Bounds and Low-Rank Approximation for Controlled Markov Processes

by

Flemming Holtorf

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Author . . . .
Department of Chemical Engineering
May 16, 2024

Certified by . . . .
Alan Edelman
Professor of Applied Mathematics, Department of Mathematics
Thesis Supervisor

Certified by . . . .
Richard D. Braatz
Edwin R. Gilliland Professor, Department of Chemical Engineering
Thesis Supervisor

Accepted by . . . .
Hadley D. Sikes
Willard Henry Dow Professor, Department of Chemical Engineering
Graduate Student Officer

Bounds and Low-Rank Approximation for Controlled Markov Processes

by

Flemming Holtorf

Submitted to the Department of Chemical Engineering
on May 16, 2024, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Stochastic processes have captivated scientific interest by balancing conceptual simplicity with the ability to model complex, poorly understood, or even entirely unknown phenomena. Still, the deployment of stochastic process models remains challenging in practice due to their intrinsically uncertain nature, complicating the computation and interpretation of model predictions. This thesis addresses two distinct challenges for the study of Markovian processes and associated control problems: certification and scale.

In the first part of this thesis, we present an algorithmic framework for conservatively answering the question: What is the best performance a controlled jump-diffusion process can attain? Answers to this question, even if conservative, shed light on fundamental limits, allowing us to distinguish situations where intrinsic noise masks poor decisions from situations where any attempt of improvement is futile. We connect infinite-dimensional linear programming over cones of occupation measures to techniques for approximating the solutions of Hamilton-Jacobi-Bellman and Kolmogorov backward equations. The result is a hierarchy of structured sum-of-squares programs that furnishes a sequence of hard, yet computable bounds for common control performance indicators encoding, for instance, operating cost or the probability of failure. These bounds in turn serve as witnesses of fundamental limitations or certificates of optimality and safety.

In the second part of this thesis, we explore the use of the framework developed in part one to shed light on the limits of quantum information technologies by quantifying the performance limits of controlled quantum devices. In the context of open-loop quantum control, our framework improves upon quantum speed limits, such as the Mandelstam-Tamm bound, and provably allows characterization of performance boundaries to arbitrary precision. For closed-loop controlled quantum systems, it constitutes the first ever approach to rigorously bound performance losses induced by continuous measurement.

The third part of the thesis is devoted to the challenge of scale as commonly encountered when studying Markov process models in high-dimensional spaces. Here, the complexity of computing and representing model predictions routinely exceeds available resources and renders interpretation challenging. We develop computational tools based on dynamical low-rank approximation that allow us to extract the dominant characteristic features of processes described by vast, nonlinear matrix-valued differential equations and track their evolution over time at a reduced cost.

The methods developed in this thesis are accompanied by software solutions exploiting the features of the Julia programming language to enable deployment of Markov process models in the context of scientific inquiry and engineering advancement more widely, with greater ease, and rigorous guarantees.

Thesis Supervisor: Alan Edelman
Title: Professor of Applied Mathematics, Department of Mathematics

Thesis Supervisor: Richard D. Braatz
Title: Edwin R. Gilliland Professor, Department of Chemical Engineering

Acknowledgments

Much like a large part of this thesis, a large part of life seems to be about dealing with the unforeseeable consequences of past decisions. Years ago, I made the choice to attend graduate school at MIT, and never before have I grown more, learned more, or found more joy in dealing with the consequences of a decision. However, I would certainly not feel this way if it had not been for the many remarkable and inspiring individuals that supported me on this journey. While a proper thank you to all of them would be longer than this thesis, I wish to express my gratitude to those who had the most significant influence on me with the following words.

To my academic mentors First and foremost, I wish to express my sincere gratitude to my advisor Alan Edelman. I know no one who sees to the essence of a problem quicker or has a better taste in scientific communication than Alan. I feel privileged to have had the chance to learn from him. Most of all, however, I thank Alan for affording me great liberty while simultaneously supporting me in finding my own research path, letting me grow into an independent thinker, but also giving me the confidence that the work I did was truly important.

I further thank the remaining members of my thesis committee: Richard Braatz, William Green, Youssef Marzouk, and Christopher Rackauckas. I thank Richard Braatz for serving as my ChemE co-advisor, for providing advice on both research and administrative matters, and for welcoming me to the Braatz group meetings, which allowed me to stay connected to process systems engineering. I thank Bill Green for engaging with my work, suggesting applications, and even taking the time to write a paper with me. I thank Youssef Marzouk for connecting me to others at MIT with similar research interests and for welcoming me to CSE events. I thank Chris Rackauckas for sharing his expertise and perspective on software development with me.

Lastly, I would like to thank my past advisors: Alexander Mitsos and Larry Biegler. One could not have asked for better academic role models in the formative undergraduate years. I am grateful for their continued support and the passion for scientific inquiry they instilled in me on my way to graduate school.

To my collaborators and colleagues Throughout my time at MIT, I have had the great fortune to work alongside some of the not only most talented and inspiring but simultaneously most supportive and generous individuals imaginable.

I am indebted to my closest collaborators: Frank Schäfer, Torkel Loman, Julian Arnold, Niels Lörch, and Avinash Subramanian. From and with you all, I have learned more about mathematics, physics, engineering, communication, and the right attitude than from anyone else. Thank you for teaching me so many things, for helping me improve my work, for sharing your ideas while making me feel like mine are heard, and most of all for inspiring me to hold myself to scientific standards as high as yours (not to mention the countless laughs we shared along the way). I owe special gratitude to Frank Schäfer. Your thoughtful feedback, contagious enthusiasm, quantum perspective, and general influence are greatly reflected in this thesis.

Beyond our close collaboration, I must thank Frank Schäfer, the undisputed Julia Lab rowing champion, and Torkel Loman, the social catalyst of the lab and best lizard finder I know, for making even the most uneventful times at work special. Thank you for the continued camaraderie, for always being open to distractions, and for engaging in our excessively long lunch debates on the most absurd and inconsequential topics one could imagine (including research at times); while they provided a fun distraction from the mundane aspects of graduate school, I would like to believe that they also led to deeper insights on occasion.

I would also like to thank Gaurav Arya, Theo Diamandis, and Shashi Gowda for the many insightful (and always fun) conversations about applied statistics, optimization, programming, and beyond. Although we never got to collaborate on any particular project very closely, I learned a great deal from all of you and feel lucky to have been in your orbit.

Lastly, I thank all past and present members of the Julia Lab for making the lab such a welcoming place. I especially thank my office mates Avik Pal, Evelyne Ringoot, Frank Schäfer, Gaurav Arya, Theo Diamandis, Torkel Loman, Utkarsh Rajput, and Vaibhav Dixit for making the office, despite its challenging architectural circumstances, such a great space to work (and not work) over the years.

To my friends The by far best part of graduate school are the lifelong friendships one forms along the way.

To Allen Jiang, Brandon Johnston, and Mohammad Alkhadra, thank you for the continued camaraderie and friendship that started on day one in the first year offices of 66, carried me through the first semester, and led to regular (extended) lunches which turned many otherwise uneventful days into a memorable time (not to mention enabling my discovery of the nectar of everlasting youth). Additional thanks to Mohammad Alkhadra – your friendship has been the most reliable source of laughter, learning, and support over the years. The late night/early morning Zoom/chess sessions will always be some of my fondest pandemic memories and I hope we will continue going on shake runs for years to come. You all are the caliber of people I had hoped to meet at MIT.

To Husain Adamji, thank you for being a fantastic roommate, squash partner, and, most of all, friend. I will always look back fondly on our laughter-filled trips, dinners, and long and honest conversations about anything from the state of academia to the most absurd topics. I hope there are many more to come.

To Frank and Nicola Schäfer, Jereon Audenaert, and Torkel Loman, thank you for the many laughs at the Muddy, national park visits, and hikes over the years (not to mention driving to Acadia, twice ... on the same weekend).

To Aditya Limaye and Matthew Van Beek, thank you for being wonderful roommates, for including me in so many social and recreational activities, and for teaching me more about the US than anyone else. Your expertise on seemingly every subject, from quantum physics to the inner workings of the world economy, will never cease to amaze me.

To Amber Phillips, Brianna Lax, Daniel and David Lundberg, Haley Beech, and Kelsey Reed, thank you for appreciating my unusal sense of humor and for the countless parties, trips, hikes, and other activities that you so kindly organized. Without you, I would have only a fraction of the stories to tell and memories to share that I have today.

To my family Finally, I owe the greatest deal of gratitude to my close family in Germany. Thank you, Stella, Bettina, and Dirk for helping me build this life, for being exceptional role models with an unbelievable work ethic, for putting up with my terrible negligence when it comes to keeping in touch, for treating me like I am on vacation whenever I visit home, and not least for being so kind as to ultimately give up on asking me what it is that I actually do for a living. Thank you for the unconditional love and support throughout all these years.

“Life can only be understood backwards, but it must be lived forwards”

– Søren Kierkegaard

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In 1827 the botanist Robert Brown documented the erratic motion of small particles suspended in a viscous medium. In the following decades, fueled by the emergence of the theory of molecules, a debate ensued among contemporary scientists about the origin of this phenomenon known today as Brownian motion. The nearly century-long debate came to its conclusion with a series of articles in the early 1900s by Albert Einstein [1], Marian von Smulochowski [2], and Paul Langevin [3]. What unites these contributions and ultimately led to the breakthrough on this problem was the probabilistic abstraction of a phenomenon simply too overwhelmingly complex to be described deterministically – the disordered collisions between a suspended particle and molecules of the surrounding medium. These ideas initially conceived by the physical community were subsequently extended and put on a firm mathematical ground by the mathematicians of the 20th century including Norbert Wiener, Andrey Kolmogorov, and Paul Lèvy [4]. Made available to the wider scientific community through this effort, today, models of stochastic processes akin to Brownian motion find applications in virtually all disciplines of science and engineering. And although the landscape of stochastic process models has become more diverse since then, the motivation to deploy them in the context of scientific inquiry and engineering advancement remains unchanged. Stochastic process models strike a delicate balance between conceptual simplicity and the capacity of describing complex, erratic behaviors. As such, they promise to be of great utility where deterministic abstractions fall short in capturing the nuances of the dynamics at play, whether that is because the governing phenomena are poorly understood, entirely unknown, or simply too overwhelmingly complex. Consider, for instance, the unpredictable motion of the stock market, the spread of a virus within a population, or the intricacies of cellular communication.

At the same time, facilitated by the steadily improving computational resources and the emergence of powerful programming languages, computational models have become an integral part of modern scientific and engineering activities. Computational models aid not only in unifying explanations for empirical observations but also guide engineering decisions and inform experimental design, paving the way for new discoveries. Unfortunately, however, the conceptual simplicity of stochastic process models rarely carries over to computation, let alone interpretation, of their predictions. The rich and complex behavior of systems typically modeled as stochastic processes gives rise to unwieldy computational representations of model predictions that do not readily compose with established decision-making routines. Additional challenges arise when stochastic models are used to guide decisions in safety- and performance-critical situations, where their inherent randomness renders it difficult to rigorously rule out erroneous conclusions or distinguish poor decisions from unavoidable, intrinsic noise as the cause of undesired outcomes.

The computational methods and tools developed in this thesis promise to enable the deployment of stochastic process models in scientific and engineering workflows more widely, and with greater ease and confidence. In safety- and performance-critical situations, they allow us to remove any doubts about our conclusions by putting hard bounds on failure probabilities or computing witnesses of fundamental limitations. In problems of challenging scale, they allow us to extract the dominant characteristic features of high-dimensional predictions, simplifying interpretation and accelerating computation.

1.1 Overview & contributions

This thesis is comprised of multiple research threads united by the common goal of advancing the computational study of controlled Markov processes through new mathematical methods, applications, and software tools. Accordingly, the contributions of this thesis are organized in three parts, each of which may be read independently. The first two parts consider new methods and applications for certifying robustness, optimality, and fundamental limits of controlled jump and diffusion processes. The third part advances techniques and presents new software tools for addressing the challenge of scale when studying high-dimensional processes governed by matrix-valued differential equations. In the following, we outline the main contents and contributions of each part in greater detail.

Part I: Local occupation measures introduces a mathematical framework for computing guaranteed bounds for the statistics of controlled jump and diffusion processes. It is an extension to the traditional approach of analyzing controlled stochastic processes through the lens of infinite-dimensional linear programming over cones of occupation measures [5, 6]. To boost the overall practicality of this traditional approach, the presented framework leverages a concept that we call local occupation measures as derived from restricting the traditional (global) notion of occupation measures to subdomains of a partition of the process' space-time domain. Through the choice of this partition, it enables fine-grained control over the construction of structured semidefinite bounding problems via the moment-sum-of-squares hierarchy.

Chapter 2 reviews the mathematical foundations on which the presented local occupation measure framework is built. It introduces essential notions of measure theory, convex optimization, moment problems, and real algebraic geometry. A particular emphasis is placed on the duality between moment problems and non-negative polynomials as well as its connection to conic linear and semidefinite programming.

In Chapter 3, we then present the local occupation measure framework for bounding the statistics of controlled jump and diffusion processes. By enabling to bound the best attainable control performance, it complements heuristics and local search techniques for controller design as the bounds shed light on fundamental performance barriers and provide optimality certificates. We show that the local occupation measure framework bridges the gap between discretization-based approximations to the solution of the Hamilton-Jacobi-Bellmann or Kolmogorov backward equations and techniques based on the traditional occupation measures framework and the moment-sum-of-squares hierarchy. With examples from population control and cellular biology, we demonstrate that it yields notable performance gains relative to the traditional approach; its additional flexibility enables the construction of tighter, numerically better conditioned bounding problems.

Chapter 4 concludes Part I with a discussion of the analysis of stochastic reaction systems within the framework of local occupation measures. Here, its bounding capabilities complement commonly employed analysis methods such as approximate sampling techniques [7] or moment closure approximations [8–11]; it enables rigorous error quantification and robustness certification, and provides side information for otherwise difficult to compute quantities such as rare event probabilities or long-term statistics. We show that a range of recently proposed moment bounding schemes for the analysis of stochastic reaction systems [12–16] are unified and extended by the framework of local occupation measures. Moreover, it is shown to bridge the gap between these schemes and the widely adopted technique of finite state projection [17, 18].

Part II: Quantifying the limits of quantum control is dedicated to the task of mapping out the performance boundaries of controlled quantum devices. Positioned as the facilitators of quantum information processing, the performance limits of such devices have immediate consequences for quantum information technologies at large, yet have remained largely unexplored. To address this gap, we leverage the occupation measure framework and its more practical extension put forward in Part I.

After a brief primer on the principles and mathematics of quantum mechanics presented in Chapter 5, we first consider the case of open-loop controlled quantum systems in Chapter 6. Through application of the occupation measure framework, we devise a hierarchy of sum-of-squares bounding problems for the optimal performance of common open-loop quantum optimal control problems. We establish practically verifiable conditions under which the bounds furnished by this hierarchy converge to the true optimal control performance and demonstrate their utility with practically relevant examples. In particular, the bounds are found to yield significantly less conservative estimates of the performance limits as those determined by known quantum speed limits such as the celebrated Mandelstam-Tamm [19] and Margolus-Levitin [20] bounds, or the recent algorithmic bounding approach by Zhang *et al.* [21].

In Chapter 7 we turn to the case of feedback-controlled quantum systems. We identify a class of quantum systems under continuous observation for which the combination of quantum filtering theory, describing the intrinsically stochastic dynamics of such systems, with the occupation measure framework yields a hierarchy of tractable sum-of-squares bounding problems for the best attainable feedback control performance. This result constitutes the first ever methodology to rigorously bound the limits of quantum feedback control and, in particular, the unavoidable losses induced by continuous observation. We further discuss practically relevant extensions of this methodology to quantum feedback control in the presence of unobserved decay channels or measurement inefficiencies, establish technical conditions under which the bounds converge to the best attainable performance, and demonstrate their utility of the method with a qubit control example.

Between the strong theoretical guarantees and empirically good performance of the proposed bounding methods, we argue they can have relevant implications for the design of the next generation of quantum devices. On the one hand, they provide access to heuristic controllers alongside performance bounds which may guide controller design or certify the optimality of a given control policy. On the other hand, by revealing fundamental limitations the bounds can inform assessments of technological potential and early-stage design decisions.

Part III: Dynamical low-rank approximation pivots from the certification questions explored in Parts I and II to tackling the issue of scale in analyzing high-dimensional Markov processes. We focus on compressing predictions of Markov processes described by extensive matrix-valued differential equations as commonly encountered in the study of phenomena governed by stochastic or multi-dimensional partial differential equations. Our approach centers on dynamical low-rank approximation, which invokes the Dirac-Frenkel variational principle to trace an approximate solution of the governing equations through the manifold of low-rank matrices. As such, it reduces the computational load associated with computing model predictions and promotes interpretability by exposing their characteristic features.

Our core contributions, presented in Chapter 8, are novel computational tools and methods for dynamical low-rank approximation under nonlinear dynamics – a feature that frustrates the state of the art. From the methodological perspective, we extend the on-the-fly sparse approximation heuristic recently proposed by Naderi and Babaee [22] to a broader context, enabling dynamical low-rank approximation in the presence of general nonlinear dynamics with local structure. Moreover, we decouple this heuristic from an explicit parameterization of the low-rank manifold, facilitating its seamless composition with robust geometric schemes for numerical dynamical low-rank approximation. Our parametrization-independent approach endows this heuristic with distinctly improved robustness in the presence of small singular values of the low-rank approximation when compared to the original proposal by Naderi and Babaee [22]. From the perspective of computational tooling, we present a performant, yet high-level ecosystem for dynamical low-rank approximation in the Julia programming language. By leveraging the features of the Julia language alongside on-the-fly sparse approximation, it enables streamlined use of dynamical low-rank approximation in a problem-agnostic, minimally intrusive way. Through this combination of methodological advancements and software tools, our efforts simplify the deployment of dynamical low-rank approximation in practice as demonstrated for the model-based uncertainty quantification in solar wind predictions.

Part I

Local occupation measures

Chapter 2

Mathematical Background

Convex conic programming, its special case of semidefinite programming, and their close connection to moment problems and non-negative polynomials lies at the heart of Parts I and II of this thesis. In the following, we introduce all four concepts and highlight key results on which the contributions of these parts are built. Along the way, we recall important foundational mathematical concepts and introduce the notation used throughout.

2.1 Notation

We largely follow the notational conventions summarized below. Occasionally, however, we take the liberty to deviate from these conventions if it avoids unnecessary clutter or improves the clarity of exposition.

General

Unless conflicting with standard notation, we use lowercase symbols to denote scalars and vectors (or scalar- and vector-valued functions), and uppercase symbols for matrices (or matrix-valued functions and operators) as well as sets. Further distinctions between these cases will be clear from context. For common (topological) spaces or sets we use the standard notation. The (non-negative) n-dimensional reals and integers are denoted by $\mathbb{R}^n$ ($\mathbb{R}^n_+$) and $\mathbb{Z}^n$ ($\mathbb{Z}^n_+$), respectively; similarly, the set of symmetric (positive semidefinite) $n \times n$ matrices is denoted by $\mathbb{S}^n$ ($\mathbb{S}^n_+$) and we use the standard shorthand notation $A \succeq B$ for $A - B \in \mathbb{S}^n_+$. The set of m-times differentiable functions on a domain $D \subset \mathbb{R}^n$ is denoted by $\mathcal{C}^m(D)$ and when D is closed differentiability shall be understood in the sense of Whitney [23]. For functions with two arguments that are differentiable n times in their first and m-times in their second argument, we write $\mathcal{C}^{n,m}(D)$. Throughout, we broadly indicate the dual of an object X (for example a vector spaces, operator, cone, ...) by X^*. A (bilinear) pairing between two vector spaces ($\mathcal{X}$, $\mathcal{Y}$) (over the reals) will be denoted by $\langle \,\cdot\,,\,\cdot\,\rangle : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$; when $\mathcal{X}$ and $\mathcal{Y}$ are dual, we refer to the pairing as duality bracket. Lastly, we abbreviate index ranges of the form $\{1, 2, \dots, m\}$ with the shorthand $[m]$.

Polynomials

We denote the ring of polynomials in the variables $x = (x_1, \dots, x_n)$ with real coefficients by $\mathbb{R}[x]$; the subset of polynomials with degree at most d is denoted by $\mathbb{R}_d[x]$. The degree of a polynomial f is referred to as $\deg f$. The addition and multiplication of polynomials (and functions in general) is denoted and defined in the usual way:

$$f = g + h \iff f, g, h \text{ satisfy } f(x) = g(x) + h(x),\ \forall x \in \mathbb{R}^n,$$
$$f = gh \iff f, g, h \text{ satisfy } f(x) = g(x)h(x),\ \forall x \in \mathbb{R}^n.$$

Furthermore, we employ the multi-index notation for monomials: given the multi-index $(j_1, \dots, j_n) \in \mathbb{Z}^n_+$, the corresponding monomial is denoted by $x^j = \prod_{i=1}^n x_i^{j_i}$; we use the shorthand notation $|j| = \sum_{i=1}^n j_i$ to denote the (total) degree of a monomial corresponding to the multi-index j.

Integration

The Lebesgue integral with respect to a measure ρ of a ρ-integrable function $f : X \to \mathbb{R}$ is denoted by $\int_X f(x)\,\mathrm{d}\rho(x)$. The Lebesgue integral with respect to the Lebesgue measure is denoted simply by $\int_X f(x)\,\mathrm{d}x$. For Lebesgue integrals with respect to probability measures, we use the expectation operator; that is, if ρ is a probability measure, we write $\mathbb{E}[f]$ in place of $\int_X f(x)\,\mathrm{d}\rho(x)$. The probability measure associated with an expectation will be clear from context. Lastly, we denote the indicator function of a set A with $\mathbb{1}_A$.

2.2 Measure theory

In preparation for the treatment of moment problems, we review in the following some quintessential notions of measure theory. Specifically, we formalize the distinction between the concept of a distribution and a measure as that will play an important role in the formulation of moment problems. For more details on these and other related measure theoretic concepts, we refer the reader to the first Chapters of the standard texts of Durrett [24] and Kowalski [25] based on which this section was developed.

Before we give a definition of a measure, we recall some other fundamental notions of measure theory. We begin with the concept of a σ-algebra.

Definition 2.1 (σ-algebra)**.** *Let X be a set and $\mathcal{S}$ a set of subsets of X. $\mathcal{S}$ is called a σ-algebra of X if it satisfies the following three conditions*

(i) *$\mathcal{S}$ contains the empty set, i.e., $\emptyset \in \mathcal{S}$*

(ii) *$\mathcal{S}$ is closed under countable union, i.e., $\{A_i \in \mathcal{S}\}_{i\geq 1} \implies \cup_{i\geq 1} A_i \in \mathcal{S}$*

(iii) *$\mathcal{S}$ is closed under complementation, i.e., $A \in \mathcal{S} \implies X \setminus A \in \mathcal{S}$*

If $A \subset X$ and $A \in \mathcal{S}$, we call A a measurable subset of X.

