . The first choice of

$$\phi_q(\bar{\rho}) = 1 - (1 - \bar{\rho})^2$$

is more tolerant than the linear function and more strict than a threshold function, except for highly correlated control variates. Another variant of  $\phi$  is given by including the lower bound  $\rho_{\min}$  and scaling the quadratic function accordingly:

$$\phi_{sq}(\bar{\rho}) = 1 - ((1 - \bar{\rho}) / (1 - \rho_{\min}))^2.$$

The different choices of  $\phi$  considered here are plotted in [Figure 4.2](#).

Now, we discuss the choice of the initial control variates. We identify control variate  $k$  by a tuple  $(m_k, \lambda_k)$  of a moment vector  $m_k$  and a time-weighting parameter  $\lambda_k$ . That is, we use  $w(t) = e^{\lambda_k(T-t)}$  and  $f(x) = x^{m_k}$  in [\(4.1\)](#). For a given set of parameters  $L$ , we use all moments up to some fixed order  $n_{\max}$  ([line 2](#)). The ideal set of parameters  $L$  is generally not known. For certain choices the correlation of the control variates and the variable of interest is higher than for others. To illustrate this, consider the above example of the birth-death process. Choosing  $\lambda = -\delta$  leads to a control variate that has a correlation of  $\pm 1$  with  $V$ . Therefore, the ideal choice of initial values for would be  $L = \{-\delta\}$ . This, however, is generally not known. Therefore, we sample a set of  $\lambda$ 's from some fixed distribution  $\pi_\lambda$  ([line 1](#)).

---

**Algorithm 3:** Estimate the mean of species  $i$  at time  $T$ 


---

**input** :  $n, d, n_{\max}, n_{\lambda}, k_{\min}$   
**output**: an estimate using linear control variates  
 $1 \quad L \leftarrow \{\lambda_i \sim \pi_{\lambda} \mid 1 \leq i < n_{\lambda}\} \cup \{0\};$   
 $2 \quad P \leftarrow \{(m, \lambda) \mid 1 \leq |m| \leq n_{\max}, \lambda \in L\};$   
 $3 \quad \textbf{for } i = 1, \dots, \lfloor n/d \rfloor \textbf{ do}$   
 $4 \quad \quad (\hat{\Sigma}_i, \hat{V}_i, \hat{Z}_i) \leftarrow \text{SSA}_{\text{CV}}(\pi_0, T, P, d);$   
 $5 \quad \quad \text{update } \hat{\Sigma}, \hat{V}, \text{ and } \hat{Z};$   
 $6 \quad \quad \rho_{\min} \leftarrow \min(0.1, \max_i \rho_{iv}/k_{\min});$   
 $7 \quad \quad P \leftarrow P \setminus \{(m_k, \lambda_k) \mid \rho_{kv} < \rho_{\min}\};$   
 $8 \quad \quad P \leftarrow P \setminus \{(m_k, \lambda_k) \mid \exists i, j, i \neq j, \phi(\bar{\rho}_{ij}) < \rho_{ij},$   
 $9 \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad k = \arg \min_{k \in \{i, j\}} \rho_{kv}\};$   
 $10 \quad \textbf{return } \hat{V} - (\hat{\Sigma}_Z^{-1} \hat{\Sigma}_{ZV})^T \hat{Z};$

---

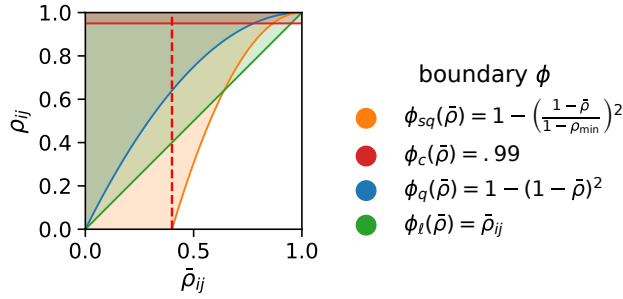
#### 4.5 CASE STUDIES

We first define a criterion of *efficiency* in order to estimate whether the reduction in variance is worth the increased cost. A natural baseline of a variance reduction is, that it is more efficient to pay for the computational overhead of the reduction than to generate more samples to achieve a similar reduction of variance. Let  $\sigma_Y^2$  be the variance of  $Y$ . The *efficiency* of the method is the ratio of the necessary cost to achieve a similar reduction with the CV estimate  $Y_{\text{CV}}$  compared to the standard estimate  $Y$  (L'Ecuyer, 1994), i.e.

$$E = \frac{c_0 \sigma_Y^2}{c_1 \sigma_{Y_{\text{CV}}}^2}. \quad (4.10)$$

This is the ratio between *slowdown*  $c_0/c_1$  and variance reduction  $\sigma_Y^2/\sigma_{Y_{\text{CV}}}^2$ . That ratio  $c_0/c_1$  depends heavily on both the specific implementation and the technical setup. The cost increase is mainly due to the computation of the integrals in (4.2). But the repeated checking of control variates for efficiency also increases the cost. The accumulation over the trajectory directly increases the cost of a single simulation which is the critical part of the estimation. To estimate the base-line cost  $c_0$ , 2000 estimations were performed without considering any control variates.





**Figure 4.2:** Different decision functions used in the redundant control variate removal. The weaker of any two control variates is removed if the pair  $(\bar{\rho}_{ij}, \rho_{ij})$  belongs to the shaded area of the considered function. The vertical dashed line indicates  $\rho_{\min}$ .

The simulation is implemented in the Rust programming language. The model description is parsed from a high level specification. Rate functions are compiled to stack programs for fast evaluation. Code is made available online (Backenköhler, 2019).

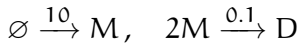
[www.rust-lang.org](http://www.rust-lang.org)

We consider four non-trivial case studies. Three models exhibit complex multi-modal behavior. We now describe the models and the estimated quantities in detail.

The first model is a simple dimerization on a countably infinite state-space.

*This model appeared earlier as [Model 1](#) on page 18.*

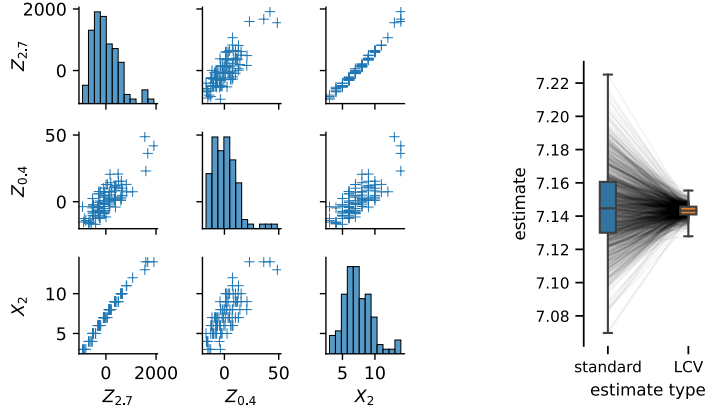
**Model 5 (Dimerization).** *We first examine a simple dimerization model on an unbounded state-space.*



with initial condition  $X_0^M = 0$ .

Despite the models simplicity, the moment equations are not closed for this system due to the second reaction which is non-linear. Therefore a direct analysis of the expected value would require a closure. For this model we will estimate  $E(X_2^M)$ .

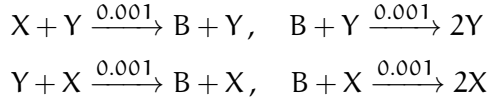
The following two models are bimodal, i.e. they each posses two stable regimes among which they can switch stochastically.



**Figure 4.3:** The effect of including **CV**s on the mean estimates  $\hat{E}(X_2^M)$  in the dimerization case study. Parameters were  $\pi_\lambda = N(0, 1)$ ,  $n_\lambda = 30$ ,  $k_{\min} = 4$ ,  $\phi(\bar{\rho}) = 1 - (1 - \bar{\rho})^2$ .

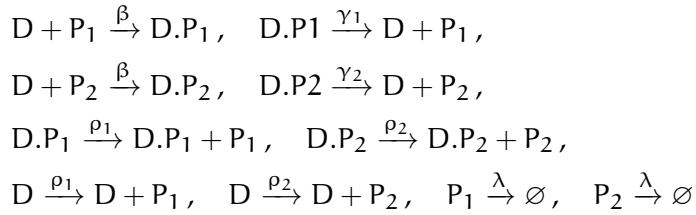
For both models we choose the initial conditions such that the process will move towards either attracting region with equal probability.

**Model 6** (Distributive Modification). *The distributive modification model was introduced in (Cardelli and Csikász-Nagy, 2012). It consists of the reactions*



with initial conditions  $X_0^X = X_0^Y = X_0^B = 100$ .

**Model 7** (Exclusive Switch). *The exclusive switch model consists of 5 species, 3 of which are typically binary (activity states of the genes) (Loinger et al., 2007).*



with initial conditions

$$X_0^D = 1 \text{ and } X_0^{D.P1} = X_0^{D.P2} = X_0^{P1} = X_0^{P2} = 0.$$

We evaluate the influence of algorithm parameters, choices of distributions to sample  $\lambda$  from, and the influence of the sample size on the efficiency of the proposed method. Note that the implementation does not simplify the constraint representations or the state space according to stoichiometric invariants or limited state spaces. [Model 6](#), for example has the invariant

$$X_t^X + X_t^Y + X_t^B = \text{const.}, \forall t \geq 0,$$

which could be used to reduce the state-space dimensionality to two. In [Model 7](#) the invariant

$$\forall t \geq 0. X_t^D, X_t^{D.P1}, X_t^{D.P2} \in \{0, 1\}$$

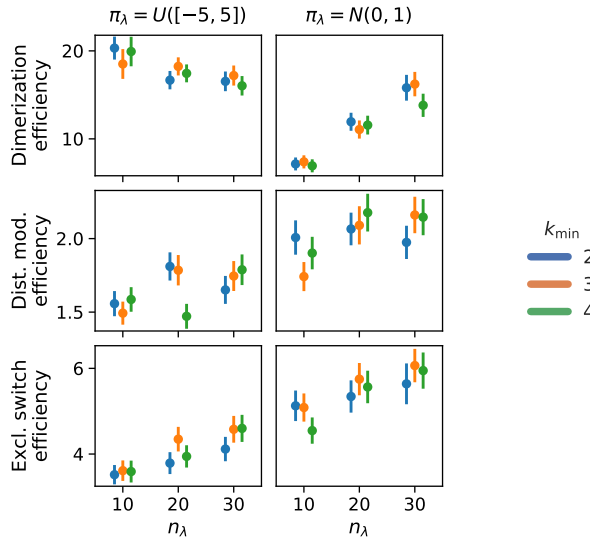
could be used to optimize the algorithm by eliminating redundant moments, e.g.  $E((X^D)^2) = E(X^D)$ . Such an optimization would further increase the efficiency of the algorithm.

We first turn to the choice of the  $\lambda$  sampling distribution. Here we consider two choices:

1. a standard normal distribution  $N(0, 1)$ ,
2. a uniform distribution on  $[-5, 5]$ .

We deterministically include  $\lambda = 0$  in the constraint set, as this parameter corresponds to a uniform weighting function. We performed estimations on the case studies using different valuations of the algorithm parameters of the minimum threshold  $k_{\min}$  and the number of  $\lambda$ -samples  $n_\lambda$ . We used samples size  $n = 10,000$  and checked the control variates every  $d = 100$  iterations for the defined criteria. For each valuation 1000 estimations were performed. In [Figure 4.4](#), we summarize the efficiencies for the arising parameter combinations on the three case studies. Most strikingly, we can note that the efficiency was consistently larger than one in all cases. Generally, the normal sampling distribution out-performed the alternative uniform distribution, except in

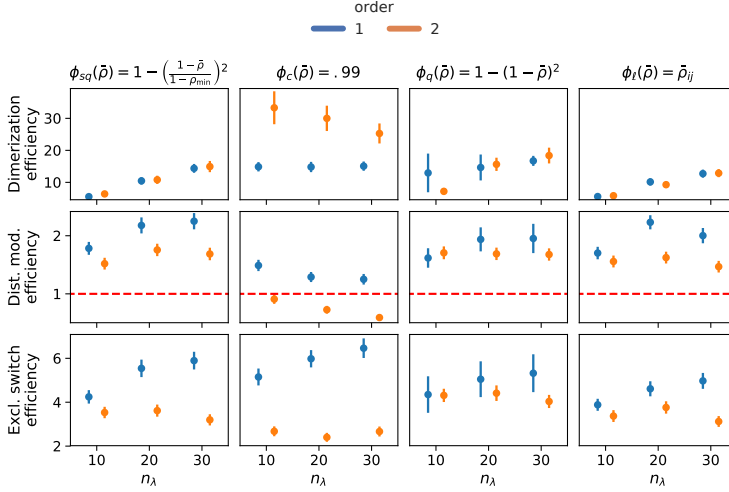
case of the dimerization. The reason for this becomes apparent, when examining Figure 4.1: In case of the dimerization model the most efficient constraints are found for  $\lambda \approx -3$ , while in case of the distributive modification they are located just above 0 (we observe a similar pattern for the exclusive switch case study). Therefore the sampling of efficient  $\lambda$  values is more likely using a uniform distribution for the dimerization case study, than it is for the others. Given that larger absolute values for  $\lambda$  seem unreasonable due their exponential influence on the weighting function and the problem of fixing a suitable interval for a uniform sampling scheme, the choice of a standard normal distribution for  $\pi_\lambda$  seems superior.



**Figure 4.4:** The efficiencies for different valuations of  $n_\lambda$  and  $k_{\min}$  and choices of  $\pi_\lambda$ . The sample size was  $n = 10,000$  in all cases with  $d = 100$ . The bars give the bootstrapped (1000 iterations) standard deviations.

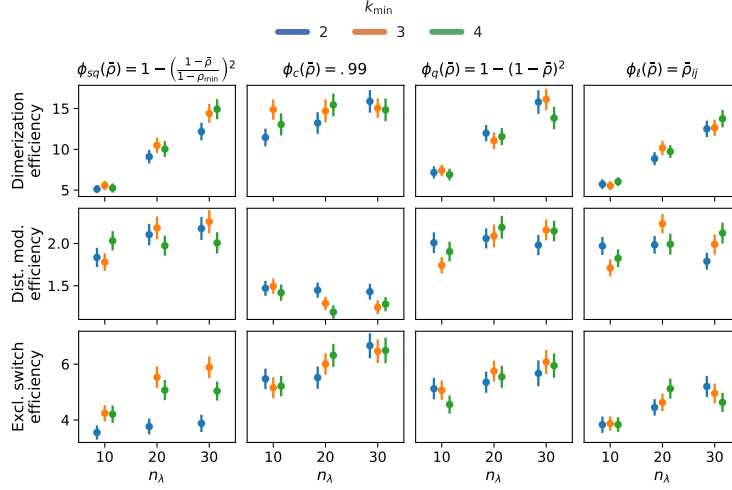
In Figure 4.5 we compare efficiencies for different maximum orders of constraints  $n_{\max}$ . This comparison is performed for different choices of the redundancy rule and initial  $\lambda$  sample sizes  $n_\lambda$ . Again, for each parameter valuation 1000 estimations were performed. With respect to the maximum constraints order

$n_{\max}$  we see a clear tendency, that the inclusion of second order constraints lessens the efficiency of the method. In case of a constant redundancy threshold it even dips below break-even for the distributive modification case study. This is not surprising, since the inclusion of second order moments increases the number of initial constraints quadratically and the incurred cost, especially of the first iterations, lessens efficiency.



**Figure 4.5:** The efficiency for different redundancy policies  $\phi$  and maximal moment orders  $n_{\max}$ . The sample size was  $n = 10,000$  in all cases with  $d = 100$ . Furthermore,  $k_{\min} = 3$ ,  $\pi_\lambda = N(0, 1)$ , and  $n_{\max} = 1$ . The bars give the bootstrapped (1000 iterations) standard deviations.

Figure 4.7 depicts the trade-off between the variance reduction  $\sigma_0^2/\sigma_1^2$  versus the cost ratio  $c_0/c_1$ . Comparing the redundancy criterion based on a constant threshold  $\phi_c$  to the others, we observe both a larger variance reduction and an increased cost. This is due to the fact, that more control variates are included throughout the simulations (Table A.1, Table A.2, Table A.3). Depending on the sample distribution  $\pi_\lambda$  and the case study, this permissive strategy may pay off. In case of the dimerization, for example, it pays off, while in case of the distributive modification it leads to a lower efficiency ratio. In the latter case the model

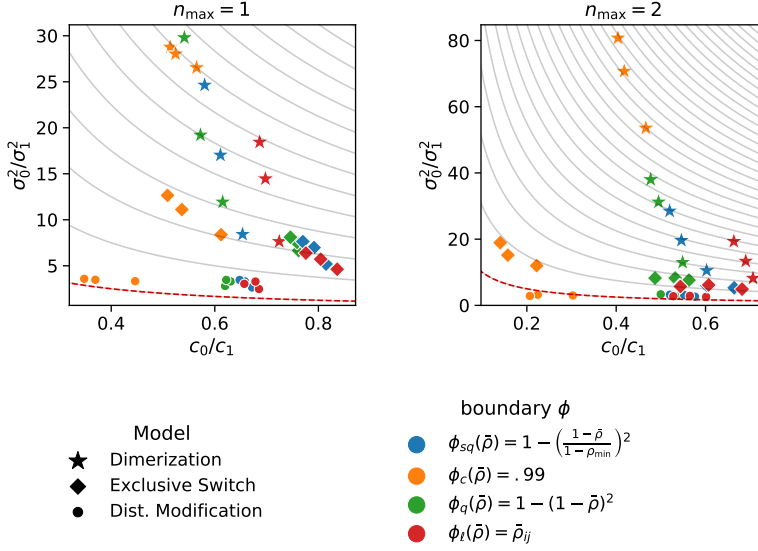


**Figure 4.6:** The empirical efficiencies for different  $n_\lambda$  and  $k_{\min}$ . On the considered case studies. The sample size was  $n = 10,000$  in all cases with  $d = 100$ . 1000 estimations were performed for each case. The bars give the bootstrapped (1000 iterations) standard deviations. The break-even  $E = 1$  is marked by the dotted red line.

is more complex, and therefore the set of initial control variates is larger. With a more permissive redundancy strategy, more control variates are kept (ca. 10 when using  $\phi_c$  vs. ca. 2 to 3 for the others). The other redundancy boundaries move the results further in the direction of less variance reduction while keeping the cost increase low. On the opposite end is the linear  $\phi_\ell$ . The quadratic versions  $\phi_q$  and  $\phi_{sq}$  can be found in the middle of this spectrum.

We also observe, that an increase of  $n_\lambda$  is particularly beneficial, if the sampling distribution  $\pi_\lambda$  does not capture the parameter region of the highest correlations well. This can be seen for the Dimerization case study, where the variance reduction increases strongly with increasing sample size (Figure 4.6, Table A.1, Table A.2, Table A.3). Since  $\pi_\lambda = N(0, 1)$ , more samples are needed to sample efficient  $\lambda$ -values (cf. Figure 4.1).

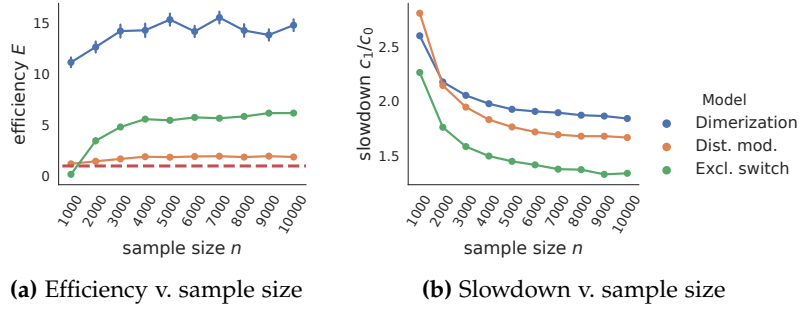
In Figure 4.6 we give detailed information on the influence of algorithm parameters  $k_{\min}$ , the number of initial  $\lambda$  values, and



**Figure 4.7:** A visualisation of the trade-off between variance reduction  $\sigma_0^2/\sigma_1^2$  and cost ratio  $c_0/c_1$ . Isolines for efficiencies are given in grey. The break-even is marked by the dashed red line. Markers of the same kind differ in  $n_\lambda$  and shift with increasing value upwards in variance reduction and lower in  $c_0/c_1$ , i.e. the shift is to the left and upwards with increasing  $n_\lambda$ . The sample size was  $n = 10,000$  in all cases with  $d = 100$ . Furthermore,  $k_{\min} = 3$  and  $\pi_\lambda = N(0, 1)$ .

different redundancy rules. The  $\lambda$  sampling distribution  $\pi_\lambda$  is a standard normal.

Finally, we discuss the effect of the sample size  $n$  on the efficiency  $E$ . In Figure 4.8 we give both the efficiencies and the slowdown for different sample sizes. As a redundancy rule we used the unscaled quadratic function, 30 initial values of  $\lambda$ , and  $k_{\min} = 3$ . With increasing sample size, the efficiency usually approaches an upper limit. This is due to the fact that most control variates are dropped early on and the control variates often remain the same for the rest of the simulations. If we assume there are no helpful control variates in the initial set and all would be removed at iteration 100, the efficiency would converge to 1 for  $n \rightarrow \infty$ .



**Figure 4.8:** The effect of sample size on the efficiency  $E$  and slowdown in the different case studies. The break-even  $E = 1$  is marked by the dashed red line. The cost increase due to the variance reduction over different sample sizes.

#### 4.6 RESAMPLING ALGORITHM

In previous work (Backenköhler, Bortolussi, and Wolf, 2019), we have proposed an algorithm that learns a set of control variates through refinement of an initial set. This initial set of control variates is based on samples of the time-weighting  $\lambda$ . Each control variate is then checked for effectiveness in isolation. Furthermore the set is refined by considering variable pairwise to determine redundancies.

In this work, we improve on the initial selection of control variates. The initial set of control variates is build using a splitting approach akin to sequential Monte Carlo methods: Over multiple rounds, new control variates are samples based on performance in prior rounds. That way, a set of variates is build up. This set is then refined in a greedy manner, taking into account the correlation between variates. This algorithm has the main benefit of needing less sensitive to user input. In particular, no heuristic redundancy threshold has to be fixed, making this approach more flexible.

As we have seen in the previous section, effective control variates have a high correlation with the target random variable. In the case of a single variate, the variance reduction is directly proportional to  $1 - \rho^2$ , where  $\rho$  is the correlation. In our case,

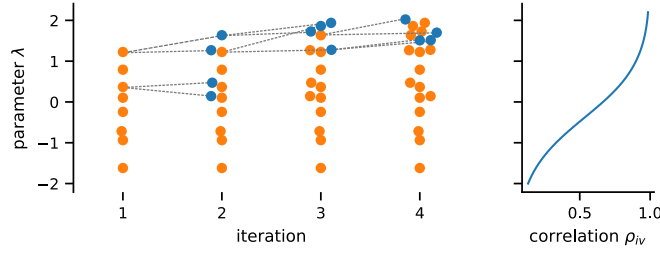


infinitely many choices of  $Z$  are available. Our goal is to choose a subset that satisfies two objectives: Firstly, every selected control variate should reduce the estimator's variance. Secondly, the subset should not be too large, i.e. we want to avoid redundancies to achieve a good overall computational efficiency of the variance reduction.

If the computation of a variate does not adequately compensate for its computation with variance reduction, we do not want to include it. Balancing both objectives is challenging because control variates often correlate with each other. Such correlations expose redundancy between different variates. This also becomes clear, when considering that the overall variance reduction depends on the coefficient of multiple correlation.

Here we follow a resampling paradigm: We start by building up a set of candidates using a particle splitting approach. After each splitting step, we generate a small number of SSA samples to estimate correlations. Promising candidates are chosen based on the *improvement* they provide and their time-weighting parameter  $\lambda$  is resampled (see Figure 4.9). The main benefit of this bottom-up approach is its lower dependence on the initial sampling distribution of  $\lambda$ . Moreover, the procedure spends less time evaluating unpromising candidates. After generating a set of control variates, the overall covariance matrix is estimated using stochastic simulations. Using this information, we construct an efficient subset using a greedy scheme, taking into account the redundancies between control variates. We discuss Algorithm 4 in more detail below.

**INITIALIZATION** A tuple  $(m_k, \lambda_k)$  of a moment vector  $m_k$  and a time-weighting parameter  $\lambda_k$  uniquely identifies a control variate  $k$ . The algorithm starts out with an initial small set of control variates. That is, we use  $w(t) = e^{\lambda_k t}$  and  $f(x) = x^{m_k}$  in (4.1). For a given set of time-weighting parameters  $L$ , we use all moments up to some fixed order  $n_{\max}$  (line 2). For a fixed moment vector  $m_k$  the time-weighting parameter  $\lambda_k$  can lead to vastly different correlations  $\rho_{kv}$  with the quantity of interest. The best choices of  $\lambda$  are usually not known beforehand. Therefore,



**Figure 4.9:** An illustration of the resampling procedure for the time-weighting parameter  $\lambda$  using [Model 6](#). Areas giving higher correlations are resampled through multiple rounds. The newly sampled values are given in blue. In each round only the new candidates are evaluated.

we sample an initial set of  $\lambda$ 's from a fixed distribution  $\pi_\lambda$  ([line 1](#)). Here, we use a standard normal distribution because its mean is the neutral weighting of  $\lambda = 0$  and extreme values are unlikely.

**RESAMPLING** Promising candidates are chosen from all control variates based on the estimated *improvement ratio* they provide, i.e.

$$\hat{\gamma}_{kv} = (1 - \hat{\rho}_{kv}^2)^{-1} \quad (4.11)$$

following [\(4.9\)](#). Specifically, control variate  $k$  is chosen with probability proportional to  $\hat{\gamma}_{kv}$  ([line 7](#)). The covariances of (only) the new variates are roughly estimated using very few (e.g.,  $d = 10$ ) [SSA](#) samples. For the selected variates  $I_{\text{cands}}$ , the time-weighting parameter is resampled using a step distribution. There is some freedom in the specifics of this resampling procedure. In particular, the number of splits  $n_c$  and descendants  $n_s$  for each candidate control the number of additional candidates. The algorithm performs  $n_r$  rounds of resampling. [Figure 4.9](#) illustrates this part of the algorithm.