The most simple examples of σ-algebras are the trivial σ-algebra $\mathcal{S} = \{\emptyset, X\}$ and the discrete σ-algebra given by the powerset $\mathcal{P}(X)$ of X, i.e., $\mathcal{P}(X) = \{A : A \subset X\}$. The powerset is the canonical choice for a σ-algebra when X is at most countable; for example in the case where X denotes all possible realizations of a discrete random variable. For uncountable subsets of $\mathbb{R}^n$, the following notion of the Borel σ-algebra is most frequently employed.

Definition 2.2 (Borel σ-algebra)**.** *The Borel σ-Algebra $\mathcal{B}(X)$ of a topological space X is the smallest σ-Algebra that contains all open subsets of X, i.e., $\mathcal{B}(X) = \cap_{\mathcal{S} \in K} \mathcal{S}$ where $K = \{\mathcal{S} : \mathcal{S}$ is a σ-Algebra of X and $\mathcal{S}$ contains all open subsets of $X\}$. The elements of $\mathcal{B}(X)$ are called Borel sets or Borel subsets of X.*

Probability theory offers a particularly intuitive interpretation to understand the significance of σ-algebras from a practical perspective. For a probabilistic experiment, say tossing a coin a certain number of times, one can think of the σ-algebra as the set of events one might be interested in studying; in this example that could be all possible sequences of heads and tails that are consistent with the number of coin tosses but it could also be simpler, for example only considering if a sequence of coin tosses included a heads or not. This view emphasizes that the choice of a σ-algebra remains in many cases a modeling choice. For example, one might only be interested in studying the probability of one specific event (and with that the probability of its complement). In that case, it is potentially advantageous to choose a σ-algebra that is simpler than the powerset of all possible outcomes of the experiment.

The concept of a σ-algebra leads to the definition of a measurable space and with that finally allows us to formally define the concepts of a measure and measured space.

Definition 2.3 (Measurable space)**.** *A measurable space is a pair of a topological space X and a σ-algebra $\mathcal{S}$ of it.*

Definition 2.4 (Measure)**.** *Let $(X, \mathcal{S})$ be a measurable space. $\rho : \mathcal{S} \to [0, +\infty]$ is a measure on $(X, \mathcal{S})$ if it is a non-negative, countably additive set function, i.e., ρ satisfies the conditions*

(i) $\rho(\emptyset) = 0$.

(ii) *If $\{A_i\}_{i \geq 1}$ is a countable collection of pairwise disjoint elements of $\mathcal{S}$, then $\rho(\cup_{i \geq 1} A_i) = \sum_{i \geq 1} \rho(A_i)$.*

If $\rho(X) < +\infty$, we call ρ finite. If further $\rho(X) = 1$, we call ρ a probability measure.

Definition 2.5 (Measured space)**.** *The triple $(X, \mathcal{S}, \rho)$ is called measured space. When ρ is a probability measure, $(X, \mathcal{S}, \rho)$ is called a probability space.*

Intuitively, a measure assigns a "size" or "mass" to a set. This idea is intimately related to the concept of a distribution as described by a density or mass function in the continuous and discrete case, respectively. Here, the result of integrating the density or summing the probability mass over a specific subset of the support of the distribution can be interpreted as the "size" or "mass" that the distribution assigns to this set. The following two examples demonstrate that, by this construction, every such distribution can be described by an associated measure.

Example 2.1 (Discrete random variable)**.** *Let x be a discrete random variable taking values in an at most countable set $X \subset \mathbb{R}^n$ and let $p : X \to \mathbb{R}$ be the probability mass function associated with the distribution of x. Then, $\rho : \mathcal{P}(X) \to \mathbb{R}$ defined by*

$$\rho(A) = \sum_{x \in A} p(x)$$

is a probability measure on $(X, \mathcal{P}(X))$. Specifically, ρ describes the distribution of x in terms of the probability of all possible events.

Example 2.2 (Continuous random variable)**.** *Let x be a continuous random variable supported on $X \subset \mathbb{R}^n$ and let $p : X \to \mathbb{R}$ be the probability density function associated with the distribution of x. Then, $\rho : \mathcal{B}(X) \to \mathbb{R}$ defined by*

$$\rho(A) = \int_A p(x) \, \mathrm{d}x$$

is a probability measure on $(X, \mathcal{B}(X))$. Specifically, ρ describes the distribution of x in terms of the probability of events in $\mathcal{B}(X)$.

The case of discrete random variables further motivates the following special classes of measures.

Definition 2.6 (Dirac measure)**.** *Consider $(X, \mathcal{S})$ and let $x \in X$. The Dirac measure at x is defined as $\delta_x(A) = \mathbb{1}_A(x)$, i.e., δ_x assigns a size of 1 to every measurable set A if it contains x, and 0 otherwise.*

Definition 2.7 (Discrete measure). *ρ is called a discrete measure if it is a finite or countable sum of non-negatively weighted Dirac measures.*

In the context of moment problems, we will further frequently encounter the class of (finite) Borel measures as defined below.

Definition 2.8 (Borel measure). *Given a topological space X, a measure $\rho : \mathcal{B}(X) \to [0, +\infty]$ is called a Borel measure. If $\rho(X) < +\infty$, ρ is called a finite Borel measure.*

There are two main advantages of describing and analyzing distributions in terms of the associated measure. On the one hand, the measure theoretic view provides a unified perspective in which discrete and continuous distributions, as well as mixtures thereof, can be treated alike. This is emphasized by the above examples as the expectation of any integrable function $f : X \to \mathbb{R}$ can concisely be written and analyzed as

$$\mathbb{E}\left[f\right] = \int_X f(x)\,\mathrm{d}\rho(x) = \begin{cases} \sum\limits_{x \in X} p(x)f(x) & \text{in Example 2.1} \\ \displaystyle\int_X f(x)p(x)\,\mathrm{d}x & \text{in Example 2.2} \end{cases}$$

irrespective of the discrete or continuous nature of the underlying distribution. On the other hand, the measure theoretic perspective provides a sound theoretical bedrock of foundational results to be leveraged for formal analysis.

Another concept we will extensively use throughout this thesis is the support of a measure.

Definition 2.9 (Support). *Given a measured space $(X, \mathcal{S}, \rho)$, the support of ρ is definded and denoted as*

$$\operatorname{supp} \rho = \operatorname{cl}\left(\{A \in \mathcal{S} : \rho(A) > 0\}\right),$$

where cl refers to the closure operation.

In the context of probability measures, the support delineates unlikely from likely outcomes. Events that do not intersect the support of the probability measure occur with zero probability. As such, the notion of support plays also an important role in describing hard constraints on random variables.

Next, we define the notion of a (canonical) moment of a measure or distribution.

Definition 2.10 (Canonical moment)**.** *Consider a measured space $(X, \mathcal{S}, \rho)$ with $X \subset \mathbb{R}^n$ and let $j \in \mathbb{Z}_+^n$. Then, we call*

$$y_j = \int_X x^j \, \mathrm{d}\rho(x)$$

the j^{th} moment of ρ. Likewise, we say μ_j is the j^{th} moment of a distribution if it is the j^{th} moment of the associated measure.

The following generalization of this notion provides additional flexibility.

Definition 2.11 (Generalized moment)**.** *Consider a measured space $(X, \mathcal{S}, \rho)$ with $X \subset \mathbb{R}^n$ and a ρ-integrable function $f : X \to \mathbb{R}$. Then, we define the generalized moment of ρ with respect to f by*

$$z_f = \int_X f(x) \, \mathrm{d}\rho(x).$$

We say z_f is generated by f.

2.3 Convex optimization

In the following, we review two classes of convex optimization problems that relate closely to moment problems: semidefinite & conic linear programming. For a broader introductory treatment of convex optimization, the reader is referred to [26].

2.3.1 Semidefinite programming

Semidefinite programming refers to optimization over the convex cone of symmetric positive semidefinite matrices. A semidefinite program (SDP) in its standard form reads

$$
\begin{aligned}
\inf_{X \in \mathbb{S}^n} \quad & \operatorname{tr}(CX) && (2.1)\\
\text{s.t.} \quad & \operatorname{tr}(A_i X) = b_i, \quad \forall i \in [m], \\
& X \succeq 0,
\end{aligned}
$$

where $C \in \mathbb{S}^n$, $A_i \in \mathbb{S}^n,\ i \in [m]$ and $b_i \in \mathbb{R},\ i \in [m]$ is considered the problem data while X denotes the decision variable of the optimization problem. Note that we deliberately use the infimum in the formulation of (2.1), indicating that the optimal value of SDPs need not be attained [27]. While every SDP can be recast in the standard form (2.1), from a practitioner's point of view the following formulation is often more natural:

$$
\begin{aligned}
\inf_{x \in \mathbb{R}^n} \quad & c^\top x \\
\text{s.t.} \quad & F_0 + \sum_{i=1}^{n} x_i F_i \succeq 0 && (2.2)\\
& Ax = b.
\end{aligned}
$$

Here, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ and $F_i \in \mathbb{S}^k, i \in [n]$ constitutes the problem data and x is the decision variable. We refer to a constraint of the form (2.2) as *linear matrix inequalitiy (LMI)*. In contrast to the standard form, the above SDP formulation provides useful insights into the geometry of the problem: from a geometric perspective, semidefinite programming can be viewed as the task of optimizing a linear function over a feasible set of the form $F = \{x \in \mathbb{R}^n : Ax = b,\ F_0 + \sum_{i=1}^{n} x_i F_i \succeq 0\}$. In words, F is the intersection of an affine subspace and the set defined by an LMI. Such sets are called *spectrahedra* and they are of remarkable versatility. For instance, spectrahedra include polyhedra and ellipsoids, or more broadly, intersections of polyhedral and second-order cones. Consequently, semidefinite programming encompasses linear and second-order cone programming as special cases. However, spectrahedra, in general, present more diverse nonlinear characteristics. Specifically, guided by Sylvester's cri-

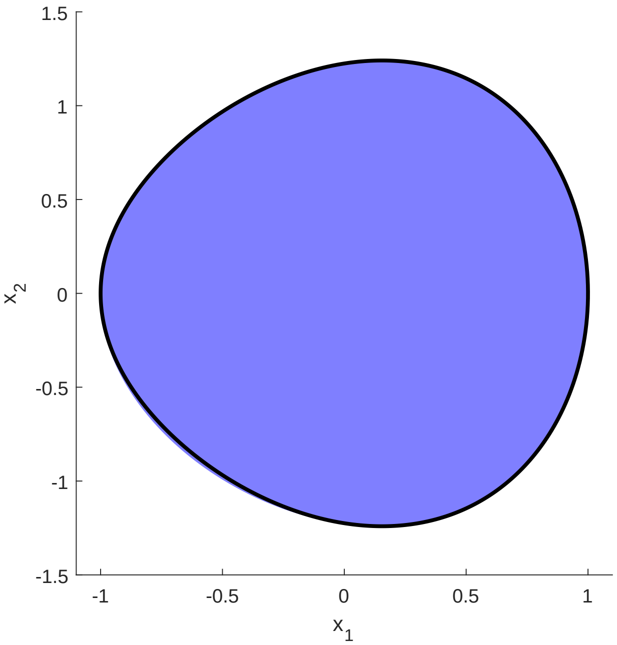


Figure 2-1: Oval of an elliptic curve described by the LMI (2.3).

terion for positive semidefiniteness, a spectrahedron can be described by a finite set of polynomial inequalities, representing the non-negativity of the principal minors of a matrix. The following examples elucidate the intricate nonlinear nature of feasible regions of SDPs.

Example 2.3 (Oval of an elliptic curve [27])**.** *Figure 2-1 shows the spectrahedron*

$$\left\{ x \in \mathbb{R}^2 : M(x_1, x_2) = \begin{bmatrix} x_1 + 1 & 0 & x_2 \\ 0 & 2 & -x_1 - 1 \\ x_2 & -x_1 - 1 & 2 \end{bmatrix} \succeq 0 \right\} \tag{2.3}$$

The boundary of this spectrahedron is given by the oval of the elliptic curve encoded by the condition $\det(M(x_1, x_2)) = 0$.

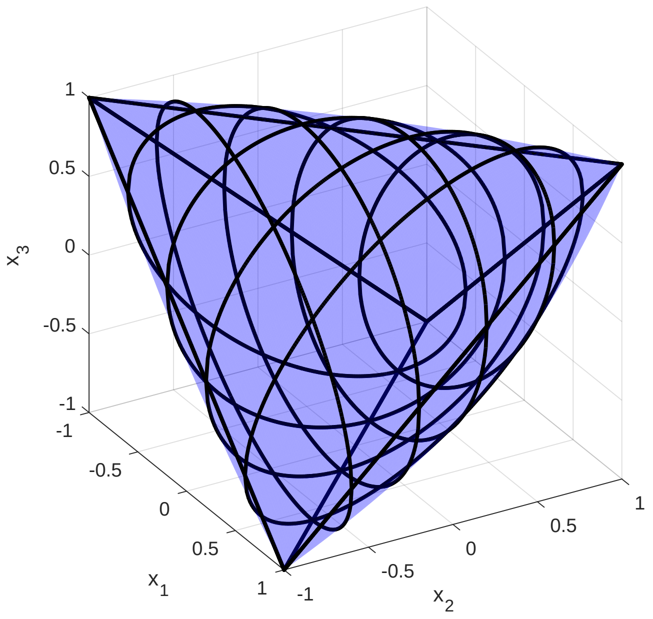


Figure 2-2: Elliptope described by the LMI (2.4).

Example 2.4 (3-D elliptope [27])**.** *Figure 2-2 shows the 3-dimensional elliptope*

$$\left\{ x \in \mathbb{R}^3 : \begin{bmatrix} 1 & x_1 & x_2 \\ x_1 & 1 & x_3 \\ x_2 & x_3 & 1 \end{bmatrix} \succeq 0 \right\} \tag{2.4}$$

The boundary of this spectrahedron is again given by a cubic surface describing the points at which the determinant of the above matrix vanishes.

We call the formulation of Problem (2.1) the *primal SDP* or sometimes simply the *primal* for short. With the primal SDP there is associated a related, often in some sense equivalent, optimization problem called the *dual SDP*:

$$\begin{aligned} \sup_{y \in \mathbb{R}^m} \quad & \sum_{i=1}^m b_i y_i \\ \text{s.t.} \quad & C - \sum_{i=1}^m y_i A_i \succeq 0 \end{aligned} \tag{2.5}$$

It is easily verified that the optimal value of (2.1) is lower bounded by the optimal value of (2.5). To see this, simply note that for any feasible point X of (2.1) and any feasible point y of (2.5), we have that

$$\operatorname{tr}\left(\left(C-\sum_{i=1}^{m} y_i A_i\right) X\right) = \operatorname{tr}(CX) - \sum_{i=1}^{m} y_i \operatorname{tr}(A_i X) = \operatorname{tr}(CX) - \sum_{i=1}^{m} b_i y_i \geq 0,$$

where the second equality and non-negativity follows from the constraints in Problem (2.1) and positive semidefiniteness of X combined with the LMI in Problem (2.5), respectively [28]. This relation between the optimal value of the primal and dual is referred to as weak duality. The argument above also shows that (2.1) and (2.5) in fact have the same optimal value if there exist feasible points y and X such that

$$\operatorname{tr}\left(\left(C-\sum_{i=1}^{m} y_i A_i\right) X\right) = 0.$$

Accordingly, $\operatorname{tr}\left(\left(C-\sum_{i=1}^{m} y_i A_i\right) X\right)$ is called the duality gap and we say strong duality holds if there exist X and y feasible for the primal and dual SDP, respectively, such that the duality gap vanishes.

To conclude this brief introduction to semidefinite programming, we wish to highlight the properties underpinning the immense practical relevance of SDPs. On one hand, semidefinite programming possesses a remarkable modeling power as spectrahedra allow for representation of a wide range of convex sets. Remarkably, this allows for many problems in operations research [26], control theory [29], algebraic geometry [27, 30] and combinatorial optimization [31] to be modeled and solved via SDPs. On the other hand, SDPs are *convex* optimization problems, thus cannot have suboptimal local optima [26]. Moreover, SDPs can in theory be solved efficiently [32] and, in practice, there exist powerful off-the-shelf available solvers [33–35] that reliably solve small to medium scale problems to high accuracy.

2.3.2 Conic linear programming

The notion of semidefinite programming and other prominent classes of convex optimization problems such as linear, second-order cone, or exponential cone programming admit natural generalization by the notion of conic linear programming. Conic linear programming refers broadly to the optimization of a linear function over the intersection of a closed convex cone and an affine set [36]. A conic linear program (CLP) in its standard form reads

$$\begin{aligned} \inf_{x \in \mathcal{X}} \quad & \langle c, x \rangle \\ \text{s.t.} \quad & Ax = b, \\ & x \in K, \end{aligned} \tag{2.6}$$

where x is the decision or optimization variable from a topological vector space $\mathcal{X}$ (over real numbers), $K \subset \mathcal{X}$ is a closed convex cone, $A : \mathcal{X} \to \mathcal{Y}$ is a linear operator encoding together with the right-hand side $b \in \mathcal{Y}$ linear constraints on the decision variable in the topological space $\mathcal{Y}$, and $c \in \mathcal{X}^*$ is an element of the continuous dual space (the space of continuous functionals on $\mathcal{X}$) defining the objective function. For optimization over subsets of finite-dimensional (euclidean) space ($\mathcal{X} \subset \mathbb{R}^n$), the CLP (2.6) includes many prominent classes of convex optimization problems such as linear programming ($K = \mathbb{R}^n_+$), second-order cone programming ($K = \{(x,t) \in \mathbb{R}^{n+1} : \|x\|_2 \leq t, t \geq 0\}$), or semidefinite programming ($K = \mathbb{S}^n_+$) as special case. Throughout this thesis, however, we will frequently consider the case where the space of decision variables and the cone K are infinite-dimensional. Most frequently, we will consider the cases of optimization over the (dual) cones of non-negative measures and continuous functions, in which case we refer to the CLP (2.6) as an infinite-dimensional linear program [37].

Analogous to the case of SDPs as reviewed earlier, any CLP (2.6) has a companion dual problem. The convex dual of (2.6) reads

$$\begin{aligned} \sup_{y \in \mathcal{Y}^*} \quad & \langle y, b \rangle \\ \text{s.t.} \quad & c - A^* y \in K^*, \end{aligned} \tag{2.7}$$

where $A^* : \mathcal{Y}^* \to \mathcal{X}^*$ is the adjoint of A and K^* the dual cone of K, i.e.,

$$K^* = \{s \in \mathcal{X}^* : \langle s, x \rangle \geq 0, \ \forall x \in K\}.$$

By an analogous argument as for SDPs, it is easily verified that weak duality holds between (2.6) and (2.7), i.e., the optimal value of (2.6) is lower bounded by the optimal value of (2.7). To see this, simply note that for any pair of primal-dual feasible points (x, y), we have that

$$\langle c, x \rangle - \langle y, b \rangle = \langle c, x \rangle - \langle y, Ax \rangle = \langle c - A^* y, x \rangle \geq 0.$$

The last inequality follows from the fact that $c - A^* y \in K^*$ and $x \in K$ must hold as x and y are feasible for (2.6) and (2.7), respectively.

2.4 Non-negative polynomials

The theory of moments draws extensively on its duality with the theory of non-negative polynomials. In preparation for the discussion of moment problems, we therefore review some key results in algebraic geometry pertaining to the non-negativity of polynomials in the following.

Definition 2.12 (Non-negative/positive polynomial)**.** *A polynomial $p \in \mathbb{R}[x]$ is called non-negative (positive) when $p(x) \geq 0$ ($p(x) > 0$) for all $x \in \mathbb{R}^n$. Similarly, we say p is non-negative (positive) on $S \subset \mathbb{R}^n$ if $p(x) \geq 0$ ($p(x) > 0$) for all $x \in S$.*

The ability to certify the non-negativity of a polynomial globally or over a subset of its domain finds a wide range of applications in the fields systems, control and optimization [27, 30]. Accordingly, this question has received substantial attention over the last century from theorists and practitioners alike. A deceptively simple case in which one can immediately conclude the global non-negativity of a polynomial is when it is a "sum of squares" as defined below.

Definition 2.13 (Sum-of-squares polynomial)**.** *A polynomial $p \in \mathbb{R}[x]$ is a sum of* *squares or a sum-of-squares polynomial if there exist finitely many polynomials $g_i \in \mathbb{R}[x]$ such that $p = \sum_{i=1}^m g_i^2$. The set of sum-of-squares polynomials of arbitrary degree is denoted by Σ and the subset of sum-of-squares polynomials of degree at most $2d$ is denoted by Σ_d.*

While, as famously noted by Hilbert [38], not all non-negative polynomials are a sum of squares, sum-of-squares polynomials serve as the primary tool for deriving computational certificates of non-negativity. This is because determining whether a polynomial is a sum of squares (up to a given maximum degree) can be reduced to solving an SDP. This is formalized in the following proposition.

Proposition 2.1. *Let d be a positive integer such that $p \in \mathbb{R}_{2d}[x]$. p is a sum of squares if and only if there exists a matrix $Z \in \mathbb{S}_+^{\binom{n+d}{n}}$ such that*

$$p(x) = b(x)^\top Z b(x), \quad \forall x \in \mathbb{R}^n,$$

where b denotes a basis of $\mathbb{R}_d[x]$.

Proof. The if direction follows immediately from the existence of an eigendecomposition of Z [39]. Conversely, if there exist m polynomials g_i such that $p = \sum_{i=1}^m g_i^2$, then clearly no g_i can be of degree greater than d such that there exist $c_i \in \mathbb{R}^{\binom{n+d}{n}}$ for which $g_i = c_i^\top b$. Thus, we can choose $Z = \sum_{i=1}^m c_i c_i^\top \in \mathbb{S}_+^{\binom{n+d}{n}}$. □

In light of Proposition 2.1, feasibility of the following problem

$$\begin{aligned} \text{find} \quad & Z \\ \text{s.t.} \quad & p(x) = b(x)^\top Z b(x), \quad \forall x \in \mathbb{R}^n, \\ & Z \in \mathbb{S}_+^{\binom{n+d}{n}}, \end{aligned}$$

provides a certificate of non-negativity of a given polynomial p of degree at most $2d$. It is worth emphasizing that the above problem translates into a finite semidefinite feasibility problem, as imposing the condition

$$p(x) = b(x)^\top Z b(x), \quad \forall x \in \mathbb{R}^n$$

is equivalent to matching the coefficients of the polynomials on both sides of the equality when expressed in a common basis. This translates into finitely many, in fact $\binom{n+d}{n}$, affine equality constraints on the entries of Z.

The idea of certifying global non-negativity of a polynomial by finding a sum-of-squares representation is easily extended to certifying non-negativity on special subsets of $\mathbb{R}^n$ called basic closed semialgebraic sets.

Definition 2.14 (Basic closed semialgebraic set)**.** *A set $S \subset \mathbb{R}^n$ is called basic closed semialgebraic if it is the intersection of finitely many 0-superlevel sets of polynomials with real coefficients, i.e., S admits a representation of the form $S = \{x \in \mathbb{R}^n : p_i(x) \geq 0,\ i \in [m]\}$ where $p_i \in \mathbb{R}[x]$.*

To that end, note that any polynomial of the form

$$p = \sigma_0 + \sum_{i=1}^{m} \sigma_i p_i, \quad \sigma_0, \dots, \sigma_m \in \Sigma \tag{2.8}$$

must necessarily be non-negative on $S = \{x \in \mathbb{R}^n : p_i(x) \geq 0,\ i \in [m]\}$. Note further that choosing $S = \mathbb{R}^n$ reduces to the previously discussed case of sum-of-squares polynomials. This observation motivates the following definition.

Definition 2.15 (Quadratic module)**.** *Given finitely many polynomials $p_1, \dots, p_m \in \mathbb{R}[x]$, the set of polynomials $Q(p_1, \dots, p_m) = \{\sigma_0 + \sum_{i=1}^m \sigma_i p_i : \sigma_0, \dots, \sigma_m \in \Sigma\}$ is called the quadratic module generated by $p_1, \dots, p_m$. Likewise, we define the truncation of the quadratic module to polynomials of degree at most $2d$ by $Q_d(p_1, \dots, p_m) = \{\sigma_0 + \sum_{i=1}^m \sigma_i p_i : \sigma_0, \dots, \sigma_m \in \Sigma \text{ such that } \deg(\sigma_0), \deg(\sigma_1 p_1), \dots, \deg(\sigma_m p_m) \leq 2d\}$.*

It is a simple corollary of Proposition 2.1 that, given polynomials $p_1, \dots, p_m$ and a degree $d \geq \max\{\deg p_1, \dots, \deg p_m\}$, the membership of p_0 in $Q_d(p_1, \dots, p_m)$ can be tested via solution of a finite SDP. As per Proposition 2.1, the sum-of-squares polynomials $\sigma_0, \ldots, \sigma_m$ in the expression

$$p = \sigma_0 + \sum_{i=1}^{m} \sigma_i p_i$$

can be parameterized by finite positive semidefinite matrices. Equality can be enforced by finitely many equality constraints after expressing all polynomials in a common basis. Feasibility of the resultant SDP certifies non-negativity of p on the basic closed semialgebraic set $\{x \in \mathbb{R}^n : p_i(x) \geq 0, \ \forall i \in [m]\}$ by construction.

So far we have only discussed conditions that are sufficient for the non-negativity of a polynomial on a basic closed semialgebraic set. Another natural and important question is addressing the converse direction: suppose p is non-negative on the basic closed semialgebraic set $\{x \in \mathbb{R}^n : p_i(x) \geq 0, \ \forall i \in [m]\}$, is $p \in Q(p_1, \ldots, p_m)$? Although it is well-known that this is not true in general, there exist practically verifiable conditions under which this implication holds. Results that establish such conditions are known in the literature as Positivstellensätze. We review one such Positivstellensatz due to Putinar next. It relies on the following technical property.

Definition 2.16 (Putinar's condition [40] & Archimedean property [41])**.** *Given polynomials $p_1, \ldots, p_m$, the associated quadratic module $Q(p_1, \ldots, p_m)$ is called Archimedean if there exists $r \in \mathbb{Z}_+$ such that $p + r \in Q(p_1, \ldots, p_m)$ for any polynomial p. Equivalently, $Q(p_1, \ldots, p_m)$ is Archimedean if there exists an integer $r \in \mathbb{Z}_+$ such that $r - \sum_{i=1}^{n} x_i^2 \in Q(p_1, \ldots, p_m)$. We say the polynomials $p_1, \ldots, p_m$ satisfy Putinar's condition if $Q(p_1, \ldots, p_m)$ is Archimedean.*

Remark 2.1. *It is clear from the definition above that if $Q(p_1, \ldots, p_m)$ is Archimedean, the set $S = \{x \in \mathbb{R}^n : p_i(x) \geq 0, \ \forall i \in [m]\}$ must be compact. While the converse conclusion does not hold in general (see Example 7.3 in [41] for a counterexample), S can then easily be modified such that the associated quadratic module is Archimedean; it suffices to include the redundant polynomial inequality constraint $r - \|x\|_2^2 \geq 0$ where $r = \left\lceil \max_{x \in S} \|x\|_2^2 \right\rceil$ in the definition of S.*

Informally, Putinar's Positivstellensatz shows that the Archimedean property guarantees the existence of a non-negativity certificates of the form (2.8) for polynomials that are strictly positive on a basic closed semialgebraic set. The formal statement is given below.

Theorem 2.1 (Putinar's Positivstellensatz [41])**.** *Let $p_1, \dots, p_m$ be polynomials on $\mathbb{R}^n$ and $Q(p_1, \dots, p_m)$ be Archimedean. Then, any polynomial p that is positive on $\{x \in \mathbb{R}^n : p_i(x) \geq 0,\ \forall i \in [m]\}$ satisfies $p \in Q(p_1, \dots, p_m)$.*

We conclude this section by reviewing a key connection between the theory of non-negative polynomials and the theory of moments – the Riesz-Haviland Theorem. To that end, we first need to introduce the so-called Riesz functional as defined in [42].

Definition 2.17 (Riesz functional)**.** *Fix $d \in \mathbb{Z}_+$ and consider a finite sequence of real numbers $\{y_j\}_{|j| \leq d}$. We define $L_y : \mathbb{R}_d[x] \to \mathbb{R}$ such that $L_y(f) = \sum_{|j|\leq d} c_j y_j$ for any polynomial $f(x) = \sum_{|j|\leq d} c_j x^j$. Likewise, for an infinite sequence $\{y_j\}_{j\in\mathbb{Z}_+^n}$ we define $L_y : \mathbb{R}[x] \to \mathbb{R}$ by $L_y(f) = \sum_{j\in\mathbb{Z}_+^n} c_j y_j$. For vector- or matrix-valued arguments, L_y shall be understood as being applied componentwise. We call L_y the Riesz functional.*

With this definition, we can state the Riesz-Haviland Theorem.

Theorem 2.2 (Riesz-Haviland Theorem [42])**.** *Let $S \subset \mathbb{R}^n$ be closed and consider a sequence of real numbers $y = \{y_j\}_{j\in Z_+^n}$. There exists a finite Borel measure ρ supported on S such that*

$$y_j = \int_S x^j \, \mathrm{d}\rho(x), \quad \forall j \in \mathbb{Z}_+^n$$

if and only $L_y(f) \geq 0$ if for any polynomial $f \in \mathbb{R}[x]$ that is non-negative on S.

In light of the Riesz-Haviland Theorem, deciding whether a sequence of real numbers can be associated with the moments of a finite Borel measure is closely tied to characterizing non-negativity of polynomials. The notion of sum-of-squares polynomials (alongside a suitable Positivstellensatz) provides such a characterization and builds the basis for tackling this question computationally through its connection to semidefinite programming.

2.5 Moment problems

In many situations, both in practice and theory, distributions and the associated measures are sufficiently characterized by their moments. In line with this practical relevance, so-called *moment problems* have a long history in mathematics [43]. Traditionally, the term moment problem refers to questions pertaining to the reconstruction of a measure solely from knowledge of its moments.

Definition 2.18 ((Real) Moment Problem). *Let S be a Borel subset of $\mathbb{R}^n$ and consider a sequence of real numbers $\{\mu_j\}_{j\in\mathbb{Z}^n_+}$. Does there exist a measure ρ supported on S such that*

$$y_j = \int_S x^j \,\mathrm{d}\rho(x)$$

holds for all $j \in \mathbb{Z}^n_+$?

With the Hamburger ($S = \mathbb{R}$), Hausdorff ($S = [0, +\infty)$) and Steltjes ($S = [0, 1]$) moment problems, one-dimensional moment problems received substantial attention throughout the 20th century and the results pertaining to existence and uniqueness of measures described by moment sequences in this case are considered rather complete [42]. Higher-dimensional moment problems are less well understood but nevertheless strong results alongside many impactful applications of moment problems and variations thereof have been established over the last years. In particular, the Generalized Moment Problem (GMP) as introduced by Lasserre [42] and its wide range of applications has received substantial attention since the early 2000s following Lasserre's landmark paper [44] and Parrilo's thesis [30].

Definition 2.19 (Primal generalized moment problem [42]). *Let $\mathcal{M}_+(S)$ be the set of finite Borel measures supported on S. Given a Borel set $S \subset \mathbb{R}^n$, an at most countable index set Γ, functions $f, h_j : S \to \mathbb{R}, j \in \Gamma$ and real numbers $\mu_j, j \in \Gamma$ that are integrable with respect to any element in $\mathcal{M}_+(S)$, the (primal) generalized* *moment problem is given by*

$$\begin{aligned} \sup_{\rho \in \mathcal{M}_+(S)} \quad & \int_S f(x) \, \mathrm{d}\rho(x) \\ \text{s.t.} \quad & \int_S h_j(x) \, \mathrm{d}\rho(x) \leq \mu_j, \quad \forall j \in \Gamma. \end{aligned} \tag{GMP}$$

In words, the feasible set of (GMP) is the set of all conceivable measures supported on S that are consistent with given information μ_j about their generalized moments as generated by the functions h_j. (GMP) therefore seeks among those measures those with maximal generalized moment as generated by f. As such, (GMP) provides a natural framework for quantifying the moments of a distribution under partial information. This situation is commonly encountered in practice when only few moments of the distribution of interest can be deduced from measurements or simulation but quantification of additional moments (or related statistics) is sought. If (GMP) is solvable it allows for computation of rigorous upper and lower bounds for the unknown moments or statistics of interest.

A key observation is that (GMP) is a conic linear program, i.e., the objective function and all constraints depend linearly on the decision variable ρ so that the feasible set of (GMP) is the intersection of the convex cone $\mathcal{M}_+(S)$ and the half spaces $H_j = \{\rho : \int_S h_j(x) \, \mathrm{d}\rho(x) \leq \mu_j\}$, $j \in \Gamma$. On the one hand, (GMP) is therefore a convex optimization problem, hence does not exhibit subobtimal local maxima. On the other hand, a crucial consequence of this property is that many instances of (GMP) that arise in practice can be readily solved or approximated by finite dimensional conic optimization problems for which powerful off-the-shelf solvers are available. For example, if S is a finite set, it follows that ρ must be a discrete measure with finitely many atoms such that the corresponding GMP is equivalent to a finite linear program. If S is infinite or even uncountable, we discuss in the following that under relatively mild assumptions (GMP) can be approximated to arbitrary accuracy by suitably constructed SDPs. More specifically, one can construct a hierarchy of SDPs whose optimal values converge from above to the true optimal value of (GMP). This hierarchy, widely referred to as the moment-sum-of-squares or Lasserre hierarchy

[44], allows a practitioner to trade off more computation for higher quality solutions.

The key insight that allows a tractable approximation of (GMP) despite its generally infinite dimensional nature is that, if Γ is finite[1], we can translate (GMP) into an optimization problem with only finitely many decision variables, namely a finite subset of the (generalized) moments of a finite Borel measure. Particularly strong results can be obtained when all the data of (GMP) is further assumed to be characterized by polynomials; that is, if we assume that the set S in Definition 2.19 is basic closed semialgebraic and the functions f and h_j, $j \in \Gamma$ are polynomials of degree at most d. Under these assumptions, (GMP) can be stated equivalently as

$$\begin{aligned} \sup_{y \in \mathbb{R}^{\binom{n+d}{n}}} \quad & L_y(f) \\ \text{s.t.} \quad & L_y(h_j) \leq \mu_j, \quad \forall j \in \Gamma, \\ & \begin{matrix} y \text{ is a truncated moment sequence} \\ \text{consistent with a measure in } \mathcal{M}_+(S) \end{matrix}, \end{aligned} \tag{2.9}$$

where L_y denotes the Riesz functional as described in Definition 2.17. Clearly the objective function and inequality constraints of the above optimization problem are linear in the decision variables. The crux lies in characterizing the condition (2.9). The concept of moment and localizing matrices enable such a characterization.

Definition 2.20 (Moment matrix). *For a non-negative integer d, define $q = \left\lfloor \frac{d}{2} \right\rfloor$ and let b be the monomial basis of $\mathbb{R}_q[x]$ arranged in a vector. Then, we call $M_d(y) = L_y\left(bb^\top\right)$ the moment matrix of degree d.*

Definition 2.21 (Localizing matrix). *For a polynomial $p \in \mathbb{R}[x]$ and a positive integer $d \geq \deg(p)$, define $q = \left\lfloor \frac{d-\deg(p)}{2} \right\rfloor$ and let b be the monomial basis of $\mathbb{R}_q[x]$ arranged in a vector. Then, we call $M_{p,d}(y) = L_y\left(pbb^\top\right)$ the localizing matrix of degree d generated by p.*

[1] If Γ is infinite, one can first relax (GMP) itself by only imposing the constraints corresponding to a finite subset of Γ.

Remark 2.2. *We remark that*

$$\mathbb{R}^{\binom{n+d}{n}} \ni y \mapsto M_d(y) \in \mathbb{S}^{\binom{n+q}{n}}$$

and

$$\mathbb{R}^{\binom{n+d}{n}} \ni y \mapsto M_{p,d}(y) \in \mathbb{S}^{\binom{n+q}{n}}$$

are linear maps between truncated moment sequences and symmetric matrices. Positive semidefiniteness of the moment and localizing matrices therefore describe convex constraints on truncated moment sequences.

As the following proposition shows, positive semidefiniteness of the moment and localizing matrices are necessary conditions for a sequence of real numbers to be associated with a finite Borel measure supported on a basic semi-algebraic set.

Proposition 2.2. *Suppose $S = \{x \in \mathbb{R}^n : p_i(x) \geq 0,\ \forall i \in [m]\}$ is basic closed semi-algebraic and consider a measure $\rho \in \mathcal{M}_+(S)$ with its associated moment sequence $\{y_j\}_{j \in \mathbb{Z}^n_+}$. For any degree $d \in \mathbb{Z}_+$, the truncated moment sequence $y = \{y_j\}_{|j| \leq d}$ satisfies*

$$M_d\left(y\right) \succeq 0$$

and

$$M_{p_i,d}\left(y\right) \succeq 0$$

for any $i \in [m]$ such that $\deg(p_i) \leq d$.

Proof. Fix $d \in \mathbb{Z}_+$ and let b be the monomial basis of $\mathbb{R}_q[x]$ arranged in a vector where $q = \left\lfloor \frac{d}{2} \right\rfloor$. It follows by Definition 2.20 that

$$\int_S b(x)b(x)^\top \,\mathrm{d}\rho(x) = M_d(y).$$

Further let $\mathbb{1}^{\infty}_{\mathbb{S}_+} : \mathbb{S}^{\binom{n+q}{n}} \to \mathbb{R} \cup \{+\infty\}$ be the extended convex indicator function of $\mathbb{S}_+^{\binom{n+q}{n}}$ given by

$$\mathbb{1}^{\infty}_{\mathbb{S}_+}(M) = \begin{cases} 0, & \text{if } M \succeq 0 \\ +\infty, & \text{otherwise} \end{cases}.$$

Since clearly $b(x)b(x)^\top \succeq 0$ for any $x \in \mathbb{R}^n \supset S$, it follows by Jensen's inequality that

$$0 \leq \mathbb{1}^{\infty}_{\mathbb{S}_+} \left(\int_S b(x)b(x)^\top \, \mathrm{d}\rho(x) \right) \leq \int_S \mathbb{1}^{\infty}_{\mathbb{S}_+} \left(b(x)b(x)^\top \right) \mathrm{d}\rho(x) = 0$$

and hence that $M_d(y) \succeq 0$. Using the fact that $p_i(x) \geq 0$ for any $x \in S$, an analogous argument shows that also $M_{p_i,d}(y) \succeq 0$ for all $i \in [m]$ that satisfy $\deg(p_i) \leq d$. □

Under additional regularity conditions, also the converse of Proposition 2.2 holds. That is, positive semidefiniteness of the moment and localizing matrices is in fact sufficient for a sequence of real numbers to be the moments associated with a finite Borel measure.

Theorem 2.3. *Suppose $S = \{x \in \mathbb{R}^n : p_i(x) \geq 0, \ \forall i \in [m]\}$ is basic closed semialgebraic and the quadratic module $Q(p_1, \ldots, p_m)$ is Archimedean. If $\{y_j\}_{j \in \mathbb{Z}^n_+}$ is a sequence of real numbers such that the finite truncations $y = \{y_j\}_{|j| \leq d}$ satisfy $M_d(y) \succeq 0$ and $M_{p_i,d}(y) \succeq 0, \ i \in [m]$ for any integer $d \geq \max_{i \in [m]} \deg(p_i)$, then there exist a measure $\rho \in \mathcal{M}_+(S)$ such that y_j is the j^{th} moment of ρ for all $j \in \mathbb{Z}^n_+$.*

Proof. By the Riesz-Haviland Theorem it suffices to show that $L_y(f) \geq 0$ for any polynomial that is non-negative on S. First let f be positive on S. By Putinar's Positivstellensatz, it follows that there exist sum-of-squares polynomials σ_i such that $f = \sigma_0 + \sum_{i \in [m]} \sigma_i p_i$. Now define $q_i = \deg(p_i)$ and choose d such that $\deg(\sigma_i) \leq 2d$ for all $i = 0, \ldots, m$. Further recall that there exist $c_i \in \mathbb{R}^{\binom{n+d}{n}}$ such that $\sigma_i(x) = c_i^\top b(x)b(x)^\top c_i$ for all $x \in \mathbb{R}^n$ where b denotes the monomial basis of $\mathbb{R}_d[x]$ arranged in a vector. By the linearity of L_y it follows that

$$\begin{aligned} L_y(f) &= L_y(\sigma_0) + \sum_{i\in[m]} L_y(p_i\sigma_i) \\ &= L_y(c_0^\top b(x)b(x)^\top c_0) + \sum_{i\in[m]} L_y\left(p_i(x)c_i^\top b(x)b(x)^\top c_i\right) \\ &= c_0^\top M_{2d}(y)c_0 + \sum_{i\in[m]} c_i^\top M_{p_i,2d+q_i}(y)c_i. \end{aligned}$$

Clearly, the above relation is non-negative if the moment and localizing matrices are positive semidefinite.

Finally, if f is non-negative on S, $f+\epsilon$ is positive on S such that $L_y(f) \geq -\epsilon$ holds by the above argument for any $\epsilon > 0$. Thus, $L_y(f) \geq 0$ must hold. □

From a practical standpoint, Proposition 2.2 establishes that SDPs of the form

$$\begin{aligned} \sup_{y\in\mathbb{R}^{\binom{n+d}{n}}} \quad & L_y(f) \\ \text{s.t.} \quad & L_y(h_j) \leq \mu_j,\ \forall j\in\Gamma, \\ & M_d(y) \succeq 0, \\ & M_{p_i,d}(y) \succeq 0,\ i\in[m], \end{aligned}$$

are tractable relaxations of (GMP). Moreover, these SDP relaxations form a hierarchy of increasingly tight relaxations of (GMP) known as the moment-sum-of-squares or Lasserre hierarchy. As the truncation order d is increased, the relaxations become strictly tighter, leaving a mechanism to balance approximation quality with computational cost. Remarkably, if the hypotheses of Theorem 2.3 are satisfied, it can be shown that the optimal value of the relaxations converges from below to the optimal value of (GMP) in the limit $d \to \infty$ [42, Theorem 4.1].

We conclude this section with a brief discussion of the dual view of (GMP) and its relation to optimization over non-negative polynomials. The dual of (GMP) is given below.

Definition 2.22 (Dual generalized moment problem [42])**.** *Let $f, \mu_j, h_j,\ j\in\Gamma$, and*

S be as in Definition 2.19. Then, the dual GMP is given by

$$\begin{aligned} \inf_{\lambda \in \mathbb{R}^m_+} \quad & \sum_{j \in \Gamma} \lambda_j \mu_j && \text{(dual GMP)} \\ s.t. \quad & \sum_{j \in \Gamma} \lambda_j h_j(x) \geq f(x), \quad \forall x \in S. \end{aligned}$$

It follows immediately from the definition of (dual GMP) that weak duality holds. To see this, let λ and ρ be feasible for the dual and primal GMP, respectively. It follows from the constraints in both problems that

$$\sum_{j \in \Gamma} \lambda_j \mu_j - \int_S f(x) \, \mathrm{d}\rho(x) \geq \sum_{j \in \Gamma} \lambda_j \int_S h_j(x) \, \mathrm{d}\rho(x) - \int_S f(x) \, \mathrm{d}\rho(x) \geq 0.$$

If the data framing the primal GMP (f, $\{h\}_{j \in \Gamma}$, $S = \{x \in \mathbb{R}^n : p_i(x) \geq 0, \ \forall i \in [m]\}$) is given only in terms of polynomials, (dual GMP) is a variant of a polynomial optimization problem in which we seek to find a polynomial p with the structure

$$p = \sum_{j \in \Gamma} \lambda_j h_j - f$$

that is non-negative on the basic closed semialgebraic set S such that a linear combination of its coefficients is minimized. From the discussions in Section 2.4, it is clear that restricting p to be an element of $Q_d(p_1, \ldots, p_m)$ for some degree d results in a tractable approximation of this problem, equivalent to a finite SDP. More precisely, it results in a valid restriction of (dual GMP), thus furnishes again a valid upper bound to the optimal value of (GMP) by weak duality. Overall, the dual view offers a different perspective through the lens of non-negative polynomials and can provide additional insights for the solution of moment problems.

Chapter 3

Stochastic optimal control via local occupation measures

The content of this chapter is based on the preprint F. Holtorf *et al.*, "Stochastic Optimal Control via Local Occupation Measures," *arXiv:2211.15652v2*, 2024

3.1 Introduction

The optimal control of stochastic processes is one of the archetypical problems of decision-making under uncertainty with a myriad of applications in science and engineering. Despite their ubiquity, however, only a small subset of such stochastic optimal control problems admits the computation of a globally optimal control policy in a tractable and certifiable manner. As a consequence, engineers are often forced to resort to one of many available heuristics for the design of control policies in practice. And although such heuristics often perform remarkably well, they seldom come with a simple mechanism to quantify rigorously the degree of suboptimality they introduce, ultimately leaving it to the engineer's intuition when the controller design process shall be terminated.

In response to this undesirable situation, the task of computing theoretically guaranteed yet informative bounds gauging the best attainable performance for various classes of stochastic optimal control and related problems has received considerable attention in the recent past; contributions range from bounding schemes for the optimal control of systems described by deterministic nonlinear ordinary [46–49] and partial differential equations [50, 51] over discrete-time Markov control problems [52, 53] to the control of diffusion and other continuous-time stochastic processes [54–58]. In particular the framework of occupation measures has proved to be a versatile and effective approach to this task. The notion of occupation measures allows for the translation of a rich class of stochastic optimal control problems into infinite-dimensional linear programs over Borel measure spaces [5, 6, 59, 60], for which a sequence of increasingly tight, tractable semidefinite programming (SDP) relaxations is readily constructed via the moment-sum-of-squares (MSOS) hierarchy [30, 44]. A key limitation of this framework, however, remains in its poor scalability. Specifically, the problem size of the SDP relaxations grows combinatorially with the hierarchy level and often high levels are necessary to establish informative bounds in practice. The notorious numerical ill-conditioning of moment problems involving high-order moments further exacerbates this limitation.