**COVARIANCE ESTIMATION** After sampling a set of candidates this way, we need to select the most promising ones. For this, we are interested in covariances between all control variates, as well. Since the resampling does not provide us with such estimates, we

**Algorithm 4:** Estimate the mean of species  $i$  at time  $T$ 


---

**input** :  $n, d, n_{\max}, n_{\lambda}, n_c, n_s, n_r$   
**output**: estimate using linear control variates

```

1  $L \leftarrow \{\lambda_i \sim \pi_{\lambda} \mid 1 \leq i < n_{\lambda}\} \cup \{0\};$  /* initialization */
2  $P \leftarrow \{(m, \lambda) \mid 1 \leq |m| \leq n_{\max}, \lambda \in L\};$ 
3  $P_{\text{all}} = \emptyset$  for  $i = 1, \dots, n_r$ ; /* resampling */
4 do
5    $(\hat{\Sigma}, \hat{V}, \hat{Z}) \leftarrow \text{SSACV}(\pi_0, T, P, d);$ 
6    $P_{\text{all}} \leftarrow P_{\text{all}} \cup P;$ 
7    $I_{\text{cands}} \leftarrow \{k \sim \hat{\gamma}_{kv} / \sum_{\ell} \hat{\gamma}_{\ell v} \mid 1 \leq k \leq |P_{\text{all}}|, j = 1, \dots, n_c\};$ 
8    $P \leftarrow \bigcup_{k \in I_{\text{cands}}} \bigcup_{l=1}^{n_s} \{(m_k, \lambda'_k) \mid \lambda'_k \sim N(\lambda_k, 0.5)\};$ 
9  $(\hat{\Sigma}, \hat{V}, \hat{Z}) \leftarrow \text{SSACV}(\pi_0, T, P_{\text{all}}, 5d);$  /* covariance */
10  $P^* = \emptyset;$ 
11 while  $\exists i : (m_i, \lambda_i) \in P_{\text{all}} \setminus P^* \wedge$ 
12    $\hat{\gamma}_{iv} \prod_{j=1; (m_j, \lambda_j) \notin P^*}^{|P_{\text{all}}|} \hat{\gamma}_{ij}^{-1} > \epsilon;$  /* selection */
13 do
14    $k \leftarrow \arg \max_i \hat{\gamma}_{iv} \prod_{j=1; (m_j, \lambda_j) \notin P^*}^{|P_{\text{all}}|} \hat{\gamma}_{ij}^{-1};$ 
15    $P^* \leftarrow P^* \cup \{(m_k, \lambda_k)\};$ 
16  $(\hat{\Sigma}, \hat{V}, \hat{Z}) \leftarrow \text{SSACV}(\pi_0, T, P^*, n);$  /* estimation */
17 return  $\hat{V} - (\hat{\Sigma}_Z^{-1} \hat{\Sigma}_{ZV})^{\top} \hat{Z}$ 

```

---

evaluate all candidates together for a fixed number of simulations (line 9).

**SELECTION** The selection part of the algorithm (line 12) proceeds in a greedy fashion wrt. the potential estimated improvement  $\hat{\gamma}_{iv}$  given by any variate. However, covariates often have high mutual correlations. For example,  $Z_{\lambda}$  and  $Z_{\lambda+\epsilon}$  for a small  $\epsilon$  are typically highly correlated — often more with each other than with the objective. We want to avoid this unnecessary computational overhead from computing nearly redundant information and numerical problems due to the covariance matrix inversion (see (4.7)). As a solution, we normalize the estimated improvement vector  $(\hat{\gamma}_{iv})_i$  by the product of the fractions of explained

variances by the already selected covariates. Therefore we choose the most promising candidate given a selection  $P^*$  as

$$\arg \max_{1 \leq i \leq |P_{\text{all}}|} \hat{\gamma}_{iv} \prod_{\substack{1 \leq j \leq |P_{\text{all}}| \\ (m_j, \lambda_j) \notin P^*}} \hat{\gamma}_{ij}^{-1} \quad (4.12)$$

in [line 14](#). This selection is done, until some lower threshold  $\epsilon$  is reached ([line 12](#)).

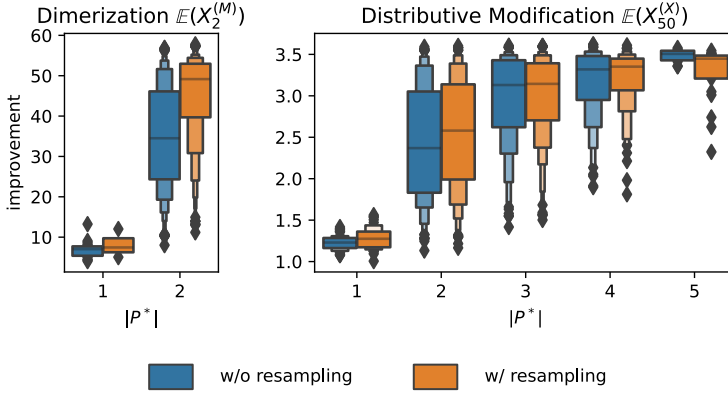
**ESTIMATION** Finally, we simulate the model  $n$  times ([line 16](#)). The resulting information enables an LCV estimation ([line 17](#)).

We applied the presented algorithm for an estimation using  $n = 10,000$  simulations. Initially  $n_\lambda = 10$  samples for the time-weighting parameter were drawn from a standard normal distribution ( $\pi_\lambda = N(0, 1)$ ). Constraints corresponding to each first-order moment, i.e. the process' expectations were generated ( $n_{\text{max}} = 1$ ). The covariance estimation during resampling used  $d = 10$  samples.

We evaluated the algorithm both with and without resampling for these first two case studies. The algorithm without resampling leaves out lines 3–8 from [Algorithm 4](#). The evaluation without resampling provides a good point of comparison to our previous heuristics performance on these cases. In the case of dimerization we observe a variance reduction of  $\approx 27.67$  compared to a best case reduction of  $\approx 28.75$  in our previous work. This close performance however has to balance very different slowdown factors: With our new heuristic the slowdown is a factor of  $\approx 1.34$  while in the previous case it was  $\approx 1.95$ . Therefore the new method clearly outperforms in terms of efficiency ( $\approx 20.5$  (new) versus  $\approx 14.86$  (old)). This is mainly due to the higher number of covariates used by the simple threshold heuristic. In contrast the new method takes into account redundancies between covariances while still retaining good performance. This becomes apparent when comparing the average number of used variates ( $\approx 3.34$  (old) versus  $\approx 1.98$  (new)).

The variance reduction factor for the distributive modification model is similar at  $\approx 2.63$  (old) versus  $\approx 2.66$  (new). Noticeably, the new method uses on average fewer CVs ( $\approx 2.74$ ) than the

previous heuristic with the best efficiency ( $\approx 3.23$ ). The overall efficiency of the new algorithm with 1.72 is slightly lower than the previous best value of 1.77, due to a higher slowdown. It is however important to note, that the trade-off differs significantly between different heuristics used in the previous algorithm. Furthermore, the lower average number of control variates would reduce the slowdown factor further, if more trajectories are generated.

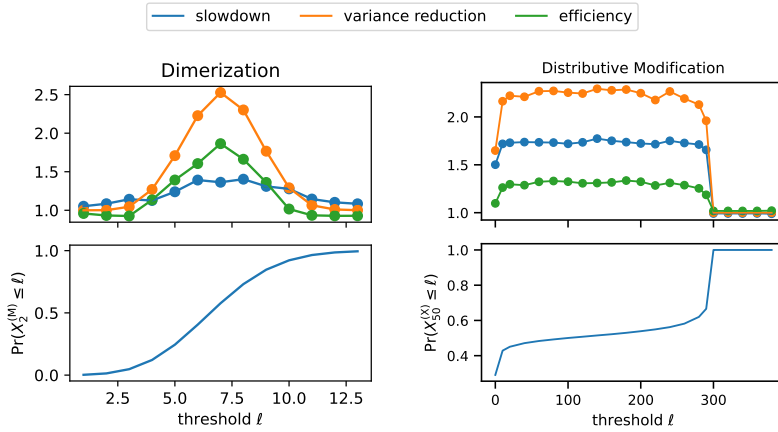


**Figure 4.10:** The variance reduction factor  $\sigma_Y^2/\sigma_{Y_{LCV}}^2$  over different numbers of selected covariates with and without the resampling procedure.

In [Figure 4.10](#) we contrast the variance improvement ratio with and without the resampling algorithm. For the dimerization model, we see a clear improvement of variance reduction. This improvement is due to the fact that the strongest correlations are present for  $\lambda \approx 2.5$  (cf. [Figure 4.9](#)). This region of the time-weighting parameter space is less likely to be sampled by the initial samples from the standard normal distribution. Therefore the resampling procedure is especially beneficial if the better parameters  $\lambda$  are farther from the origin. In case of the distributive modification case study, we see a slight improvement. Here, the best parameters  $\lambda$  are close to zero and thereby more likely to be sampled by a standard normal. Still, the resampling improves covariate performance for the most frequent cases of 2–4 covariates being selected (the case of 5 covariates has only a few instances).

Note, that the additional cost incurred by the resampling procedure is comparatively small, because at most 4 candidates are evaluated in each iteration.

Next, we turn to the estimation of probabilities. In particular, we consider the event of a species being below a threshold  $\ell$  at time  $t$  (species  $M$  for the dimerization and  $X$  for the distributive modification). In Figure 4.11 we summarize the results of this study for varying levels  $\ell$ . In both case studies we observe that control variates are efficient for probabilities not close to either zero or one. In this case control variates are able to reduce the variance of the estimated probabilities whilst maintaining a beneficial reduction-slowdown trade-off. This region is larger for the distributive modification model because of its bimodal behavior. If the probability to be estimated is close to either one or zero, the event occurs too rarely or too often, respectively, to adequately explain variance using linear correlations. We note, that the worst case efficiency is close to one. This is due to the algorithm throwing out all covariate candidates leaving us with a standard estimation. Only the initial covariate evaluation and resampling causes a slowdown, driving efficiency slightly below one. Naturally this cost decreases with more samples  $n$ .

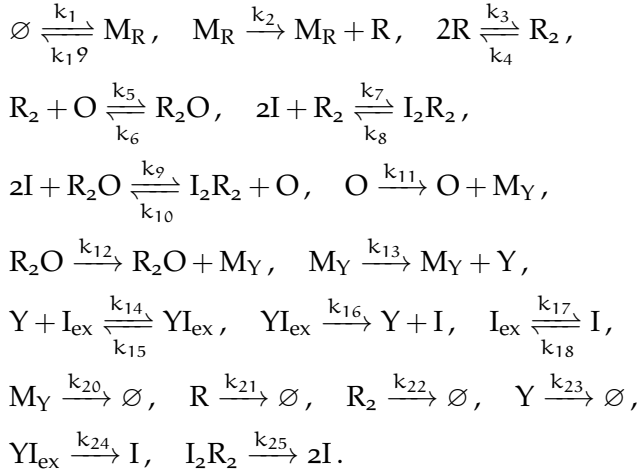


**Figure 4.11:** The methods efficiency for the estimation of threshold probabilities. For each threshold  $\ell$  at least 200 estimations were performed.

Control variates based on test functions restricted to intervals did not lead to an improvement (data not shown).

Finally, with the lac operon model we consider a larger case study. This model consists of 11 species and 25 – partly non-linear – reactions.

**Model 8** (Lac operon). *This is a well-known model of genetic regulation with positive feedback (Stamatakis and Mantzaris, 2009). Its reactions are*



Initially,  $X_0^{(O)} = 1$  and  $X^{(I_{\text{ex}})} = 48,177$  while all other abundancies are zero. The parameters are  $k_1 = 0.111$ ,  $k_2 = 15.0$ ,  $k_3 = 103.8$ ,  $k_4 = 0.001$ ,  $k_5 = 1992.7$ ,  $k_6 = 2.4$ ,  $k_7 = k_9 = 1.293 \times 10^{-7}$ ,  $k_8 = 12$ ,  $k_{10} = 9963.2$ ,  $k_{11} = 0.5$ ,  $k_{12} = 0.01$ ,  $k_{13} = 30.0$ ,  $k_{14} = 0.249$ ,  $k_{15} = 0.1$ ,  $k_{16} = 6.0 \times 10^4$ ,  $k_{17} = k_{18} = 0.92$ ,  $k_{19} = k_{20} = 0.462$ ,  $k_{21} = k_{22} = k_{23} = k_{24} = k_{25} = 0.2$ .

We estimate the abundancy of LacY after one time unit, i.e.  $E(X_1^{(Y)})$ . It is encoded by Y and facilitates the lactose import via reactions 14 and 16. A typical simulation of the system up to time-horizon  $T = 1$  takes well above one minute of computational time. Therefore we reduce the number of used trajectories to  $n = 1000$ . The other settings remain as above.

Despite the high dimensionality, we observe a good efficiency value of  $E \approx 4.85$ . The slowdown caused by the method is approximately 1.98. A big part of this slowdown is due to the initial

search of covariates. Initially 10 covariates are generated for each first order moment, i.e. each of the 11 species. The number of additionally resampled covariates is similar to previous case studies. Thus the main cost of the initial resampling and selection is due to the first iteration of the resampling loop and the simulation loop of the selection procedure. This part naturally has still potential for optimization: Not all known covariates need to be reconsidered at the selection stage. Instead, unpromising candidates could be discarded prior to that stage.

Still, the high variance reduction by a factor of approx.  $9.64^{-1}$  more than compensates for this increase in computational cost, leading to the good overall efficiency. This shows that, even for more complex models, the method is applicable and can extremely be beneficial for Monte Carlo estimation.

#### 4.7 CONCLUSION

In the context of Monte Carlo simulation, variance reduction techniques offer an elegant way of improving the performance without introducing approximation errors in addition to the statistical uncertainty. In this work we have shown that known constraints on the moment dynamics can be successfully leveraged in simulation-based estimation of expected values. The empirical results indicate that the supplementing a standard [SSA](#) estimation with moment information can drastically reduce the estimators' variance. This reduction is paid for by accumulating information on the trajectory during simulation. However, the reduction is able to compensate for this increase. This means that for fixed costs, using estimates with control variates is more beneficial than using estimates without control variates.

In particular, we improve an initial subset by selecting particularly effective variates and removing redundant variates. By resampling the time-weighting parameter  $\lambda$  we ensure that appropriate values are flexibly explored. In the worst case, all variates are dropped and the performance approaches the standard [SSA](#). In most cases, however, a suitable subset is found together with the corresponding choices of  $\lambda$ .



We analyze the performance of the method when estimating event probabilities and not only average molecule counts. Our largest case study has 11 species and 24 reactions.

Another open question regarding this work is its performance when multiple quantities instead of a single quantity are to be estimated. In such a case, constraints would be particularly beneficial, if they lead to improvements as many estimation targets as possible.

In the future, we will further explore the algorithmic design space. For example, the resampling distribution could be adjusted using decaying standard deviations. Furthermore, we will look at different test functions weighting the state space more flexibly. Different choices of  $f$  and  $w$  in (4.1) may improve efficiency further. These choices become particularly interesting when moving from the estimation of simple first order moments to more complex queries such as probabilities. In such cases, one might even attempt to find efficient control variate functions using machine learning methods.

Another worthwhile direction is the combination of CV with other Monte Carlo techniques. In particular, importance sampling might benefit from this. Control variates of the biased process could be used to improve estimation of the likelihood ratio.



## Part III

### AGGREGATION & REFINEMENT

We present a state-space lumping scheme that aggregates states in a grid structure. Approximations based on this lumping are used to iteratively refine relevant and truncate irrelevant parts of the state-space. This way, the algorithm learns a well-justified finite-state projection for different scenarios.



# 5

---

## STATE-SPACE AGGREGATION

---

In this part, we propose a method to identify a truncation that optimizes the trade-off between the size of the considered state-space and the approximation error due to the finite state projection (FSP). To this end, we start with a very coarse-grained model abstraction that we refine iteratively. The coarse-grained model is based on an grid-shaped aggregation (i.e. lumping) scheme that identifies a set of macro-states. These macro-states can be used to compute an interim model solution that guides the refinement in the next step. We perform refinements until the approximation arrives at the resolution of the original model (i.e. each macro-state has only one constituent) such that the aggregation introduces no approximation error.

↪ page 19

### 5.1 RELATED WORK

Aggregation-based numerical methods for the analysis of discrete or continuous-time Markov chains have been studied in previous work. Popular approaches rely on an alternation of aggregation and disaggregation of the state-space (Schweitzer, 1991; Stewart, 1994). In the case of stiff chains, such aggregations are typically based on a separation of time-scales (Cao and Stewart, 1985). However, these methods have been developed for finite chains with arbitrary structure and are motivated by numerical issues of standard methods such as the power method or Jacobi iteration (Stewart, 1994). They do not consider a truncation of irrelevant

states, while here our aggregation approach is used to determine the most relevant states under stationary conditions in large or infinite chains with population structure.

More recently, a scheme based on the Kullback-Leibler divergence between the approximate and the original system has been proposed for DTMCs (Geiger et al., 2014).

## 5.2 MACRO-STATES

A macro-state is a collection of micro-states (or simply states) treated as one state in the aggregated model, which can be seen as an abstraction of the original model. The aggregation scheme defines a partitioning of the state-space. We choose a scheme based on a grid structure. That is, each macro-state is a hypercube in  $\mathbb{Z}_{\geq 0}^{n_s}$ .

Hence, each macro-state  $\bar{x}_i(\ell^{(i)}, u^{(i)})$  (denoted by  $\bar{x}_i$  for notational ease) can be identified using two vectors  $\ell^{(i)}$  and  $u^{(i)}$ . The vector  $\ell^{(i)}$  gives the corner closest to the origin, while  $u^{(i)}$  gives the corner farthest from the origin. Formally,

$$\bar{x}_i = \bar{x}_i(\ell^{(i)}, u^{(i)}) = \{x \in \mathbb{N}^{n_s} \mid \ell^{(i)} \leq x \leq u^{(i)}\}, \quad (5.1)$$

where ' $\leq$ ' denotes element-wise comparison.

In order to solve the aggregated model, we need to define transition rates between macro-states. Therefore, we assume that, given that the system is in a particular macro-state, all constituent states are equally likely (uniformity assumption). This assumption is the reason why the aggregated model provides only a coarse-grained approximation.

The uniformity assumption is a modeling choice yielding significant advantages. Firstly, it eases the computation of the rates between macro-states and, therefore, makes a fast solution of the aggregated model possible. Secondly, even though it induces an approximation error, it provides suitable guidance as uniformity assumption spreads out the probability mass conservatively. Hence, it becomes less likely that regions of interest are disregarded. Lastly, the uniformity assumption is theoretically well-founded, as it stems from the maximum entropy principle: In the

absence of concrete knowledge about the probability distribution inside a macro-state, we assume the distribution with the highest uncertainty, i.e., the uniform distribution.

### 5.3 CONSTRUCTION

The grid structure makes the computation of transition rates between macro-states particularly convenient and computationally simple. Mass-action reaction rates can be given in a closed-form, due to the Faulhaber formulae (Knuth, 1993) and more complicated rate functions such as Hill-functions can often be handled as well by taking appropriate integrals (see page 145).

Suppose, we are interested in the transition rate from macro-state  $\bar{x}_i$  to macro-state  $\bar{x}_k$  according to reaction  $j$ . Using the uniformity assumption, this is simply the mean rate of the states in  $\bar{x}_i$  that go to  $\bar{x}_k$  using  $j$ . However, only a small subset of constituents in  $\bar{x}_i$  are actually relevant for this transition. Hence, we identify the subset of states of  $\bar{x}_i$  that lie at the border to  $\bar{x}_k$  and in such a way that applying reaction  $j$  shifts them to a state in  $\bar{x}_k$ . Then, we sum up the corresponding rates of these states. Lastly, we normalize according to the number of states inside of  $\bar{x}_i$ .

It is easy to see that the relevant set of border states is itself an interval-defined macro-state  $\bar{x}_{i \rightarrow k}^j$ . To compute this macro-state we can simply shift  $\bar{x}_i$  by  $v_j$ , take the intersection with  $\bar{x}_k$  and project this set back. Formally,

$$\bar{x}_{i \rightarrow k}^j = ((\bar{x}_i + v_j) \cap \bar{x}_k) - v_j, \quad (5.2)$$

where the additions are applied element-wise to all states making up the macro-states. For ease of notation, we also define a general exit state

$$\bar{x}_{i \rightarrow}^j = ((\bar{x}_i + v_j) \setminus \bar{x}_i) - v_j. \quad (5.3)$$

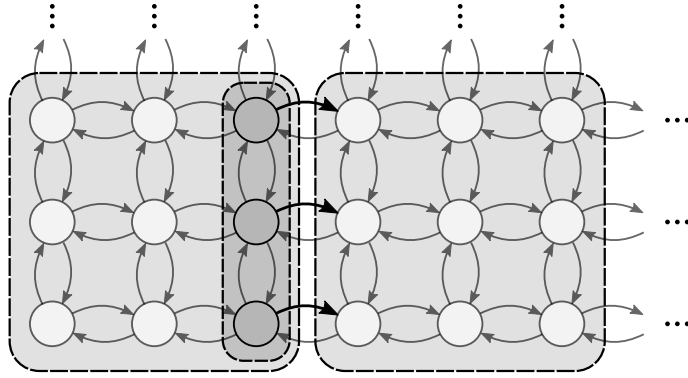
This state captures all micro-states inside  $\bar{x}_i$  that can leave the state via reaction  $j$ .

A particularly convenient feature of the transition states is, that they also are macro-states. That means, they also are specified

by independent intervals in each dimension as in (5.1). This holds because all operations in both (5.2) and (5.3) preserve this structure.

*Model 9 on page 122 gives this structure.*

**EXAMPLE** In Figure 5.1 we give an example of two adjacent macro states and the transition state from the left to the right via one reaction. As such it illustrates the result of the computation given in (5.2): The left state is shifted along the reaction vector, intersected with the right macro-state, and finally shifted back.  $\diamond$



**Figure 5.1:** Two macro-states and a transition state from the left to the right.

This uniformity assumption gives rise to the following  $Q$ -matrix of the aggregated model:

$$\bar{Q}_{\bar{x}_i, \bar{x}_k} = \begin{cases} \sum_{j=1}^{n_R} \bar{\alpha}_j \left( \bar{x}_{i \rightarrow k}^j \right) / |\bar{x}_i|, & \text{if } \bar{x}_i \neq \bar{x}_k \\ -\sum_{j=1}^{n_R} \bar{\alpha}_j \left( \bar{x}_{i \rightarrow}^j \right) / |\bar{x}_i|, & \text{otherwise} \end{cases} \quad (5.4)$$

where

$$\bar{\alpha}_j(\bar{x}) = \sum_{x \in \bar{x}} \alpha_j(x). \quad (5.5)$$

is the sum of all rates belonging to reaction  $j$  in  $\bar{x}$ . In particular, the division by  $|\bar{x}_i|$  in (5.4) is due to the uniformity assumption: According to the assumption, given the process is in  $\bar{x}_i$  it is in each constituent micro-state of the transition state  $\bar{x}_{i \rightarrow k}^j$  with



probability  $1/|\bar{x}_i|$ . Therefore, each of the added micro-state rates in (5.5) is divided by the state volume  $|\bar{x}_i|$ .

The aggregated CME becomes

$$\begin{aligned} \frac{d}{dt} \hat{\pi}(\bar{x}_i, t) = \sum_{j=1}^{n_R} \left( \sum_{i_j} \bar{\alpha}_j(\bar{x}_{i_j \rightarrow i}) \hat{\pi}(\bar{x}_{i_j}, t) / |\bar{x}_{i_j}| \right) \\ - \bar{\alpha}_j(\bar{x}_{i \rightarrow}) \hat{\pi}(\bar{x}_i, t) / |\bar{x}_i| \end{aligned} \quad (5.6)$$

and the associated backwards equation

$$\begin{aligned} \frac{d}{dt} \hat{\beta}(\bar{x}_i, t) = \sum_{j=1}^{n_R} \hat{\beta}(\bar{x}_i, t) \left( \sum_{i_j} \bar{\alpha}_j(\bar{x}_{i_j \rightarrow i}) \hat{\beta}(\bar{x}_{i_j}, t) / |\bar{x}_{i_j}| \right) \\ - \bar{\alpha}_j(\bar{x}_{i \rightarrow}) \hat{\beta}(\bar{x}_i, t) / |\bar{x}_i| \end{aligned} \quad (5.7)$$

Under the assumption of polynomial rates, as is the case for mass-action systems, we can compute the sum of rates over this transition set efficiently using Faulhaber's formula.

**EXAMPLE** Consider the following mass-action reaction  $2X \xrightarrow{c} \emptyset$ . For macro-state  $\bar{x} = \{0, \dots, n\}$  we can compute the corresponding lumped transition rate

$$\begin{aligned} \bar{\alpha}(\bar{x}) &= \frac{c}{2} \sum_{i=1}^n i(i-1) = \frac{c}{2} \left( \sum_{i=1}^n i^2 - \sum_{i=1}^n i \right) \\ &= \frac{c}{2} \left( \frac{2n^3 + 3n^2 + n}{6} - \frac{n^2 + n}{2} \right) \end{aligned}$$

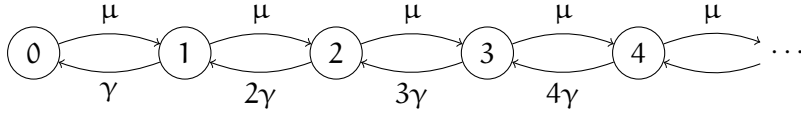
eliminating the explicit summation in the lumped propensity function.  $\diamond$

Interestingly, the lumped distribution tends to be less concentrated. This is due to the assumption of a uniform distribution inside macro-states. This effect is illustrated by the example of a birth-death process below. Due to this effect, an iterative refinement typically keeps an over-approximation in terms of state-space area. This is a desirable feature since relevant regions are less likely to be pruned due to lumping approximations.

EXAMPLE We illustrate the scheme on the birth-death process, i.e. [Model 2](#). Its CTMC has the following generator matrix

$$Q = \begin{bmatrix} -\mu & \mu & 0 & & & \cdots \\ \gamma & -(\mu + \gamma) & \mu & 0 & & \cdots \\ 0 & 2\gamma & -(\mu + 2\gamma) & \mu & 0 & \cdots \\ 0 & 0 & 3\gamma & -(\mu + 3\gamma) & \mu & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

The structure is more obvious in the graph visualization:



Now we lump states in groups of 5 states. The states are constructed as

*We omit the vector notation here for clarity because the process has a single dimension.*

$$\bar{x}_k(5k, (5+1)k-1)$$

The transition states for the birth reaction  $\emptyset \rightarrow S$  is

$$\bar{x}_{k \xrightarrow{1} k+1} = \{5(k+1)-1\}.$$

The lumped transition rate

$$\bar{\alpha}_1 \left( \bar{x}_{k \xrightarrow{1} k+1} \right) = \mu.$$

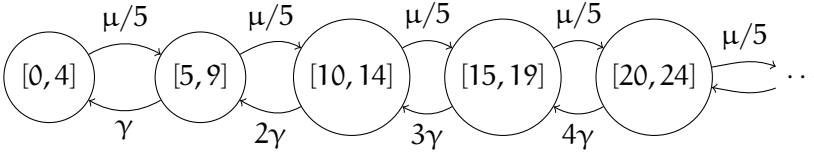
Similarly, the transition states for the death reaction  $S \rightarrow \emptyset$  is

$$\bar{x}_{k \xrightarrow{2} k+1} = \{5(k+1)\}.$$

The lumped transition rate

$$\bar{\alpha}_2 \left( \bar{x}_{k+1 \xrightarrow{1} k} \right) = 5k\gamma.$$

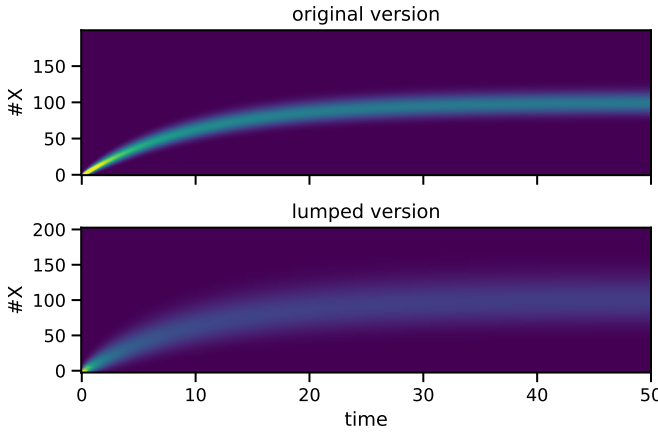
Using [\(5.4\)](#) the aggregated transitions are as follows.