In this paper, we set out to improve the practicality of the occupation measure approach to stochastic optimal control by proposing a simple modification of the traditional framework. To that end, we introduce a localized notion of occupation measures based on partitioning of the state space of the controlled process and the control horizon. Analogous to its traditional counterpart, the resultant local occupation measure framework enables the construction of SDP relaxations for a large class of stochastic optimal control problems via the MSOS hierarchy. In contrast to the traditional approach, however, the resultant relaxations can be tightened without increasing the hierarchy level, but instead by simply refining the spatio-temporal partition of the problem domain. Such a “tightening-by-refinement” provides two major practical advantages:

1. It avoids numerical ill-conditioning originating from high-order moments which in practice often prohibits the accurate solution of SDP relaxations furnished

by high levels of the MSOS hierarchy.

2. It provides more fine-grained and easily interpretable control over tightening of the SDP relaxations when compared to increasing the level in the MSOS hierarchy.

As we demonstrate with examples, these advantages hold the potential to construct equally or even tighter relaxations that can be solved notably faster than those derived with the traditional approach. Another potential advantage worth mentioning yet beyond the scope of this work is that the proposed approach is similar in spirit to a wide range of numerical approximation techniques for the solution of partial differential equations (PDEs); as such, the resultant moment-sum-of-squares relaxations exhibit a benign, weakly-coupled block structure akin that of discretized PDEs which may be exploited further, for example by distributed optimization techniques [61, 62].

A partitioning approach closely related to the here proposed local occupation measure framework has recently been studied by Cibulka *et al.* [63] in the context of approximating the region of attraction for deterministic control systems via sum-of-squares programming. In another related work, Holtorf and Barton [16] have used temporal partitioning in order to improve MSOS bounding schemes for trajectories of stochastic chemical systems modeled by jump processes. Both works report significant computational merits of the respective modifications. Here, we unify and extend these contributions by introducing the notion of local occupation measures which applies beyond deterministic control problems to jump and diffusion control problems alike. The resulting framework is independent from and can be complemented by other approaches aimed at improving the tractability and practicality of the MSOS hierarchy, such as symmetry reduction [64, 65], sparsity exploitation [66–68], and linear/second-order cone programming hierarchies [69–72].

The remainder of this chapter is structured as follows: In Section 3.2, we review the concept of occupation measures and show how it enables the construction of tractable convex relaxations for a large class of stochastic optimal control problems with embedded diffusion processes. In Sections 3.3 and 3.4, we introduce the notion of local occupation measures and study its interpretation in the context of stochastic optimal control from the primal (moment) and dual (polynomial) perspective, respectively. Section 3.5 is dedicated to highlight the advantages of the proposed framework for the construction of high quality relaxations with regard to the scaling properties and structure of the resultant optimization problems. In Section 3.6, we demonstrate the practical advantages of the proposed approach with an example problem from population control. In Section 3.7, we discuss the extension of the proposed local occupation measure framework to discounted infinite horizon control problems as well as the control of jump processes, supported with an example from systems biology. We conclude with some final remarks in Section 3.8.

3.2 Problem description & preliminaries

We consider a continuous-time diffusion process x_t in $\mathbb{R}^{n_x}$ driven by a standard $\mathbb{R}^m$-Brownian motion b_t and controlled by a non-anticipative control process u_t in $\mathbb{R}^{n_u}$,

$$\mathrm{d}x_t = f(x_t, u_t)\,\mathrm{d}t + g(x_t, u_t)\,\mathrm{d}b_t, \tag{3.1}$$

and study the associated finite horizon optimal control problem

$$\begin{aligned} J := \inf_{u_t} \quad & \mathbb{E}_{\nu_0}\left[\int_{[0,T]} \ell(x_t, u_t)\,\mathrm{d}t + \phi(x_T)\right] \\ \text{s.t.} \quad & x_t \text{ satisfies (3.1) on } [0,T] \text{ with } x_0 \sim \nu_0, \\ & (x_t, u_t) \in X \times U \text{ on } [0,T], \\ & u_t \text{ is non-anticipative.} \end{aligned} \tag{OCP}$$

Here, $\mathbb{E}_{\nu_0}$ denotes the expectation with respect to the probability measure $\mathbb{P}_{\nu_0}$ over the paths of the diffusion process (3.1). The subscript ν_0 indicates the dependence on the distribution of the initial state, which we assume to be known. Throughout, we further assume that all problem data is described in terms of polynomials in the following sense.

Assumption 3.1. *The drift coefficient $f : X \times U \to \mathbb{R}^{n_x}$, diffusion matrix $gg^\top : X \times U \to \mathbb{R}^{n_x \times n_x}$, stage cost $l : X \times U \to \mathbb{R}$ and terminal cost $\phi : X \times U \to \mathbb{R}$ are componentwise polynomial functions jointly in both arguments. The state space X and the set of admissible control inputs U are basic closed semialgebraic sets.*

We say a control process u_t is admissible if the the controlled process (x_t, u_t) satisfies the constraints in Problem (OCP). Furthermore, we make the following well-posedness assumption that ensures that the optimal value of (OCP) is finite.

Assumption 3.2. *The controlled diffusion process* (3.1) *has finite moments for any admissible control process, i.e., $\mathbb{E}_{\nu_0}[w(x_t, u_t)]$ is finite for all polynomials w and $t \in [0, T]$.*

Note that this assumption does not impose strong practical restrictions as it is for instance implied if the distribution of the controlled process has exponentially decaying tails or if X and U are compact.

The key insight enabling the construction of convex relaxations of (OCP) is that the controlled process described by (3.1) admits a weak-form characterization in terms of a pair of occupations measures: the (terminal) instantaneous and expected state-action occupation measure [5, 6, 60]. This characterization endows the control problem with a convex, albeit infinite-dimensional, geometry, sidestepping the nonlinear dependence of the paths of the diffusion process (3.1) on the control process.

The (terminal) instantaneous occupation measure ν is given by the probability to observe the process state in Borel set $B \subset X$ at the terminal time T. Formally, we define

$$\nu(B) \coloneqq \mathbb{P}_{\nu_0}[x_T \in B].$$

or equivalently,

$$\langle w, \nu \rangle \coloneqq \mathbb{E}_{\nu_0}[w(T, x_T)]$$

for every continuous test function $w \in \mathcal{C}([0,T] \times X)$, where

$$\langle w, \nu \rangle := \int_X w(T,x)\,\mathrm{d}\nu(x)$$

denotes the standard duality bracket between continuous functions and finite measures.

The expected state-action occupation measure ξ is defined as the average time the controlled process (t, x_t, u_t) remains in a Borel subset of $[0,T] \times X \times U$; formally, we define

$$\xi(B_T \times B_X \times B_U) := \mathbb{E}_{\nu_0}\left[\int_{[0,T]\cap B_T} \mathbb{1}_{B_X \times B_U}((x_t,u_t))\,\mathrm{d}t\right]$$

for any Borel subsets $B_T \subset [0,T]$, $B_X \subset X$, $B_U \subset U$; or equivalently,

$$\langle w, \xi \rangle := \mathbb{E}_{\nu_0}\left[\int_{[0,T]} w(t,x_t,u_t)\,\mathrm{d}t\right]$$

for any continuous test function $w \in \mathcal{C}([0,T] \times X \times U)$. The instantaneous and expected state-action occupation measures are finite, non-negative measures by construction. The notions of instantaneous and expected occupation measures are graphically illustrated in Figure 3-1 for an uncontrolled process. The intuition is analogous for the controlled case.

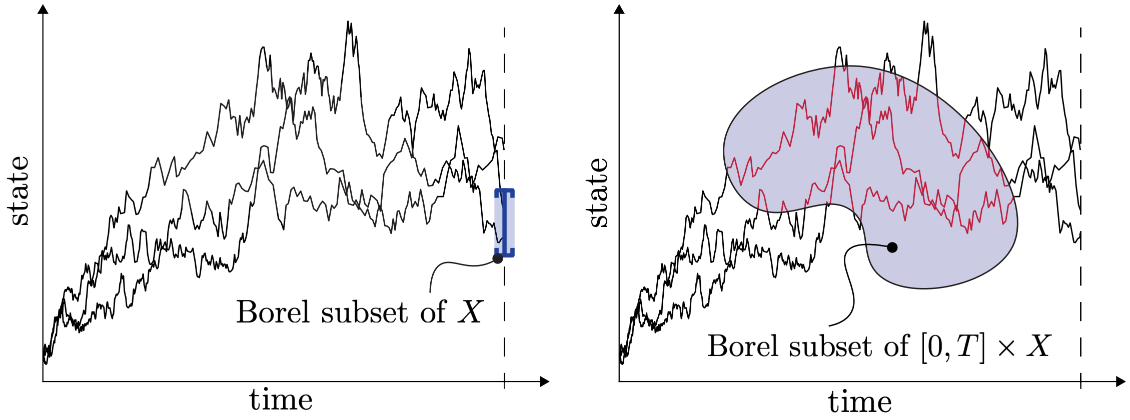
The occupation measure pair (ν, ξ) characterizes the expected time evolution of sufficiently smooth observables[1] $w \in \mathcal{C}^{1,2}([0,T] \times X)$ of the controlled process (t, x_t) by Dynkin's formula [73, Theorem 1.24],

$$\mathbb{E}_{\nu_0}\left[w(T,x_T)\right] = \mathbb{E}_{\nu_0}\left[w(0,x_0)\right] + \mathbb{E}_{\nu_0}\left[\int_{[0,T]} \mathcal{A}w(s,x_s,u_s)\,\mathrm{d}s\right],$$

or equivalently,

$$\langle w, \nu \rangle = \langle w, \nu_0 \rangle + \langle \mathcal{A}w, \xi \rangle\,, \tag{3.2}$$

[1] that is functions on the domain $[0,T] \times X$ with continuous first and second derivatives (in the sense of Whitney [23]) in the first and second argument, respectively.



(a) instantaneous occupation measure

(b) expected state occupation measure

Figure 3-1: Illustration of expected occupation measure pair associated with a stochastic process. Pane (a) illustrates the instantenous occupation measure. The instantaneous occupation measure of a Borel set $B \subset X$ (shown in blue) corresponds intuitively to the fraction of trajectories terminating in B. Pane (b) illustrates the expected occupation measure of an uncontrolled process. The expected state occupation measure of a Borel subset of $S \subset [0,T] \times X$ (shown in blue) corresponds with the average time the process remains in S.

where $\mathcal{A} : \mathcal{C}^{1,2}([0,T] \times X) \to \mathcal{C}([0,T] \times X \times U)$ denotes the (extended) infinitesimal generator of the diffusion process (3.1) [73], i.e.,

$$\mathcal{A} : w(t,x) \mapsto \frac{\partial w}{\partial t}(t,x) + f(x,u)^\top \nabla_x w(t,x) + \frac{1}{2}\mathrm{tr}\left(gg^\top(x,u)\nabla_x^2 w(t,x)\right).$$

Conversely, we say that a measure pair (ν, ξ) is a weak solution to (3.1) on the interval $[0,T]$ if it satisfies Equation (3.2) for all test functions $w \in \mathcal{C}^{1,2}([0,T] \times X)$. This notion of weak solutions to (3.1) motivates the following weak form of (OCP) [5]:

$$\begin{aligned} J^* := \inf_{\nu,\xi} \quad & \langle \ell, \xi \rangle + \langle \phi, \nu \rangle && \text{(weak-OCP)} \\ \text{s.t.} \quad & \langle w, \nu \rangle - \langle \mathcal{A}w, \xi \rangle = \langle w, \nu_0 \rangle, \quad \forall w \in \mathcal{C}^{1,2}([0,T] \times X), \\ & \nu \in \mathcal{M}_+(X), \\ & \xi \in \mathcal{M}_+([0,T] \times X \times U). \end{aligned}$$

where $\mathcal{M}_+(Y)$ denotes the cone of finite, positive Borel measures supported on the set

Y. Problem (weak-OCP) is an infinite-dimensional linear program [37] and generally a relaxation of (OCP), albeit conditions for their equivalence can be established (see for example [6, Section 4]).

From a practical perspective, (weak-OCP) remains intractable as an infinite dimensional linear program; however, Assumption 3.1 enables the construction of a sequence of increasingly tight SDP relaxations via the MSOS hierarchy [30, 44]. To that end, (weak-OCP) is relaxed to the optimization over moment sequences of the measures ν and ξ truncated at finite order d. For polynomial test functions, constraints of the form (3.2) reduce to affine constraints on the moment sequence as $\mathcal{A}$ maps polynomials to polynomials under Assumption 3.1. Similarly, the conic constraints $\nu \in \mathcal{M}_+(X)$ and $\xi \in \mathcal{M}_+([0, T] \times X \times U)$ can be relaxed to positive semidefiniteness constraints of certain moment and localizing matrices, which under Assumption 3.1 reduce to linear matrix inequalities [44]; see Chapter 2 for a detailed discussion on this construction.

The infinite-dimensional linear programming dual [37] to (weak-OCP) has an informative interpretation that serves as motivation for the partitioning strategy presented in the next section. The dual reads

$$\begin{aligned}
\sup_{w} \quad & \int_X w(0,x)\,\mathrm{d}\nu_0(x) && \text{(sub-HJB)}\\
\text{s.t.} \quad & \mathcal{A}w + \ell \geq 0, \ \text{ on } [0,T] \times X \times U, && (3.3)\\
& w(T,\cdot) \leq \phi, \ \text{ on } X, && (3.4)\\
& w \in \mathcal{C}^{1,2}([0,T] \times X),
\end{aligned}$$

where the decision variable w can be interpreted as a smooth underestimator of the value function associated with the control problem (OCP). The following Corollary formalizes this claim.

Corollary 3.1. *Let w be feasible for* (sub-HJB) *and let δ_z denote the Dirac measure* *centered at z. Then, w underestimates the value function*

$$\begin{aligned} V(t,z) \coloneqq \inf_{u_s} \; & \mathbb{E}_{\delta_z}\left[\int_t^T \ell(x_s,u_s)\,\mathrm{d}s + \phi(x_T)\right] \qquad (3.5)\\ \text{s.t. } & x_s \text{ satisfies (3.1) on } [t,T] \text{ with } x_t \sim \delta_z,\\ & (x_s,u_s) \in X \times U \text{ on } [t,T],\\ & u_s \text{ is non-anticipative.} \end{aligned}$$

for any $(t,z) \in [0,T] \times X$.

Proof. Let $z \in X$ and $0 \le t \le T$ and fix any admissible control policy u_s, i.e., a control policy such that the path of the stochastic process (x_s,u_s) remains in $X \times U$ on $[t,T]$. Then, Constraints (3.3) and (3.4) imply that

$$\mathbb{E}_{\delta_z}\left[-\int_t^T \mathcal{A}w(s,x_s,u_s)\,\mathrm{d}s + w(T,x_T)\right] \le \mathbb{E}_{\delta_z}\left[\int_t^T \ell(x_s,u_s)\,\mathrm{d}s + \phi(x_T)\right].$$

The left-hand-side coincides with $w(t,z)$ by Dynkin's formula. The result follows by minimizing over all admissible control policies. □

Remark 3.1. *It is worth emphasizing the interpretation of Corollary 3.1 for the special case of an uncontrolled process ($n_u = 0$) and vanishing stage cost $\ell(x,u) \equiv 0$. In this case, Problems* (weak-OCP) *and* (sub-HJB) *bound the terminal expectations $\mathbb{E}_{\nu_0}[\phi(x_T)]$ of the process. More specifically in light of Corollary 3.1, any feasible point of* (sub-HJB) *bounds the conditional expectation*

$$V(t,z) = \mathbb{E}\left[\phi(x_T) | x_t = z\right]$$

from below. Moreover, for any sufficiently smooth observable $\phi \in \mathcal{C}^2(X)$, it is well-known that V as defined above is the unique solution of the Kolmogorov backward equation [74, Theorem 8.1.1]

$$\begin{cases} \mathcal{A}V = 0, \text{ on } [0,T] \times X \\ V(T,\cdot) = \phi \text{ on } X. \end{cases}$$

Thus it holds in particular that V is feasible for (sub-HJB) *and hence*

$$\mathbb{E}_{\nu_0}[\phi(x_T)] \geq J^* \geq \int_X V(0,x)\,\mathrm{d}\nu_0(x) = \mathbb{E}_{\nu_0}[\phi(x_T)]$$

implies that strong duality holds between (weak-OCP) *and* (sub-HJB) *and in fact* $J^* = \mathbb{E}_{\nu_0}[\phi(x_T)]$.

Analogous to its primal counterpart, the MSOS hierarchy gives rise to a sequence of increasingly tight SDP restrictions of (sub-HJB) by restricting w to be a polynomial of degree at most d and imposing the non-negativity constraints by means of sufficient sum-of-squares conditions [30, 44]. The restriction is weakened by increasing the degree d yielding a monotonically increasing sequence of lower bounds for the optimal value J^* of (weak-OCP). The following theorem establishes a set of easily verifiable conditions under which this sequence converges from below to J^* (implying also strong duality between (sub-HJB) and (weak-OCP)).

Theorem 3.1. *Let J_d be the optimal value of the d^{th} level MSOS restriction of* (sub-HJB) *(resp. relaxation of* (weak-OCP)*). If Assumption 3.1 holds and moreover X and U are represented as*

$$X = \{x : p_i(x) \geq 0,\ i = 1, \dots, v,\ R_X - \|x\|_2^2 \geq 0\},$$
$$U = \{u : q_i(x) \geq 0,\ i = 1, \dots, w,\ R_U - \|u\|_2^2 \geq 0\},$$

with suitable polynomials p_i and q_i, and sufficiently large R_X and R_U, then $J_d \uparrow J^*$.

Proof. First note that under the given assumptions, the set $[0,T] \times X \times U$ is compact. Thus, it suffices to impose condition (3.2) for all polynomial test functions in (weak-OCP) as a dense subset of $\mathcal{C}^{1,2}([0,T] \times X)$. Further observe that constraint (3.2) implies that every feasible pair (ν, ξ) has constant mass. Specifically, for test functions $w(t,x) \equiv 1$ and $w(t,x) = t$, constraint (3.2) reduces to $\langle 1, \nu \rangle = 1$ and $\langle 1, \xi \rangle = T$, respectively. The result thus follows from [75, Corollary 8]. □

Remark 3.2. *The condition imposed by Theorem 3.1 on the representation of X and U is only marginally stronger than imposing their compactness in addition to*

Assumption 3.1. If X and U are compact basic closed semialgebraic sets, one may add redundant ball constraints $R_X - \|x\|_2^2 \geq 0$ and $R_U - \|u\|_2^2 \geq 0$ to their description to enforce the hypotheses of Theorem 3.1.

3.3 The dual perspective revisited: piecewise polynomial approximation

In order to construct improved approximations to the value function in the spirit of (sub-HJB), we consider a generalization of problem (sub-HJB) that seeks a *piecewise smooth underapproximation* of the value function over the problem's space-time domain $[0,T]\times X$. To that end, we consider a discretization $0 = t_0 < t_1 < \cdots < t_{n_T} = T$ of the control horizon and a collection of state space restrictions $X_1, \ldots, X_{n_X} \subset X$ which satisfy the following assumption and hence form a partition of X.

Assumption 3.3. *The collection $X_1, \ldots, X_n \subset \mathbb{R}^{n_X}$ satisfies*

(i) $X = \cup_{k=1}^{n_X} X_k$,

(ii) $X_i \cap X_j = \emptyset$ *for all* $1 \leq i \neq j \leq n_X$.

(iii) *the closure $\bar{X}_i$ and boundary ∂X_i are each the union of finitely many basic closed semialgebraic sets for all $i = 1, \ldots, n_X$*

Remark 3.3. *A partition that satisfies Assumption 3.3 is in practice easily constructed from a collection of disjoint interval boxes $I_1, \ldots, I_{n_X}$ whose union covers the entire state space X. The partition elements may then simply be chosen as $X_k = X \cap I_k$, $k = 1, \ldots, n_X$. This construction is illustrated in Figure 3-2.*

The elements $[t_{i-1}, t_i]\times X_k$ then form a partition of the problem's entire space-time

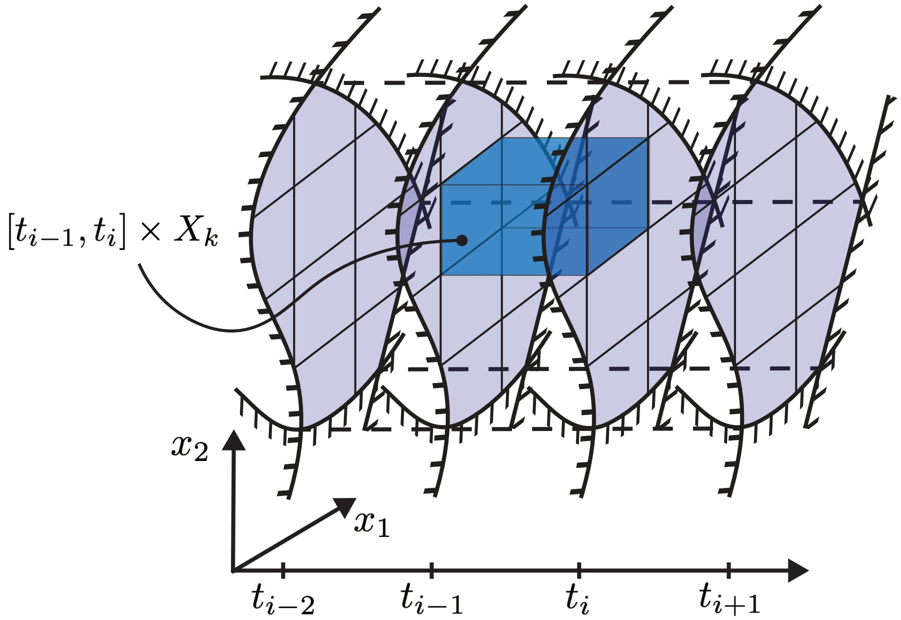


Figure 3-2: Spatio-temporal partition generated by an interval box cover of X. The light blue shaded area are slices of the space-time tube corresponding to a two-dimensional state space X.

domain and we can formulate the following natural generalization of (sub-HJB):

$$\sup_{w} \quad \sum_{k=1}^{n_X} \int_{X_k} w_{1,k}(0,x)\, \mathrm{d}\nu_0(x) \tag{pw-sub-HJB}$$

$$\text{s.t.} \quad \mathcal{A} w_{i,k} + \ell \geq 0 \text{ on } [t_{i-1}, t_i] \times X_k \times U, \quad \forall (i,k) \in P, \tag{3.6}$$

$$w_{i,k}(t_{i-1}, \cdot) \geq w_{i-1,k}(t_{i-1}, \cdot) \text{ on } X_k, \quad \forall (i,k) \in P^\circ, \tag{3.7}$$

$$w_{i,k} = w_{i,j} \text{ on } [t_{i-1}, t_i] \times (\partial X_j \cap \partial X_k), \quad \forall (i,j,k) \in \partial P, \tag{3.8}$$

$$w_{n_T,k}(T, \cdot) \leq \phi \text{ on } X_k, \quad \forall k \in \{1, \ldots, n_X\}, \tag{3.9}$$

$$w_{i,k} \in \mathcal{C}^{1,2}([0,T] \times X_k), \quad \forall (i,k) \in P, \tag{3.10}$$

with the index sets

$$\begin{aligned}
&P \coloneqq \left\{(i,k) \in \mathbb{Z}_+^2 : 1 \leq i \leq n_T, 1 \leq k \leq n_X\right\},\\
&P^\circ \coloneqq \{(i,k) \in \mathbb{Z}_+^2 : 2 \leq i \leq n_T, 1 \leq k \leq n_X\},\\
&\partial P \coloneqq \{(i,j,k) \in \mathbb{Z}_+^3 : 1 \leq i \leq n_T, 1 \leq k \neq j \leq n_X\}.
\end{aligned}$$

The constraints in Problem (pw-sub-HJB) ensure that a valid underestimator of the value function can be constructed from the function pieces $\{w_{i,k} : (i,k) \in P\}$ for all elements of the partition. As such, Problem (pw-sub-HJB) yields a lower bound for the optimal value of (OCP). This is formalized in the following Corollary.

Corollary 3.2. *Let* $\{w_{i,k} : (i,k) \in P\}$ *be feasible for* (pw-sub-HJB) *and define*

$$w(t,x) = w_{i(t),k(x)}(t,x) \text{ where } i(t) = \max\{j : t \in [t_{j-1}, t_j]\} \\ \text{and } k(x) \text{ such that } x \in X_{k(x)}. \quad (3.11)$$

Then, w *underestimates the value function* V *as defined in Equation* (3.5).

Proof sketch. The proof proceeds by splitting the paths of the process (t, x_t, u_t) up into pieces during which it remains confined to a single subdomain $[t_{i-1}, t_i] \times X_k \times U$. For each of resultant pieces an analogous argument as in Corollary 3.1 implies that $w_{i,k}$ underestimates the value function upon confinement of the process to the partition element $[t_{i-1}, t_i] \times X_k \times U$. Additionally, Constraints (3.7) and (3.8) ensure conservative underestimation of the value function when the process crosses between different time intervals and subdomains of the state space, respectively. Specifically, Constraint (3.7) enforces that $w(t, x_t)$ can at most decrease when traced backward in time across the boundary between the intervals $[t_i, t_{i+1}]$ and $[t_{i-1}, t_i]$, ensuring that w cannot cross V at such time points. Similarly, Constraint (3.8) imposes spatial continuity and thus enforces that w cannot cross V when the process crosses spatial boundaries between partition elements. The formal argument is presented in Appendix A.1. □

Remark 3.4. *Contrasting the monotonicity condition* (3.7) *enforced between subsequent time intervals, the stronger continuity requirement* (3.8) *at the boundary between spatial subdomains is necessary as the process may cross the boundary in any direction due to stochastic fluctations. In case of a deterministic process ($g = 0$) this condition may be further relaxed as we only require that $w(t, x_t)$ must at most increase for all trajectories of the system when crossing the boundary between two subdomains. Cibulka* et al. *[63] show in a similar argument that in this case it suffices to impose that*

$$(w_{i,k}(t,x) - w_{i,j}(t,x))n_{j,k}^\top f(x,u) \geq 0, \;\; \forall (t,x,u) \in [t_{i-1}, t_i] \times (\partial X_j \cap \partial X_k) \times U,$$

where $n_{j,k}$ denotes the normal vector of the boundary between X_j and X_k pointing from X_j to X_k.

Remark 3.5. *Valid MSOS restrictions of* (pw-sub-HJB) *are readily obtained simply by restricting each function piece $w_{i,j}$, $(i,j) \in P$ to be a degree-d polynomial and imposing non-negativity constraints through sufficient sum-of-squares constraints on the closure of the respective sets. Note that such sufficient sum-of-squares constraints are indeed well-posed due to Condition (iii) in Assumption* (3.3).

3.4 The primal perspective revisited: local occupation measures

In this section, we discuss the primal counterpart of the construction presented in the previous section. By infinite-dimensional linear programming duality, the primal counterpart of (pw-sub-HJB) reads

$$\begin{aligned} \inf_{\nu,\xi,\pi} \quad & \sum_{(i,k)\in P} \langle \ell, \xi_{i,k}\rangle + \sum_{k=1}^{n_X} \langle \phi, \nu_{n_T,k}\rangle \qquad \text{(pw-weak-OCP)} \\ \text{s.t.} \quad & \langle w, \nu_{i,k}\rangle - \langle w, \nu_{i-1,k}\rangle = \langle \mathcal{A} w, \xi_{i,k}\rangle + \sum_{j\neq k} \langle w, \pi_{i,j,k}\rangle \,, \\ & \qquad\qquad \forall w \in \mathcal{C}^{1,2}([t_{i-1}, t_i]\times X_k),\ \forall (i,k)\in P, \\ & \nu_{i,k} \in \mathcal{M}_+(X_k), \quad \forall (i,k) \in P, \\ & \xi_{i,k} \in \mathcal{M}_+([t_{i-1}, t_i]\times X_k \times U), \quad \forall (i,k)\in P, \\ & \pi_{i,j,k} = -\pi_{i,k,j} \in \mathcal{M}([t_{i-1}, t_i]\times (\partial X_j \cap \partial X_k)), \quad \forall (i,j,k) \in \partial P, \end{aligned}$$

where $\mathcal{M}(Y)$ refers to the space of signed measures supported on Y. The decision variables in (pw-weak-OCP) can be interpreted as localized generalization of the occupation measure pair introduced in Section 3.2. Specifically, the restriction of the expected state-action occupation measures ξ to a subdomain $[t_{i-1}, t_i] \times X_k \times U$ from the partition generates the local state-action occupation measure $\xi_{i,k}$:

$$\xi_{i,k}(B_T \times B_X \times B_U) = \xi((B_T \cap [t_{i-1}, t_i]) \times (B_X \cap X_k) \times B_U).$$

Likewise, the local instantaneous occupation measures with respect to different time points t_i and subdomains X_k are given by the restriction of the instantaneous occupation measure at time t_i to X_k, i.e.,

$$\nu_{i,k}(B) = \mathbb{P}_{\nu_0}(x_{t_i} \in B \cap X_k).$$

The measure $\pi_{i,j,k}$ in (pw-weak-OCP) takes the role of a slack variable and accounts for transitions of the process between the spatial subdomains X_j and X_k in the time interval $[t_{i-1}, t_i]$. Formally, $\pi_{i,j,k}$ can be defined by

$$\langle w, \pi_{i,j,k}\rangle \coloneqq \mathbb{E}_{\nu_0}\left[\sum_{n=1}^{N_+^{jk}} w\left(\tau_{n+}^{jk}, x_{\tau_{n+}^{jk}}\right) - \sum_{n=1}^{N_-^{jk}} w\left(\tau_{n-}^{jk}, x_{\tau_{n-}^{jk}}\right)\right],$$

where τ_{n+}^{jk} and τ_{n-}^{jk} denote the n^{th} time points in $[t_{i-1}, t_i]$ at which the process transitions from subdomain X_j into X_k and vice versa, respectively. With these interpretations, we can observe that the equality constraints in (pw-weak-OCP) reduce to Dynkin's formula applied between the stopping times of leaving and entering a given subdomain X_k in the time interval $[t_{i-1}, t_i]$ (see Appendix A.1 for a more detailed derivation).

Finally, it is important to emphasize that the above interpretation of the decision variables in (pw-weak-OCP) as local occupation measures shows immediately that every feasible point for (pw-weak-OCP) generates a feasible point for (weak-OCP) via the assignment $\xi = \sum_{(i,k)\in P} \xi_{i,k}$ and $\nu = \sum_{k=1}^{n_X} \nu_{n_T,k}$ with equal objective value. Analogously, any smooth function w that is feasible for (sub-HJB) generates upon restriction to the individual subdomains of the partition a feasible point for (pw-sub-HJB) with equal objective value. This property carries over directly to the MSOS restrictions and relaxations of (pw-sub-HJB) and (pw-weak-OCP), respectively, as long as the closure $\bar{X}_k$ of each subdomain is represented in terms of a strictly greater set of polynomial inequalities than X is. This condition, which is easily obeyed in practice (see Remark 3.3), therefore guarantees that MSOS restrictions and relaxations of (pw-sub-HJB) and (pw-weak-OCP), respectively, furnish bounds that are at least as tight as those obtained from the traditional formulation.

3.5 Moment-sum-of-squares approximations: structure & scaling

The construction of tractable relaxations of the problems (sub-HJB) or (weak-OCP) relies on the restriction to optimization over polynomials of fixed degree d or the relaxation to optimization over moment sequences truncated at order d, respectively. Increasing this approximation order d has traditionally been the only mechanism used to weaken the restriction, respectively strengthen the relaxation, to improve the resultant bounds to a desired level. The main motivation behind the proposed partitioning approach lies in circumventing the limited practicality and interpretability of this tightening mechanism. With the proposed notion of local occupation measures, refinement of the space-time domain partition serves as an additional bound tightening mechanism. Table 3.1 summarizes how the MSOS SDP restrictions and relaxations of (pw-sub-HJB) and (pw-weak-OCP) scale in size with respect to the different tightening mechanisms of increasing n_X, n_T (refining the partition), or d (increasing the approximation order). The linear scaling of the SDP sizes with respect to n_X and n_T underlines the fine-grained control over the tightening process via refinement of the partition. In particular, it opens the door to exploit problem specific insights such as the knowledge of critical parts of the (extended) state space $[0, T] \times X$ to be resolved more finely than others, to construct tighter relaxations without incurring a combinatorial increase in the number of partition elements. This flexibility and interpretability is in stark contrast to tightening the bounds by increasing the approximation order d; translating such insights into specific moments to be constrained or polynomial basis elements to be considered for the value function approximator is significantly less straightforward. It is further worth emphasizing that not only the linear scaling with respect to n_T and n_X is desirable but in particular that the invariance of the linear matrix inequality (LMI) dimension promotes practicality due to the unfavorable scaling (worse than cubic) of interior point algorithms with respect to this quantity [76].

Table 3.1: Scaling of problem size of the MSOS SDP approximations for (pw-sub-HJB) and (pw-weak-OCP) with respect to different tightening mechanisms.[2]

	#variables	**# LMIs**	**dimension of LMIs**
d	$O\left(\binom{n_x+1+d}{d}\right)$	$O(1)$	$O\left(\binom{n_x+1+\lfloor (d+c)/2 \rfloor}{\lfloor (d+c)/2 \rfloor}\right)$
n_T	$O(n_T)$	$O(n_T)$	$O(1)$
n_X	$O(n_X)$	$O(n_X)$	$O(1)$

Additionally, the problems (pw-sub-HJB) and (pw-weak-OCP) give rise to highly structured SDPs. Specifically, all constraints involve only variables corresponding to

[2] here, $c = \max\{\deg_u f + \deg_x f - 1, \deg_u g + \deg_x g - 2\}$.

adjacent subdomains. As a consequence, the structure of the constraints is analogous to those arising from discretized PDEs and may be exploited with suitable distributed optimization algorithms and computing architectures.

3.6 Example: population control

3.6.1 Control problem

We demonstrate the computational merits of the proposed local occupation measure framework with an example problem from the field of population control. The objective is to control the population size of a primary predator and its prey in a noisy ecosystem featuring the prey species, primary predator species as well as a secondary predator species. The problem is adapted from Savorgnan *et al.* [53] where it has been studied in a discrete time, infinite horizon setting.

The interactions between the primary predator and prey population are described by a standard Lotka-Volterra model, while the effect of the secondary predator species is modeled by a Brownian motion. The population sizes are assumed to be controlled via hunting of the primary predator species. The controlled evolution of the population sizes is thus described by the diffusion process

$$\begin{aligned}
\mathrm{d}(x_t)_1 &= (\gamma_1 (x_t)_1 - \gamma_2 (x_t)_1 (x_t)_2)\,\mathrm{d}t + \gamma_5 (x_t)_1\,\mathrm{d}b_t,\\
\mathrm{d}(x_t)_2 &= (\gamma_4 (x_t)_1 (x_t)_2 - \gamma_3 (x_t)_2 - (x_t)_2 u_t)\,\mathrm{d}t,
\end{aligned}$$

where x_1, x_2, and u refer to the prey species, predator species, and hunting effort, respectively. The model parameters $\gamma = (1, 2, 1, 2, 0.025)$ and initial condition $x_0 \sim \delta_{(1,0.25)}$ are assumed to be known deterministically. Moreover, we assume that the admissible hunting effort is confined to $U = [0, 1]$. Under these assumptions, it is easily verified that the process state x_t evolves by construction within the non-negative orthant $X = \mathbb{R}^2_+$ for any admissible control policy. For the control problem

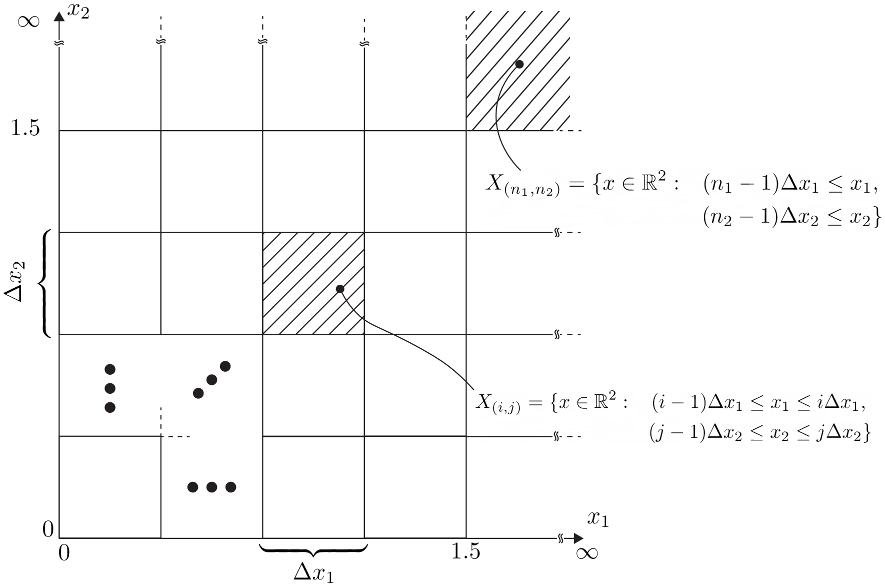

∞
x_2
1.5
$X_{(n_1,n_2)} = \{x \in \mathbb{R}^2 : \quad (n_1-1)\Delta x_1 \le x_1,$
$(n_2-1)\Delta x_2 \le x_2\}$
Δx_2
$X_{(i,j)} = \{x \in \mathbb{R}^2 : \quad (i-1)\Delta x_1 \le x_1 \le i\Delta x_1,$
$(j-1)\Delta x_2 \le x_2 \le j\Delta x_2\}$
0
x_1
0
1.5
∞
Δx_1

Figure 3-3: Spatial partition of $X = \mathbb{R}^2_+$ with $n_1 \times n_2$ interval boxes.

we further choose a time horizon of $T = 10$ and stage cost

$$\ell(x,u) = (x_1 - 0.75)^2 + \frac{(x_2 - 0.5)^2}{10} + \frac{(u - 0.5)^2}{10}$$

penalizing variations from the target population sizes.

3.6.2 Partition of problem domain

In order to investigate the effect of different spatio-temporal partitions on bound quality and computational cost, we utilize a simple grid partition of the state space X as parameterized by the number of grid cells n_1 and n_2 in the x_1 and x_2 direction, respectively. As the state space is the non-negative orthant in this example, and hence semi-infinite, we choose to discretize the compact interval box $[0, 1.5] \times [0, 1.5]$ with a uniform grid of $(n_1 - 1) \times (n_2 - 1)$ cells and cover the remainder of X with appropriately chosen semi-infinite interval boxes. This choice is motivated by the insight that the uncontrolled system resides with high probability in $[0, 1.5] \times [0.1.5]$. The resultant grid is illustrated in Figure 3-3.

The temporal domain is partitioned uniformly into n_T subintervals, i.e., $t_i = i\Delta t$ with $\Delta t = T/n_T$. Throughout, we refer to a specific partition with the associated triple (n_1, n_2, n_T). The computational experiments are conducted for all partitions corresponding to the triples $\{(n_1, n_2, n_T) \in \mathbb{Z}_+^3 : 1 \leq n_1, n_2 \leq 5, 1 \leq n_T \leq 10\}$.

3.6.3 Evaluation of bound quality

In order to assess the tightness of the bounds obtained with different approximation orders and partitions, we compare the relative optimality gap $(\bar{J} - \underline{J})/\bar{J}$, where $\underline{J}$ and $\bar{J}$ refer to the lower bound furnished by an instance of the sum-of-squares restriction of (pw-sub-HJB) and to the control cost associated with the *best known* admissible control policy, respectively. The best known control policy was constructed from the approximate value function w^* obtained as the solution of the sum-of-squares restriction of (pw-sub-HJB) with approximation order $d = 4$ on the grid described by $n_1 = n_2 = 4$ and $n_T = 10$. To that end, we employed the following control law mimicking a one-step model-predictive controller

$$u_t^* \in \operatorname*{arg\,min}_{u \in U} \mathcal{A}w^*(t, x_t, u) + \ell(x_t, u)$$

and estimated the associated control cost

$$\bar{J} = \mathbb{E}_{\nu_0}\left[\int_{[0,T]} \ell(x_t, u_t^*)\,\mathrm{d}t\right]$$

by the ensemble average over $100,000$ sample trajectories generated with the Euler-Maruyama scheme and a step size of 1×10^{-3}.

3.6.4 Computational aspects

All computational experiments presented in this section were conducted on a MacBook M1 Pro with 16GB unified memory. All sum-of-squares programs and the corresponding SDPs were constructed using our custom developed and publicly available

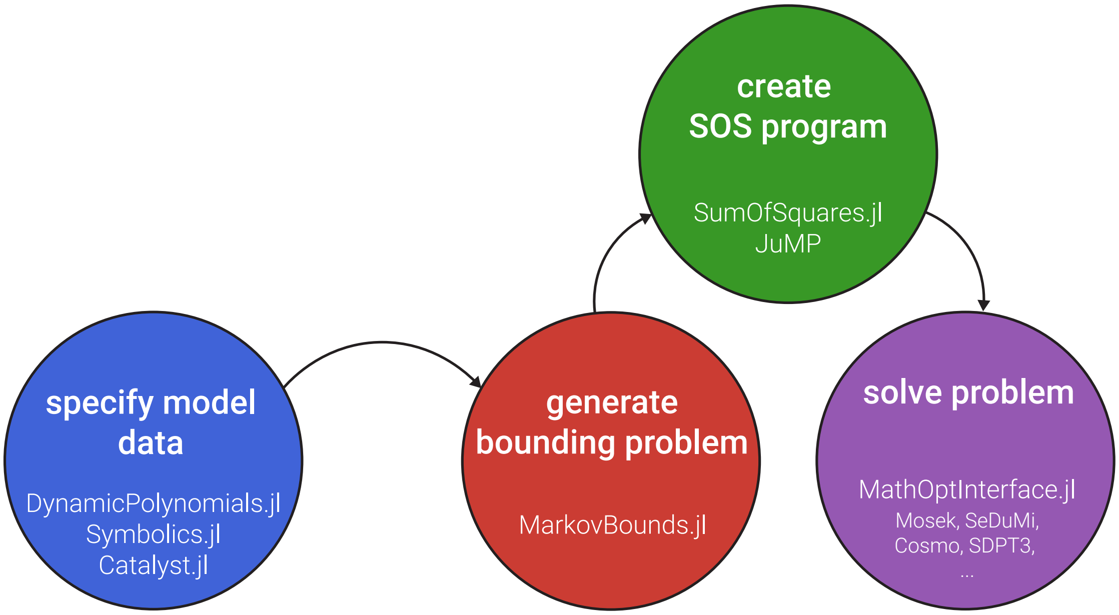


Figure 3-4: Workflow for computing performance bounds for controlled jump-diffusion processes via `MarkovBounds.jl`.

package `MarkovBounds.jl`[3] built on top of `SumOfSquares.jl` [77] and the `MathOptInterface` [78]. Figure 3-4 illustrates how these packages compose to facilitate a largely automated workflow from model specification via symbolic modeling tools to computation of bounds with general purpose SDP solvers. In the following, all SDPs were solved using Mosek v10.

3.6.5 Results

We put special emphasis on investigating the effect of refining the discretization of the problem domain on bound quality and computational cost. Focusing on the effect on computational cost in isolation first, Figure 3-5 confirms that the computational cost for the solution of sum-of-squares restrictions of (pw-sub-HJB) scales approximately linearly with the number of elements $n_1 \times n_2 \times n_T$ of the spatio-temporal partition as discussed in Section 3.5. Moreover, Figure 3-5 also indicates that increasing the approximation order d results in a notably steeper increase in computational cost. These results are in line with the discussion in Section 3.5 (see Table 3.1.

[3] https://github.com/fholtorf/MarkovBounds.jl

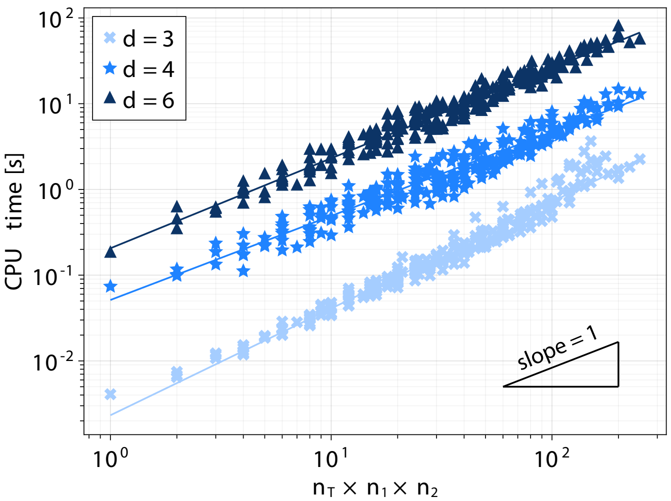
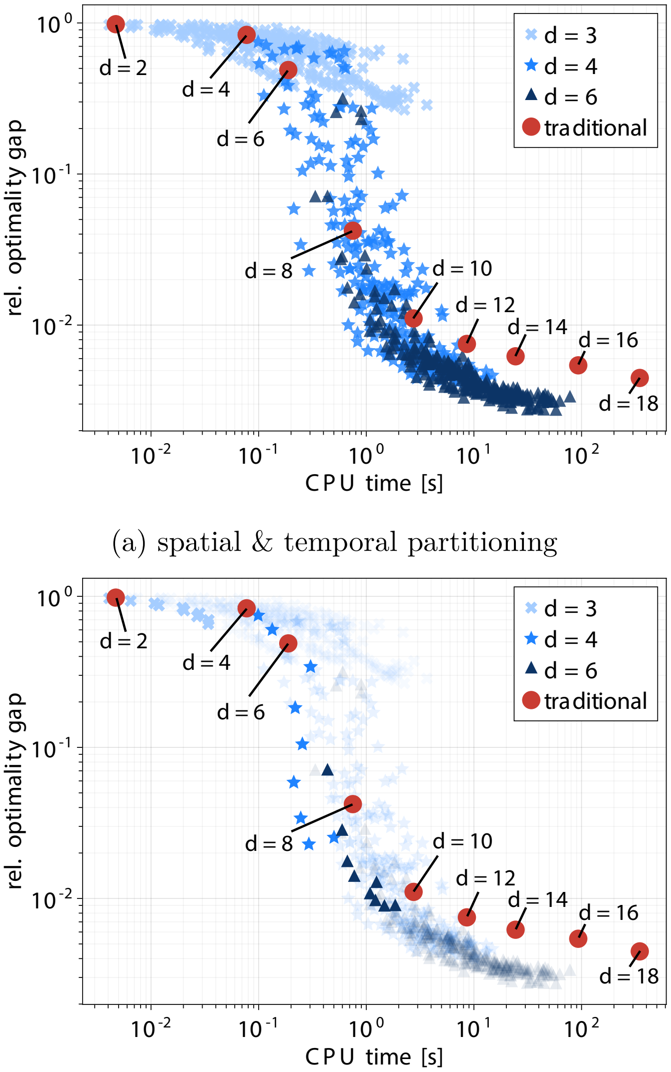


Figure 3-5: Computational cost of solving (pw-sub-HJB) with increasing number of partition elements.