In this example the rates remain in effect the same. But the same number of macro-states covers more micro-states.

A forward integration of both, the model at original granularity and the aggregated version is shown in Figure 5.2.

◇ The integration is done using an ad-hoc FSP on  $[0, 200]$ .



**Figure 5.2:** A lumping approximation of Model 2 on the state-space truncation to  $[0, 200]$  on  $t \in [0, 50]$ . On the left-hand side solutions of a regular truncation approximation and a lumped truncation (macro-state size is 5) are given.

## 5.4 APPROXIMATION FEATURES

Considering the previous example, we observe that the lumped version shows a similar temporal dynamic, but the distributions are spread out more. This is a desirable feature because it indicates the location of the main probability mass with significantly less states.

These features are not valid in general and the aggregation scheme remains a heuristic approach. This is partly due to the MPM formalism allowing for arbitrary propensity functions. One

can construct a process such that a significant change in dynamics inside a macro-state would be missed. Mostly we encounter this phenomenon in models that exhibit a significant change near the zero-boundary for some species. Examples include the toggle switch with Hill-functions ([Model 14](#)) and die-out in epidemics models ([Model 15](#)). Other cases are rather rare in the standard model repertoire, but some awareness is necessary.

Fortunately, by refining the aggregation down to “full resolution”, we can gain the guarantees inherent to [FSP](#).

# 6

---

## TRUNCATIONS FOR STATIONARY DISTRIBUTIONS

---

An important part of the analysis of such models concerns their long-run behavior. Given an ergodic underlying Markov chain, the chain's stationary distribution characterizes this behavior. For some special model classes, such as zero-deficiency networks (Anderson and Kurtz, 2011), analytical solutions for the stationary distribution are known. However, most models require numerical approaches, often based on some form of approximation to guarantee tractability. Those approaches can be based on stochastic simulation (Gillespie, 1977) (which for steady-state analysis tends to be slow and inaccurate) or moment-bounds via mathematical programming (Kuntz, Juan And Thomas, Philipp And stan and Barahona, 2019). Here, we draw on numerical approaches based on state-space truncation, which represent a viable option to approximate stationary distributions (Kuntz et al., 2021a). Truncation-based approaches have the benefit of describing the complete dynamics within a finite subset of the typically very large or infinite state-space. As such, they enable the approximation of complex distributions that are not well-described by low-order moments.

The main step in the computation of such an approximation is the identification of a suitable truncation, i.e. a subset of the state-space encompassing most of the stationary probability mass. Existing methods typically rely on Foster-Lyapunov drift conditions to define such subsets (Dayar et al., 2011). While these

truncations come with bounds on the contained stationary probability mass, they typically are far larger than necessary. The truncation is usually strongly constrained by the form of the chosen Lyapunov function (Dayar et al., 2011; Gupta, Briat, and Khammash, 2014). Optimizing over possible functions to identify efficient truncations is technically challenging and, to our knowledge, has not been demonstrated for general reaction networks (Miliadis-Argeitis and Khammash, 2014).

In this work, we address the identification of suitable truncations by using an aggregation-refinement scheme. Initially, a Lyapunov analysis yields a set containing at least  $1 - \epsilon$  of the stationary probability mass. On this subset of the state-space, we apply an aggregation scheme that groups together states in hypercube macro-states. Throughout each of these macro-states, we assume a uniform distribution among its constituent micro-states. This allows us to roughly analyze large portions of the state-space with exponentially fewer variables. We then iteratively truncate and refine the approximation based on the stationary distribution of this aggregated Markov chain. We keep only the most relevant macro-states and continue this scheme until the macro-states contain a single original state. In this way, we arrive at an effective truncation to compute an approximation of the stationary distribution.

We investigate the approximation results on case studies with known stationary distributions and complex models with intricate stationary distributions. We evaluate the truncation quality by assessing the stationary probability mass captured. To this end, we use analytical solutions and bounds given by a Lyapunov analysis. Further, we explore the control of the truncation size through the truncation parameter. Finally, we demonstrate the method on the p53 oscillator model exhibiting a complex stationary distribution.

## 6.1 RELATED WORK

**ANALYTICAL SOLUTIONS** For some specific models, analytical solutions for the stationary distribution have been found

(Kurasov et al., 2018; Mélykúti, Hespanha, and Khammash, 2014). For the class of zero-deficiency networks, the stationary distribution is known to have a Poisson product form (Anderson, Craciun, and Kurtz, 2010). Monomolecular reaction networks can be solved explicitly, as well (Jahnke and Huisinga, 2007).

**TRUNCATION-BASED ANALYSIS** The analysis of countably infinite-sized state-spaces is often handled by pre-defined truncations (Kwiatkowska, Norman, and Parker, 2011). Sophisticated state-space truncations for the (unconditioned) forward analysis have been developed that give lower bounds. They typically provide a trade-off between computational load and tightness of the bound (Andreychenko et al., 2011; Henzinger, Mateescu, and Wolf, 2009; Lapin, Mikeev, and Wolf, 2011; Mikeev et al., 2013; Munsky and Khammash, 2006). Such methods cannot be directly applied to the estimation of stationary distributions because the approximation usually introduces a sink-state.

Truncations for stationary distributions often involve re-direction schemes for transitions leaving and entering the subset. A comprehensive survey of such state-space truncation methods can be found in (Kuntz et al., 2021b). A popular method of identifying truncations is the construction of a suitable Lyapunov function. Beyond their use for establishing ergodicity (Dayar et al., 2011; Gupta, Briat, and Khammash, 2014; Meyn and Tweedie, 1993), these functions can be used to obtain truncations, guaranteed to contain a certain amount of stationary probability mass (Dayar et al., 2011). Using Lyapunov functions for the construction of truncations often leads to very conservative sets (Miliadis-Argeitis and Khammash, 2014). Different approaches have been employed to find truncations: In Gupta, Mikelson, and Khammash (2017) SSA estimates are used to set up an increasing family of truncations.

**MOMENT-BASED APPROXIMATION** Apart from approaches based on state-space truncations, moment-based approaches have been particularly popular recently (Dowdy and Barton, 2018a; Ghusinga et al., 2017; Kuntz, Juan And Thomas, Philipp And stan

and Barahona, 2019; Sakurai and Hori, 2017). Such approaches are based on the fact that particular matrices of distributional moments such as mean and variance are positive semi-definite. Along with linear constraints stemming from the Kolmogorov equations (Backenköhler, Bortolussi, and Wolf, 2016), a semi-definite program can be formulated and solved using existing tools. While this method is suited to compute bounds on both moments and subsets of the state-space, its application is limited, due to numerical issues inherent in the formulation (Dowdy and Barton, 2018a).

An approach where quantities are only described in terms of their magnitude has been proposed in Češka and Kretínský (2019). This allows for an efficient qualitative analysis of both dynamic and transient behavior.

## 6.2 TRUNCATION-BASED APPROXIMATION

In many relevant cases, the state-space is huge or infinite and therefore the stationary solution cannot be computed directly. To make such a computation possible we have to restrict ourselves to a finite manageable subset of the state-space and assume the majority of the probability mass is concentrated within that finite subset. The main problem is to deal with the transitions leading to and from the truncated set (cf. Figure 6.2). In forward analysis, the outgoing transitions are simply redirected into a sink-state. This way, a forward analysis provides lower bounds since mass leaving the truncation does not re-enter. This approach, however, is unsuitable for the computation of stationary distributions because mass would accumulate in the sink-state leading to a distribution assigning all mass to it. Therefore, transitions leaving the truncation need to be redirected back into the truncation.

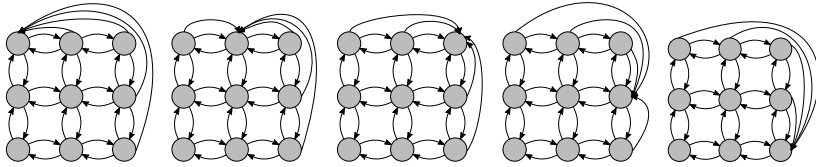
The process' dynamics outside the truncation are defined by the *stochastic complement* (Spieler, 2014). If its behavior was known, one could redirect outgoing to incoming transitions optimally and preserve the correct stationary distribution. However, this reentry distribution is typically unknown in most relevant cases. Many different reentry distributions have been used, such as



redirecting to some internal state or states with incoming transition from outside the truncation. Kuntz et al. (2021a) provides a comprehensive review of such methods.

The most natural choice is to pick a reentry distribution that redirects mass to states with incoming transitions from truncated states (cf. Figure 6.2 (center)).

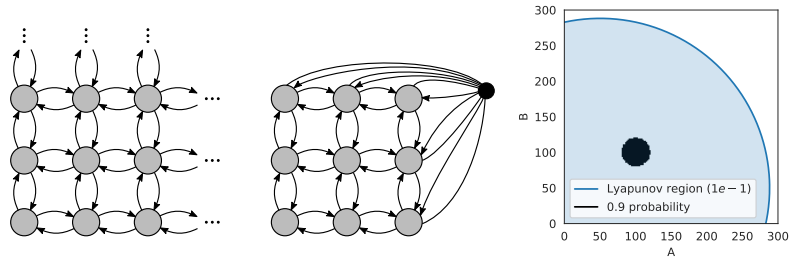
Using varying redirections, we can compute bounds on the stationary probability conditioned on a truncation (Spieler, 2014, Thm. 14). To do this, one has to compute the stationary distribution for every possible way of connecting all outgoing to a single incoming transition. Naturally, such an algorithm is rather expensive since one has to solve a linear system for each combination. Therefore this method of computing bounds is costly on very large truncations, often given by Lyapunov functions. In Figure 6.1 an enumeration of all such redirections is shown on a small  $3 \times 3$ -truncation.



**Figure 6.1:** Enumeration and solution of all single-state redirections yields upper and lower bounds on the stationary distribution, conditioned on being inside the truncation.

When computing an approximation instead of bounds, we employ a uniform redirection scheme: Outgoing transitions are split uniformly among incoming transitions. Due to the threshold-based truncation scheme, we are likely to end up with a somewhat uniform distribution over in-boundary states (see Section 6.2.2).

The identification of good truncations remains a major task in such approximations. Using approaches such as Lyapunov functions (Section 2.7) (Dayar et al., 2011) or moment-bounds (Kuntz et al., 2021a) can provide a good initial estimate, but typically the resulting truncations are far larger than necessary. This leads to dramatically increased computational costs, especially



**Figure 6.2:** (left) A countably infinite state-space. (center) Outgoing transitions are re-directed (according to the reentry distribution) to states that have incoming transitions from outside the truncation. (right) A comparison of the area prescribed by a Lyapunov analysis using Geobound and threshold 0.1 and the minimal area containing 0.9 stationary probability mass. The model is a parallel birth-death process (Model 9).

when bounding methods mentioned above are performed. Until a system for a larger truncation is solved, the precise location of most of the probability mass is often unknown. Instead of solving the full system for such a large space, we employ an aggregation scheme to cover large areas of the state-space with exponentially fewer variables.

Error bounds have been derived for increasing truncation sets in the case of linear Lyapunov functions (Gupta, Mikelson, and Khammash, 2017). However, until now it has not been shown that these bounds are applicable in practice (Meyn, Tweedie, et al., 1994). Alternatively, one can monitor the product of the probability-outflow rate and the maximum L1-norm. This bounds the approximation error up to a constant  $M > 0$ , assuming a linear Lyapunov function exists (Gupta, Mikelson, and Khammash, 2017).

### 6.2.1 Initial Aggregation

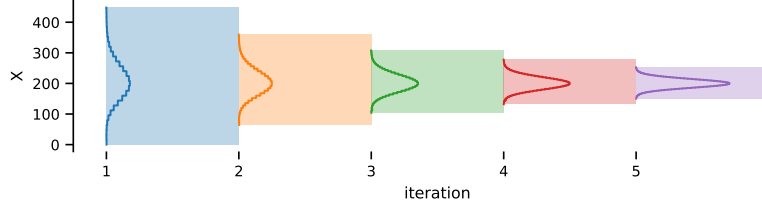
The initial aggregated space  $\hat{\mathcal{S}}^{(0)}$  should encompass all regions of the state-space that could contain significant mass because states outside this initial area will not be refined. In principle,

multiple approaches could be used to identify such a region. One possibility is the computation of moment bounds for the stationary distribution (Dowdy and Barton, 2018a; Ghusinga et al., 2017). Based on these bounds on expectations and covariances, an initial truncation could be fixed. The approach we use here is to identify such a region by a Lyapunov analysis (Dayar et al., 2011). This way, we obtain a polynomial describing a semi-algebraic subset of the entire state-space containing  $1 - \epsilon_\ell$  of the mass, where  $\epsilon_\ell > 0$  can be fixed arbitrarily. These sets usually are far larger than a minimal set containing  $1 - \epsilon_\ell$  of stationary probability mass would be. As an initial aggregation, we build an aggregation on a subset  $[0..n]^{n_s} \subset \mathcal{S}$  containing the set prescribed by the Lyapunov analysis. We also employ this approach to estimate errors in the evaluation. Specifically, we employ the tool Geobound (Spieler, 2010) with L2-norm as function  $g$  implementing techniques presented in Dayar et al. (2011).

In many cases, simple choices of  $g$  such as the L1- or L2- norm are sufficient. However, the sets resulting from such functions are often very conservative. Consider Figure 6.2 (right) as an example, where the Lyapunov truncation with  $\epsilon_\ell = 0.1$  for two parallel birth-death processes (Model 9) is compared to the smallest set containing 0.9 of stationary probability. Clearly, the area given by the Lyapunov function is magnitudes larger than necessary to capture probability mass consistent with  $\epsilon_\ell$ .

### 6.2.2 Iterative Refinement Algorithm

The refinement algorithm (Algorithm 5) starts with a set of large macro-states that are iteratively refined, based on approximate stationary distributions. We start by constructing square macro-states of size  $2^m$  in each dimension for some  $m \in \mathbb{N}$  such that they form a large-scale grid  $\mathcal{S}^{(0)}$ . Hence, each initial macro-state has a volume of  $(2^m)^{n_s}$ . This choice of grid size is convenient because we can halve states in each dimension. Moreover, this choice ensures that all states have an equal volume and we end up with unit-sized macro-states, equivalent to a truncation of the original non-lumped state-space.



**Figure 6.3:** The state-space refinement algorithm on a birth-death process. From left to right the state size is halved and states with low probability are removed from the truncation. The final truncation is a typical truncation with states of size 1 and the initial states are of size  $2^4$ .

---

**Algorithm 5:** Approximating the stationary distribution

---

**input :** Initial partitioning  $\mathcal{S}^{(0)}$ , truncation threshold  $\epsilon$

**output:** approximate stationary distribution  $\hat{\pi}_\infty$

---

1 **for**  $i = 1, \dots, m$  **do**

2      $\hat{\pi}_\infty^{(i)} \leftarrow$  approximate stationary distribution on  $\mathcal{S}^{(i)}$ ;

3      $\mathcal{R} \leftarrow$  smallest  $\mathcal{R}' \subseteq \mathcal{S}^{(i)}$  s. t.  $\sum_{\bar{x} \in \mathcal{R}'} \hat{\pi}_\infty^{(i)}(\bar{x}) \geq 1 - \epsilon$ ;

4      $\mathcal{S}^{(i+1)} \leftarrow \bigcup_{\bar{x} \in \mathcal{R}} \text{split}(\bar{x})$ ;

5     update  $\hat{Q}$ -matrix;

6 **return**  $\hat{\pi}_\infty^{(m)}$ ;

---

An iteration of the state-space refinement starts by computing the stationary distribution, using the lumped  $\hat{Q}$ -matrix. Based on a threshold parameter  $\epsilon > 0$  states are either removed or split (line 4), depending on the mass assigned to them by the approximate stationary probabilities  $\hat{\pi}_\infty^{(i)}$ . Thus, each macro-state is either split into  $2^{n_s}$  new states or removed entirely. The result forms the next lumped state-space  $\mathcal{S}^{(i+1)}$ . The  $\hat{Q}$ -matrix is updated (line 5) using (5.4) to calculate the transition rates of the next aggregated truncation  $\mathcal{S}^{(i+1)}$ . Entries of truncated states are removed from the updated transition matrix. Transitions leading to them are re-directed according to the re-entry matrix (Section 6.2). After  $m$  iterations (we started with states of side lengths  $2^m$ ) we have a

standard FSP scheme on the original model tailored to computing an approximation of the stationary distribution.

This way, the refinement algorithm focuses only on those parts of the state-space contributing most to the stationary distribution. For instance, in Figure 6.3 the stationary probability mass mostly concentrates around  $\#S = 200$ . Therefore, states that are further away from this area can be dropped in further refinement. This filtering (line 3 in Algorithm 5) ensures that states contributing significantly to  $\hat{\pi}_{\infty}^{(i)}$  will be kept and refined in the next iteration. The selection of states is done by sorting states in descending order according to their approximate probability mass. This ensures the construction of the smallest possible subset chosen for refinement according to the approximation. Then states are collected until their overall approximate mass is above  $1 - \epsilon$ .

An interesting feature of the aggregation scheme is that the distribution tends to spread out more. This is due to the assumption of a uniform distribution inside macro-states. To gain an intuition, consider a macro-state that encompasses a peak of the stationary distribution. If we re-distribute the actual probability mass inside this macro-state uniformly, a higher probability is assigned to states at the macro-state's border. When plugging such macro-states together, this increased mass away from the peak will increase the mass assigned to adjacent macro-states. This effect is illustrated by the example of a birth-death process in Figure 6.3. Due to this effect, an iterative refinement typically keeps an over-approximation in terms of state-space area. This is a desirable feature since relevant regions are less likely to be pruned due to lumping approximations.

## 6.3 RESULTS

A prototype was implemented in Rust 1.50 and Python 3.8. The linear systems were solved either using Numpy (Harris et al., 2020) for up to 5000 states, or the sparse linear solver as available through Scipy (Virtanen et al., 2020), or the iterative biconjugate gradient stabilized algorithm (Van der Vorst, 1992) (up to 10,000 iterations and absolute tolerance  $10^{-16}$ ).

The examples that we consider in the sequel are typical benchmarks for the analysis of MPMs. For most of them, appropriate Lyapunov functions have been determined using Geobound (Spieler, 2010, 2014). However, the corresponding Lyapunov sets containing at least  $1 - \epsilon_\ell$  of the stationary probability mass are very large for typical choices of  $\epsilon_\ell$  (e.g.  $\epsilon_\ell \in \{0.1, 0.05, 0.001\}$ ). Even for extremely large  $\epsilon_\ell$ , say  $\epsilon_\ell = 0.8$ , the remaining state-space may still be huge (e.g. 15,198 states).

### 6.3.1 Parallel Birth-Death Process

We first examine the algorithm on the simple example of two parallel, uncoupled birth-death processes.

**Model 9** (Parallel Birth-Death Process). *Two uncoupled parallel birth-death processes result in a simple stationary distribution that is given by a product of two Poisson distributions.*

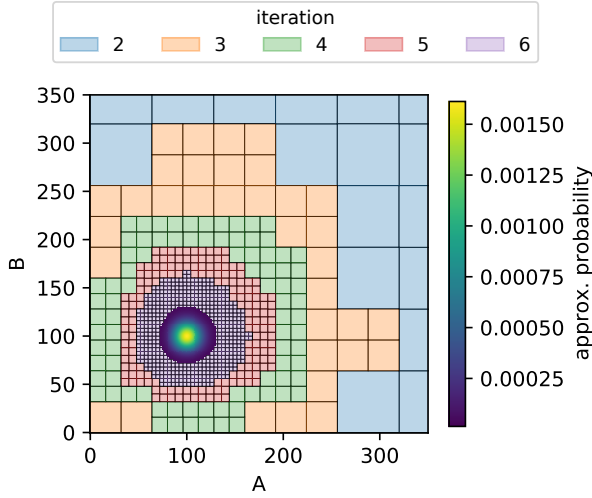
$$\emptyset \xrightarrow{\rho} A, \quad A \xrightarrow{\delta} \emptyset, \quad \emptyset \xrightarrow{\rho} B, \quad B \xrightarrow{\delta} \emptyset.$$

As a parameterization we choose  $\rho = 100$  and  $\delta = 1$ .

For this model, the stationary distribution is known to be the product of two Poisson distributions with rate  $\rho/\delta$ .

According to the Lyapunov analysis with a  $1 \times 10^{-4}$  bound, we fix the initial truncation to a  $70 \times 70$  grid of macro-states with size  $2^7$  in each dimension. This implies 8 iterations of the algorithm to arrive at a truncation with the original granularity. In Figure 6.4, we illustrate the truncations of different iterations. Over the iterations, the covered area decreases, while the aggregation granularity increases. The final truncation distribution approximation is also depicted and covers  $1 - 1.27 \times 10^{-2}$  of the true stationary distribution (cf. Table 6.1).

For this case study, we also compute state-wise bounds on the probabilities conditioned on the truncation as discussed in Section 6.2. In Figure 6.5, we present the difference between upper and lower bound for  $\epsilon = 0.1$ . We observe intervals that are narrowest in the truncation's interior near the distribution's mode. The largest intervals or the largest absolute uncertainty is present



**Figure 6.4:** Truncations of different iterations are layered on top of each other. At higher iterations, truncations cover less area but increase in detail, due to the refinement of macro-states. The final approximation is indicated by its approximate probabilities.

in the boundary states. This indicates, that the specific reentry distribution has little effect on the main approximate stationary mass. More detailed results on the intervals' magnitudes are given in [Table 6.1](#).

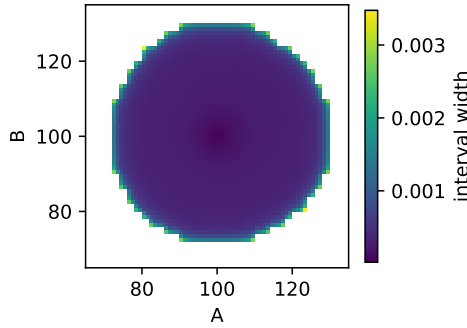
### 6.3.2 Exclusive Switch

The exclusive switch (Barzel and Biham, 2008) has three different modes of operation, depending on the DNA state, i.e. on whether a protein of type one or two is bound to the DNA.

**Model 10** (Exclusive Switch). *The exclusive switch model consists of a promoter region that can express both proteins  $P_1$  and  $P_2$ . Both can bind to the region, suppressing the expression of the other protein.*

	threshold parameter $\epsilon$			
	$1 \times 10^{-1}$	$1 \times 10^{-2}$	$1 \times 10^{-3}$	$1 \times 10^{-4}$
total width	1.2336	$3.09 \times 10^{-2}$	$5.39 \times 10^{-4}$	$8.12 \times 10^{-6}$
max. width	$3.47 \times 10^{-3}$	$9.29 \times 10^{-5}$	$4.04 \times 10^{-7}$	$4.65 \times 10^{-9}$
outside mass	$1.27 \times 10^{-2}$	$1.05 \times 10^{-4}$	$1.05 \times 10^{-6}$	$1.06 \times 10^{-8}$

**Table 6.1:** Results for [Model 9](#) : The characteristics of the lower-upper bound intervals on the conditional probability and the mass not contained in the truncation are given.



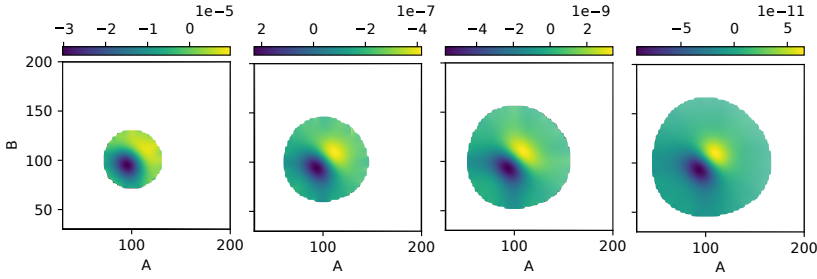
**Figure 6.5:** Results for [Model 9](#) with truncation threshold  $\epsilon = 0.1$ . The difference between the upper and lower bounds on the probability conditioned on the truncation.

*For certain parameterizations, this leads to a bi-modal or even tri-modal behavior.*

$$\begin{aligned}
 D + P_1 &\xrightarrow{\beta} D.P_1, & D.P_1 &\xrightarrow{\gamma_1} D + P_1, \\
 D + P_2 &\xrightarrow{\beta} D.P_2, & D.P_2 &\xrightarrow{\gamma_2} D + P_2, \\
 D.P_1 &\xrightarrow{\rho_1} D.P_1 + P_1, & D.P_2 &\xrightarrow{\rho_2} D.P_2 + P_2, \\
 D &\xrightarrow{\rho_1} D + P_1, & D &\xrightarrow{\rho_2} D + P_2, & P_1 &\xrightarrow{\lambda} \emptyset, & P_2 &\xrightarrow{\lambda} \emptyset
 \end{aligned}$$

We choose parameter values  $\rho_1 = 0.7$ ,  $\rho_2 = 0.6$ ,  $\lambda = 0.02$ ,  $\beta = 0.005$ ,  $\gamma_1 = 0.06$ , and  $\gamma_2 = 0.05$ .





**Figure 6.6:** Errors of the stationary distribution (Model 9) approximations for increasing ( $1 \times 10^{-1}$ ,  $1 \times 10^{-2}$ ,  $1 \times 10^{-3}$ ,  $1 \times 10^{-4}$ ) truncation thresholds.

Since the exclusive switch models mutually exclusive binding of proteins to a single genetic locus, we know a priori that there are exactly three distinct operating modes. In particular are  $D$ ,  $D.P_1$ , and  $D.P_2$  mutually exclusive such that

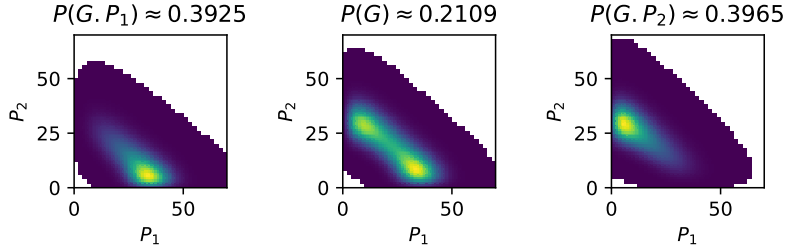
$$X_D(t) + X_{D.P_1}(t) + X_{D.P_2}(t) = 1, \quad \forall t \geq 0.$$

This model characteristic often leads to bi-modal stationary distributions, where one or the other protein is more abundant depending on the genetic state.

Accordingly, we adjust the initial truncation: The state-space for the DNA states is not lumped. Instead we “stack” lumped approximations of the  $P_1$ - $P_2$  plane upon each other. Such special treatment of DNA states is common for such models (Lapin, Mikeev, and Wolf, 2011). Using Lyapunov analysis for threshold 0.001, we fix an initial state-space of  $63 \times 63$  macro-states with size  $2^7$ . Detailed results for different parameters  $\epsilon$  are presented Table A.5. We compute error bounds using a worst-case analysis based on reference solutions provided by Geobound with  $\epsilon_\ell = 0.01$ . We observe a strong decrease in both upper bounds on the total absolute and maximal absolute error in the final iteration. Interestingly, the errors between different thresholds are very close in earlier iterations. This is mainly due to the usage of absolute errors which causes probabilities close to the mode dominate.

↪ page 180

Using Geobound we observe that our final truncation captures the stationary mass very well (cf. Table 6.2 and Figure 6.7). We



**Figure 6.7:** The approximate stationary distribution of the exclusive switch (Model 10) obtained with  $\epsilon = 1 \times 10^{-4}$ .

use the Geobound’s lower bounds with  $\epsilon_\ell = 1 \times 10^{-2}$  and find that the uncovered mass by the aggregation-based truncation is magnitudes lower than  $\epsilon$  or close to it (for  $\epsilon = 0.1$ ). While they capture the mass well, they are much smaller than the Geobound truncation ( $\epsilon_\ell = 0.1$ ) with 16,780 states, regardless of the threshold parameter  $\epsilon$ .

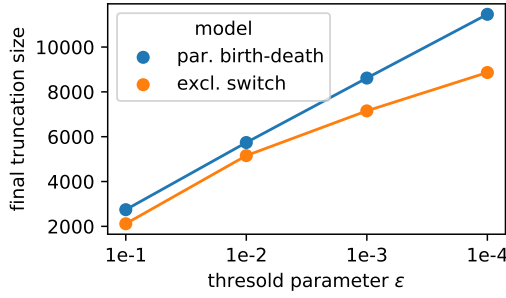
	threshold parameter $\epsilon$			
	$1 \times 10^{-1}$	$1 \times 10^{-2}$	$1 \times 10^{-3}$	$1 \times 10^{-4}$
total width	5.5171	1.5559	$2.89 \times 10^{-2}$	$3.71 \times 10^{-4}$
max. width	$1.58 \times 10^{-1}$	$3.30 \times 10^{-3}$	$3.47 \times 10^{-5}$	$3.84 \times 10^{-7}$
outside mass $\leq$	$1.52 \times 10^{-1}$	$1.29 \times 10^{-3}$	$2.02 \times 10^{-5}$	$2.72 \times 10^{-7}$

**Table 6.2:** Results for Model 10: The characteristics of the lower-upper bound intervals on the conditional probability and the upper bound on mass not contained in the truncation are given.

In Figure 6.8, we show the effect of the threshold parameter  $\epsilon$  on the size of the final truncation. We observe a roughly linear increase in size with an exponential decrease of  $\epsilon$ .

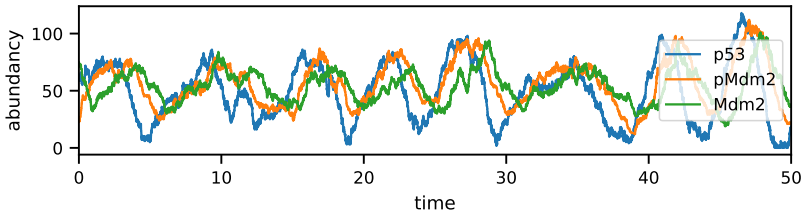
### 6.3.3 p53 Oscillator

We now consider a model of the interactions of the tumor suppressor p53 (Geva-Zatorsky et al., 2006). The system describes the negative feedback loop between p53 and the oncogene Mdm2.



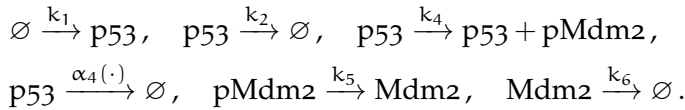
**Figure 6.8:** The sizes of the final truncation vs. the threshold parameter  $\epsilon$  (Model 2 and Model 10).

Species pMdm2 models a precursor to Mdm2. This model is particularly interesting due to its complex three-dimensional oscillatory behavior. The model is ergodic with a unique stationary distribution (Gupta, Briat, and Khammash, 2014).



**Figure 6.9:** A sample trajectory of Model 11 illustrating the oscillatory long-run behavior.

**Model 11** (p53 Oscillator).



The non-polynomial degradation reaction rate

$$\alpha_4(x) = k_3 x_{\text{Mdm2}} \frac{x_{\text{p53}}}{x_{\text{p53}} + k_7}.$$

The parameterization based on (Ale, Kirk, and Stumpf, 2013) is  $k_1 = 90$ ,  $k_2 = 0.002$ ,  $k_3 = 1.7$ ,  $k_4 = 1.1$ ,  $k_5 = 0.93$ ,  $k_6 = 0.96$ , and  $k_7 = 0.01$ .

With the exception of propensity function  $\alpha_4$ , we can compute the transition rates  $\bar{\alpha}_i$  using the Faulhaber formulae, as discussed in [Chapter 5](#). We consider  $\alpha_4$  separately, because it is non-polynomial and therefore, we have to make an approximation. The fraction occurring in the non-linear propensity function  $\alpha_4$  can roughly be characterized as an activation function: Due to the low value of parameter  $k_7 = 0.01$  we can approximate

*Note, that  $\sum_{i=0}^n i/(i+k_7)$  can be solved analytically. However, the approximation presented above is much simpler to compute.*

$$\frac{x_{p53}}{x_{p53} + k_7} \approx \begin{cases} 0 & \text{if } x_{p53} = 0 \\ 1 & \text{otherwise} \end{cases}.$$

We use this approximation at the coarser levels of aggregation to efficiently compute the approximate transition rate  $\bar{\alpha}_4$ . At the finest granularity we switch back to exact propensity function  $\alpha_4$ .

We now derive Lyapunov-sets for the p53 oscillator case study ([Model 11](#)). Let the Lyapunov function

$$g(x) = 120x_{p53} + 0.2x_{pMdm2} + 0.1x_{Mdm2}. \quad (6.1)$$

Then the drift

$$\begin{aligned} d(x) &= -\frac{k_3 x_{Mdm2} x_{p53}}{x_{p53} + k_7} - 0.1k_6 x_{Mdm2} + 120k_1 \\ &\quad - 120k_2 x_{p53} + 0.2k_4 x_{p53} - 0.1k_5 x_{pMdm2} \\ &= -\frac{204x_{Mdm2} x_{p53}}{x_{p53} + 0.01} - 0.096x_{Mdm2} - 0.02x_{p53} \\ &\quad - 0.0093x_{pMdm2} + 10,800. \end{aligned} \quad (6.2)$$

Clearly,  $c = \sup_{x \in S} d(x) = 10,800$ . In particular, the supremum  $c$  is at the origin since all non-constant terms are negative. The slowest rate of decrease for (6.2) is  $x_{p53}$  with  $x_{Mdm2} = x_{pMdm2} = 0$ . We are content with a superset of a Lyapunov set (2.31) for some threshold  $\epsilon_\ell$ . Therefore taking (2.31), we can solve the inequality

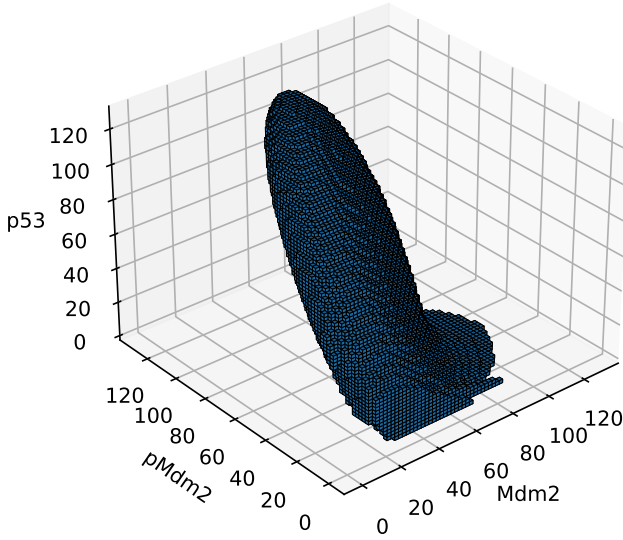
$$\frac{\epsilon_\ell}{c} (c - 0.02x_{p53}) > \epsilon_\ell - 1$$

for  $x_{p53}$  and

$$\frac{c}{0.02\epsilon_\ell} < x_{p53}.$$

Therefore

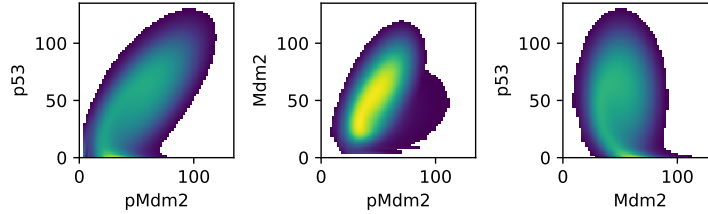
$$\pi_{\infty} \left( \left\{ x \in \mathcal{S} \mid \frac{c}{0.2\epsilon_{\ell}} < \|x\| \right\} \right) > 1 - \epsilon_{\ell} .$$



**Figure 6.10:** The final truncation at original granularity derived for the p53 oscillator.

Due to the exponential increase stemming from the three-dimensional nature of this model, we only evaluated with parameter  $\epsilon = 0.1$ . According to the Lyapunov analysis shown above, the area covered by an  $6 \times 6 \times 6$  macro-states with size  $2^{20}$ , covers 0.9 of stationary mass. A truncation of this same area would consist of 226,492,416 states instead of the 216 macro-states. The model has a striking oscillatory behavior (cf. [Figure 6.9](#)) that is reflected in its stationary distribution. This feature is well-captured in the approximate distribution, where the oscillatory behavior leads to a complex stationary distribution (cf. [Figure 6.11](#)). This distri-

bution leads to a non-trivial truncation (357,488 states) which is tailored to the main stationary mass (Figure 6.10).



**Figure 6.11:** The approximate marginal distributions of the stationary distribution based on the truncation derived with  $\epsilon = 0.1$ .

#### 6.4 CONCLUSION

State-of-the-art methods for numerically calculating the stationary distribution of Markovian population models rely on coarse truncations of irrelevant parts of large or infinite discrete state-spaces. These truncations are either obtained from the stationary statistical moments of the process or from Lyapunov theory. They are limited in shape because these methods do not take into account the detailed steady-state flow within the truncated state-space but only consider the average drift or stationary moments.

Here, we propose a method to find a tight truncation that is not limited in its shape and iteratively optimizes the set based on numerically cheap solutions of abstract intermediate models. It captures the main portion of probability mass even in the case of complex behaviors efficiently. In particular, the method represents another option, where Lyapunov analysis leads to forbiddingly large truncations.

# 7

---

## ANALYSIS UNDER TERMINAL CONSTRAINTS

---

Many tasks, such as the analysis of rare events or the inference of agent counts under partial observations naturally introduce terminal constraints on the system. In these cases, the system's initial state is known, as well as the system's (partial) state at a later time-point. The probabilities corresponding to this so-called *bridging problem* are often referred to as *bridging probabilities* (Golightly and Sherlock, 2019; Golightly and Wilkinson, 2011). For instance, if the exact, full state of the process  $X_t$  has been observed at time  $t = 0$  and  $t = T$ , the bridging distribution is given by

$$\Pr(X_t = x \mid X_0 = x_0, X_T = x_g)$$

for all states  $x$  and times  $t \in [0, T]$ . Often, the condition is more complex, such that in addition to an initial distribution, a terminal distribution is present. Such problems typically arise in a Bayesian setting, where the a priori behavior of a system is filtered such that the posterior behavior is compatible with noisy, partial observations (Broemeling, 2017; Huang et al., 2016). For example, time-series data of protein levels is available while the mRNA concentration is not (Adan et al., 2017; Huang et al., 2016). In such a scenario our method can be used to identify a good truncation to analyze the probabilities of mRNA levels.

Bridging probabilities also appear in the context of rare events. Here, the rare event is the terminal constraint because we are only interested in paths containing the event. Typically researchers

have to resort to Monte Carlo simulations in combination with variance reduction techniques in such cases (Daigle Jr et al., 2011; Kuwahara and Mura, 2008).

Efficient numerical approaches that are not based on sampling or ad-hoc approximations have rarely been developed.

Here, we combine state-of-the-art truncation strategies based on a forward analysis (Andreychenko et al., 2011; Lapin, Mikeev, and Wolf, 2011) with a refinement approach that starts from an abstract MPM with lumped states. We base this lumping on a grid-like partitioning of the state-space. Throughout a lumped state, we assume a uniform distribution that gives an efficient and convenient abstraction of the original MPM. Note that the lumping does not follow the classical paradigm of Markov chain lumpability (Buchholz, 1994) or its variants (Dayar and Stewart, 1997). Instead of an approximate block structure of the transition-matrix used in that context, we base our partitioning on a segmentation of the molecule counts. Moreover, during the iterative refinement of our abstraction, we identify those regions of the state-space that contribute most to the bridging distribution. In particular, we refine those lumped states that have a bridging probability above a certain threshold  $\delta$  and truncate all other macro-states. This way, the algorithm learns a truncation capturing most of the bridging probabilities. This truncation provides guaranteed lower bounds because it is at the granularity of the original model.

## 7.1 RELATED WORK

**TERMINAL CONSTRAINTS** The problem of endpoint constrained analysis occurs in the context of Bayesian estimation (Särkkä, 2013). For MPMs, this problem has been addressed by Huang et al. (2016) using moment closure approximations and by Wildner and Koeppl (2019) further employing variational inference. Golightly and Sherlock modified stochastic simulation algorithms to approximatively augment generated trajectories (Golightly and Sherlock, 2019). Since a statistically exact augmentation is only possible for few simple cases, diffusion approximations (Golightly and Wilkinson, 2005) and moment approximations



(Milner, Gillespie, and Wilkinson, 2013) have been employed. Such approximations, however, do not give any guarantees on the approximation error and may suffer from numerical instabilities (Schnoerr, Sanguinetti, and Grima, 2014).

Another manifestation of the bridging problem is found during the estimation of first passage times and rare event analysis. Approaches for first-passage times are often of heuristic nature (Bortolussi and Lanciani, 2014; Hayden, Stefanek, and Bradley, 2012; Schnoerr et al., 2017). Rigorous approaches yielding guaranteed bounds are currently limited by the performance of state-of-the-art optimization software (Backenköhler, Bortolussi, and Wolf, 2020).

**RARE EVENT PROBABILITIES** The analysis of rare events is a special case of the terminal constrained Markov processes considered in this chapter. Most methods for the estimation of rare event probabilities rely on importance sampling (Daigle Jr et al., 2011; Kuwahara and Mura, 2008). For other queries, alternative variance reduction techniques such as control variates are available (Backenköhler, Bortolussi, and Wolf, 2019). Apart from sampling-based approaches, dynamic finite-state projections have been employed by Mikeev, Sandmann, and Wolf (2013), but are lacking automated truncation schemes.

**STATE-SPACE TRUNCATION** The analysis of countably infinite state-spaces is often handled by a pre-defined truncation (Kwiatkowska, Norman, and Parker, 2011). Sophisticated state-space truncations for the (unconditioned) forward analysis have been developed to give lower bounds and rely on a trade-off between computational load and tightness of the bound (Andreychenko et al., 2011; Henzinger, Mateescu, and Wolf, 2009; Lapin, Mikeev, and Wolf, 2011; Mikeev et al., 2013; Munsky and Khammash, 2006).

**VERIFICATION** The problem considered in this chapter can be interpreted as an instance of a model checking problem (Bentriou, Ballarini, and Cournède, 2021; Brim, Češka, and Šafránek, 2013).

Reachability – relevant in the context of probabilistic verification (Bortolussi and Lanciani, 2014; Neupane et al., 2019) – is a bridging problem where the endpoint constraint is the visit of a set of goal states. Backward probabilities are commonly used to compute reachability likelihoods (Amparore and Donatelli, 2013; Zapreev and Katoen, 2006). Approximate techniques for reachability, based on moment closure and stochastic approximation, have also been developed in (Bortolussi and Lanciani, 2014; Bortolussi, Lanciani, and Nenzi, 2018), but lack error guarantees.

**HIDDEN MARKOV MODELS** There is also a conceptual similarity between computing bridging probabilities and the forward-backward algorithm for computing state-wise posterior marginals in hidden Markov models (HMMs) (Rabiner and Juang, 1986). Like MPMs, HMMs are a generative model that can be conditioned on observations. We only consider two observations (initial and terminal state) that are not necessarily noisy but the forward and backward probabilities admit the same meaning.

## 7.2 BACKWARDS PROBABILITIES

Let  $x_g \in \mathcal{S}$  be a fixed goal state. Given the terminal constraint

$$\Pr(X_T = x_g) = 1 \text{ for some } T \geq 0,$$

we are interested in the so-called backward probabilities

$$\beta(x_i, t) = \Pr(X_T = x_g \mid X_t = x_i), \quad t \leq T. \quad (7.1)$$

Note that  $\beta(\cdot, t)$  is a function of the conditional event and thus is no probability distribution over the state-space. Instead  $\beta(\cdot, t)$  gives the reaching probabilities for all states over the time span of  $[t, T]$ . To compute these probabilities, we can employ the Kolmogorov backward equation

$$\frac{d}{dt}\beta(t) = Q\beta(t), \quad (7.2)$$

where we use the same vectorization to construct  $\beta(t)$  as we used for  $\pi(t)$ . The above equation is integrated backwards in time and

yields the reachability probability for each state  $x_i$  and time  $t < T$  of ending up in  $x_g$  at time  $T$ . Similar to the CME (2.18), we can state a backward chemical master equation

$$\frac{d\beta}{dt}(x, t) = \sum_{j=1}^{n_R} (\beta(x, t) - \beta(x + v_j, t)) \alpha_j(x). \quad (7.3)$$

The state-space of many MPMs, even simple ones, is countably infinite. In this case, we have to truncate the state-space to a *reasonable* finite subset. The choice of this truncation heavily depends on the goal of the analysis. If one is interested in the most “common” behavior, for example, a dynamic mass-based truncation scheme is most appropriate (Mikeev and Sandmann, 2019). Such a scheme truncates states with small probability during the numerical integration. However, common mass-based truncation schemes are not as useful for the bridging problem. This is because trajectories that meet the specific terminal constraints can be far off the main bulk of the probability mass. We solve this problem by a state-space lumping in connection with an iterative refinement scheme.

### 7.3 BRIDGING DISTRIBUTION

The process’ probability distribution given both initial and terminal constraints is formally described by the conditional probabilities

$$\gamma(x_i, t) = \Pr(X_t = x_i \mid X_0 = x_0, X_T = x_g), \quad (7.4)$$

for  $0 \leq t \leq T$ . for fixed initial state  $x_0$  and terminal state  $x_g$ . We call these probabilities the *bridging probabilities*. It is straightforward to see that  $\gamma$  admits the factorization

$$\gamma(x_i, t) = \pi(x_i, t) \beta(x_i, t) / \pi(x_g, T) \quad (7.5)$$

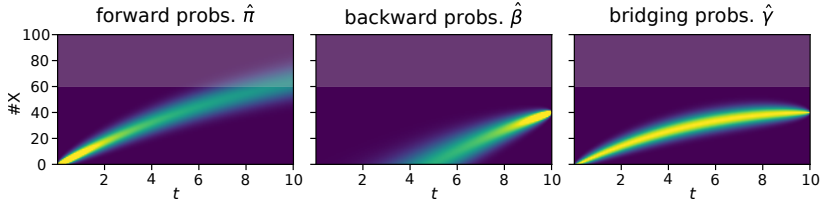
due to the Markov property. The normalization factor, given by the reachability probability

$$\pi(x_g, T) = \beta(x_0, 0),$$

ensures that  $\gamma(\cdot, t)$  is a distribution for all time points  $t \in [0, T]$ . We call each  $\gamma(\cdot, t)$  a *bridging distribution*. From the Kolmogorov equations (2.17) and (7.2) we can obtain both the forward probabilities  $\pi(\cdot, t)$  and the backward probabilities  $\beta(\cdot, t)$  for  $t < T$ . ↪ page 17

We can easily extend this procedure to deal with hitting times constrained by a finite time-horizon by making the goal state  $x_g$  absorbing.

↪ page 20 **EXAMPLE** In Figure 7.1 we plot the forward, backward, and bridging probabilities for Model 2. The probabilities are computed on a  $[0, 100]$  state-space truncation. The approximate forward solution  $\hat{\pi}$  shows how the probability mass drifts upwards towards the stationary distribution  $\text{Poisson}(100)$ . The backward probabilities are highest for states below the goal state  $x_g = 40$ . This is expected because upwards drift makes reaching  $x_g$  more probable for “lower” states. Finally, the approximate bridging distribution  $\hat{\gamma}$  can be recognized to be proportional to the product of forward  $\hat{\pi}$  and backward probabilities  $\hat{\beta}$ . ◇



**Figure 7.1:** Forward, backward, and bridging probabilities for Model 2 with initial constraint  $X_0 = 0$  and terminal constraint  $X_{10} = 40$  on a truncated state-space. Probabilities over 0.1 in  $\hat{\pi}$  and  $\hat{\beta}$  are given full intensity for visual clarity. The lightly shaded area ( $\geq 60$ ) indicates a region being more relevant for the forward than for the bridging probabilities.

#### 7.4 BRIDGE TRUNCATION VIA LUMPING APPROXIMATION

We first discuss the truncation of countably infinite state-spaces to analyze backward and forward probabilities (Section 7.4.1). To identify effective truncations we employ a lumping scheme.

Finally, in [Section 7.4.2](#) we present an iterative refinement algorithm yielding a suitable truncation for the bridging problem.

#### 7.4.1 Finite State Projection

Even in simple models such as a birth-death Process ([Model 2](#)), the reachable state-space is countably infinite. Direct analyzes of backward ([7.1](#)) and forward equations ([2.16](#)) are often infeasible.  $\hookrightarrow$  page [17](#) Instead, the integration of these differential equations requires working with a finite subset of the infinite state-space (Munsky and Khammash, 2006). If states are truncated, their incoming transitions from states that are not truncated can be re-directed to a *sink state*. The accumulated probability in this sink state is then used as an error estimate for the forward integration scheme. Consequently, many truncation schemes, such as dynamic truncations (Andreychenko et al., 2011), aim to minimize the amount of “lost mass” of the forward probability. We use the same truncation method but base the truncation on bridging probabilities rather than the forward probabilities.

#### 7.4.2 Iterative Refinement Algorithm

The iterative refinement algorithm ([Algorithm 6](#)) starts with a set of large macro-states that are iteratively refined, based on approximate solutions to the bridging problem. We start by constructing square macro-states of size  $2^m$  in each dimension for some  $m \in \mathbb{N}$  such that they form a large-scale grid  $\mathcal{S}^{(0)}$ . Hence, each initial macro-state has a volume of  $(2^m)^{n_s}$ . This choice of grid size is convenient because we can halve states in each dimension. Moreover, this choice ensures that all states have equal volume and we end up with states of volume  $2^0 = 1$  which is equivalent to a truncation of the original non-lumped state-space.

An iteration of the state-space refinement starts by computing both the forward and backward probabilities ([line 2](#) and [line 3](#)) via integration of ([2.17](#)) and ([7.2](#)), respectively, using the lumped  $\hat{Q}$ -matrix. Based on the resulting approximate forward  $\hookrightarrow$  page [17](#)

and backward probabilities, we compute an approximation of the bridging distributions (line 4). This is done for each time-point in an equispaced grid on  $[0, T]$ . The time grid granularity is a hyper-parameter of the algorithm. If the grid is too fine, the memory overhead of storing backward  $\hat{\beta}^{(i)}$  and forward solutions  $\hat{\pi}^{(i)}$  increases. We denote the approximations with a hat (e.g.  $\hat{\pi}$ ) rather than a bar (e.g.  $\bar{\pi}$ ) to indicate that not only the lumping approximation but also a truncation is applied and similarly for the Q-matrix. If, on the other hand, the granularity is too low, too much of the state-space might be truncated. Based on a threshold parameter  $\delta > 0$  states are either removed or split (line 7), depending on the mass assigned to them by the approximate bridging probabilities  $\hat{\gamma}_t^{(i)}$ . A state can be split by the `split`-function which halves the state in each dimension. Otherwise, it is removed. Thus, each macro-state is either split into  $2^{n_s}$  new states or removed entirely. The result forms the next lumped state-space  $\mathcal{S}^{(i)}$ . The Q-matrix is adjusted (line 10) such that transition rates for  $\mathcal{S}^{(i)}$  are calculated according to (5.4). Entries of truncated states are removed from the transition matrix. Transitions leading to them are re-directed to a sink state (Section 7.4.1). After  $m$  iterations (we started with states of side lengths  $2^m$ ) we have a standard FSP scheme on the original model tailored to computing an approximation of the bridging distribution.

In Figure 7.2 we give a demonstration of how Algorithm 6 works to refine the state-space iteratively. Starting with an initial lumped state-space  $\mathcal{S}^{(0)}$  covering a large area of the state-space, repeated evaluations of the bridging distributions are performed. After five iterations the remaining truncation includes all states that significantly contribute to the bridging probabilities over the times  $[0, T]$ .