Figure 3-6 shows the trade-off between bound quality and computational cost for different approximation orders and partitions. First, it is worth noting that the proposed partitioning strategy enables the computation of overall tighter bounds with an approximation order of only up to $d = 6$ when compared to the traditional formulation with an approximation order of up to $d = 18$. It is further worth emphasizing that beyond $d = 18$, numerical issues prohibited an accurate solution of the SDPs arising from the traditional formulation such that no tighter bounds could be obtained this way. Furthermore, upon choice of a suitable partition, the proposed local occupation measure framework enables a notable speed-up over the traditional occupation measure framework across the entire accuracy range. Lastly, the results indicate that a careful choice of partitioning is crucial to achieve good performance. Figure 3-6b suggests that for this example particularly good performance is achieved when only the time domain is partitioned; additionally partitioning the spatial domain becomes an effective means of bound tightening only after the time domain has been resolved sufficiently finely.



(a) spatial & temporal partitioning

(b) exclusively temporal partitions highlighted ($n_1 = n_2 = 1$)

Figure 3-6: Trade-off between computational cost and bound quality for different approximation orders d and spatio-temporal partitions (n_1, n_2, n_T). The red markers correspond to MSOS restrictions of the labeled approximation order for the traditional formulation (sub-HJB).

3.7 Extensions

Before we close, we briefly discuss two direct extensions to the described local occupation measure framework showcasing its versatility.

3.7.1 Discounted infinite horizon problems

Consider the following discounted infinite horizon stochastic optimal control problem with discount factor $\rho > 0$:

$$\begin{aligned}
\inf_{u_t} \quad & \mathbb{E}_{\nu_0}\left[\int_{[0,\infty)} e^{-\rho t}\ell(x_t, u_t)\,\mathrm{d}t\right] \\
\text{s.t.} \quad & x_t \text{ satisfies (3.1) on } [0,\infty) \text{ with } x_0 \sim \nu_0, \\
& (x_t, u_t) \in X \times U, \text{ on } [0,\infty), \\
& u_t \text{ is non-anticipative.}
\end{aligned}$$

The construction of a weak formulation of this problem akin (weak-OCP) can be done in full analogy to Section 3.2. To that end, note that the infinitesimal generator $\mathcal{A}$ maps functions of the form $\hat{w}(t,x) = e^{-\rho t}w(t,x)$ to functions of the same form, i.e.,

$$\mathcal{A}\hat{w}(t,x,u) = e^{-\rho t}(\mathcal{A}w(t,x,u) - \rho w(t,x,u)).$$

By analogous arguments as in Section 3.2, it therefore follows that any function $w \in C^{1,2}([0,\infty) \times X)$ that satisfies

$$\mathcal{A}w(t,x,u) - \rho w(t,x,u) + \ell(x,u) \geq 0, \quad \forall (t,x,u) \in [0,\infty) \times X \times U$$

generates a valid subsolution $\hat{w}(t,x) = e^{-\rho t}w(t,x)$ of the value function associated with the infinite horizon problem. Since the proposed partitioning approach does neither rely on boundedness of the state space nor control horizon in order to establish valid bounds, it follows that it readily extends to the infinite horizon setting.

3.7.2 Jump processes with discrete state space

Many application areas ranging from chemical physics to queuing theory call for models that describe stochastic transitions between discrete states. In those cases, jump processes are a common modeling choice [79, 80]. In the following, we show that the local occupation measure framework extends with only minor modifications to stochastic optimal control of a large class of such jump processes. Specifically, we consider controlled, continuous-time jump processes driven by m independent Poisson counters $n_i(t)$ with associated propensities $a_i(x_t, u_t)$:

$$\mathrm{d}x_t = \sum_{i=1}^{m} \left(h_i(x_t, u_t) - x_t\right) \, \mathrm{d}n_{i,t}. \tag{3.12}$$

We will again assume that the process can be fully characterized by polynomials, but additionally impose the assumption that the state space of the process is discrete.

Assumption 3.4. *The jumps $h_i : X \times U \to X$, propensities $a_i : X \times U \to \mathbb{R}_+$, stage cost $l : X \times U \to \mathbb{R}$ and terminal cost $\phi : X \times U \to \mathbb{R}$ are polynomial functions jointly in both arguments. The state space is a discrete, countable set and the set of admissible control inputs U is basic closed semialgebraic.*

The local occupation measure framework outlined previously for diffusion processes can be extended for computing lower bounds on the best attainable control performance for such jump processes:

$$\begin{aligned} \inf_{u_t} \quad & \mathbb{E}_{\nu_0}\left[\int_{[0,T]} \ell(x_t, u_t)\, \mathrm{d}t + \phi(x_T)\right] \qquad \text{(jump OCP)} \\ \text{s.t.} \quad & x_t \text{ satisfies (3.12) on } [0,T] \text{ with } x_0 \sim \nu_0, \\ & (x_t, u_t) \in X \times U, \text{ on } [0,T], \\ & u_t \text{ is not anticipative.} \end{aligned}$$

Given the extended infinitesimal generator $\mathcal{A} : \mathcal{C}^{1,0}([0,T] \times X) \to \mathcal{C}([0,T] \times X \times U)$ associated with the process (3.12),

$$\mathcal{A}w \mapsto \frac{\partial w}{\partial t}(t,x) + \sum_{i=1}^{m} a_i(x,u)\left(w(t, h_i(x,u)) - w(t,x)\right),$$

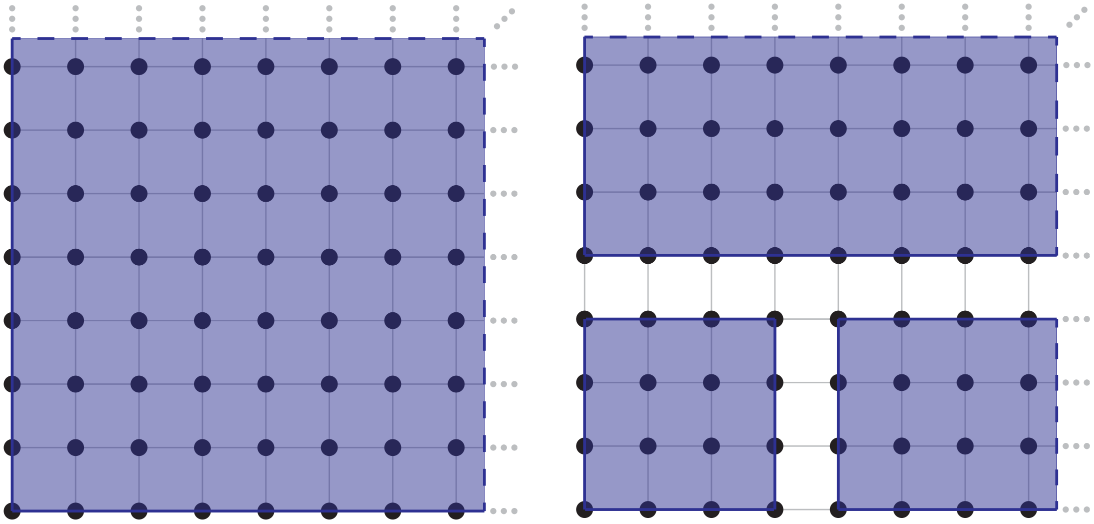
the weak form of (jump OCP) and its dual are analogous to (weak-OCP) and (sub-HJB), respectively. Further, given a partition of the problem's space-time domain as introduced in Section 3.3, the analog of Problem (pw-sub-HJB) seeking a piecewise smooth subsolution of the value function takes the form

$$
\begin{aligned}
\sup_{w_{i,k}:(i,k)\in P} \quad & \sum_{k=1}^{n_X} \int_{X_k} w_{1,k}(0,\cdot)\,\mathrm{d}\nu_0 && \text{(jump pw-subHJB)}\\
\text{s.t.} \quad & \mathcal{A}w_{i,k} + \ell \geq 0 \text{ on } [t_{i-1}, t_i] \times X_k \times U, \quad \forall (i,k) \in P,\\
& w_{i,k}(t_{i-1},\cdot) \geq w_{i-1,k}(t_{i-1},\cdot) \text{ on } X_k, \quad \forall (i,k) \in P^{\circ},\\
& w_{i,k} = w_{i,j} \text{ on } [t_{i-1}, t_i] \times N_{kj}, \quad \forall (i,j,k) \in \partial P,\\
& w_{n_T,k}(T,\cdot) \leq \phi \text{ on } X_k, \quad \forall k \in \{1,\dots,n_X\},\\
& w_{i,k} \in \mathcal{C}^{1,0}([0,T]\times X_k), \quad \forall (i,k) \in P,
\end{aligned}
$$

where N_{kj} denotes the "neighborhood" of X_k in X_j defined as all states in X_j which have a non-zero transition probability into X_k; formally,

$$N_{kj} = \{x \in X_j : \exists u \in U \text{ such that } h_i(x,u) \in X_k \text{ for some } i \text{ and } a_i(x,u) > 0\}.$$

Note that under Assumption 3.4, the extended infinitesimal generator $\mathcal{A}$ associated with a jump process again maps polynomials to polynomials laying the basis for the application of the MSOS hierarchy to construct tractable relaxations of (jump OCP). In contrast to the discussion in Section 3.2, however, the state space X of a jump process is closed basic semialgebraic if and only if it is finite. Thus, the MSOS hierarchy provides finite SDP relaxations of the weak form of (jump OCP) only in the case of a finite state space X. Moreover, even if X is finite but of large cardinality, these relaxations may not be practically tractable due to the large number or high degree of the polynomial inequalities needed to describe such a set. If X is infinite (or of sufficiently large cardinality), tractable MSOS relaxations can only be constructed at the price of introducing additional conservatism. From the dual perspective, this additional conservatism is introduced by imposing the non-negativity conditions in (sub-HJB) on a basic semialgebraic overapproximation of X; in particular polyhedral overapproximations are a common choice [12–14, 16, 81]. The


(a) global convex overapproximation

(b) union of convex overapproximations

Figure 3-7: Tightening of bounding problems for discrete jump problems on a 2-D lattice by partitioning of the state space into unions of convex overapproximations. Overapproximations are shown in blue, lattice nodes are shown with black circles.

framework of local occupation measures provides a way to reduce this conservatism. While the construction of tractable MSOS restrictions of (jump pw-subHJB) requires basic semialgebraic overapproximation of each infinite (or sufficiently large) partition element, the union of such overapproximations will generally be less conservative than a global semialgebraic overapproximation (see Figure 3-7).

3.7.3 Example: optimal gene regulation for protein expression

We demonstrate the efficacy of the local occupation measure framework for the control of jump processes with an example from cellular biology. Specifically, we consider the problem of optimal regulation of protein expression through actuation of the promoter kinetics in the biocircuit illustrated in Figure 3-8. The associated jump process has three states encoding the molecular counts of protein (x_1), active promoter (x_2), and inactive promoter (x_3). The process undergoes jump transitions in response to the

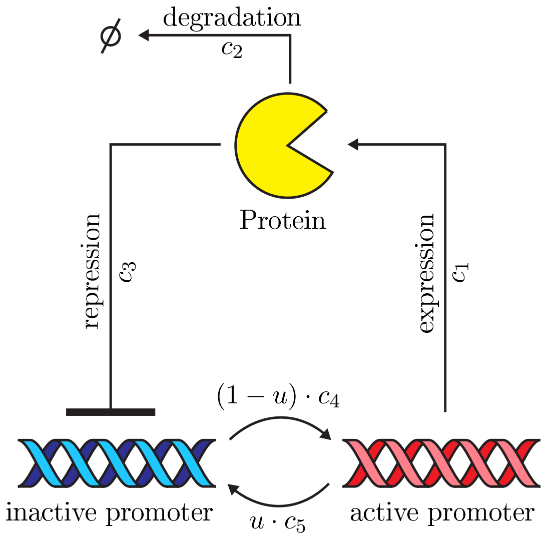


Figure 3-8: Biocircuit model for gene regulation.

following chemical reactions with associated rates:

$$h_1 : (x_1, x_2, x_3) \mapsto (x_1 + 1, x_2, x_3), \quad a_1(x, u) = 10x_2 \tag{expression}$$

$$h_2 : (x_1, x_2, x_3) \mapsto (x_1 - 1, x_2, x_3), \quad a_2(x, u) = 0.1x_1 \tag{degradation}$$

$$h_3 : (x_1, x_2, x_3) \mapsto (x_1, x_2 - 1, x_3 + 1), \quad a_3(x, u) = 0.1x_1x_2 \tag{repression}$$

$$h_4 : (x_1, x_2, x_3) \mapsto (x_1, x_2 + 1, x_3 - 1), \quad a_4(x, u) = 10(1 - u)x_3 \tag{activation}$$

$$h_5 : (x_1, x_2, x_3) \mapsto (x_1, x_2 - 1, x_3 + 1), \quad a_5(x, u) = 10ux_2 \tag{inactivation}$$

The expression of protein can be controlled indirectly via the activation and inactivation rates of the promoter. Admissible control actions u are constrained to lie within the interval $U = [0, 1]$. Moreover, we assume a deterministic initial condition $x_0 \sim \delta_{(0,1,0)}$ and exploit that due to the reaction invariant $(x_t)_2 + (x_t)_3 = (x_0)_2 + (x_0)_3$ the state space X is effectively two-dimensional, i.e., we eliminate $(x_t)_3 = 1 - (x_t)_2$. It can be easily verified that, after elimination of the reaction invariant, the state space of the jump process is given by

$$X = \{x \in \mathbb{Z}^2_+ : x_2 \in \{0, 1\}\}$$

such that Assumption 3.4 is satisfied.

The goal of the control problem is stabilization the protein level in the cell at a desired target value of 10 molecules. To that end, we seek to minimize the stage cost

$$\ell(x, u) = (x_1 - 10)^2 + 10(u - 0.5)^2$$

over the horizon $[0, 10]$.

In order to investigate the effect of different partitions of the problem domain on bound quality and computational cost, we discretize the time horizon uniformly into n_T intervals and partition the state space into $2n_X$ singletons

$$X_i = \begin{cases} \{(i-1, 0), & i \leq n_X \\ \{(i - n_X - 1, 1)\}, & i > n_X \end{cases} \quad \text{for } i = 1, \dots, 2n_X$$

and lump the remaining part of the state space in the last partition element

$$X_{2n_X+1} = \{x \in \mathbb{Z}_+^2 : x_1 \geq n_X, x_2 \in \{0, 1\}\}.$$

We explore the partitions corresponding to all combinations of $n_T \in \{2, 4, \dots, 18, 20\}$ and $n_X \in \{0, 8, \dots, 32, 40\}$.

Note that the partition elements $X_1, \dots, X_{2n_X}$ are already basic closed semialgebraic such that no overapproximation is required for the construction of valid MSOS restriction of the non-negativity constraints in (jump pw-subHJB). In contrast, the partition element X_{2n_X+1} is discrete and infinite, hence not basic closed semialgebraic. We therefore strengthen the formulation of the MSOS restriction of (jump pw-subHJB) by imposing the non-negativity conditions on the polyhedral convex hull of X_{2n_X+1}, thereby recovering tractability.

Figure 3-9 shows the trade-off between computational cost and bound quality achieved by different choices for the partition of the problem domain and approximation order. The bound quality is again measured by the relative optimality gap, estimated as described in Section 3.6.3. Analogous to the diffusion control example

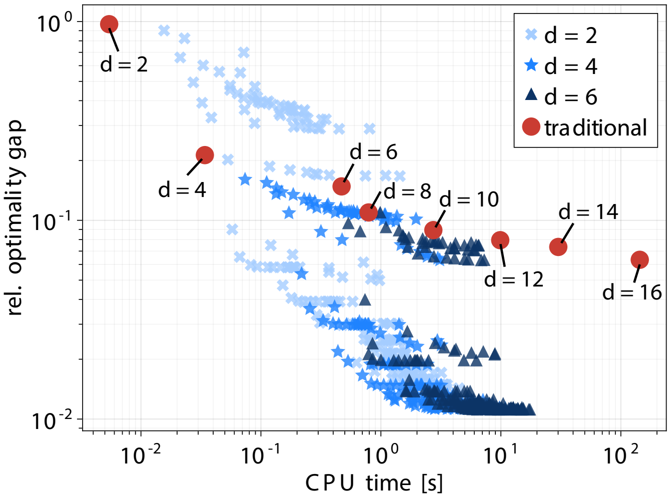


Figure 3-9: Trade-off between computational cost and bound quality for different approximation orders d and domain partitions. The red markers correspond to MSOS restrictions of the labeled approximation order for the traditional formulation (sub-HJB).

considered in Section 3.6, the results demonstrate that an adequate partitioning of the problem domain substantially reduces the cost of computing bounds of a given quality when compared to the traditional approach. Moreover, notably tighter bounds could be computed overall due to a less conservative overapproximation of the process' infinite state space in the formulation of the bounding problems.

3.8 Conclusion

We have proposed a simple partitioning strategy for improving the practicality of MSOS relaxations for stochastic optimal control problems with polynomial data. From the primal perspective, this strategy can be interpreted as constructing the MSOS relaxation for a linear program over finitely many occupation measures "localized" on elements of a partition of the control problem's space-time domain. From the dual perspective, the bounding problems seek a maximal piecewise-polynomial underestimator to the value function via sum-of-squares programming.

The key advantage of this framework over application of the MSOS hierarchy to the traditional occupation measure formulation for stochastic optimal control is that it offers a flexible and interpretable mechanism to tighten the obtained semidefinite bounding problems without degree augmentation – simple refinement of the problem domain partition. On the one hand, this enables tightening of the bounding problems at as benign as linearly increasing cost, contrasting the combinatorial scaling incurred by naive degree augmentation. On the other hand, it promotes practicality by providing a way to avoid high degree sum-of-squares constraints and their notorious implications for poor numerical conditioning. As demonstrated with two examples, these advantages can lead to notable improvements in practical utility of the occupation measure approach to stochastic optimal control.

In future work, we will investigate the use of distributed optimization techniques to further improve efficacy of the proposed framework by exploiting the weakly-coupled block structure of the bounding problems.

Chapter 4

Analysis of stochastic reaction systems via local occupation measures

The content of this chapter is an extension of the publication F. Holtorf and P. I. Barton, “Tighter bounds on transient moments of stochastic chemical systems,” *Journal of Optimization Theory and Applications*, vol. 200, no. 1, pp. 104–149, 2024. The essential ideas were conceived under supervision of Paul I. Barton.

4.1 Introduction

The model-based analysis of reacting systems at the microscopic scale is garnering increasing interest in the pursuit of understanding the functioning of living cells [82]. As continuum assumptions underpinning the classical deterministic description of chemical reaction kinetics begin to break down at the length scales and low molecular counts present in singular cells, however, it becomes essential to account for the randomness originating from the complex and chaotic motion of molecules in this regime. In fact, the intrinsically noisy nature of microscopic reaction systems is found to play a key role in facilitating biological processes as general as cellular decision-making, gene expression, and enzymatic regulation [83–87]. While a probabilistic description of such systems is therefore crucial to develop a complete understanding of their behavior, it also complicates associated analysis and inference tasks dramatically.

Stochastic reaction systems as found at the microscopic and in particular cellular level are canonically modeled as continuous-time Markov chains on discrete state spaces (jump processes) [82]. The associated Kolmogorov forward equation, or in this context more commonly known as the chemical master equation (CME), therefore governs the dynamics of the law of stochastic reaction systems. And although it reduces to a linear ordinary differential equation (ODE) due to the discrete nature of the state space, its solution remains out of computational reach for all but the simplest systems. The dimension of the CME coincides with the number of reachable states which routinely exceeds millions if it is finite at all. In practice, the computational analysis of stochastic reaction systems therefore has traditionally relied on Monte Carlo techniques [7, 88, 89], finite state projection [17], or moment closure approximations [8–11]. Here, we focus on a fourth more recently considered approach: moment bounding schemes [12–14, 81, 90]. Compared to the traditional approaches, moment bounding schemes combine the advantages of a low-dimensional moment-based description of the system's statistics with a mechanism for rigorous error control. On one hand, this alleviates the shortcomings of moment-closure approximations which are generally based on unverifiable assumptions and known to introduce severe errors as these assumptions break down [91–93]. On the other hand, it complements Monte Carlo approaches by providing additional side information for quantities with poor sample complexity such as rare event probabilities, quantities of large variance, or statistics of the system's long-term behavior.

In recent years, several researchers have proposed convex optimization-based bounding routines for the moments (and related statistics) of stochastic reaction networks; such bounding schemes have been proposed for statistics of stationary [12–14, 90], transient [16, 94–96], and exit time distributions [15] of such networks. While derived independently, these techniques can be unified within the occupation measure framework. As such, they stand to be generalized by the spatio-temporal partitioning approach and the associated notion of local occupation measures put forward in Chapter 3. We show in the following that viewing stochastic reaction systems through the lens of local occupation measures in fact not only unifies and generalizes moment bounding schemes for stochastic reaction systems but also that it bridges the gap to finite state projection [17] and related truncation-based analysis techniques [81]. In this context, we show further that the notion of local occupation measures enables a practical method to approximate the stationary distribution of stochastic reaction systems with potentially unbounded state space by invoking the maximum entropy principle – an approach that has previously been found to be greatly effective for the construction of moment closure approximations [11].