It is important to realize that determining the most relevant states is *the* main challenge. The above algorithm solves this problem by considering only those parts of the state-space that contribute most to the bridging probabilities. The truncation is tailored to this condition and might ignore regions that are likely in the unconditioned case. For instance, in Figure 7.1 the bridging

**Algorithm 6:** Iterative refinement for the bridging problem

---

**input** : initial partitioning  $\mathcal{S}^{(0)}$ , truncation threshold  $\delta$   
**output**: approximate bridging distribution  $\hat{\gamma}$

```

1 for  $i = 1, \dots, m$  do
2    $\hat{\pi}_t^{(i-1)} \leftarrow$  approximate forward equation on  $\mathcal{S}^{(i)}$ ;
3    $\hat{\beta}_t^{(i-1)} \leftarrow$  approximate backward equation on  $\mathcal{S}^{(i)}$ ;
4    $\hat{\gamma}_t^{(i)} \leftarrow \hat{\beta}_t^{(i)} \hat{\pi}_t^{(i)} / \hat{\pi}(x_g, T)$ ;    /* approx. bridging */
5    $\mathcal{S}^{(i)} \leftarrow \emptyset$ ;
6   foreach  $\bar{x} \in \mathcal{S}^{(i)}$  do
7     if  $\exists t. \hat{\gamma}_t^{(i)}(\bar{x}) \geq \delta$ ;                /* refinement */
8     then
9        $\mathcal{S}^{(i)} \leftarrow \mathcal{S}^{(i)} \cup \text{split}(\bar{x})$ ;
10  update  $\hat{Q}$ -matrix;
11 return  $\hat{\gamma}^{(i)}$ ;

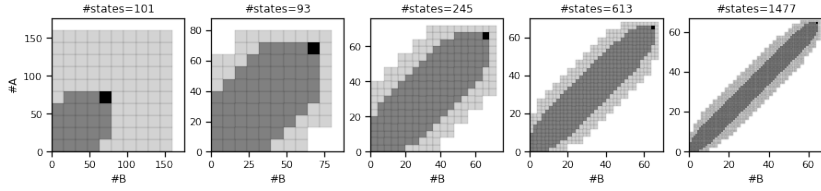
```

---

probabilities mostly remain below a population threshold of  $\#X = 60$  (as indicated by the lighter/darker coloring), while the forward probabilities mostly exceed this bound. Hence, in this example a significant portion of the forward probabilities  $\hat{\pi}_t^{(i)}$  is captured by the sink state. However, the condition in [line 7](#) of [Algorithm 6](#) ensures that states contributing significantly to  $\hat{\gamma}_t^{(i)}$  will be kept and refined in the next iteration.

## 7.5 RESULTS

We present four examples in this section to evaluate our proposed method. A prototype was implemented in Python 3.8. For numerical integration we used the Scipy implementation (Virtanen et al., 2020) of the implicit method based on backward-differentiation formulas (Byrne and Hindmarsh, 1975). The analysis as a Jupyter notebook is made available online (Backenköhler, 2020).



**Figure 7.2:** The state-space refinement algorithm on two parallel unit-rate arrival processes. The bridging problem from  $(0,0)$  to  $(64,64)$  and  $T = 10$  and truncation threshold  $\delta = 5 \times 10^{-3}$ . States with a bridging probability below  $\delta$  are light grey. The macro-state containing the goal state is marked in black. The initial macro-states are of size  $16 \times 16$ .

### 7.5.1 Bounding Rare Event Probabilities

We consider a simple model of two parallel Poisson processes describing the production of two types of agents. The corresponding probability distribution has Poisson product form at all time points  $t \geq 0$  and hence we can compare the accuracy of our numerical results with the exact analytic solution. We use the proposed approach to compute lower bounds for rare event probabilities. These bounds are rigorous up to the approximation error of the numerical integration scheme. However, the forward solution could be replaced by an uniformization approach for a more rigorous error control.

↪ page 14

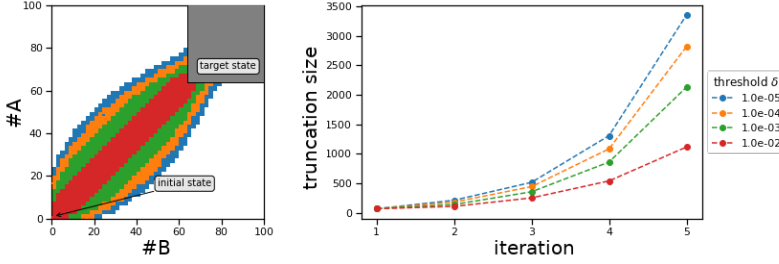
**Model 12** (Parallel Poisson Processes). *The model consists of two parallel independent Poisson processes with unit rates.*

$$\emptyset \xrightarrow{1} A \quad \text{and} \quad \emptyset \xrightarrow{1} B.$$

*The initial condition  $X_0 = (0,0)$  holds with probability one. After  $t$  time units each species abundance is Poisson distributed with rate  $\lambda = t$ .*

We consider the final constraint of reaching a state where both processes exceed a threshold of 64 at time 20. Without prior knowledge, a reasonable truncation would have been  $160 \times 160$ . But our analysis shows that just 20% of the states are necessary to capture over 99.6% of the probability mass reaching the target





**Figure 7.3:** State-space truncation for varying values of the threshold parameter  $\delta$ : Two parallel Poisson processes under terminal constraints  $X_{20}^{(A)} \geq 64$  and  $X_{20}^{(B)} \geq 64$ . The initial macro-states are  $16 \times 16$  such that the final states are regular micro states.

	threshold $\delta$			
	$1 \times 10^{-2}$	$1 \times 10^{-3}$	$1 \times 10^{-4}$	$1 \times 10^{-5}$
$ \mathcal{S}^{(m)} $	1154	2354	3170	3898
$\sum_m  \mathcal{S}^{(m)} $	2074	3546	4586	5450
estimate	$8.88 \times 10^{-30}$	$1.85 \times 10^{-29}$	$1.86 \times 10^{-29}$	$1.86 \times 10^{-29}$
rel. error	$5.22 \times 10^{-1}$	$3.66 \times 10^{-3}$	$3.74 \times 10^{-5}$	$9.52 \times 10^{-8}$

**Table 7.1:** Estimated reachability probabilities based on varying truncation thresholds  $\delta$ : The true probability is  $1.8625 \times 10^{-29}$ . We also report the size of the final truncation  $|\mathcal{S}^{(m)}|$  and the accumulated size of all truncations during refinement iterations (overall states)  $\sum_m |\mathcal{S}^{(m)}|$ .

event (cf. Table 7.1). Decreasing the threshold  $\delta$  leads to a larger set of states retained after truncation as more of the bridging distribution is included (cf. Figure 7.3). We observe an increase in truncation size that is approximately logarithmic in  $\delta$ , which, in this example, indicates robustness of the method with respect to the choice of  $\delta$ .

**COMPARISON TO OTHER METHODS** The truncation approach that we apply is similar to the one used by Mikeev, Sandmann, and Wolf (2013) for rare event estimation. However, they used a

given linearly biased MPM model to obtain a truncation. A general strategy to compute an appropriate biasing was not proposed. It is possible to adapt our truncation approach to the dynamic scheme in Mikeev, Sandmann, and Wolf (2013) where states are removed in an on-the-fly fashion during numerical integration.

*The goal is not treated as a single state. Otherwise, it consisted of 24,130 states.*

A finite state-space truncation covering the same area as the initial lumping approximation would contain 25,600 states. The standard approach would be to build up the entire state-space for such a model (Kwiatkowska, Norman, and Parker, 2011). Even using a conservative truncation threshold  $\delta = 1 \times 10^{-5}$ , our method yields an accurate estimate using only about a fifth (5450) of this accumulated over all intermediate lumped approximations.

### 7.5.2 Mode Switching

*This often is an instance of rare-event analysis.*

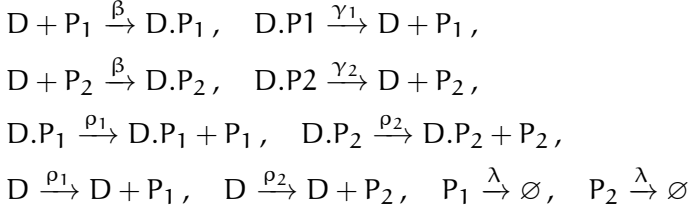
Mode switching occurs in models exhibiting *multi-modal* behavior when a trajectory traverses a potential barrier from one mode to another. Often, mode switching is a rare event and occurs in the context of gene regulatory networks where a mode is characterized by the set of genes being currently active (Loinger et al., 2007). Similar dynamics also commonly occur in queuing models where a system may for example switch its operating behavior stochastically if traffic increases above or decreases below certain thresholds. Using the presented method, we can get both a qualitative and quantitative understanding of switching behavior without resorting to Monte-Carlo methods such as importance sampling.

#### *Exclusive Switch*

The exclusive switch (Barzel and Biham, 2008) has three different modes of operation, depending on the DNA state, i.e. on whether a protein of type one or two is bound to the DNA.

**Model 13** (Exclusive Switch). *The exclusive switch model consists of a promoter region that can express both proteins  $P_1$  and  $P_2$ . Both can bind to the region, suppressing the expression of the other protein.*

For certain parameterizations, this leads to a bi-modal or even tri-modal behavior.



The parameter values are  $\rho = 1 \times 10^{-1}$ ,  $\lambda = 1 \times 10^{-3}$ ,  $\beta = 1 \times 10^{-2}$ , and  $\gamma = 8 \times 10^{-3}$ .

Since we know a priori of the three distinct operating modes, we adjust the method slightly: The state-space for the DNA states is not lumped. Instead we “stack” lumped approximations of the  $P_1$ - $P_2$  phase space upon each other. Special treatment of DNA states is common for such models (Lapin, Mikeev, and Wolf, 2011).

To analyze the switching, we choose the transition from

$$x_1 = (32, 0, 0, 0, 1) \quad \text{to} \quad x_2 = (0, 32, 0, 1, 0)$$

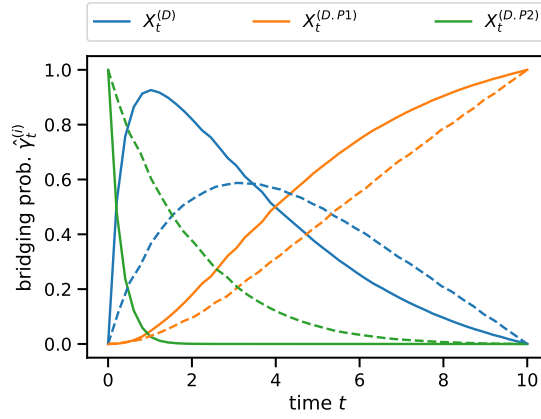
Variable order:  $P_1$ ,  
 $P_2$ ,  $D$ ,  $D.P_1$ ,  $D.P_2$ .

over the time interval  $t \in [0, 10]$ . The initial lumping scheme covers up to 80 molecules of  $P_1$  and  $P_2$  for each mode. Macro-states have size  $8 \times 8$  and the truncation threshold is  $\delta = 1 \times 10^{-4}$ .

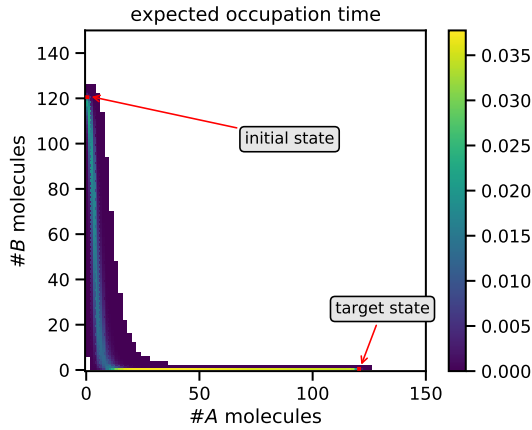
In the analysis of biological switches, not only the switching probability but also the switching dynamics is a central part of understanding the underlying biological mechanisms. In [Figure 7.4](#), we therefore plot the time-varying probabilities of the gene state conditioned on the mode. We observe a rapid unbinding of  $P_2$ , followed by a slow increase of the binding probability for  $P_1$ . These dynamics are already qualitatively captured by the first lumped approximation (dashed lines).

### Toggle Switch

Next, we apply our method to a toggle switch model exhibiting non-polynomial rate functions. This well-known model considers two proteins A and B inhibiting the production of the respective other protein (Lipshtat et al., 2006).

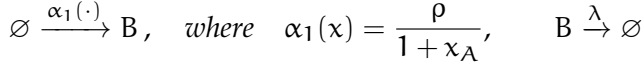
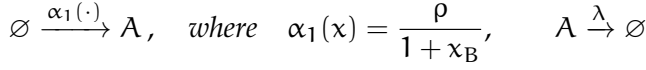


**Figure 7.4:** Mode probabilities of the exclusive switch bridging problem over time for the first lumped approximation (dashed lines) and the final approximation (solid lines) with constraints  $X_0 = (32, 0, 0, 1, 0)$  and  $X_{10} = (0, 32, 0, 0, 1)$ .



**Figure 7.5:** The expected occupation time (excluding initial and terminal states) for the switching problem of the toggle switch using Hill-type functions. The bridging problem is from initial  $(0, 120)$  to a first passage of  $(120, 0)$  in  $t \in [0, 10]$ .

**Model 14** (Toggle Switch using Hill functions). We have population types A and B with the following reactions and reaction rates.



The parameterization is  $\rho = 10, \lambda = 0.1$ .

Due to the non-polynomial rate functions  $\alpha_1$  and  $\alpha_2$ , the transition rates between macro-states are approximated by using the continuous integral

$$\bar{\alpha}_1(\bar{x}) \approx \int_{a-0.5}^{b+0.5} \frac{\rho}{1+x} dx = \rho (\log(b+1.5) - \log(a+0.5))$$

for a macro-state  $\bar{x} = \{a, \dots, b\}$ .

We analyze the switching scenario from  $(0, 120)$  to the first visit of state  $(120, 0)$  up to time  $T = 10$ . The initial lumping scheme covers up to 352 molecules of A and B and macro-states have size  $32 \times 32$ . The truncation threshold is  $\delta = 1 \times 10^{-4}$ . The resulting truncation is shown in [Figure 7.5](#). It also illustrates the kind of insights that can be obtained from the bridging distributions. For an overview of the switching dynamics, we look at the expected occupation time under the terminal constraint of having entered state  $(120, 0)$ . Letting the corresponding hitting time be

$$\tau = \inf\{t \geq 0 \mid X_t = (120, 0)\},$$

the expected occupation time for some state  $x$  is

$$\mathbb{E} \left( \int_0^\tau 1_{=x}(X_t) dt \mid \tau \leq 10 \right).$$

We observe that in this example the switching behavior seems to be asymmetrical. The main mass seems to pass through an area where initially a small number of A molecules is produced followed by a total decay of B molecules.

### 7.5.3 Recursive Bayesian Estimation

We now turn to the method's application in recursive Bayesian estimation. This is the problem of estimating the system's past, present, and future behavior under given observations. Thus, the [MPM](#) becomes a hidden Markov model ([HMM](#)). The observations in such models are usually noisy, meaning that we cannot infer the system state with certainty.

This estimation problem entails more general distributional constraints on terminal  $\beta(\cdot, T)$  and initial  $\pi(\cdot, 0)$  distributions than the point mass distributions considered up until now. We can easily extend the forward and backward probabilities to more general initial distributions and terminal distributions  $\beta(T)$ . For the forward probabilities we get

$$\pi(x_i, t) = \sum_j \Pr(X_t = x_i \mid X_0 = x_j) \pi(x_j, 0), \quad (7.6)$$

and similarly the backward probabilities are given by

$$\beta(x_i, t) = \sum_j \Pr(X_T = x_j \mid X_t = x_i) \beta_T(x_j). \quad (7.7)$$

We apply our method to an susceptible-exposed-infected-removed ([SEIR](#)) model. This is widely used to describe the spreading of an epidemic such as the current COVID-19 outbreak (Großmann, Backenköhler, and Wolf, 2020; He, Peng, and Sun, 2020). Temporal snapshots of the epidemic spread are mostly only available for a subset of the population and suffer from inaccuracies of diagnostic tests. Bayesian estimation can then be used to infer the spreading dynamics given uncertain temporal snapshots.

**Model 15** (Epidemic Spread). *A population of susceptible individuals can contract a disease from infected agents. In this case, they are exposed, meaning they will become infected but cannot yet infect others. After being infected, individuals change to the removed state. The mass-action reactions are as follows.*



The parameter values are  $\lambda = 0.5$ ,  $\mu = 3$ ,  $\rho = 3$ . Due to the stoichiometric invariant  $X_t^{(S)} + X_t^{(E)} + X_t^{(I)} + X_t^{(R)} = \text{const.}$ , we can eliminate  $R$  from the system.

We consider the following scenario: There are  $N$  individuals. We know that initially ( $t = 0$ ) one individual is infected and the rest is susceptible. At time  $t = 0.3$  all individuals are tested for the disease. The test, however, only identifies infected individuals with probability  $p_{tp} = 0.99$ . Moreover, the probability of a false positive is  $p_{fp} = 0.05$ . The random variable  $Y_t$  models the measurement likelihood at time  $t$ . Based on the description above

*The false positive probability is the same for all non-infected individuals.*

$$\begin{aligned} \Pr(Y_t = \hat{n}_I \mid X_t^{(I)} = n_I) \\ = \sum_{k=0}^{\hat{n}_I} B(k; n_I, p_{tp}) B(\hat{n}_I - k; N - n_I, p_{fp}) \end{aligned}$$

where  $B$  is the binomial probability mass function and

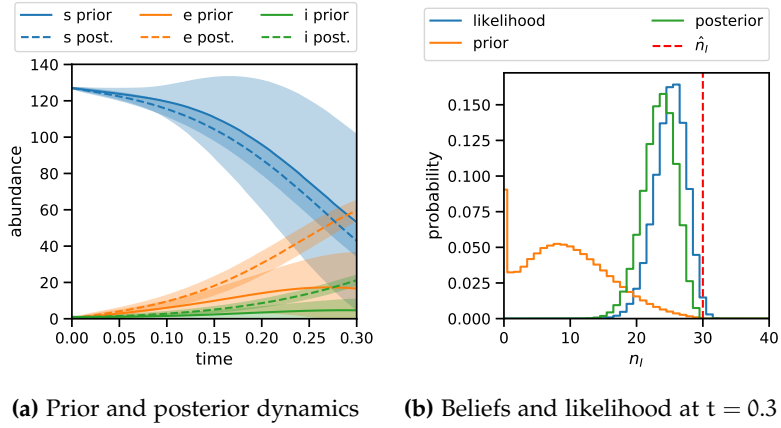
$$B(k; n, p) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

We like to identify the distribution given both the initial state and the measurement at time  $t = 0.3$ . In particular, we want to infer the distribution over the latent counts of  $S$  and  $E$  by *recursive Bayesian estimation*.

The posterior for  $n_I$  infected individuals at time  $t$ , given measurement  $Y_t = \hat{n}_I$  can be computed using Bayes' rule

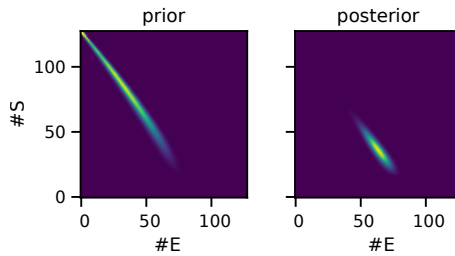
$$\begin{aligned} \Pr(X_t^{(I)} = n_I \mid Y_t = \hat{n}_I, X_0 = x_0) \\ \propto \Pr(Y_t = \hat{n}_I \mid X_t^{(I)} = n_I) \Pr(X_t^{(I)} = n_I \mid X_0 = x_0). \quad (7.8) \end{aligned}$$

This problem is an extension of the bridging problem discussed up until now. The difference is that the terminal posterior is estimated it using the result of the lumped forward equation and the measurement distribution using (7.8). Based on this estimated terminal posterior, we compute the bridging probabilities and refine the truncation tailored to the location of the posterior distribution. In [Figure 7.6a](#), we illustrate the bridging distribution



**Figure 7.6:** (a) A comparison of the prior dynamics and the posterior smoothing (bridging) dynamics. (b) The prior, likelihood, and posterior of the number of infected individuals  $n_I$  at time  $t = 0.3$  given the measurement  $\hat{n}_I = 30$ .

between the terminal posterior and initial distribution. In the context of filtering problems this is commonly referred to as smoothing. Using the learned truncation, we can obtain the posterior distribution for the number of infected individuals at  $t = 0.3$  (Figure 7.6b). Moreover, can we infer a distribution over the unknown number of susceptible and exposed individuals (Figure 7.7).



**Figure 7.7:** The prior and posterior distribution over the latent types  $E$  and  $S$  at time  $t = 0.3$ .



## 7.6 CONCLUSION

The analysis of [MPMs](#) with constraints on the initial and terminal behavior is an important part of many probabilistic inference tasks such as parameter estimation using Bayesian or maximum likelihood estimation, inference of latent system behavior, the estimation of rare event probabilities, and reachability analysis for the verification of temporal properties. If endpoint constraints correspond to atypical system behaviors, standard analysis methods fail as they have no strategy to identify those parts of the state-space relevant for meeting the terminal constraint.

Here, we proposed a method that is not based on stochastic sampling and statistical estimation but provides a direct numerical approach. It starts with an abstract lumped model, which is iteratively refined such that only those parts of the model are considered that contribute to the probabilities of interest. In the final step of the iteration, we operate at the granularity of the original model and compute lower bounds for these bridging probabilities that are rigorous up to the error of the numerical integration scheme.

Our method exploits the population structure of the model, which is present in many important application fields of [MPMs](#). Based on experience with other work based on truncation, the approach can be expected to scale up to at least a few million states (Mikeev, Sandmann, and Wolf, [2011](#)). Compared to previous work, our method neither relies on approximations of unknown accuracy nor additional information such as a suitable change of measure in the case of importance sampling. It only requires a truncation threshold and an initial choice for the macro-state sizes.

In future work, we plan to extend our method to hybrid approaches, in which a moment representation is employed for large populations while discrete counts are maintained for small populations. Moreover, we will apply our method to model checking where constraints are described by some temporal logic (Hajnal et al., [2019](#)).



# 8

---

## RARE EVENT PROBABILITIES

---

Rare events are events that occur with very low probability. Such events can be, for example, the die-out of some population or the switching of a multimodal system across some potential barrier. In biological applications, rare events of interest are typically related to the reachability of certain thresholds on molecule counts or mode switching (Strasser, Theis, and Marr, 2012). By their nature, most standard methods focus on regions with high probability. As an example consider the standard SSA: Trajectories are generated according to the processes' density. Therefore unlikely events are exactly as unlikely to be sampled using the direct method. Accordingly, lots of samples are necessary to even sample a rare event at all and many more to get an estimator with a reasonable variance. Therefore, standard Monte Carlo estimation falls short for such use cases and typically analysis is done using variance reduction methods tailored to them.

The arguably most widely used method for rare event analysis is importance sampling (IS). This variance reduction method is very well-suited to the analysis of such events. In a nutshell, this method alters the model's dynamics and keeps track of the *likelihood ratio* between this altered and the original model. The alteration of the model is referred to as the *change of measure*. The ratio between the changed and the original measure can be used to correct the weight of each run. The main challenge is to find a "good" way to alter the model. Ideally, such an altered model should generate trajectories conditioned on the rare event with

the same distribution found in the original model. At the same time, finding such a change of measure should be sufficiently convenient.

*In Backenköhler (2019), we provide an implementation.*

A popular approach is found in dwSSA (Daigle Jr et al., 2011; Kuwahara and Mura, 2008). Therein each reaction rate is altered by some constant scalar factor. These biasing values are identified by using SSA pilot runs and adjusting biases to according to a cross-entropy objective. We iterate on that approach by changing constant biases across fixed time-intervals. Additionally, these constant biases are state-dependent.

*If the backward probabilities are known, the event probability is known.*

If the complete backward probabilities are available, we would have the optimal biasing. Since they are not known in most interesting cases, we use the aggregation-based approximation discussed in the previous chapters. Here, we study using the approximate backward probabilities studies in the previous chapter as a proxy for the *optimal* change of measure. In particular, the ratios of adjacent macro-states act as a bias for the respective reaction propensities. To deal with continuous time, we record backward probabilities at a pre-determined number of time points. This way, the biases remain constant within the resulting time-intervals and sampling the jump time becomes easier. We illustrate the method and its characteristics on a set of case studies along with a comparison to the dwSSA method.

## 8.1 RELATED WORK

**IMPORTANCE SAMPLING** Most methods for the estimation of rare event probabilities rely on importance sampling using constant reaction biases (Chong, Saglam, and Zuckerman, 2017; Daigle Jr et al., 2011; Kuwahara and Mura, 2008). This method has been extended to be state-dependent in Roh et al. (2011).

**IMPORTANCE SPLITTING** Importance Splitting is an alternative Monte Carlo estimation strategy. In this setting *promising* paths are split – initiating new simulation runs. These runs are then reweighted according to the specific splitting algorithm. An early proposal of the splitting technique is the repetitive

simulation trials after reaching thresholds ([RESTART](#)) algorithm (Villén-Altamirano and Villén-Altamirano, 1994). Further popular iterations of this idea where, for example, fixed effort and fixed success splitting (Garvels and Kroese, 1998). For more recent developments, we refer to (Budde, D’Argenio, and Hartmanns, 2017; Jégourel, Legay, and Sedwards, 2013). A central problem in importance splitting is determining what constitutes a *promising* path. Algorithms rely on an *importance function* as an oracle. The (approximate) backward probabilities presented here could be used as such an importance function, as well.

**FINITE STATE PROJECTION** Apart from sampling-based approaches, dynamic finite-state projections have been employed by Mikeev, Sandmann, and Wolf (2013), but are lacking automated truncation schemes. In [Chapter 7](#), we presented a solution to this shortcoming.

## 8.2 IMPORTANCE SAMPLING

Importance sampling ([IS](#)) is a popular variance reduction technique. Typically it is applied for the Monte Carlo estimation of rare event probabilities. The main idea is to sample from a different distribution, the [IS](#) density, and adjust samples using the likelihood ratio. Let  $f$  be the original density and the goal is to estimate

*This explanation follows Kroese, Taimre, and Botev (2011, Chapter 9.7).*

$$E(\theta(X)) = \int \theta(x)f(x) dx.$$

Now let  $g$  be another density, dominating  $\theta f$ , i.e.

$$g(x) \Rightarrow \theta(x)f(x) = 0.$$

Then we can re-write the above as

$$E_f(\theta(X)) = \int \theta(x)f(x) dx = \int \theta(x) \frac{f(x)}{g(x)} g(x) dx = E_g \theta(X) \frac{f(X)}{g(X)}.$$

Therefore, we can replace the estimate using sampling from  $f$  by an estimate using the density  $g$  instead. According to the

right-hand side of this equation, the estimate using i.i.d. samples  $X_i$ ,  $1 \leq i \leq N$

$$\hat{E}(\theta(X)) = N^{-1} \sum_{k=1}^N \theta(X_k) \frac{f(X_k)}{g(X_k)}. \quad (8.1)$$

The term factor

$$W(x) = \frac{f(x)}{g(x)}$$

is called the *likelihood ratio*.

Thus, the method hinges on finding a density  $g^*$ , that has a computable likelihood ratio and minimizes the variance of the estimator (8.1). If  $\theta(x)$  is an event, the perfect [IS](#) (Kroese, Taimre, and Botev, [2011](#), Chapter 9.7.1)

$$g^*(x) = \frac{\theta(x)f(x)}{E(\theta(x))}.$$

Therefore the ideal [IS](#) distribution is the conditional density

$$g^*(x) = f(x \mid \theta(X) = 1).$$

### 8.3 NEAR-OPTIMAL BIASING

An [MPM](#) can be modified to fulfill terminal constraints. Previously, in [Section 7.3](#), we have seen how the endpoint constrained process can be described using the backward probabilities  $\beta$ . The bridging distribution can be either computed using both backward and forward probabilities, but for us it is more instructive to consider, the bridging [CME](#). This is the endpoint constrained version of the [CME](#). It depends on ratios of the backwards probabilities which act as factors to the propensity values. Taking the derivative of  $\gamma(x, t) = \pi(x, t)\beta(x, t)$ , yields the bridging [CME](#) (see also Huang et al. ([2016](#)))

$$\frac{d\gamma}{dt}(x, t) = \sum_{j=1}^{n_R} (\tilde{\alpha}_j(x - v_j)\gamma(x - v_j, t) - \tilde{\alpha}_j(x)\gamma(x, t)) , \quad (8.2)$$

where the propensities

$$\tilde{\alpha}_j(x, t) = \alpha_j(x) \phi_j(x, t). \quad (8.3)$$

The time-dependent predilection factor

$$\phi_j(x, t) = \beta(x + v_j, t) / \beta(x, t). \quad (8.4)$$

Equation (8.2) reveals the optimal biasing scheme for IS. Since, we know that the ideal IS distribution is the conditional distribution  $\gamma$ , the ratios give a perfect biasing. We use approximations of these ratios as a time and state dependent predilection functions  $\phi_j$  during the stochastic simulation.

Naturally, computing (8.2) requires full knowledge of all backward probabilities  $\beta(x, t)$  for all states  $x$  and times  $t$ . In particular, the event probability is precisely  $\beta(x_0, 0)$ . To make this approach feasible, we will use the aggregation-based approximation  $\hat{\beta}$  of the backward probabilities  $\beta$ . Thus, we will compute and store backward probabilities for the aggregated system for discrete time points up to  $T$ .

$$\bar{\phi}(i, j; t) = \hat{\beta} \left( \bar{x}_j, \Delta \left\lfloor \frac{t}{\Delta} \right\rfloor \right) / \hat{\beta} \left( \bar{x}_i, \Delta \left\lfloor \frac{t}{\Delta} \right\rfloor \right)$$

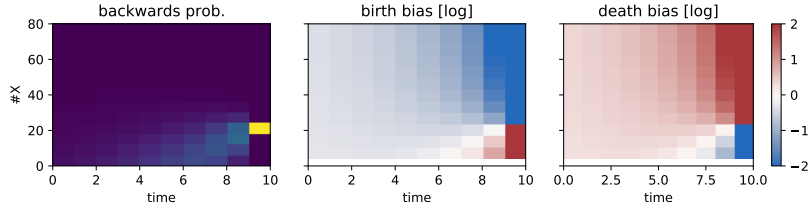
The biases of the aggregated system serve as a basis for the biases at the original state granularity. To transfer the macro-state biases, we again employ the assumption of a uniform distribution. Under the assumption it is reasonable to take the average approximate bias weighted according to the size of the transition states. Therefore

$$\hat{\phi}_j(x, t) = \left| \bar{x}_{i \rightarrow j} \right|^{-1} \sum_k \bar{\phi}(i, j; t) \left| \bar{x}_{i \rightarrow k} \right|,$$

where  $x \in \bar{x}_i$ . Accordingly, we can define the biased propensity function

$$\hat{\alpha}_j = \alpha_j(x) \hat{\phi}_j(x, t).$$

The discrete steps in the time domain are kept. This enables the application of an adjusted SSA that can deal with piecewise



**Figure 8.1:** The approximate dynamic biasing (values cut to  $[e^{-2}, e^2]$ ) illustrated for a birth death process with macro-state size 8 and target state  $[24, 31]$  at  $T = 10$  for 10 time points.

constant rate functions in the waiting time distributions. This algorithm is discussed in the following section.

In [Figure 8.1](#), we illustrate this scheme for a birth-death process ([Model 2](#)) with goal macro-state  $[24, 31]$  at  $T = 10$  and  $X_0 = 0$ . [Figure 8.1](#) clearly illustrates how the biases increase towards the end increasing the push towards the goal state the farther time progresses. Since by the approximation assumption, all constituent states of a macro-state are treated as equivalent, the macro-state biases are transferred to the micro-states. Depending on the scenario, this interpolation could be replaced by other schemes, such as nearest neighbor interpolations.

*In fact, such an approach might be better suited in the example above, where we aim to push the process to a specific micro-state.*

#### 8.4 NON-HOMOGENEOUS STOCHASTIC SIMULATION

We are changing the rate bias dynamically over fixed intervals of the time-domain. Therefore we cannot use the default [SSA](#). With [Algorithm 7](#) we present a version of the stochastic simulation algorithm that simulates trajectories of a system with such dynamically changing biases. The main change is the handling of the time-discrete changes in the loop in [line 8](#). Here it is tested, in which time interval the sampled jump will take place. Therefore rates are recomputed each time the algorithm jumps forward one time-interval.



Let us first consider how exactly the jump time distribution changes. A time-inhomogeneous exponential has the cdf

$$F(t) = 1 - \exp \left( - \int_0^t \lambda(s) ds \right).$$

Accordingly, the pdf

$$f(t) = \lambda(t) \exp \left( - \int_0^t \lambda(s) ds \right).$$

Since we compute the biases for fixed intervals, we have an exponential model with a piecewise constant rate function. In [Figure 8.2a](#), we give an example of such a distribution. Assume, we have an increasing sequence of time points  $\tau_0 = 0, \tau_1, \tau_2, \tau_3, \dots$  with corresponding rates  $\lambda_i > 0$ . The piecewise constant hazard function

$$\lambda(t) = \begin{cases} \lambda_1, & \text{if } \tau_0 \leq t \leq \tau_1 \\ \lambda_2, & \text{if } \tau_1 < t \leq \tau_2 \\ \lambda_3, & \text{if } \tau_2 < t \leq \tau_3 \\ \vdots & \end{cases}.$$

We can give the pdf using a case distinction as

$$f(t) = \begin{cases} f_1(t), & \text{if } \tau_0 \leq t \leq \tau_1 \\ f_2(t), & \text{if } \tau_1 < t \leq \tau_2 \\ f_3(t), & \text{if } \tau_2 < t \leq \tau_3 \\ \vdots & \end{cases}, \quad (8.5)$$

where, letting  $\Delta_k = \tau_k - \tau_{k-1}$ , the piecewise densities

$$f_k(t) = \lambda_k \exp \left( - \sum_{i=1}^{k-1} \Delta_i \lambda_i - \lambda_k (t - \tau_{k-1}) \right).$$

Similarly, the cdf

$$F(t) = 1 - \begin{cases} S_1(t), & \text{if } \tau_0 \leq t \leq \tau_1 \\ S_2(t), & \text{if } \tau_1 < t \leq \tau_2 \\ S_3(t), & \text{if } \tau_2 < t \leq \tau_3 \\ \vdots & \end{cases}, \quad (8.6)$$

where the component survival functions

$$S_k(t) = \exp \left( - \sum_{i=1}^{k-1} \Delta_i \lambda_i - \lambda_k (t - \tau_{k-1}) \right).$$

The sampling from this density – necessary in the stochastic simulation – uses an inverse transform. This transform is most concisely expressed in lines 3–22 of the pseudocode in [Algorithm 7](#). The uniform random sample  $X$  is checked against the probability mass of the current time interval. The mass follows an exponential distribution with the current exit rate  $\alpha_0$ . In each iteration we recompute the rate  $\alpha_0$  and advance the time until the correct interval is identified. Finally, in [line 22](#) the jump time is computed. In [Figure 8.2a](#), we illustrate the probability density function of an exponential distribution with a piecewise constant rate function. Along with the pdf, an empirical distribution is presented, which samples were generated by the algorithm described here.

Since the jump time is determined after this part, the reaction selection uses the rates of this time-interval. The probability of a reaction  $j$  being selected is proportional to its biased reaction rate  $\hat{\alpha}_i$  ([line 18](#)).

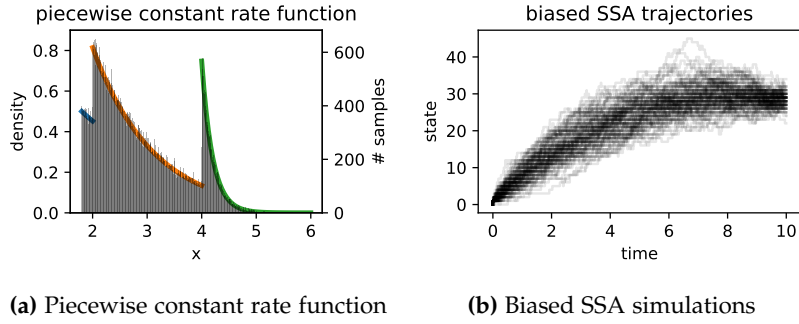
The sampling of successive reactions is performed until the predefined termination function  $\Theta$  is true. Typically this means reaching some time-horizon  $T > 0$ , i.e.  $\Theta(s, t) = t > T$ .

In [Figure 8.2b](#), we illustrate the algorithms result using the biases discussed in the previous section.

```

input :  $\pi_0, \theta, \Theta$ 
output: sample weighted by the likelihood ratio
1  $s \sim \pi_0; t \leftarrow 0; j \leftarrow 1; w \leftarrow 1;$ 
2 loop
3    $X \sim \mathcal{U}[0, 1];$ 
4    $a_0 \leftarrow \sum_j \hat{\alpha}_j(s);$  /* exit rate */
5    $\delta \leftarrow t_{j+1} - t;$  /* rest of time interval */
6    $\Delta \leftarrow 0;$  /* time offset */
7    $\tau \leftarrow t;$  /* start of the current component */
8   while  $X > 1 - \exp(-a_0\delta - \Delta);$  /* find interval */
9   do
10     $\Delta \leftarrow \Delta + a_0\delta;$ 
11     $\tau \leftarrow \tau + \delta;$ 
12     $j \leftarrow j + 1;$ 
13     $\delta \leftarrow \text{time-interval width};$ 
14     $a_0 \leftarrow \sum_j \hat{\alpha}_j(s);$  /* exit rate */
15     $\delta \leftarrow -(\log(1 - X) + \Delta)/a_0;$ 
16    if  $\Theta(s, \tau + \delta)$  then
17      return  $\theta(s, \tau)w$ 
18     $k \leftarrow \text{sample } j \text{ with probability } \hat{\alpha}_j(s) / \sum_i \hat{\alpha}_i(s);$ 
19     $\ell_1 \leftarrow \hat{\alpha}_k(s) \exp(-\Delta - \delta a_0);$ 
20     $\ell_0 \leftarrow \alpha_k(s) \exp(-(\tau + \delta - t) \sum_i \alpha_i(s));$ 
21     $w \leftarrow w \ell_0 / \ell_1;$  /* update likelihood ratio */
22     $t \leftarrow \tau + \delta;$  /* update time */
23     $s \leftarrow s + v_k;$  /* update state */

```



**Figure 8.2:** Importance Sampling using piecewise constant jump distributions: (a) A piecewise constant rate function is induced by the biases changing at discrete time points. (b) We can simulate the system using piecewise constant biases.

## 8.5 CASE STUDIES

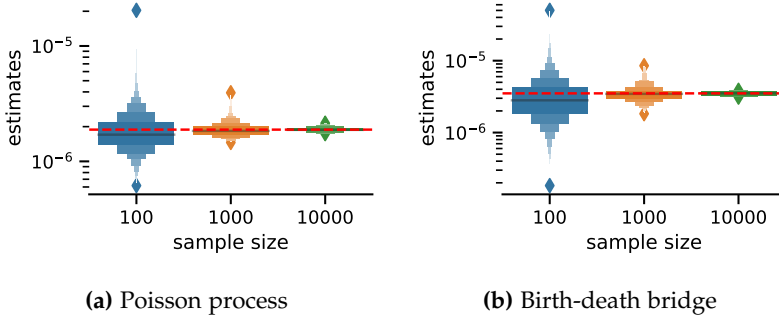
We implemented the methods in Python and evaluated the method on three case studies. The first two are rather simple, but have the advantage of a known analytical distribution. Thus, we have a reference to compare the Monte Carlo results. In the second part, we take a look at a more challenging model – the toggle switch. In this example, we face non-polynomial rate functions and a bimodal behavior.

### 8.5.1 Two Simple Examples

For the Poisson process we look at the event of exceeding a threshold of 150 before a time-horizon of  $T = 10$ . The aggregation scheme groups 10 states together and records the approximate backward probabilities for 10 time points. In [Figure 8.3a](#), we summarize the estimates for different sample sizes. In this study, we observe no finite sample bias and a quick convergence to the analytical result.

In case of the birth-death process, we focus on the event of being in state 30 at time  $T = 10$ . The aggregation scheme groups 10 states together and records the approximate backward prob-

abilities for 10 time points. In [Figure 8.3b](#), we summarize the estimates for different sample sizes. In this study observe no finite sample bias and a quick convergence to the analytical result. The switching between the two distinct meta-stable regions is of particular interest. At the same time, it is hard to analyze using standard Methods because this event typically is rare.

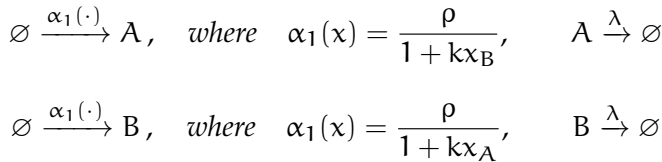


**Figure 8.3:** Estimate distribution for different sample sizes.

### 8.5.2 Toggle Switch

We return to the example of the toggle switch, known from [Chapter 7](#). It notably includes two non-polynomial propensity functions. This model exhibits a bimodal behavior. Either mostly agents of type A are produced while the counts of B-type agents remains low or vice versa.

**Model 16** (Toggle Switch using Hill functions (Lipshtat et al., 2006)). We have population types A and B with the following reactions and reaction rates.



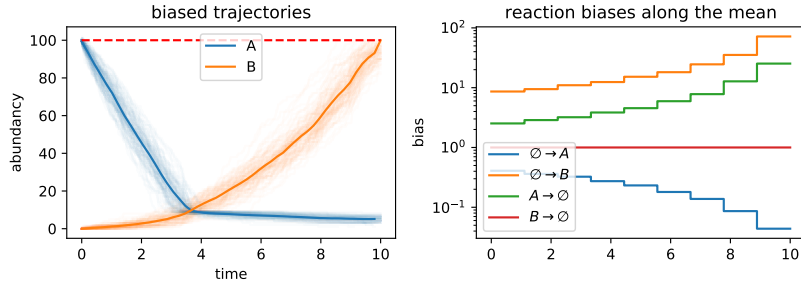
The parameterization is  $\rho = 10$ ,  $\lambda = 0.1$ , and  $k = 1.5$ .

*This approximation is different from the previous version on page 145, because we previously had  $k = 1$ .*

The sums of non-polynomial propensity functions have no simple analytical solution as the Faulhaber formulae. Instead, we approximate the discrete sums via an integral:

$$\begin{aligned}\bar{\alpha}_1(\bar{x}) &\approx \int_{a-0.5}^{b+0.5} \frac{\rho}{1+kx} dx \\ &= \frac{\rho}{k} \left( \log \left( k \left( b - \frac{1}{2} \right) + 1 \right) - \log \left( k \left( a + \frac{1}{2} \right) + 1 \right) \right)\end{aligned}$$

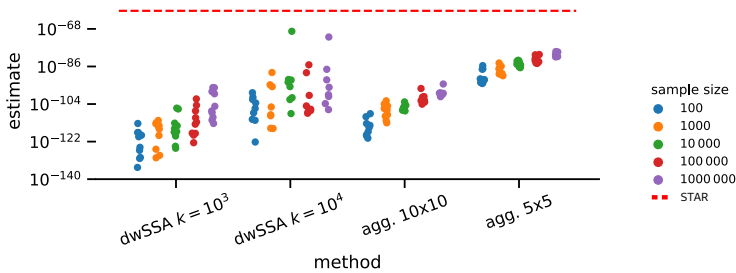
In this case study, we use both, a aggregation of  $5 \times 5$  and  $10 \times 10$  states. As an initial state we fix  $(100, 0)$  and the target event is  $\{X_t^{(B)} \geq 100 \mid t \leq 10\}$ . In Figure 8.4, we illustrate the sample trajectories and their mean under the biasing of the approximate backward probabilities. On the left-hand side, the mean bias is given over time and per reaction. The biasing strength increases towards the time-horizon.



**Figure 8.4:** The dynamic biasing of a toggle switch based on a aggregation of  $10 \times 10$  states.

In Figure 8.5, we provide estimates for this case study using the aggregation-based approach presented here and the popular dwSSA method as a point of comparison. dwSSA (Daigle Jr et al., 2011) essentially derives a constant bias coefficient for each reaction. The biases are optimized using a cross-entropy objective. This optimization, however, needs a large number of runs in each iteration. We used ensemble sizes of 1000 and 10,000 for each iteration. This imposes a large additional cost, since in almost all instances more than 10 optimization iterations were necessary, until a fraction of 0.1 ensemble trajectories reached the rare event

region. In some instances no change of measure was found using the cross-entropy optimization.

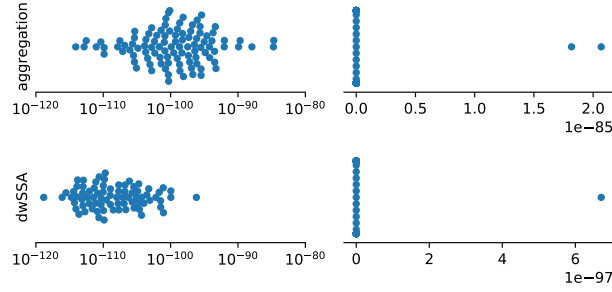


**Figure 8.5:** Rare event estimates (toggle switch with Hill functions) using different methods and sample sizes. For each scenario 10 estimations were performed. To the left are the dwSSA results for comparison. We provide results for estimates using 1000 and 10,000 estimates in each iteration of the prior parameter optimization rounds. On the right, the results for the aggregation method presented in this chapter are given for grids with  $5 \times 5$  and  $10 \times 10$  macro states.

With both methods, we observe that with smaller sample sizes, the estimate tends to be smaller as well, regardless of the method. This is due to the (non-zero) weights not being normally distributed. There tend to be more samples with lower weight, which are compensated by few samples with a marginally larger weight. In other words, there seems to be a long tail towards higher weights in most cases. The theoretical estimate is still unbiased. But the estimates which are much larger (“correcting” the low estimates) are very infrequent. These long-tailed distributions are illustrated in Figure 8.6. This issue is present regardless of the chosen method. The most clear difference is that the estimates using aggregation are spread less. This effect is likely due to the aggregation-based biasing steering the process with more detail. Therefore trajectories tend to be more similar — an effect that is also apparent in the samples of Figure 8.7.

As a reference and baseline we consider the dynamic truncation result obtained by the STAR tool (Lapin, Mikeev, and Wolf, 2011). We ran the analysis using a relative tolerance of

*mosi.uni-saarland.de/tools/shave/*



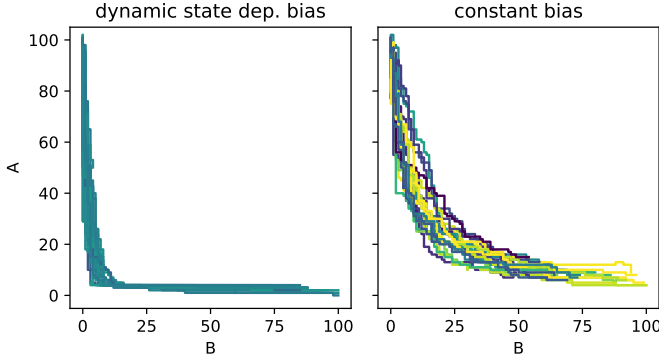
**Figure 8.6:** Comparison of 100 non-zero weights using importance sampling based on aggregation and on dwSSA. On the left, the samples are plotted on a log-scale and on the right on a linear scale.

$1 \times 10^{-15}$  and an absolute tolerance of  $1 \times 10^{-90}$ . This yielded  $\Pr(X_{10}^{(B)} \geq 100) \approx 4.64 \times 10^{-60}$ . Note, that this is a lower bound to the probability since we only consider the probability at  $t = 10$ , whereas the stopping time above considers all  $t \leq 10$ . Considering this baseline, it is clear, that both methods have only slowly approach the probability's true magnitude with increasing sample size.

In Figure 8.7, we compare sample paths using the technique presented here with a constant biasing scheme. The constant biases are obtained using a cross-entropy optimization using  $10^4$  SSA simulations in each iteration until 0.1 of the trajectories reach the rare event region. It is immediately clear, that the sampled trajectories differ in the paths that are sampled. We further observe a far larger range of likelihood ratios in the constant biasing case.

This case study showcases the major difficulty in finding an efficient change of measure that one can encounter. Ideally, sampled trajectories should follow the original process, conditioned on the rare event. As in this example, constant biases might not be enough to reliably alter the dynamics in this way. However, the approximate biasing scheme presented in this chapter, while being based on an approximation of the ideal biasing, seems to over-constrain the process and thus not explore the trajectory





**Figure 8.7:** Biased sample paths of the toggle switch using (left) the aggregation based sampling and (right) a constant biasing, determined by cross-entropy optimization (Daigle Jr et al., 2011). The constant bias vector is  $\approx (0.873, 26.248, 2.751, 0.00159)^\top$ . The path color represents the relative log-weights of each trajectory (darker – higher weight).

space sufficiently. Consequently, both methods tend to underestimate in a majority of cases.

## 8.6 CONCLUSION

In this chapter, we presented a novel change of measure using approximate backward probabilities. This approach avoids the necessity of a significant number of pilot runs. In the last example, the number of these pilot runs often exceeded the number of runs used for the estimate. Therefore, the approach presented is a feasible alternative for cases in which the state-space size allows for an approximate solution.

We further showcased a challenge that is often emphasized too little. This is the issue of changes of measures, that steer the process to the target region along trajectories that are significantly less likely in the original system than the conditioned paths would be. This leads to a case where the estimate distribution is very different from a normal distribution. Consequently, the

estimates are too low in most cases, and a lot too large in very few cases.

↪ page 131 This issue might be tackled by, using state space refinement, similar to the previous chapter. Also, combining backward probability approximations of differing granularity might be a viable strategy.

## Part IV

### CONCLUDING REMARKS



# 9

---

## CONCLUSION

---

While Markovian population models (MPMs) are a powerful modeling framework, their analysis – especially wrt. their stochasticity – poses a significant challenge. This thesis has proposed several techniques to tackle some parts of this challenge. All of these techniques use the uniform structure of the CTMCs underlying an MPM.

This structure enables us to derive moment equations that yield constraints that have shown to be useful in two contexts. In the context of mean first-passage time bounds, they formed linear constraints on the moments of the process and the exit location measure. In combination with semi-definite constraints, this yields a hierarchy of convex optimization problems. This hierarchy yields convergent bounds on the reaching probability and the mean first-passage time in a wide class of problems. The other application of these moments constraints has been demonstrated in the context of Monte Carlo estimation: When used as control variates (CVs) such constraints yield an alternative unbiased estimator with lower variance. Using both heuristics and sequential Monte Carlo methods, we demonstrated that efficient constraint sets can be constructed.

The aggregation approach takes advantage of the fact that often propensity functions result in a smooth “landscape of transition rates” across the state-space. This motivates an aggregation scheme that approximates a uniform distribution inside macro-states. Deriving closed forms for the transition rates be-

tween those states significantly reduces the computational load compared to the original model. We have demonstrated the usefulness of this aggregation scheme in the domains of stationary distributions, bridging distributions, and the use in rare event sampling.

### 9.1 FUTURE WORK

All ideas presented in this thesis provide starting points for future work.

**MEAN FIRST-PASSAGE TIMES VIA SEMI-DEFINITE PROGRAMMING** The [SDP](#) approach for [MFPTs](#) presented in [Chapter 3](#) improved in terms of scalability. Here, scalability mainly refers to the problem of large populations (or time horizons). Such large values increase the moments which tend to grow exponentially with their order. The resulting stiffness can be problematic for off-the-shelf [SDP](#) solvers. Better performance could be achieved by a suitable re-scaling of the model or stronger cooperation with domain experts in mathematical optimization to pinpoint problems more precisely. Another interesting direction is the study of Hausdorff constraints sketched in [Section 3.8](#). Since these are linear moment constraints, they can easily be incorporated into the [SDP](#) formulation.

**CONTROL VARIATES** Control variates for Monte Carlo ([Chapter 4](#)) estimation offer a tremendous design space, both algorithmically and in terms of the constraints themselves. Depending on the quantity of interest, non-polynomial test functions might provide better correlations. The control variate optimization, based on sequential Monte Carlo, could perhaps be replaced by a search using approximations such as moment closures. A challenge in designing [CVs](#) and their algorithmic optimization is always elegance. Keeping the algorithm flexible, while keeping the number of parameters low is difficult, but crucial.

**AGGREGATION** Similarly to the control variates, the aggregation scheme offers quite a few possibilities for further experimentation. For example, a dynamic aggregation-disaggregation scheme could be used for forward analysis similar to dynamic truncations (Andreychenko et al., 2011). Another possible modification would be to define states more flexible than hyper cubes. Other shapes might be better suited for this kind of approximation.

**GENERAL** The main challenge in stochastic reaction networks today is the connection of methods to practitioners and their problems. A large area body of research is devoted to methods precisely analyzing SRNs. It is a danger that the practical value of proposed methods is a secondary concern. For these methods to be useful, more dialog is needed with domain experts to understand the challenges. Only then can the new approaches – such as the ones presented in this thesis – develop practical uses.





## Part V

## APPENDIX





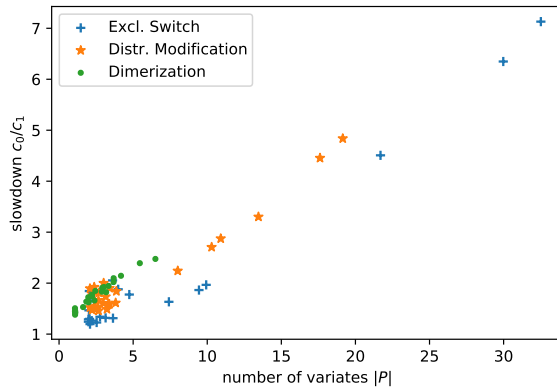
---

## ADDITIONAL DATA

---

### A.1 CONTROL VARIATE RESULTS

Detailed results for control variate presented in [Chapter 4](#) are given in [Table A.1](#), [Table A.2](#), and [Table A.3](#).



**Figure A.1:** The slowdown  $c_0/c_1$  v. the number of control variates  $|P|$ .

### A.2 AGGREGATION FOR STATIONARY DISTRIBUTIONS

Experimental results of aggregation for stationary distribution approximation presented in [Chapter 6](#) are given in [Table A.4](#) and [Table A.5](#).