The remainder of this chapter is organized as follows. In Section 4.2, we briefly review the modeling framework of stochastic chemical kinetics and discuss essential assumptions. In Section 4.3, we adapt the local occupation measure framework to derive tractable bounding problems for transient and stationary statistics of stochastic reaction systems. In Sections 4.4 and 4.5 we establish formal connections between the proposed bounding problems and previously proposed moment bounding schemes and truncation-based analysis techniques for stochastic reaction systems, respectively. In Section 4.6, we discuss the combination of local occupation measures and maximum entropy regularization to approximate stationary distributions of stochastic reaction systems before we conclude in Section 4.7.

4.2 Stochastic chemical kinetics

We consider throughout a reaction system featuring n chemical species $S_1, \dots, S_n$ undergoing n_R different reactions. The system state x is assumed to be encoded entirely by the molecular counts of the individual species, i.e., $x = [x_1 \, \dots \, x_n]^\top \in \mathbb{Z}_+^n$. Upon the event of chemical reaction $r \in [n_R]$ occurring, the system state changes according to the stoichiometry

$$\gamma_{1,r}^{-} S_1 + \cdots + \gamma_{n,r}^{-} S_N \to \gamma_{1,r}^{+} S_1 + \cdots + \gamma_{n,r}^{+} S_n$$

leading to a discrete transition from state x to state $x + \gamma_r$, where $\gamma_r = [\gamma^+_{1,r} - \gamma^-_{1,r} \ \dots \ \gamma^+_{n,r} - \gamma^-_{n,r}]^\top \in \mathbb{Z}^n$. We will restrict ourselves to the framework of stochastic chemical kinetics for modeling such systems.

The notion of stochastic chemical kinetics treats the position and velocities of all molecules in the system as random variables; reactions are assumed to occur at collisions with a prescribed probability. Consequently, the evolution of the system state is a continuous-time Markov chain, i.e., a continuous-time jump process with discrete state space. Accordingly, we assume that the dynamics of the distribution of the system state is described by the Kolmogorov forward equation which is more commonly known as the chemical master equation (CME) in this context.

Assumption 4.1. *Let $p(t, x)$ be the probability to observe the system in state x at time t given the distribution ν_0 of the initial state of the system. Then, $p(t, x)$ satisfies*

$$\begin{cases} \dfrac{\partial p}{\partial t}(t,x) = \displaystyle\sum_{R_+(x)} a_r(x-\gamma_r)p(t,x-\gamma_r) - \sum_{R_-(x)} a_r(x)p(t,x), \ (t,x) \in [0,T] \times X, \\ p(0,x) = \nu_0(x), \ x \in X, \end{cases} \tag{CME}$$

where a_r denotes the propensity of reaction r, i.e., $a_r(x)\,\mathrm{d}t$ quantifies the probability that reaction r occurs in $[0, \mathrm{d}t)$ as $\mathrm{d}t \to 0$ assuming the system is in state x initially. $R_+(x)$ and $R_-(x)$ denote the sets of reactions with non-zero propensity into and out of state x, respectively; i.e., $R_+(x) = \{r \in [n_R] : a_r(x - \gamma_r) > 0\}$ and $R_-(x) = \{r \in [n_R] : a_r(x) > 0\}$.

Throughout, our constructions will rely on explicit knowledge of the system's state space defined as follows.

Definition 4.1. *The state space of a stochastic reaction system is the set of reachable states. A state $x \in \mathbb{Z}^n_+$ is called reachable on the time-horizon $[0, T]$ (with $T \in \mathbb{R}_+ \cup \{+\infty\}$) if there is a non-zero probability to observe the system in state x at some time point in $[0, T]$. Accordingly, the set of reachable states is formally given by*

$$X = \{x \in \mathbb{Z}^n_+ : \exists t \in [0,T] \text{ such that } p(t,x) > 0\}.$$

Note in particular that the set of reachable states is generally fully characterized by conservation of mass and the stoichiometry of the chemical reactions. It is hence easily determined by intersecting of the lattice of non-negative integers $\mathbb{Z}_+^n$ with reaction invariants implied by the stoichiometry.

Furthermore, we will restrict our considerations to the case of polynomial reaction propensities.

Assumption 4.2. *The reaction propensities* a_r*,* $r \in [n_R]$ *in* (CME) *are polynomials and non-negative on the reachable set* X*.*

To ensure the moments of the CME solution remain finite and well-defined at all times, we will further assume that the jump process describing the reaction system is regular, i.e., it does not explode in finite time. The following assumption ensures regularity [97].

Assumption 4.3. *The number of reaction events occurring in the system within finite time is finite with probability 1.*

We wish to emphasize that Assumptions 4.1 – 4.3 are rather weak; Assumptions 4.1 and 4.2 are in line with widely accepted microscopic models [79] while Assumption 4.3 should intuitively be satisfied for any practically relevant system for which the CME is a reasonable modeling approach. Furthermore, Assumption 4.3 is formally necessary for the CME to hold on an indefinite time horizon [97]. For a detailed, physically motivated derivation of the CME alongside discussion of the underlying assumptions and potential relaxations thereof, the interested reader is referred to Gillespie [79].

4.3 Local occupation measures for stochastic reaction systems

4.3.1 Transient problems

It is easily verified that the CME described in Assumption 4.1 is the Kolmogorov forward equation associated with the jump process

$$\mathrm{d}x_t = \sum_{r=1}^{n_R} \gamma_r \, \mathrm{d}n_t^r, \tag{4.1}$$

where the Poisson counter n_t^r fires at rate of the reaction propensity $a_r(x_t)$. As such, the algorithmic machinery discussed in Chapter 3, Section 3.7.2 in principle applies to bounding the statistics of stochastic reaction systems without modification. (Note that the absence of a control action is simply a special case of what is discussed in Chapter 3, Section 3.7.2.) However, the simple structure of the jump process (4.1) enables additional simplifications that were not employed in Chapter 3. We therefore begin by briefly revisiting the key concepts of the traditional occupation measure framework from the perspective of stochastic reaction systems and subsequently derive its localized generalization tailored to such systems.

The instantaneous and expected occupation measures associated with a stochastic reaction system are defined as

$$\nu(B) = \mathbb{E}_{\nu_0}\left[\mathbb{1}_B(x_T)\right] = \sum_{x \in B} p(T, x)$$

and

$$\xi(A \times B) = \mathbb{E}_{\nu_0}\left[\int_{A \cap [0,T]} \mathbb{1}_B(x_t)\, \mathrm{d}t\right] = \int_{A \cap [0,T]} \sum_{x \in B} p(t, x)\, \mathrm{d}t,$$

respectively, for any discrete subset of reachable states $B \subset X$ and Borel subset A of the time horizon $[0, T]$. (We used Tonelli's theorem to exchange limits for the expected occupation measure.)

The instantaneous and expected occupation measures are related by Dynkin's formula [74] according to

$$\langle \mathcal{A}w, \xi \rangle = \langle w, \nu \rangle - \langle w, \nu_0 \rangle$$

for any sufficiently smooth test function $w \in \mathcal{C}^{1,0}([0,T] \times X)$. Here, $\mathcal{A}$ refers to the extended infinitesimal generator of the stochastic reaction system (4.1), and $\langle \cdot, \xi \rangle$ and $\langle \cdot, \nu \rangle$ denote the duality brackets

$$\langle w, \xi \rangle = \mathbb{E}_{\nu_0}\left[\int_{[0,T]} w(t, x_t) \, \mathrm{d}t \right] \text{ and } \langle w, \xi \rangle = \mathbb{E}_{\nu_0}\left[w(T, x_T) \right]$$

as defined for continuous observables $w \in \mathcal{C}([0,T] \times X)$. The extended infinitesimal generator $\mathcal{A}$ acts according to

$$\mathcal{A}w(t,x) = \frac{\partial w}{\partial t}(t,x) + \sum_{r \in R_-(x)} a_r(x) \left(w(t, x + \gamma_r) - w(t,x) \right) \tag{4.2}$$

and describes via Dynkin's formula how expectations of the system state of sufficiently smooth observables evolve over time.

The following, infinite-dimensional linear program over non-negative measures therefore bounds the expectation of an observable $\phi \in \mathcal{C}(X)$ of the state of the stochastic reaction system (4.1) at time T from below:

$$\begin{aligned} J^*_{\mathrm{OM}} = \inf_{\nu, \xi} \quad & \langle \phi, \nu \rangle \\ \text{s.t.} \quad & \langle \mathcal{A}w, \xi \rangle = \langle w, \nu \rangle - \langle w, \nu_0 \rangle, \quad \forall w \in \mathcal{C}^{1,0}([0,T] \times X), \\ & \xi \in \mathcal{M}_+([0,T] \times X), \\ & \nu \in \mathcal{M}_+(X), \end{aligned} \tag{OM}$$

where $\mathcal{M}_+(Y)$ denotes to the cone of non-negative measures supported on Y. (Note that by construction the true occupation measures as defined above are feasible for (OM).) The dual of (OM) is an infinite-dimensional linear program over continuous functions and reads

$$
\begin{aligned}
J^*_{\text{subKBE}} = \sup_{w} \quad & \int_X w(0,\cdot)\,\mathrm{d}\nu_0 \\
\text{s.t.} \quad & \mathcal{A}w \geq 0, \ \text{on } [0,T]\times X, \\
& \phi - w(T,\cdot) \geq 0, \ \text{on } X, \\
& w \in \mathcal{C}^{1,0}([0,T]\times X).
\end{aligned} \qquad \text{(sub-KBE)}
$$

Analogous to the controlled case treated in Section 3, the dual (sub-KBE) admits an informative interpretation as seeking the maximal smooth subsolution of the Kolmogorov backward equation. This claim is formalized in the following proposition.

Proposition 4.1. *Let $\phi \in \mathcal{C}(X)$ and V be the solution to the Kolmogorov backward equation*

$$
\begin{cases} \mathcal{A}V(t,x) = 0, & (t,x) \in [0,T)\times X \\ V(T,x) = \phi(x), & x \in X. \end{cases} . \tag{4.3}
$$

Further let w be feasible for (sub-KBE)*. It follows that $w(t,x) \leq V(t,x)$ for all $(t,x) \in [0,T]\times X$.*

Proof. First recall that $V(s,z) = \mathbb{E}\left[\phi(x_T) | x_s = z\right]$ is the unique solution of Equation (4.3) [74, Theorem 8.1.1].[1] From the endpoint constraint in (sub-KBE) it further follows immediately that

$$
V(T,x) = \phi(x) \geq w(T,x)
$$

holds for all $x \in X$. Now let $s \in [0,T]$ and $z \in X$. Then, Dynkin's formula implies

[1] The statement and proof of Theorem 8.1.1 [74] is presented for diffusion processes but holds almost verbatim for any process that admits an infinitesimal generator and for which Dynkin's formula holds.

that

$$
\begin{aligned}
V(s,z) = \mathbb{E}\left[\phi(x_T)|x_s = z\right] &\geq \mathbb{E}\left[w(T,x_T)|x_s = z\right] \\
&= w(s,z) + \mathbb{E}\left[\int_{[s,T]} \mathcal{A}w(t,x_t)\,\mathrm{d}t \,\middle|\, x_s = z\right] \geq w(s,z),
\end{aligned}
$$

where the second inequality follows from the non-negativity of $\mathcal{A}w$ on $[0,T] \times X$. □

Corollary 4.1. *Strong duality holds between* (OM) *and* (sub-KBE)*. Moreover,* $J^*_{\mathrm{OM}} = J^*_{\mathrm{subKBE}} = \mathbb{E}_{\nu_0}\left[\phi(x_T)\right]$*.*

Proof. $V(s,z) = \mathbb{E}_{\nu_0}\left[\phi(x_T)|x_s = z\right]$ is the unique solution to Equation (4.3) [74, Theorem 8.1.1]. As a consequence $V \in \mathcal{C}^{1,0}([0,T] \times X)$ (note that Equation (4.3) implies that $s \mapsto V(s,z)$ is differentiable for all $z \in X$). It follows by Proposition 4.1 that $J^*_{\mathrm{subKBE}} = \int_X V(0,\cdot)\,\mathrm{d}\nu_0 = \mathbb{E}_{\nu_0}[\phi(x_T)]$. The result thus follows from weak duality in infinite dimensional linear programming [37] and by noting that $J^*_{\mathrm{OM}} \leq \mathbb{E}_{\nu_0}[\phi(x_T)]$ holds by construction:

$$
\mathbb{E}_{\nu_0}[\phi(x_T)] \geq J^*_{\mathrm{OM}} \geq J^*_{\mathrm{subKBE}} = \mathbb{E}_{\nu_0}[\phi(x_T)].
$$

□

Remark 4.1. *As the solution of the backward equation* (4.3) *is given by* $V(t,z) = \mathbb{E}\left[\phi(x_T)|x_t = z\right]$*, any feasible point of* (sub-KBE) *enables the computation of lower bounds for the conditional expectations*

$$
\int_X w(s,\cdot)\,\mathrm{d}\rho \leq \mathbb{E}\left[\phi(x_T)|x_s \sim \rho\right]
$$

for arbitrary times $s \in [0,T]$ *and distributions* ρ *supported on the reachable set* X*.*

Proposition 4.1 and Corollary 4.1 establish that feasible points of (OM) and (sub-KBE) furnish bounds on the statistics of stochastic reaction systems. Moreover, under Assumption 4.2 both problems admit tractable moment-sum-of-squares (MSOS) approximations to compute such bounds in practice. The only complicating factor in practice arises from the fact that the state space X of stochastic reaction systems is discrete and typically vast. To recover practically tractable MSOS approximations, the state space X must be overapproximated by a “simple” semialgebraic set; see Chapter 3, Section 3.7.2 for a detailed discussion of this issue and its practical implications. Here, we simply note that for stochastic reaction systems such semialgebraic overapproximations are readily and easily constructed. In particular, the state space of stochastic reaction systems is by construction contained in the non-negative orthant. Further refinements of this crude basic semialgebraic overapproximation can be obtained by considering reaction invariants and isolating particular integer states through low-degree polynomial inequalities [12–14, 16, 81].

Another natural approach to address complications arising from the discrete state space of stochastic reaction systems is the local occupation measure framework described in Chapter 3. In the following, we adapt and tailor this framework to stochastic reaction systems. To that end, let us consider a state space partition $X_1, \dots, X_{n_X}$ such that $X_i \cap X_j = \emptyset$ for $i \neq j$ and $\cup_{i=1}^{n_X} X_i = X$. Further consider a discretization of the time horizon $0 = t_0 < t_1 < \dots < t_{n_T} = T$. The sets $[t_{i-1}, t_i] \times X_k$ thus partition the entire space-time domain $[0, T] \times X$. When restricting the global occupation measures introduced in the beginning of this section to the subdomains $[t_{i-1}, t_i] \times X_k$ we obtain the localized occupation measures

$$\nu_{ik}(B) = \sum_{x \in B \cap X_k} p(t_i, x)$$

and

$$\xi_{ik}(A \times B) = \xi(A \cap [t_{i-1}, t_i] \times B \cap X_k) = \int_{[0,T] \cap [t_{i-1}, t_i] \cap A} \sum_{x \in B \cap X_k} p(t, x) \, \mathrm{d}t.$$

We further introduce “exchange” measures π_{ijk} which capture the probability mass concentrated in states of X_j from which the process can transition into X_k during the time interval $[t_{i-1}, t_i]$. The exchange measure π_{ijk} is thus supported on the “relative neighborhood” of X_k in X_j defined as follows.

Definition 4.2 (Relative neighborhood)**.** *The relative neighborhood N_{kj} of X_k in X_j consists of all states in X_j that transition with non-zero rate into X_k, i.e.,*

$$N_{kj} = \{x \in X_j : R_k(x) \neq \emptyset\} \text{ where } R_k(x) = \{r \in [n_R] : x + \gamma_r \in X_k \text{ and } a_r(x) > 0\}.$$

Formally, π_{ijk} is defined as

$$\pi_{ijk}(A \times B) = \int_{[t_{i-1},t_i]\cap A} \sum_{x \in B \cap N_{kj}} p(t,x)\,\mathrm{d}t$$

for Borel subsets $A \subset [t_{i-1}, t_i]$ and $B \subset N_{kj}$. By definition, the exchange measure π_{ijk} characterize the effect of transitions of the process from X_j to X_k in the time interval $[t_{i-1}, t_i]$. The effect of these transitions on observables is described by the flux operator

$$\mathcal{F}_k w(t,x) = \sum_{r \in R_k(x)} w(t, x + \gamma_r) a_r(x)$$

which encodes the rate of change of the observable w due to transitions into X_k in expectation.

Overall, the following conservation equation encodes how the expectations of a sufficiently smooth observable $w \in \mathcal{C}^{1,0}([0,T] \times X)$ of the process evolves over $[t_{i-1}, t_i] \times X_k$:

$$\langle w, \nu_{ik} \rangle - \langle w, \nu_{(i-1)k} \rangle = \langle \mathcal{A}w, \xi_{ik} \rangle + \sum_{j \neq k} \langle \mathcal{F}_k w, \pi_{ijk} \rangle - \langle \mathcal{F}_j w, \pi_{ikj} \rangle. \tag{4.4}$$

A detailed, constructive derivation of this relation from the governing CME is provided in Appendix B.1. This constructive derivation further justifies the following intuition: the left-hand side of Equation 4.4 quantifies the change in expectation of an observable w that vanishes outside of the subdomain $X_k \times [t_{i-1}, t_i]$ induced by

- dynamics of the process while residing in X_k (first term of the right-hand side),
- transitions of the process from X_j to X_k (second term of the right-hand side),

- and transitions of the process from X_k to X_j (third term of the right-hand side).

When finally taking the non-negativity and support of the localized occupation and exchange measures into account, the following generalization of the infinite-dimensional linear program (OM) is obtained.

$$\begin{aligned}
\inf_{\nu,\xi,\pi} \quad & \sum_{k=1}^{n_X} \langle \phi, \nu_{n_T k} \rangle && \text{(local-OM)}\\
\text{s.t.} \quad & \langle w, \nu_{ik} \rangle - \left\langle w, \nu_{(i-1)k} \right\rangle = \langle \mathcal{A} w, \xi_{ik} \rangle + \sum_{j \neq k} \langle \mathcal{F}_k w, \pi_{ijk} \rangle - \langle \mathcal{F}_j w, \pi_{ikj} \rangle \,, \\
& \qquad\qquad\qquad\qquad \forall w \in \mathcal{C}^{1,0}([t_{i-1}, t_i] \times \bar{X}_k),\ \forall (i,k) \in P, \\
& \nu_{ik} \in \mathcal{M}_+(X_k), \quad \forall (i,k) \in P, \\
& \xi_{ik} \in \mathcal{M}_+([t_{i-1}, t_i] \times X_k), \quad \forall (i,k) \in P, \\
& \pi_{ijk} \in \mathcal{M}_+([t_{i-1}, t_i] \times N_{kj}), \quad \forall (i,j,k) \in \partial P,
\end{aligned}$$

with the index sets

$$\begin{aligned}
P &:= \left\{ (i,k) : 1 \leq i \leq n_T, 1 \leq k \leq n_X \right\}, \\
P^\circ &:= \{ (i,k) : 2 \leq i \leq n_T, 1 \leq k \leq n_X \}, \\
\partial P &:= \{ (i,j,k) : 1 \leq i \leq n_T, 1 \leq k \neq j \leq n_X \}.
\end{aligned}$$

The extension of the test function domain by all states that transition into X_k, i.e.,

$$\bar{X}_k = \left(\cup_{j=1}^{n_X} N_{kj} \right) \cup X_k,$$

ensures that $\mathcal{A} w$ and $\mathcal{F}_j w$ are well-defined on the support of ξ_{ik} and π_{ikj}, respectively.

We remark that while (local-OM) and (OM) are equivalent, their MSOS relaxations are not. In general, MSOS relaxations of (local-OM) require semialgebraic overapproximation of the subdomains $X_1, \dots, X_{n_X}$ which can give rise to significantly tighter relaxations than obtained for (local-OM) for which a global semialgebraic overapproximation of X is required. This holds particular potential when the probability mass of the system is concentrated in few states which then can be resolved by the partition as singletons. Analogous advantages are obtained for sum-of-squares restrictions of the dual of (local-OM) which in full analogy to Chapter 3 admits interpretation as seeking the maximal piesewise smooth subsolution to the Kolmogorov backward equation (4.3).

4.3.2 Stationary problems

The occupation measure framework and its local generalization extend naturally to the analysis of stationary statistics of stochastic reaction systems. To that end, let p_∞ be a stationary distribution of the system and $\mu(A) = \sum_{x \in A} p_\infty(x)$ for any $A \subset X$ be the associated stationary (occupation) measure. Then, p_∞ is invariant under the CME, i.e.,

$$\sum_{R_+(x)} a_r(x - \gamma_r) p_\infty(x - \gamma_r) - \sum_{R_-(x)} a_r(x) p_\infty(x) = 0 \text{ for all } x \in X.$$

Consequently, Dynkin's formula implies that for any observable $w \in \mathcal{C}(X)$

$$\langle \mathcal{A} w, \mu \rangle = 0$$

must hold. To derive an analogous problem to (local-OM) for bounding stationary averages, we thus define for any discrete set $A \subset X$ the localized stationary occupation and exchange measures as

$$\mu_k(A) = \sum_{x \in A \cap X_k} p_\infty(x) \qquad \text{and} \qquad \pi_{jk}(A) = \sum_{x \in A \cap N_{kj}} p_\infty(x), \tag{4.5}$$

respectively. By analogy to Equation (4.4) the localized stationary occupation and exchange measures satisfy the conservation relation

$$\langle \mathcal{A} w, \mu_k \rangle + \sum_{j \neq k} \langle \mathcal{F}_k, \pi_{jk} \rangle - \langle \mathcal{F}_j w, \pi_{kj} \rangle = 0.$$

Taking further the non-negativity and support of the localized stationary occupation and exchange measures into account, the following infinite-dimensional linear program bounds the stationary averages of an observable $\phi \in \mathcal{C}(X)$ from below:

$$\begin{aligned}
\inf_{\mu,\pi} \quad & \sum_{k=1}^{n_X} \langle \phi, \mu_k \rangle && \text{(local-OM}_\infty\text{)} \\
\text{s.t.} \quad & \langle \mathcal{A}w, \mu_k \rangle + \sum_{j \neq k} \langle \mathcal{F}_k w, \pi_{jk} \rangle - \langle \mathcal{F}_j w, \pi_{kj} \rangle = 0, \quad \forall w \in \mathcal{C}(\bar{X}_k), \ \forall k \in P, \\
& \mu_k \in \mathcal{M}_+(X_k), \quad \forall k \in P, \\
& \pi_{jk} \in \mathcal{M}_+(N_{kj}), \quad \forall (j,k) \in \partial P,
\end{aligned}$$

with index sets

$$P := \{k \in \mathbb{Z}_+ : 1 \leq k \leq n_X\} \text{ and } \partial P := \{(j,k) \in \mathbb{Z}_+^2 : 1 \leq k \neq j \leq n_X\}.$$

We finally remark that, unlike in the transient case discussed previously, the stationary distribution of a stochastic reaction system need not be unique. The optimal value of (local-OM$_\infty$) consequently bounds the minimal average of ϕ as generated among all possible stationary distributions from below.