$n_{\max}$	$n_\lambda$	$\phi$	$1 - \frac{\sigma_1^2}{\sigma_0^2}$	slowdown	efficiency	$ P $
1	10	$\phi_{sq}$	0.807184	1.227471	4.239255	2.536479
		$\phi_c$	0.880285	1.633530	5.135205	7.411732
		$\phi_q$	0.849082	1.312416	5.067770	3.639250
		$\phi_\ell$	0.783459	1.195821	3.874778	2.090101
	20	$\phi_{sq}$	0.856593	1.263340	5.539683	2.206154
		$\phi_c$	0.910480	1.864405	6.011256	9.441336
		$\phi_q$	0.867987	1.317958	5.765884	3.140806
		$\phi_\ell$	0.825518	1.243075	4.627662	1.981143
	30	$\phi_{sq}$	0.869165	1.298893	5.905196	2.059415
		$\phi_c$	0.921019	1.966191	6.461331	9.928998
		$\phi_q$	0.876822	1.340409	6.079876	2.762449
		$\phi_\ell$	0.843288	1.288925	4.968796	1.983174
2	10	$\phi_{sq}$	0.811956	1.505521	3.544783	2.323999
		$\phi_c$	0.916866	4.507566	2.681363	21.692390
		$\phi_q$	0.868874	1.776190	4.309354	4.739893
		$\phi_\ell$	0.795802	1.466579	3.353046	2.016196
	20	$\phi_{sq}$	0.832562	1.657484	3.617313	2.085711
		$\phi_c$	0.934280	6.348223	2.406431	29.976320
		$\phi_q$	0.878944	1.879341	4.416281	3.990881
		$\phi_\ell$	0.837922	1.647329	3.759896	1.978017
	30	$\phi_{sq}$	0.829427	1.844766	3.190308	2.043201
		$\phi_c$	0.947324	7.130628	2.673225	32.513670
		$\phi_q$	0.878830	2.053317	4.034987	3.611746
		$\phi_\ell$	0.824936	1.838879	3.118728	1.978836

**Table A.1:** Variance reduction results for up to second order moments with parameters  $n_{\max} = 2$ ,  $n = 10,000$ ,  $d = 100$ ,  $k_{\min} = 3$ . Exclusive switch.

$n_{\max}$	$n_\lambda$	$\phi$	$1 - \frac{\sigma_1^2}{\sigma_0^2}$	slowdown	efficiency	P
1	10	$\phi_{sq}$	0.619560	1.488483	1.770218	3.232575
		$\phi_c$	0.700255	2.241695	1.492171	8.008607
		$\phi_q$	0.643550	1.613001	1.743500	3.817641
		$\phi_\ell$	0.596650	1.459405	1.703170	2.657000
	20	$\phi_{sq}$	0.697414	1.519425	2.181687	2.631677
		$\phi_c$	0.713445	2.706546	1.292838	10.295856
		$\phi_q$	0.697654	1.585313	2.092817	3.398235
		$\phi_\ell$	0.695846	1.473976	2.235418	2.226530
	30	$\phi_{sq}$	0.712941	1.543068	2.263644	2.378037
		$\phi_c$	0.721354	2.874249	1.252541	10.910880
		$\phi_q$	0.711877	1.607712	2.164485	2.979704
		$\phi_\ell$	0.669963	1.522184	1.996300	2.085473
2	10	$\phi_{sq}$	0.619450	1.734737	1.519168	3.148184
		$\phi_c$	0.665361	3.301159	0.909443	13.456259
		$\phi_q$	0.680592	1.840457	1.705876	3.864240
		$\phi_\ell$	0.612674	1.662962	1.556868	2.659592
	20	$\phi_{sq}$	0.684789	1.811408	1.755652	2.687379
		$\phi_c$	0.689835	4.455005	0.726640	17.609554
		$\phi_q$	0.687665	1.901258	1.688449	3.413595
		$\phi_\ell$	0.651262	1.770238	1.623924	2.266729
	30	$\phi_{sq}$	0.690602	1.922217	1.686011	2.375455
		$\phi_c$	0.649191	4.837419	0.591701	19.145054
		$\phi_q$	0.701253	2.001179	1.677062	3.007525
		$\phi_\ell$	0.639123	1.894074	1.467403	2.086275

**Table A.2:** Variance reduction results for up to second order moments with parameters  $n_{\max} = 2$ ,  $n = 10,000$ ,  $d = 100$ ,  $k_{\min} = 3$ . Distributive modification.

$n_{\max}$	$n_{\lambda}$	$\phi$	$1 - \frac{\sigma_1^2}{\sigma_0^2}$	slowdown	efficiency	$ P $
1	10	$\phi_{sq}$	0.881641	1.530137	5.558692	1.621917
		$\phi_c$	0.965224	1.945588	14.859417	3.338501
		$\phi_q$	0.916445	1.625232	7.409904	1.997045
		$\phi_\ell$	0.868288	1.380344	5.529745	1.081152
	20	$\phi_{sq}$	0.941153	1.637272	10.437978	1.842971
		$\phi_c$	0.964204	1.907999	14.747328	2.915082
		$\phi_q$	0.947984	1.747519	11.072422	2.227250
		$\phi_\ell$	0.931030	1.433401	10.169570	1.088572
	30	$\phi_{sq}$	0.959517	1.723449	14.404936	1.972426
		$\phi_c$	0.962514	1.770936	15.142156	2.216103
		$\phi_q$	0.966216	1.847441	16.117387	2.446661
		$\phi_\ell$	0.945724	1.456432	12.710196	1.084188
2	10	$\phi_{sq}$	0.905254	1.659799	6.402491	2.388472
		$\phi_c$	0.987526	2.474939	33.074955	6.501180
		$\phi_q$	0.923063	1.822654	7.195544	3.179257
		$\phi_\ell$	0.878232	1.415909	5.830248	1.092264
	20	$\phi_{sq}$	0.949038	1.831995	10.797164	2.890898
		$\phi_c$	0.985710	2.391457	29.704344	5.450299
		$\phi_q$	0.968076	2.021487	15.662368	3.681229
		$\phi_\ell$	0.925413	1.449386	9.298961	1.072761
	30	$\phi_{sq}$	0.964855	1.924268	14.911787	3.026275
		$\phi_c$	0.981507	2.144089	25.520987	4.179125
		$\phi_q$	0.973902	2.095985	18.507746	3.685851
		$\phi_\ell$	0.948349	1.507425	12.904707	1.074538

**Table A.3:** Variance reduction results for up to second order moments with parameters  $n_{\max} = 2$ ,  $n = 10,000$ ,  $d = 100$ ,  $k_{\min} = 3$ . Dimerization.

$\epsilon$	iteration $i$							
	1	2	3	4	5	6	7	8
$1 \times 10^{-1}$	$ S^{(i)} $	4900	28	112	232	472	960	1932
	tot. error	1.91	1.84	1.55	1.29	$9.35 \times 10^{-1}$	$4.88 \times 10^{-1}$	$3.54 \times 10^{-2}$
	max. error	$3.15 \times 10^{-3}$	$3.13 \times 10^{-3}$	$3.08 \times 10^{-3}$	$2.77 \times 10^{-3}$	$2.38 \times 10^{-3}$	$1.57 \times 10^{-3}$	$6.04 \times 10^{-5}$
$1 \times 10^{-2}$	$ S^{(i)} $	4900	52	208	464	988	2008	4052
	tot. error	1.91	1.84	1.56	1.30	$9.46 \times 10^{-1}$	$5.01 \times 10^{-1}$	$6.22 \times 10^{-4}$
	max. error	$3.15 \times 10^{-3}$	$3.13 \times 10^{-3}$	$3.08 \times 10^{-3}$	$2.78 \times 10^{-3}$	$2.39 \times 10^{-3}$	$1.59 \times 10^{-3}$	$8.33 \times 10^{-7}$
$1 \times 10^{-3}$	$ S^{(i)} $	4900	84	152	300	652	1440	6068
	tot. error	1.91	1.83	1.73	1.56	1.30	$9.46 \times 10^{-1}$	$5.01 \times 10^{-1}$
	max. error	$3.15 \times 10^{-3}$	$3.13 \times 10^{-3}$	$3.08 \times 10^{-3}$	$2.98 \times 10^{-3}$	$2.78 \times 10^{-3}$	$2.39 \times 10^{-3}$	$1.14 \times 10^{-8}$
$1 \times 10^{-4}$	$ S^{(i)} $	4900	116	212	400	848	1872	8060
	tot. error	1.91	1.83	1.73	1.56	1.30	$9.46 \times 10^{-1}$	$5.01 \times 10^{-1}$
	max. error	$3.15 \times 10^{-3}$	$3.13 \times 10^{-3}$	$3.08 \times 10^{-3}$	$2.98 \times 10^{-3}$	$2.78 \times 10^{-3}$	$2.39 \times 10^{-3}$	$1.83 \times 10^{-10}$

**Table A.4:** Detailed results for Model 9. The errors are computed wrt. the reference Poissonian product. The total absolute error and the maximum absolute errors are given.

$\epsilon$	iteration i								
	1	2	3	4	5	6	7	8	
$1 \times 10^{-1}$	$ \mathcal{S}^{(1)} $	11,907	20	32	60	140	340	840	2116
	tot. error $\leq$	$1.86 \times 10^0$	$1.85 \times 10^0$	$1.45 \times 10^0$	$1.18 \times 10^0$	$9.31 \times 10^{-1}$	$6.41 \times 10^{-1}$	$4.67 \times 10^{-1}$	$4.89 \times 10^{-1}$
	max. error $\leq$	$1.63 \times 10^{-3}$	1.63e-3	$1.55 \times 10^{-3}$	$1.40 \times 10^{-3}$	$1.22 \times 10^{-3}$	$9.36 \times 10^{-4}$	$8.40 \times 10^{-4}$	$1.40 \times 10^{-3}$
$1 \times 10^{-2}$	$ \mathcal{S}^{(1)} $	11,907	48	112	148	300	720	1892	5156
	tot. error $\leq$	$1.86 \times 10^0$	$1.84 \times 10^0$	$1.44 \times 10^0$	$1.21 \times 10^0$	$9.56 \times 10^{-1}$	$6.65 \times 10^{-1}$	$3.41 \times 10^{-1}$	$3.31 \times 10^{-2}$
	max. error $\leq$	$1.63 \times 10^{-3}$	$1.62 \times 10^{-3}$	$1.53 \times 10^{-3}$	$1.39 \times 10^{-3}$	$1.20 \times 10^{-3}$	$9.59 \times 10^{-4}$	$5.86 \times 10^{-4}$	$5.37 \times 10^{-5}$
$1 \times 10^{-3}$	$ \mathcal{S}^{(1)} $	11,907	84	192	244	488	1084	2692	7152
	tot. error $\leq$	$1.86 \times 10^0$	$1.83 \times 10^0$	$1.46 \times 10^0$	$1.22 \times 10^0$	$9.63 \times 10^{-1}$	$6.67 \times 10^{-1}$	$3.37 \times 10^{-1}$	$8.01 \times 10^{-4}$
	max. error $\leq$	$1.63 \times 10^{-3}$	$2.95 \times 10^{-2}$	$1.54 \times 10^{-3}$	$1.39 \times 10^{-3}$	$1.20 \times 10^{-3}$	$9.51 \times 10^{-4}$	$5.79 \times 10^{-4}$	$1.09 \times 10^{-6}$
$1 \times 10^{-4}$	$ \mathcal{S}^{(1)} $	11,907	124	324	352	672	1436	3408	8864
	tot. error $\leq$	$1.86 \times 10^0$	$1.83 \times 10^0$	$1.46 \times 10^0$	$1.22 \times 10^0$	$9.63 \times 10^{-1}$	$6.67 \times 10^{-1}$	$3.37 \times 10^{-1}$	$1.12 \times 10^{-5}$
	max. error $\leq$	$1.63 \times 10^{-3}$	$3.19 \times 10^{-2}$	$1.54 \times 10^{-3}$	$1.39 \times 10^{-3}$	$1.20 \times 10^{-3}$	$9.51 \times 10^{-4}$	$5.79 \times 10^{-4}$	$1.28 \times 10^{-8}$

**Table A.5:** Detailed results for [Model 10](#). Upper bounds on the total absolute error and the maximum absolute error are given. The worst-case errors are computed wrt. the reference Geobound solution with  $\epsilon_\ell = 1 \times 10^{-2}$ .



---

## BIBLIOGRAPHY

---

- Adan, Aysun, Günel Alizada, Yağmur Kiraz, Yusuf Baran, and Ayten Nalbant (2017). "Flow Cytometry: Basic Principles and Applications." In: *Critical Reviews in Biotechnology* 37.2, pp. 163–176.
- Ale, Angelique, Paul Kirk, and Michael PH Stumpf (2013). "A General Moment Expansion Method for Stochastic Kinetic Models." In: *The Journal of Chemical Physics* 138.17, p. 174101.
- Amparore, Elvio Gilberto and Susanna Donatelli (2013). "Backward Solution of Markov Chains and Markov Regenerative Processes: Formalization and Applications." In: *Electronic Notes on Theoretical Computer Science* 296, pp. 7–26.
- Anderson, David F, Gheorghe Craciun, and Thomas G Kurtz (2010). "Product-Form Stationary Distributions for Deficiency Zero Chemical Reaction Networks." In: *Bulletin of Mathematical Biology* 72.8, pp. 1947–1970.
- Anderson, David F and Thomas G Kurtz (2011). "Continuous Time Markov Chain Models for Chemical Reaction Networks." In: *Design and Analysis of Biomolecular Circuits*. Springer, pp. 3–42.
- Anderson, David F and Chaojie Yuan (2018). "Low Variance Couplings for Stochastic Models of Intracellular Processes with Time-Dependent Rate Functions." In: *Bulletin of Mathematical Biology*, pp. 1–29.
- Anderson, William J (2012). *Continuous-Time Markov Chains: An Applications-Oriented Approach*. Springer Science & Business Media.
- Andreychenko, Aleksandr, Luca Bortolussi, Ramon Grima, Philipp Thomas, and Verena Wolf (2017). "Distribution Approximations for the Chemical Master Equation: Comparison of the

- Method of Moments and the System Size Expansion." In: *Modeling Cellular Systems*. Springer, pp. 39–66.
- Andreychenko, Aleksandr, Linar Mikeev, David Spieler, and Verena Wolf (2011). "Parameter Identification for Markov Models of Biochemical Reactions." In: *International Conference on Computer Aided Verification*. Vol. 6806. Lecture Notes in Computer Science. Springer, pp. 83–98.
- Andreychenko, Aleksandr, Linar Mikeev, and Verena Wolf (2015). "Reconstruction of Multimodal Distributions for Hybrid Moment-Based Chemical Kinetics." In: *Journal of Coupled Systems and Multiscale Dynamics* 3.2, pp. 156–163.
- Asgari-Targhi, Ameneh and Elizabeth B Klerman (2019). "Mathematical Modeling of Circadian Rhythms." In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 11.2, e1439.
- Aziz, Adnan, Kumud Sanwal, Vigyan Singhal, and Robert Brayton (1996). "Verifying Continuous Time Markov Chains." In: *International Conference on Computer Aided Verification*. Vol. 1102. Lecture Notes in Computer Science. Springer, pp. 269–276.
- Backenköhler, Michael (2019). *Stochastic Simulation and LCV Estimation Code*. URL: <https://github.com/mbackenkoehler/cme-simulation>.
- Backenköhler, Michael (2020). *MJP Aggregation and Bridging Code*. URL: [https://www.github.com/mbackenkoehler/mjp\\_bridging](https://www.github.com/mbackenkoehler/mjp_bridging).
- Backenköhler, Michael, Luca Bortolussi, Gerrit Großmann, and Verena Wolf (2021). "Abstraction-Guided Truncations for Stationary Distributions of Markov Population Models." In: *18th International Conference on Quantitative Evaluation of Systems*. Vol. 12846. Lecture Notes in Computer Science. Springer, pp. 351–371.
- Backenköhler, Michael, Luca Bortolussi, Gerrit Großmann, and Verena Wolf (2021). "Analysis of Markov Jump Processes under Terminal Constraints." In: *27th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Vol. 12651. Lecture Notes in Computer Science. Springer, pp. 210–229.

- Backenköhler, Michael, Luca Bortolussi, and Verena Wolf (2016). "Generalized Method of Moments for Stochastic Reaction Networks in Equilibrium." In: *International Conference on Computational Methods in Systems Biology*. Vol. 9859. Lecture Notes in Computer Science. Springer, pp. 15–29.
- Backenköhler, Michael, Luca Bortolussi, and Verena Wolf (2018). "Moment-Based Parameter Estimation for Stochastic Reaction Networks in Equilibrium." In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 15.4, pp. 1180–1192.
- Backenköhler, Michael, Luca Bortolussi, and Verena Wolf (2019). "Control Variates for Stochastic Simulation of Chemical Reaction Networks." In: *17th International Conference on Computational Methods in Systems Biology*. Vol. 11773. Lecture Notes in Computer Science. Springer, pp. 42–59.
- Backenköhler, Michael, Luca Bortolussi, and Verena Wolf (2020). "Bounding Mean First Passage Times in Population Continuous-Time Markov Chains." In: *17th International Conference on Quantitative Evaluation of SysTems*. Vol. 12289. Lecture Notes in Computer Science. Springer, pp. 155–174.
- Backenköhler, Michael, Luca Bortolussi, and Verena Wolf (2021). "Variance Reduction in Stochastic Reaction Networks using Control Variates." In: *Festschrift anlässlich Thomas Henzingers 60. Geburtstags*. Vol. to appear. Lecture Notes in Computer Science. Springer.
- Backenköhler, Michael and Gerrit Großmann (2020). "Birth-Death Processes Reproduce the Infection Footprint of Complex Networks." In: *7th International Workshop on Hybrid Systems and Biology*.
- Baier, Christel, Boudewijn Haverkort, Holger Hermanns, and Joost-Pieter Katoen (2000). "Model Checking Continuous-Time Markov Chains by Transient Analysis." In: *International Conference on Computer Aided Verification*. Vol. 1855. Lecture Notes in Computer Science. Springer, pp. 358–372.
- Baier, Christel, Boudewijn Haverkort, Holger Hermanns, and Joost-Pieter Katoen (2003). "Model-Checking Algorithms for

- Continuous-Time Markov Chains." In: *IEEE Transactions on Software Engineering* 29.6, pp. 524–541.
- Barzel, Baruch and Ofer Biham (2008). "Calculation of Switching Times in the Genetic Toggle Switch and other Bistable Systems." In: *Physical Review E* 78.4, p. 041919.
- Beentjes, Casper (2021). "Variance Reduction Techniques for Chemical Reaction Network Simulation." PhD thesis. University of Oxford.
- Beica, Andreea, Jérôme Feret, and Tatjana Petrov (2020). "Tropical Abstraction of Biochemical Reaction Networks with Guarantees." In: *Electronic Notes in Theoretical Computer Science* 350, pp. 3–32.
- Bel, Golan, Brian Munsky, and Ilya Nemenman (2009). "The Simplicity of completion Time Distributions for Common Complex Biochemical Processes." In: *Physical Biology* 7.1, p. 016003.
- Bentriou, Mahmoud, Paolo Ballarini, and Paul-Henry Cournède (2021). "Automaton-ABC: A Statistical Method to Estimate the Probability of Spatio-Temporal Properties for Parametric Markov Population Models." In: *Theoretical Computer Science* 893, pp. 191–219.
- Bogomolov, Sergiy, Thomas A Henzinger, Andreas Podelski, Jakob Ruess, and Christian Schilling (2015). "Adaptive Moment Closure for Parameter Inference of Biochemical Reaction Networks." In: *International Conference on Computational Methods in Systems Biology*. Vol. 9308. Lecture Notes in Computer Science. Springer, pp. 77–89.
- Bortolussi, Luca, Jane Hillston, Diego Latella, and Mieke Massink (2013). "Continuous Approximation of Collective System Behaviour: A Tutorial." In: *Performance Evaluation* 70.5, pp. 317–349.
- Bortolussi, Luca and Roberta Lanciani (2013). "Model Checking Markov Population Models by Central Limit Approximation." In: *International Conference on Quantitative Evaluation of Systems*. Springer, pp. 123–138.

- Bortolussi, Luca and Roberta Lanciani (2014). "Stochastic Approximation of Global Reachability Probabilities of Markov Population Models." In: *Computer Performance Engineering - 11th European Workshop*, pp. 224–239.
- Bortolussi, Luca, Roberta Lanciani, and Laura Nenzi (2018). "Model Checking Markov Population Models by Stochastic Approximations." In: *Information and Computation* 262, pp. 189–220.
- Bortolussi, Luca, Dimitrios Milios, and Guido Sanguinetti (2015). "Efficient Stochastic Simulation of Systems with Multiple Time Scales via Statistical Abstraction." In: *International Conference on Computational Methods in Systems Biology*. Vol. 9308. Lecture Notes in Computer Science. Springer, pp. 40–51.
- Breuer, Lothar (2003). *From Markov Jump Processes to Spatial Queues*. Kluwer Academic Publishers.
- Brim, Luboš, Milan Česka, and David Šafránek (2013). "Model checking of biological systems." In: *International School on Formal Methods for the Design of Computer, Communication and Software Systems*. Vol. 7938. Lecture Notes in Computer Science. Springer, pp. 63–112.
- Broemeling, Lyle D (2017). *Bayesian Inference for Stochastic Processes*. CRC Press.
- Buchholz, Peter (1994). "Exact and Ordinary Lumpability in Finite Markov Chains." In: *Journal of Applied Probability*, pp. 59–75.
- Budde, Carlos E, Pedro R D'Argenio, and Arnd Hartmanns (2017). "Better Automated Importance Splitting for Transient Rare Events." In: *International Symposium on Dependable Software Engineering: Theories, Tools, and Applications*. Springer, pp. 42–58.
- Byrne, George D. and Alan C. Hindmarsh (1975). "A Polyalgorithm for the Numerical Solution of Ordinary Differential Equations." In: *ACM Transactions on Mathematical Software* 1.1, pp. 71–96.

- Cao, Wei-Lu and William J Stewart (1985). "Iterative Aggregation/Disaggregation Techniques for Nearly Uncoupled Markov Chains." In: *Journal of the ACM* 32.3, pp. 702–719.
- Cao, Yang, Daniel T Gillespie, and Linda R Petzold (2005). "The Slow-Scale Stochastic Simulation Algorithm." In: *The Journal of Chemical Physics* 122.1, p. 014116.
- Cardelli, Luca and Attila Csikász-Nagy (2012). "The Cell Cycle Switch Computes Approximate Majority." In: *Scientific reports* 2, p. 656.
- Češka, Milan and Jan Kretínský (2019). "Semi-quantitative Abstraction and Analysis of Chemical Reaction Networks." In: *International Conference on Computer Aided Verification*. Vol. 11561. Lecture Notes in Computer Science. Springer, pp. 475–496.
- Chen, Taolue, Marco Diciolla, Marta Kwiatkowska, and Alexandru Mereacre (2011). "Time-Bounded Verification of CTMCs Against Real-Time Specifications." In: *International Conference on Formal Modeling and Analysis of Timed Systems*. Springer, pp. 26–42.
- Chen, Taolue, Tingting Han, Joost-Pieter Katoen, and Alexandru Mereacre (2009). "Quantitative Model Checking of Continuous-Time Markov Chains Against Timed Automata Specifications." In: *2009 24th IEEE Symposium on Logic In Computer Science*. IEEE, pp. 309–318.
- Cheng, Russell CH (1978). "Analysis of Simulation Experiments Under Normality Assumptions." In: *Journal of the Operational Research Society* 29.5, pp. 493–497.
- Chong, Lillian T, Ali S Saglam, and Daniel M Zuckerman (2017). "Path-Sampling Strategies for Simulating Rare Events in Biomolecular Systems." In: *Current Opinion in Structural Biology* 43, pp. 88–94.
- Cranmer, Miles (2020). *PySR: Fast & Parallelized Symbolic Regression in Python/Julia*. URL: <http://doi.org/10.5281/zenodo.4041459>.
- Daigle Jr, Bernie J, Min K Roh, Dan T Gillespie, and Linda R Petzold (2011). "Automated Estimation of Rare Event Probabil-

- ities in Biochemical Systems." In: *The Journal of Chemical Physics* 134.4, 01B628.
- David, Alexandre, Kim G Larsen, Axel Legay, Marius Mikučionis, Danny Bøgsted Poulsen, and Sean Sedwards (2015). "Statistical Model Checking for Biological Systems." In: *International Journal on Software Tools for Technology Transfer* 17.3, pp. 351–367.
- Dayar, Tuğrul, Holger Hermanns, David Spieler, and Verena Wolf (2011). "Bounding the Equilibrium Distribution of Markov Population Models." In: *Numerical Linear Algebra with Applications* 18.6, pp. 931–946.
- Dayar, Tuğrul and William J Stewart (1997). "Quasi Lumpability, Lower-Bounding Coupling Matrices, and Nearly Completely Decomposable Markov Chains." In: *SIAM Journal on Matrix Analysis and Applications* 18.2, pp. 482–498.
- Dehnert, Christian, Sebastian Junges, Joost-Pieter Katoen, and Matthias Volk (2017). "A Storm is Coming: A Modern Probabilistic Model Checker." In: *International Conference on Computer Aided Verification*. Vol. 10427. Lecture Notes in Computer Science. Springer, pp. 592–600.
- Diamond, Steven and Stephen Boyd (2016). "CVXPY: A Python-Embedded Modeling Language for Convex Optimization." In: *Journal of Machine Learning Research* 17.83, pp. 1–5.
- Dowdy, Garrett R and Paul I Barton (2018a). "Bounds on Stochastic Chemical Kinetic Systems at Steady State." In: *The Journal of Chemical Physics* 148.8, p. 084106.
- Dowdy, Garrett R and Paul I Barton (2018b). "Dynamic Bounds on Stochastic Chemical Kinetic Systems using Semidefinite Programming." In: *The Journal of Chemical Physics* 149.7, p. 074103.
- Engblom, Stefan (2006). "Computing the Moments of High Dimensional Solutions of the Master Equation." In: *Applied Mathematics and Computation* 180.2, pp. 498–515.
- Ethier, Stewart N and Thomas G Kurtz (2009). *Markov Processes: Characterization and Convergence*. Vol. 282. John Wiley & Sons.

- Feller, William (1971). *An Introduction to Probability Theory and Its Applications*. Wiley & Sons.
- Garvels, Marnix JJ and Dirk P Kroese (1998). "A Comparison of RESTART Implementations." In: *1998 Winter Simulation Conference*. Vol. 1. IEEE, pp. 601–608.
- Gast, Nicolas, Luca Bortolussi, and Mirco Tribastone (2019). "Size Expansions of Mean Field Approximation: Transient and Steady-State Analysis." In: *Performance Evaluation* 129, pp. 60–80.
- Geiger, Bernhard C, Tatjana Petrov, Gernot Kubin, and Heinz Koepl (2014). "Optimal Kullback-Leibler Aggregation via Information Bottleneck." In: *IEEE Transactions on Automatic Control* 60.4, pp. 1010–1022.
- Geva-Zatorsky, Naama, Nitzan Rosenfeld, Shalev Itzkovitz, Ron Milo, Alex Sigal, Erez Dekel, Talia Yarnitzky, Yuval Liron, Paz Polak, Galit Lahav, et al. (2006). "Oscillations and Variability in the p53 System." In: *Molecular Systems Biology* 2.1, pp. 2006–0033.
- Ghusinga, Khem Raj, Andrew Lamperski, and Abhyudai Singh (2018). "Estimating Stationary Characteristic Functions of Stochastic Systems via Semidefinite Programming." In: *European Control Conference*. IEEE, pp. 2720–2725.
- Ghusinga, Khem Raj, Cesar A Vargas-Garcia, Andrew Lamperski, and Abhyudai Singh (2017). "Exact Lower and Upper Bounds on Stationary Moments in Stochastic Biochemical Systems." In: *Physical Biology* 14.4, 04LT01.
- Gihman, Iosif I and Anatoli V Skorohod (1975). *The Theory of Stochastic Processes II*. Springer.
- Gillespie, Daniel T (1977). "Exact Stochastic Simulation of Coupled Chemical Reactions." In: *The Journal of Physical Chemistry* 81.25, pp. 2340–2361.
- Gillespie, Daniel T (2000). "The Chemical Langevin Equation." In: *The Journal of Chemical Physics* 113.1, pp. 297–306.



- Gillespie, Daniel T (2001). "Approximate Accelerated Stochastic Simulation of Chemically Reacting Systems." In: *The Journal of Chemical Physics* 115.4, pp. 1716–1733.
- Glasserman, Paul and Bin Yu (2005). "Large Sample Properties of Weighted Monte Carlo Estimators." In: *Operations Research* 53.2, pp. 298–312.
- Golightly, Andrew and Chris Sherlock (2019). "Efficient Sampling of Conditioned Markov Jump Processes." In: *Statistics and Computing* 29.5, pp. 1149–1163.
- Golightly, Andrew and Darren J Wilkinson (2005). "Bayesian Inference for Stochastic Kinetic Models using a Diffusion Approximation." In: *Biometrics* 61.3, pp. 781–788.
- Golightly, Andrew and Darren J Wilkinson (2011). "Bayesian Parameter Inference for Stochastic Biochemical Network Models using Particle Markov Chain Monte Carlo." In: *Interface focus* 1.6, pp. 807–820.
- Großmann, Gerrit, Michael Backenköhler, and Verena Wolf (2020). "Importance of Interaction Structure and Stochasticity for Epidemic Spreading: A COVID-19 Case Study." In: *17th International Conference on Quantitative Evaluation of SysTems*. Vol. 12289. Lecture Notes in Computer Science. Springer, pp. 211–229.
- Großmann, Gerrit, Michael Backenköhler, and Verena Wolf (2021). "Heterogeneity Matters: Contact Structure and Individual Variation Shape Epidemic Dynamics." In: *PLoS one* 16.7, e0250050.
- Gupta, Ankit, Corentin Briat, and Mustafa Khammash (2014). "A Scalable Computational Framework for Establishing Long-Term Behavior of Stochastic Reaction Networks." In: *PLoS Computational Biology* 10.6, e1003669.
- Gupta, Ankit, Jan Mikelson, and Mustafa Khammash (2017). "A Finite State Projection Algorithm for the Stationary Solution of the Chemical Master Equation." In: *The Journal of Chemical Physics* 147.15, p. 154101.
- Gurobi Optimization, LLC (2021). *Gurobi Optimizer Reference Manual*. URL: <https://www.gurobi.com>.

- Hajnal, Matej, Morgane Nouvian, David Šafránek, and Tatjana Petrov (2019). "Data-Informed Parameter Synthesis for Population Markov Chains." In: *International Workshop on Hybrid Systems Biology*. Springer, pp. 147–164.
- Harris, Charles R et al. (2020). "Array Programming with NumPy." In: *Nature* 585, pp. 357–362.
- Hasenauer, Jan, Philipp Rumschinski, Steffen Waldherr, Steffen Borchers, Frank Allgöwer, and Rolf Findeisen (2009). "Guaranteed Steady-State Bounds for Uncertain Chemical Processes." In: *7th IFAC Symposium on Advanced Control of Chemical Processes*. Vol. 42. 11. Elsevier, pp. 643–648.
- Hasenauer, Jan, Verena Wolf, Atefeh Kazerooni, and Fabian J Theis (2014). "Method of Conditional Moments (MCM) for the Chemical Master Equation." In: *Journal of Mathematical Biology* 69.3, pp. 687–735.
- Hausdorff, Felix (1921). "Summationsmethoden und Momentfolgen." In: *Mathematische Zeitschrift* 9, pp. 74–109.
- Hayden, Richard A, Anton Stefanek, and Jeremy T Bradley (2012). "Fluid Computation of Passage-Time Distributions in Large Markov Models." In: *Theoretical Computer Science* 413.1, pp. 106–141.
- He, Shaobo, Yuexi Peng, and Kehui Sun (2020). "SEIR Modeling of the COVID-19 and its Dynamics." In: *Nonlinear Dynamics*, pp. 1–14.
- Helmes, Kurt and Stefan Röhl (2008). "A Geometrical Characterization of Multidimensional Hausdorff Polytopes with Applications to Exit Time Problems." In: *Mathematics of Operations Research* 33.2, pp. 315–326.
- Helmes, Kurt, Stefan Röhl, and Richard H Stockbridge (2001). "Computing Moments of the Exit Time Distribution for Markov Processes by Linear Programming." In: *Journal of Operations Research* 49.4, pp. 516–530.
- Henzinger, Thomas A, Maria Mateescu, and Verena Wolf (2009). "Sliding Window Abstraction for Infinite Markov Chains." In:

- International Conference on Computer Aided Verification*. Vol. 5643. Lecture Notes in Computer Science. Springer, pp. 337–352.
- Hespanha, João P (2008). “Moment Closure for Biochemical Networks.” In: *2008 3rd International Symposium on Communications, Control and Signal Processing*. IEEE, pp. 142–147.
- Hinton, Andrew, Marta Kwiatkowska, Gethin Norman, and David Parker (2006). “PRISM: A Tool for Automatic Verification of Probabilistic Systems.” In: *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, pp. 441–444.
- Holtorf, Flemming and Paul I Barton (2021). “Tighter Bounds on Transient Moments of Stochastic Chemical Systems.” In: *arXiv preprint arXiv:2104.01309*.
- Huang, Lirong, Loïc Paulevé, Christoph Zechner, Michael Unger, Anders Hansen, and Heinz Koeppel (2016). “Reconstructing Dynamic Molecular States from Single-Cell Time Series.” In: *Journal of The Royal Society Interface* 13.122, p. 20160533.
- Iyer-Biswas, Srividya and Anton Zilman (2016). “First-Passage Processes in Cellular Biology.” In: *Advances in Chemical Physics* 160, pp. 261–306.
- Jahnke, Tobias and Wilhelm Huisinga (2007). “Solving the Chemical Master Equation for Monomolecular Reaction Systems Analytically.” In: *Journal of Mathematical Biology* 54.1, pp. 1–26.
- Jégourel, Cyrille, Axel Legay, and Sean Sedwards (2013). “Importance Splitting for Statistical Model Checking Rare Properties.” In: *International Conference on Computer Aided Verification*. Springer, pp. 576–591.
- Kampen, Nicolaas Godfried Van (1992). *Stochastic Processes in Physics and Chemistry*. Vol. 1. Elsevier.
- Kashima, Kenji and Reiichiro Kawai (2009). “Polynomial Programming Approach to Weak Approximation of Lévy-Driven Stochastic Differential Equations with Application to Option Pricing.” In: *International joint Conference. Institute of Control,*

- Robotics and Systems and Society of Instrument and Control Engineers*. IEEE, pp. 3902–3907.
- Kazeroonian, Atefeh, Fabian J Theis, and Jan Hasenauer (2014). “Modeling of Stochastic Biological Processes with Non-Polynomial Propensities Using Non-Central Conditional Moment Equation.” In: *International Federation of Automatic Control Proceedings Volumes* 47.3, pp. 1729–1735.
- Kim, Jinsu, Jason Dark, German Enciso, and Suzanne Sindi (2020). “Slack Reactants: A State-Space Truncation Framework to Estimate Quantitative Behavior of the Chemical Master Equation.” In: *The Journal of Chemical Physics* 153.5, p. 054117.
- Knuth, Donald E (1993). “Johann Faulhaber and Sums of Powers.” In: *Mathematics of Computation* 61.203, pp. 277–294.
- Kroese, Dirk P, Thomas Taimre, and Zdravko I Botev (2011). *Handbook of Monte Carlo Methods*. Vol. 706. Wiley Series in Probability and Statistics. John Wiley & Sons.
- Kuntz, Juan And Thomas, Philipp And stan, Guy-Bart and Mauricio Barahona (2019). “Bounding the Stationary Distributions of the Chemical Master Equation via Mathematical Programming.” In: *The Journal of Chemical Physics* 151.3, p. 034109.
- Kuntz, Juan, Philipp Thomas, Guy-Bart Stan, and Mauricio Barahona (2019). “The Exit Time Finite State Projection Scheme: Bounding Exit distributions and Occupation Measures of Continuous-time Markov Chains.” In: *SIAM Journal on Scientific Computing* 41.2, A748–A769.
- Kuntz, Juan, Philipp Thomas, Guy-Bart Stan, and Mauricio Barahona (2021a). “Approximations of Countably Infinite Linear Programs over Bounded Measure Spaces.” In: *SIAM Journal on Optimization* 31.1, pp. 604–625.
- Kuntz, Juan, Philipp Thomas, Guy-Bart Stan, and Mauricio Barahona (2021b). “Stationary Distributions of Continuous-Time Markov Chains: A Review of Theory and Truncation-Based Approximations.” In: *SIAM Review* 63.1, pp. 3–64.

- Kurasov, Pavel, Alexander Lück, Delio Mugnolo, and Verena Wolf (2018). "Stochastic Hybrid Models of Gene Regulatory Networks – A PDE Approach." In: *Mathematical Biosciences* 305, pp. 170–177.
- Kurtz, Thomas G (1981). *Approximation of Population Processes*. Vol. 36. Regional Conference Series in Applied Mathematics. SIAM.
- Kuwahara, Hiroyuki and Ivan Mura (2008). "An Efficient and Exact Stochastic Simulation Method to Analyze Rare Events in Biochemical Systems." In: *The Journal of Chemical Physics* 129.16, 10B619.
- Kwiatkowska, Marta, Gethin Norman, and David Parker (2011). "PRISM 4.0: Verification of Probabilistic Real-Time Systems." In: *International Conference on Computer Aided Verification*. Vol. 2324. Lecture Notes in Computer Science. Springer, pp. 585–591.
- L'Ecuyer, Pierre (1994). "Efficiency Improvement and Variance Reduction." In: *26th Conference on Winter Simulation*. Society for Computer Simulation International, pp. 122–132.
- Lapin, Maksim, Linar Mikeev, and Verena Wolf (2011). "SHAVE: Stochastic Hybrid Analysis of Markov Population Models." In: *14th International Conference on Hybrid Systems: Computation and Control*, pp. 311–312.
- Lasserre, Jean-Bernard (2010). *Moments, Positive Polynomials and their Applications*. Vol. 1. World Scientific.
- Lasserre, Jean-Bernard, Tomás Prieto-Rumeau, and Mihail Zervos (2006). "Pricing a Class of Exotic Options via Moments and SDP Relaxations." In: *Mathematical Finance* 16.3, pp. 469–494.
- Lavenberg, Stephen S, Thomas L Moeller, and Peter D Welch (1982). "Statistical Results on Control Variables with Application to Queueing Network Simulation." In: *Operations Research* 30.1, pp. 182–202.
- Lipshtat, Azi, Adiel Loinger, Nathalie Q Balaban, and Ofer Biham (2006). "Genetic Toggle Switch without Cooperative Binding." In: *Physical Review Letters* 96.18, p. 188101.

- Loinger, Adiel, Azi Lipshtat, Nathalie Q Balaban, and Ofer Biham (2007). "Stochastic Simulations of Genetic Switch Systems." In: *Physical Review E* 75.2, p. 021904.
- Lotka, Alfred James (1925). *Elements of Physical Biology*. Williams & Wilkins.
- MOSEK ApS (2018). *MOSEK Optimizer API for C 8.1.0.67*. URL: <https://docs.mosek.com/8.1/capi/index.html>.
- Mateescu, Maria, Verena Wolf, Frederic Didier, and Thomas A Henzinger (2010). "Fast Adaptive Uniformisation of the Chemical Master Equation." In: *IET Systems Biology* 4.6, pp. 441–452.
- Mélykúti, Bence, João P Hespanha, and Mustafa Khammash (2014). "Equilibrium Distributions of Simple Biochemical Reaction Systems for Time-Scale Separation in Stochastic Reaction Networks." In: *Journal of The Royal Society Interface* 11.97, p. 20140054.
- Meurer, Aaron et al. (2017). "SymPy: Symbolic Computing in Python." In: *PeerJ Computer Science* 3, e103. ISSN: 2376-5992.
- Meyn, Sean P and Richard L Tweedie (1993). "Stability of Markovian Processes III: Foster-Lyapunov Criteria for Continuous-Time Processes." In: *Advances in Applied Probability*, pp. 518–548.
- Meyn, Sean P and Richard L Tweedie (2012). *Markov Chains and Stochastic Stability*. Springer.
- Meyn, Sean P, Richard L Tweedie, et al. (1994). "Computable Bounds for Geometric Convergence Rates of Markov Chains." In: *The Annals of Applied Probability* 4.4, pp. 981–1011.
- Mikeev, Linar, Martin R Neuhäüßer, David Spieler, and Verena Wolf (2013). "On-the-Fly Verification and Optimization of DTA-Properties for Large Markov Chains." In: *Formal Methods in System Design* 43.2, pp. 313–337.
- Mikeev, Linar and Werner Sandmann (2019). "Approximate Numerical Integration of the Chemical Master Equation for Stochastic Reaction Networks." In: *arXiv preprint arXiv:1907.10245*.

- Mikeev, Linar, Werner Sandmann, and Verena Wolf (2011). "Efficient Calculation of Rare Event Probabilities in Markovian Queueing Networks." In: *5th International ICST Conference on Performance Evaluation Methodologies and Tools*, pp. 186–196.
- Mikeev, Linar, Werner Sandmann, and Verena Wolf (2013). "Numerical Approximation of Rare Event Probabilities in Biochemically Reacting Systems." In: *International Conference on Computational Methods in Systems Biology*. Vol. 8130. Lecture Notes in Computer Science. Springer, pp. 5–18.
- Miliadis-Argeitis, Andreas and Mustafa Khammash (2014). "Optimization-Based Lyapunov Function Construction for Continuous-Time Markov Chains with Affine Transition Rates." In: *53rd IEEE Conference on Decision and Control*. IEEE, pp. 4617–4622.
- Milner, Peter, Colin S Gillespie, and Darren J Wilkinson (2013). "Moment Closure Based Parameter Inference of Stochastic Kinetic Models." In: *Statistics and Computing* 23.2, pp. 287–295.
- Mode, Charles J and Candace K Sleeman (2000). *Stochastic Processes in Epidemiology: HIV/AIDS, Other Infectious Diseases, and Computers*. World Scientific.
- Mollison, Denis (1991). "Dependence of Epidemic and Population Velocities on Basic Parameters." In: *Mathematical biosciences* 107.2, pp. 255–287.
- Munsky, Brian and Mustafa Khammash (2006). "The Finite State Projection Algorithm for the Solution of the Chemical Master Equation." In: *The Journal of Chemical Physics* 124.4, p. 044104.
- Munsky, Brian, Ilya Nemenman, and Golan Bel (2009). "Specificity and Completion Time Distributions of Biochemical Processes." In: *The Journal of Chemical Physics* 131.23, 12B616.
- Nelson, Barry L (1990). "Control Variate Remedies." In: *Operations Research* 38.6, pp. 974–992.
- Neupane, Thakur, Chris J Myers, Curtis Madsen, Hao Zheng, and Zhen Zhang (2019). "STAMINA: STochastic Approximate Model-Checker for INfinite-state Analysis." In: *International*

- Conference on Computer Aided Verification*. Vol. 11561. Lecture Notes in Computer Science. Springer, pp. 540–549.
- O'Donoghue, B., E. Chu, N. Parikh, and S. Boyd (Nov. 2017). *SCS: Splitting Conic Solver, version 2.1.0*. URL: <https://github.com/cvxgrp/scs>.
- Pardoux, Étienne (2008). *Markov Processes and Applications: Algorithms, Networks, Genome and Finance*. Vol. 796. John Wiley & Sons.
- Parrilo, Pablo A (2003). "Semidefinite Programming Relaxations for Semialgebraic Problems." In: *Mathematical Programming* 96.2, pp. 293–320.
- Rabiner, Lawrence R and Biing-Hwang Juang (1986). "An Introduction to Hidden Markov Models." In: *IEEE ASSP Magazine* 3.1, pp. 4–16.
- Roh, Min K, Bernie J Diagle Jr, Dan T Gillespie, and Linda R Petzold (2011). "State-Dependent Doubly Weighted Stochastic Simulation Algorithm for Automatic Characterization of Stochastic Biochemical Rare Events." In: *The Journal of Chemical Physics* 135.23, p. 234108.
- Sakurai, Yuta and Yutaka Hori (2017). "A Convex Approach to Steady State Moment Analysis for Stochastic Chemical Reactions." In: *56th Conference on Decision and Control*. IEEE, pp. 1206–1211.
- Sakurai, Yuta and Yutaka Hori (2019). "Bounding Transient Moments of Stochastic Chemical Reactions." In: *IEEE Control Systems Letters* 3.2, pp. 290–295.
- Särkkä, Simo (2013). *Bayesian Filtering and Smoothing*. Vol. 3. Cambridge University Press.
- Schnoerr, David, Botond Cseke, Ramon Grima, and Guido Sanguinetti (Nov. 2017). "Efficient Low-Order Approximation of First-Passage Time Distributions." In: *Physical Review Letters* 119 (21), p. 210601.
- Schnoerr, David, Guido Sanguinetti, and Ramon Grima (2014). "Validity Conditions for Moment Closure Approximations in



- Stochastic Chemical Kinetics." In: *The Journal of Chemical Physics* 141.8, 08B616\_1.
- Schnoerr, David, Guido Sanguinetti, and Ramon Grima (2015). "Comparison of Different Moment-Closure Approximations for Stochastic Chemical Kinetics." In: *The Journal of Chemical Physics* 143.18, p. 185101.
- Schnoerr, David, Guido Sanguinetti, and Ramon Grima (2017). "Approximation and Inference Methods for Stochastic Biochemical Kinetics—A Tutorial Review." In: *Journal of Physics A: Mathematical and Theoretical* 50.9, p. 093001.
- Schweitzer, Paul J (1991). "A Survey of Aggregation-Disaggregation in Large Markov Chains." In: *Numerical Solution of Markov Chains* 8, pp. 63–88.
- Singh, Abhyudai and João P Hespanha (2006). "Lognormal moment closures for Biochemical reactions." In: *45th IEEE Conference on Decision and Control*. IEEE, pp. 2063–2068.
- Spieler, David (2010). *Geobound*. URL: <https://mosi.uni-saarland.de/tools/geobound>.
- Spieler, David (2014). "Numerical Analysis of Long-Run Properties for Markov Population Models." PhD thesis. Saarland University.
- Spieler, David, Ernst Moritz Hahn, and Lijun Zhang (2014). "Model Checking CSL for Markov Population Models." In: *Twelfth International Workshop on Quantitative Aspects of Programming Languages and Systems*. Vol. 154. EPTCS, pp. 93–107.
- Stamatakis, Michail and Nikos V Mantzaris (2009). "Comparison of Deterministic and Stochastic Models of the Lac Operon Genetic Network." In: *Biophysical Journal* 96.3, pp. 887–906.
- Stekel, Dov J and Dafyd J Jenkins (2008). "Strong Negative Self Regulation of Prokaryotic Transcription Factors Increases the Intrinsic Noise of Protein Expression." In: *BMC Systems Biology* 2.1, p. 6.
- Stewart, William J (1994). *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press.

- Stewart, William J (2009). *Probability, Markov Chains, Queues, and Simulation: the Mathematical Basis of Performance Modeling*. Princeton University Press.
- Strasser, Michael, Fabian J Theis, and Carsten Marr (2012). "Stability and Multiattractor Dynamics of a Toggle Switch Based on a Two-Stage Model of Stochastic Gene Expression." In: *Biophysical Journal* 102.1, pp. 19–29.
- Szechtman, Roberto (2003). "Control Variate Techniques for Monte Carlo Simulation." In: *35th Conference on Winter Simulation: Driving Innovation*. Winter Simulation Conference, pp. 144–149.
- Tweedie, Richard L (1975). "Sufficient Conditions for Regularity, Recurrence and Ergodicity of Markov Processes." In: *Mathematical Proceedings of the Cambridge Philosophical Society* 78.1, pp. 125–136.
- Ullah, Mukhtar and Olaf Wolkenhauer (July 2009). "Stochastic Approaches for Systems Biology." In: *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* 2, pp. 385–97.
- Van der Vorst, Henk A (1992). "Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of Nonsymmetric Linear Systems." In: *SIAM Journal on scientific and Statistical Computing* 13.2, pp. 631–644.
- Vandenbergh, Lieven (2010). *The CVXOPT Linear and Quadratic Cone Program Solvers*. URL: <http://cvxopt.org/documentation/coneprog.pdf>.
- Villén-Altamirano, Manuel and José Villén-Altamirano (1994). "RESTART: A Straightforward Method for Fast Simulation of Rare Events." In: *Proceedings of Winter Simulation Conference*. IEEE, pp. 282–289.
- Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." In: *Nature Methods* 17, pp. 261–272.
- Wildner, Christian and Heinz Koepl (2019). "Moment-Based Variational Inference for Markov Jump Processes." In: *International Conference on Machine Learning*. PMLR, pp. 6766–6775.

- Wilkinson, Darren J (2018). *Stochastic Modelling for Systems Biology*. CRC press.
- Wilson, James R (1984). "Variance Reduction Techniques for Digital Simulation." In: *American Journal of Mathematical and Management Sciences* 4.3-4, pp. 277-312.
- Zapreev, Ivan and Joost-Pieter Katoen (2006). "Safe On-the-Fly Steady-State Detection for Time-Bounded Reachability." In: *Third International Conference on the Quantitative Evaluation of Systems*. IEEE, pp. 301-310.



---

## LIST OF FIGURES

---

1.1	Chapter dependencies.	7
2.1	Moments and probability distribution $\pi(t)$	25
2.2	Augmented Foster-Lyapunov function	32
2.3	Augmented v. proposal Lyapunov sets	33
2.4	Subset sizes using Lyapunov-Augmentation	34
3.1	Occupation measure $\xi$ and exit location probability measures $v_1$ and $v_2$	49
3.2	Moment matrix scaling	57
3.3	FPT and MFPT distribution and bounds	58
3.4	MFPTs up to a varying time-horizon	59
3.5	MFPT bound convergence	63
4.1	CV correlation characteristics	81
4.2	CV redundancy heuristics	85
4.3	CV mean estimates v. std. mean estimates	86
4.4	Estimates for varying $n_\lambda$ and $k_{\min}$	88
4.5	Estimates for varying redundancy heuristics and moment orders	89
4.6	Varying redundancy heuristics and $k_{\min}$	90
4.7	Cost & variance reduction trade-off	91
4.8	Effect of sample size on control variate performance	92
4.9	Resampling algorithm	94
4.10	Number of CV v. variance reduction	97
4.11	CV for probability estimation	98
5.1	Macro-state transition	108
5.2	Lumping approximation of Model 2	111
6.1	Redirections for bounds	117
6.2	Stationary FSP	118
6.3	State-space refinement algorithm	120
6.4	Overview of truncation refinements	123
6.5	Probability bound widths	124
6.6	Stationary distribution errors	125

6.7	Approximate stationary distribution of the exclusive switch	126
6.8	The sizes of the final truncation v. the threshold parameter $\epsilon$	127
6.9	Sample trajectory (p53)	127
6.10	The final truncation at original granularity derived for the p53 oscillator.	129
6.11	approximate marginal distributions of the p53 stationary distribution	130
7.1	Forward, backward, and bridging probabilities for <a href="#">Model 2</a>	136
7.2	State-space refinement algorithm on two parallel unit-rate arrival processes	140
7.3	State-space truncation for varying values of the threshold parameter $\delta$	141
7.4	Mode probabilities of the exclusive switch bridging problem	144
7.5	Expected occupation time	144
7.6	Bayesian estimation on the <a href="#">SEIR</a> model	148
7.7	Prior and posterior ( <a href="#">SEIR</a> )	148
8.1	Approximate biasing	156
8.2	Piecewise constant jump distributions & biased SSA	160
8.3	Estimates for different sample sizes	161
8.4	Dynamic biasing for the toggle switch	162
8.5	Rare event estimates (toggle switch)	163
8.6	Weight distribution of different <a href="#">IS</a> methods	164
8.7	Comparison of biased sample paths	165
A.1	The slowdown $c_0/c_1$ v. the number of control variates $ P $ .	175

---

## LIST OF TABLES

---

3.1	SDP MFPT bounds	62
3.2	LP MFPT bounds for a Poisson process	67
3.3	LP MFPT bounds for a birth-death process	67
3.4	LP MFPT bounds for Model 1	68
6.1	Probability bound properties for approximation of the stationary distribution	124
6.2	Characteristics of the lower-upper bound intervals	126
7.1	Rare event analysis on the Model 2	141
A.1	Variance reduction results for up to second order moments	176
A.2	Variance reduction results for up to second order moments	177
A.3	Variance reduction results for up to second order moments	178
A.4	Stationary distribution approximation results for Model 9	179
A.5	Stationary distribution approximation results for Model 10	180

---

## LIST OF MODELS

---

1, 5	Dimerization	18, 85
2	Birth-Death Process	20
3	Parallel independent dimerization	60
4	Negative self-regulated gene expression	62
6	Distributive Modification	86
7, 10, 13	Exclusive Switch	86, 123, 142
8	Lac Operon	99
9	Parallel Birth-Death Process	122
11	p53 Oscillator	127
12	Parallel Poisson Process	140
14, 16	Toggle Switch	145, 161
15	Epidemics Model	146



---

## ACRONYMS

---

CDF	cumulative density function
CME	chemical master equation
CRN	chemical reaction network
CV	control variate
CTMC	continuous-time Markov chain
DTMC	discrete-time Markov chain
FPT	first-passage time
FSP	finite state projection
GMP	generalized moment problem
HMM	hidden Markov model
IS	importance sampling
IVP	initial value problem
LP	linear program
MCM	method of conditional moments
MFPT	mean first-passage time
MPM	Markovian population model
ODE	ordinary differential equation
pCTMC	population CTMC
RESTART	repetitive simulation trials after reaching thresholds
SEIR	susceptible-exposed-infected-removed
SDP	semi-definite program
SSA	stochastic simulation algorithm
SRN	stochastic reaction network