4.4 Connections to moment bounding schemes: necessary moment conditions revisited

In recent years, several algorithmic schemes for putting hard bounds on the moments of stochastic reaction systems have emerged as a counterpart to heuristic moment closure approximations with rigorous error control [12–16, 81, 94, 95]. These moment bounding schemes share a common basis that is unified and extended by the local occupation measure framework discussed in the previous section. In broad strokes, they identify moment bounds by searching the extreme points of a set of truncated moment sequences confined by so-called necessary moment conditions, i.e., conditions that the moments of the solution of the CME must satisfy. In particular moment conditions in the form of affine constraints (reflecting the system's dynamics) and linear matrix inequalities (LMIs) (reflecting the support of the underlying distribution) have been found to enable a computationally efficient search via convex optimization while providing often remarkably accurate bounds for low-order moments of interest [12–16, 81, 94, 95]. We show here that combining the MSOS hierarchy with the traditional occupation measure framework outlined in Section 4.3 gives rise to analogous moment conditions and bounding problems as derived by these contributions.

Although this observation comes with little surprise as all of the preceding moment bounding schemes leverage the MSOS hierarchy (with or without explicit mention) and some even use the concept of occupation measures directly [15], it establishes that the occupation measure framework provides a unifying perspective. Furthermore, this perspective will underline the advantages and promises of employing the localized occupation measure framework for the construction of more stringent bounding problems.

In the following, we focus on the moment bounding scheme for transient stochastic reaction systems by Dowdy and Barton [94], Sakurai and Hori [95], and Holtorf and Barton [16]. Adapting the discussion to moment bounding schemes for the stationary case [12–14] or the case of exit time distributions [15] is straightforward.

4.4.1 Affine constraints

In contrast to the discussion in Section 4.3, the moment bounding schemes of Dowdy and Barton [94], Sakurai and Hori [95] and Holtorf and Barton [16] do not directly rely on the characterization of the system dynamics through Dynkin's formula. Instead, they leverage the differential analog of Dynkin's formula. Specifically, by exchanging expectation and integration in Dynkin's formula via Tonelli's theorem, it is easily established that the dynamics of the expectation of an observable $w \in \mathcal{C}(X)$ satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{\nu_0}[w(x_t)] = \mathbb{E}_{\nu_0}\left[\mathcal{A}w(x_t)\right].$$

The previous contributions therefore derived necessary moment conditions by integrating

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(g(t)\mathbb{E}_{\nu_0}[w(x_t)]\right) = \frac{\mathrm{d}g}{dt}(t)\mathbb{E}_{\nu_0}[w(x_t)] + g(t)\mathbb{E}_{\nu_0}\left[\mathcal{A}w(x_t)\right].$$

for suitable test functions $g \in \mathcal{C}^1([0,T])$ and $w \in \mathcal{C}(X)$ [94, 95]. The resultant conditions read

$$g(T)\mathbb{E}_{\nu_0}[w(x_T)] - g(0)\mathbb{E}_{\nu_0}[w(x_0)] = \int_0^T \mathbb{E}_{\nu_0}\left[\frac{\mathrm{d}g}{\mathrm{d}t}(t)w(x_t) + g(t)\mathcal{A}w(x_t)\right]\mathrm{d}t. \tag{4.6}$$

By defining $\tilde{w}(t,x) = g(t)w(x)$, however, it is easily verified that the above relation is equivalent to the condition

$$\langle \tilde{w}, \nu\rangle - \langle \tilde{w}, \nu_0\rangle = \langle \mathcal{A}\tilde{w}, \xi\rangle.$$

To see this, simply note that expectation and integration in (4.6) may be exchanged by Tonelli's theorem and that

$$\mathcal{A}\tilde{w}(t,x) = \frac{\mathrm{d}g}{\mathrm{d}t}(t)w(x) + g(t)\mathcal{A}w(x).$$

holds by the definition of the infinitesimal generator (cf. Equation (4.2)).

The above relation makes clear that upon the choice of monomial test functions g and w, the so-derived moment conditions coincide with constraints generated by the standard MSOS hierarchy when applied to Problem (OM). While all previous contributions rely on monomials for the spatial test functions w, different choices have been explored for the temporal test functions g. Dowdy and Barton [14] argue that exponential test functions g are better suited due to their relation to eigenfunctions of the infinitesimal generator $\mathcal{A}$ associated with linear reaction networks.[2] Sakurai and Hori [95] contrast Dowdy and Barton's proposal with monomial test functions and

[2] Dowdy and Barton [14] do not make the connection to the infinitesimal generator explicit. Instead, they analyze the coefficient matrix representing the action of $\mathcal{A}$ on monomial test functions [98, Appendix C.3].

find empirically that neither choice yields a clear advantage. In the conceptual predecessor to the local occupation measure framework, Holtorf and Barton [16] discuss the choice of temporal test functions in a more abstract setting and argue based on smoothness and causality of moment trajectories in favor of piecewise temporal test functions. By employing piecewise monomial and exponential temporal test functions, they demonstrate notable advantages over the proposal by Dowdy and Barton [98] as well as the use of monomial test functions [95]. As a generalization of that, the local occupation measure framework generates strictly more stringent affine moment constraints via MSOS relaxations of Problem (local-OM) by effectively employing piecewise monomial test functions for both the temporal and spatial domain.

4.4.2 Linear matrix inequalities

The fact that the solution of the CME, $p(t,\cdot)$, is at every instant $t \in [0,T]$ a non-negative measure supported only on the reachable set $X \subset \mathbb{Z}_+^n$ implies that its truncated moment sequences satisfy certain LMIs. To see this, let f be a polynomial that is non-negative on X and b be the monomial basis of $n-$variate polynomials in x up to degree d arranged in a vector. The fact that the law $p(t,\cdot)$ of a stochastic reaction system is a non-negative measure directly implies that the moment matrix $\mathbb{E}_{\nu_0}\left[f(x_t)b(x_t)b(x_t)^\top\right]$ is positive semidefinite for any $t \in [0,T]$. By convexity of the cone of positive semidefinite matrices, it further follows that

$$\int_0^T g(t)\mathbb{E}_{\nu_0}\left[f(x_t)b(x_t)b(x_t)^\top\right] \mathrm{d}t \succeq 0$$

holds for any non-negative test function $g(t)$ on $[0,T]$. The moment bounding schemes of Dowdy and Barton [14] and Sakurai and Hori [95] accordingly rely necessary moment conditions of the form

$$\mathbb{E}_{\nu_0}\left[f(x_T)b(x_T)b(x_T)^\top\right] \succeq 0 \text{ and } \mathbb{E}_{\nu_0}\left[\int_0^T g(t)f(x_t)b(x_t)b(x_t)^\top \,\mathrm{d}t\right] \succeq 0$$

for suitable test functions g. While Sakurai and Hori [95] leverage polynomial test functions of the form $g(t) = t^\alpha (T-t)^\beta$ with $\alpha \in \mathbb{Z}_+$ and $\beta \in \{0,1\}$ for this purpose, Dowdy and Barton [14] use only the exponential test functions as discussed in the previous section. The latter choice, however, has the notable disadvantage of not capturing the support on the time interval $[0,T]$ explicitly (exponential test functions are non-negative on $\mathbb{R}$, not only $[0,T]$). Holtorf and Barton [16] instead employ a combination of both test functions restricted to successive subintervals of $[0,T]$.

The above conditions are in one-to-one correspondence with the positive definiteness of analogous moment matrices of the occupation measures:

$$\begin{aligned} &\langle f b b^\top, \nu \rangle \succeq 0, \\ &\langle g f b b^\top, \xi \rangle \succeq 0, \end{aligned}$$

where the duality bracket shall be understood as being evaluated component-wise. When g is chosen as a polynomial of the form $g(t) = t^\alpha (T-t)^\beta$ with $\alpha \in \mathbb{Z}_+$ and $\beta \in \{0,1\}$, these conditions are a subset of the standard MSOS outer approximation of the cones $\mathcal{M}_+(X)$ and $\mathcal{M}_+([0,T] \times X)$ through the positive semidefiniteness of moment and localizing matrices; see Chapter 2 or [42] for details. These conditions are therefore also implied by the stronger MSOS relaxations of (local-OM).

4.5 Connection to truncation-based approximation schemes

Viewing stochastic reaction systems through the lens of local occupation measures is not only related to moment bounding schemes but also similar in spirit to truncation-based analysis techniques. Notably, finite state projection (FSP) algorithms [17, 18] and Kuntz *et al.*'s [81] recently proposed bounding scheme for stationary statistics share common features with the local occupation measure approach: they come with mechanisms to bound approximation errors rigorously and rely explicitly on a partitioning of the reachable state space X. In the following, we discuss their connections,

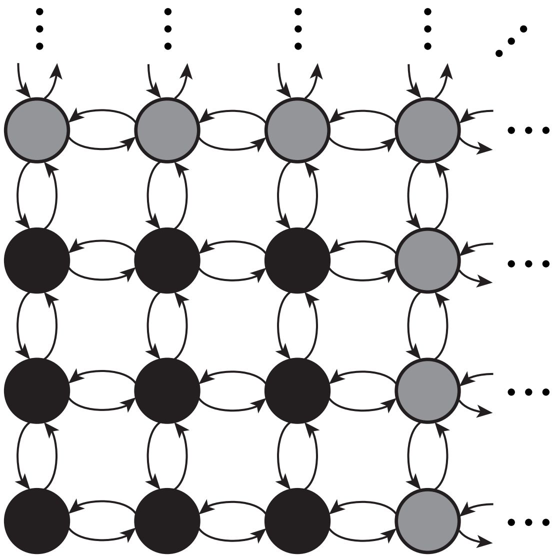
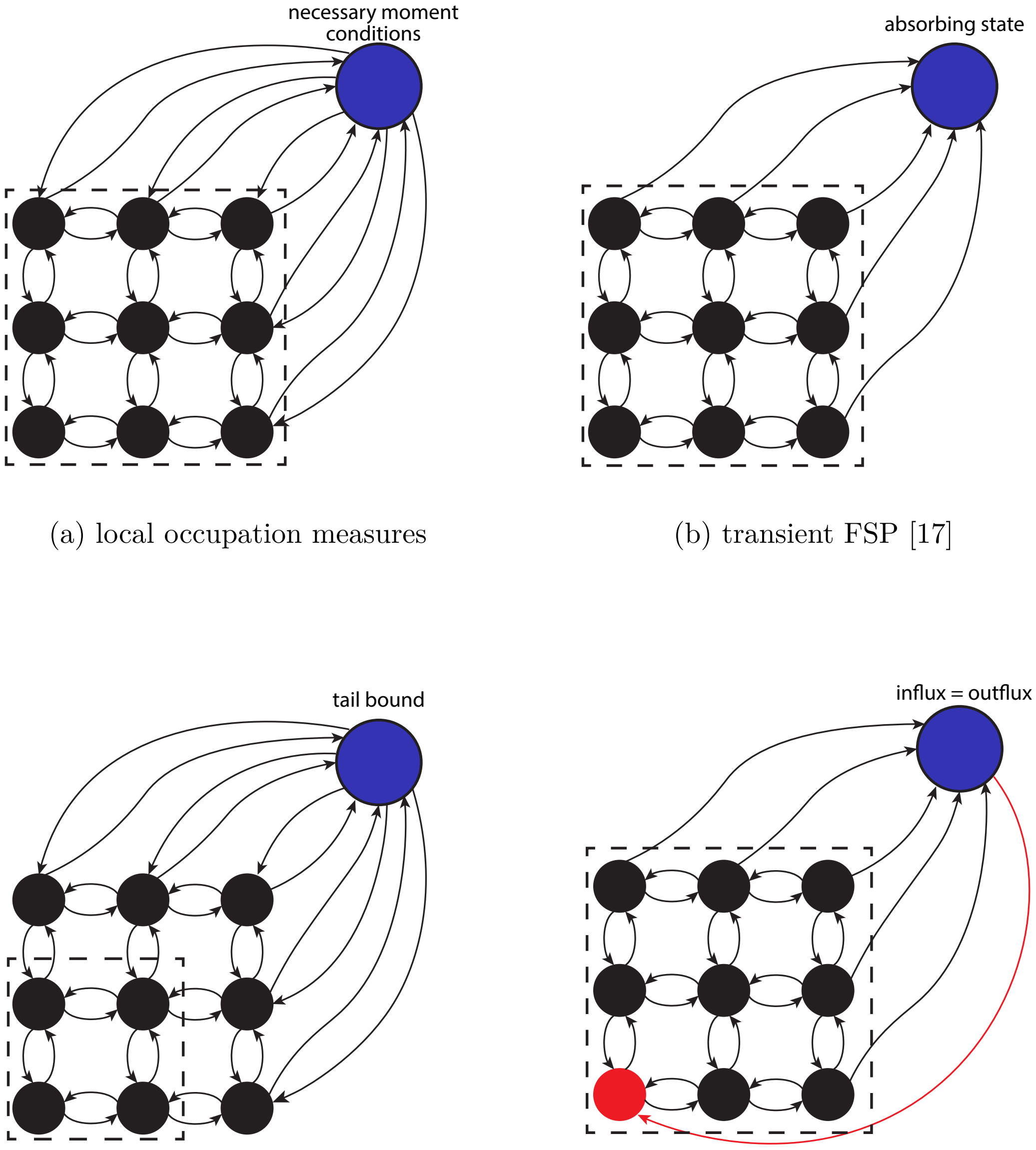


Figure 4-1: State space partition of an open system with two chemical species. Singleton elements of the partition are shown in black. The remainder is shown in gray. Arrows indicate possible transitions due to reaction.

similarities, and differences by contrasting the transient FSP algorithm [17], its stationary counterpart [99], and the bounding scheme of Kuntz *et al.* [81] with the local occupation measure framework.

The first notable difference between these approaches is that the FSP algorithms and Kuntz *et al.*'s bounding scheme are less flexible than the local occupation measure with regard to the state space partitions they accommodate. Specifically, they only consider state space truncations, i.e., partitions of the state space into $n_X - 1$ singletons $X_i = \{x_i \in X\}$, $i = [n_X - 1]$ and a remainder $X_{n_X} = X \setminus \cup_{i=1}^{n_X - 1} X_i$ lumping the remaining reachable states. Figure 4-1 illustrates such a partition for a two-dimensional state space. Beyond this difference, the methods may be distinguished by how probability fluxes in the truncated part of the state space X_{n_X} are approximated. Figure 4-2 provides an overview of these differences. In the following sections, we discuss the relations between all four methods in greater detail.



(a) local occupation measures

(b) transient FSP [17]

(c) truncation-based LP [100]

(d) stationary FSP [18]

Figure 4-2: Overview of truncation-based approximation & bounding schemes for stochastic reaction systems. The singleton elements $X_1, \dots, X_{n_X - 1}$ of the partition are depicted as black circles. The remainder partition element $X_{n_X} = X \setminus \cup_{i=1}^{n_X - 1} X_i$, covering all truncated states, is illustrated in blue and labeled according to the method of approximation. The "designated state" of the stationary FSP algorithm [18] is highlighted in red. Arrows indicate possible transitions due to reaction events. States with probability mass dynamics explicitly constrained by the CME balance equation (in weak/integral or strong/differential form) are enclosed within a dashed box.

4.5.1 The transient case

The transient FSP algorithm proposed by Munsky and Khammash [17] approximates the solution of the CME on a finite time horizon. As illustrated in Figure 4-2b, FSP relies on a finite approximation of the stochastic reaction system by treating the collection of truncated states X_{n_X} as an absorbing state. Formally, an approximation $\tilde{p}$ for the solution of the CME is obtained by solving the following $n_X - 1$ coupled ODEs

$$\begin{cases} \dfrac{\partial \tilde{p}}{\partial t}(t,x) = \displaystyle\sum_{\substack{r \in R_+(x) \\ x-\gamma_r \notin X_{n_X}}} a_r(x-\gamma_r)\tilde{p}(t,x-\gamma_r) - \sum_{r \in R_-(x)} a_r(x)\tilde{p}(t,x), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (t,x) \in [0,T] \times X \setminus X_{n_X}, \\ \tilde{p}(0,x) = \nu_0(x), \quad x \in X \setminus X_{n_X}. \end{cases} \tag{4.7}$$

This construction results in general in an irreversible outflux of probability mass from the collection of states $X_1, \dots, X_{n_X - 1}$ into the absorbing state X_{n_X}. The cumulative probability mass over these states therefore decays in general strictly, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t} \sum_{x \in X \setminus X_{n_X}} \tilde{p}(t,x) < 0.$$

A consequence of this irreversible outflux of probability mass into X_{n_X} is that $\tilde{p}(t,x)$ bounds the true solution $p(t,x)$ of the CME from below for all $x \in \cup_{i=1}^{n_X - 1} X_i$ and $t \in [0,T]$. On one hand, this implies hard bounds on the approximation error [17, Theorems 2.1 & 2.2]. On the other hand, however, it results in a steady deterioration of the approximation as time increases.

Similar to FSP, MSOS relaxations of (local-OM) may also be used to approximate the solution of the CME on the chosen partition with guaranteed error bounds. Specifically, one may compute a lower/upper bound for the probability mass $p(T, x_k)$ by solving MSOS relaxations of (local-OM) for the (piecewise constant) observable $\phi(x) = \pm \mathbb{1}_{X_k}(x)$. The optimal point of such relaxations in turn provides a proxy for the solution of the CME at the boundary points of the spatio-temporal partition elements due to the correspondence with the instantaneous occupation measures, i.e., $p(t_i, x_j) \approx \nu_{ij}(\{x_j\})$ for $i \in [n_T]$ and $j \in [n_X - 1]$. Note that this approximation is readily computable from solutions of MSOS relaxations of (local-OM) as it is fully characterized by the zeroth order moments of the measures ν_{ij}. In contrast to FSP, however, the local occupation measure approach does not treat X_{n_X} as an absorbing state (see Figure 4-2a). Instead, MSOS relaxations of (local-OM) impose necessary moment conditions on the probability mass flowing into X_{n_X}, reflecting its support and dynamics, as well as the interactions with the neighboring partition elements through the exchange measures π_{ijn_X} and π_{in_Xj}, $j \in [n_X - 1]$. A clear advantage of this construction over FSP is therefore that this approximation and any bounds computed from it do not necessarily deteriorate as the time horizon T increases; see for instance [16] for the several examples of such behavior.

Finally we wish to emphasize another connection between FSP and the linear program (local-OM). As shown in Appendix B.1, the equality constraints in (local-OM) are simply weak form equivalents of the possibly infinite ODE system (CME). Consequently, the equality constraints in (local-OM) are weak form analogs of the ODE governing the evolution of $\tilde{p}(t, x)$ for any state x without direct transitions into X_{n_X}.

4.5.2 The stationary case

Treatment of truncated states as absorbing state renders the transient FSP algorithm [17] fundamentally incapable of predicting the long-term statistics of stochastic reaction systems. In the common case of irreducible systems, i.e., systems in which every state can be reached from any another state by a finite sequence of reactions with non-zero propensity, the system will absorb in the truncated part of the state space with probability one in the limit of long times. In other words, the only stationary solution to the FSP ODE system (4.7) for irreducible systems is $\tilde{p}_\infty(x) = 0$, $x \in X \setminus X_{n_X}$. For reducible systems, the situation is typically not much better as any irreducible component of the state space with a non-zero propensity transition into X_{n_X} will suffer the same fate.

To address this limitation, Gupta *et al.* [18] present an FSP approximation scheme for stationary solutions of the CME. The state space partition underlying their approximation is illustrated in Figure 4-2d. In this partition, the truncated part of the state space is crucially no longer treated as an absorbing state but instead as an instant redirection into a "designated state" (marked in red in Figure 4-2d). As a consequence, probability mass no longer accumulates in the truncated states X_{n_X} but rather remains conserved in $X \setminus X_{n_X}$. Remarkably, for growing state space truncations approaching the full (possibly infinite) reachable set in the limit, convergence of this scheme to a unique stationary solution of the CME can be established under appropriate Foster-Lyapunov conditions [18, Theorem 3.1]. Unlike its transient counterpart, however, the stationary FSP algorithm does not provide hard error bounds. The bounding scheme of Kuntz *et al.* [81] alleviates this shortcoming of the stationary FSP algorithm at the cost of relying on an optimization-based routine. In the following, we discuss the close relationship of this bounding scheme with the local occupation measure framework. We show that under mild assumptions it can in fact be viewed as a linear programming relaxation of certain MSOS relaxations of (local-OM$_\infty$). A notable consequence of this observation is that the rich set of convergence guarantees of Kuntz *et al.*'s approach under growing state space truncations are inherited by the MSOS relaxations of (local-OM$_\infty$).

In the spirit of moment bounding schemes, Kuntz *et al.* [81] characterize a set of candidate stationary solutions for the CME by finitely many conditions imposed on a state space partition as illustrated in Figure 4-2c. Concretely, Kuntz *et al.* [81] use that any stationary solution of the CME is a member of the bounded polyhedron

$$\mathcal{P}_\infty = \left\{ \tilde{p}_\infty \;\middle|\; \begin{array}{l} \sum\limits_{r \in R_+(x)} a_r(x-\gamma_r)\tilde{p}_\infty(x-\gamma_r) - \sum\limits_{r \in R_-(x)} a_r(x)\tilde{p}_\infty(x) = 0, \quad \forall x \in X^\circ, \\ \tilde{p}_\infty(x) \geq 0, \quad \forall x \in X \setminus X_{n_X} \\ \sum\limits_{x \in X \setminus X_{n_X}} \tilde{p}_\infty(x) \geq 1 - \epsilon \end{array} \right\}$$

where

$$X^\circ = \{x \in X : x \notin X_{n_X} \text{ and } x \notin \cup_{i=1}^{n_X - 1} N_{n_X i}\}$$

denotes all states without transitions into the set of truncated states X_{n_X}. (X° is illustrated by the dashed box in Figure 4-2c. Recall that $N_{n_X i}$ denotes the states in X_i that transition with non-zero rate into X_{n_X}.)

The set of candidate stationary solutions $\mathcal{P}_\infty$ is framed by three kinds of constraints. The first one is the obvious stationarity condition

$$\sum_{r \in R_+(x)} a_r(x-\gamma_r)\tilde{p}_\infty(x-\gamma_r) - \sum_{r \in R_-(x)} a_r(x)\tilde{p}_\infty(x) = 0 \tag{4.8}$$

as implied by the CME. The second kind are non-negativity constraints on the probability masses assigned to states in the truncation, and the third kind is a tail bound imposed on the probability mass in the truncated part of the state space X_{n_X}:

$$\sum_{x \in X_{n_X}} \tilde{p}_\infty(x) \le \epsilon \iff \sum_{x \in X \setminus X_{n_X}} \tilde{p}_\infty(x) \ge 1-\epsilon.$$

As these conditions are in general not fully determinant of stationary solutions of the CME, Kuntz *et al.* [81] proposed to identify approximate stationary distributions as the extreme points of $\mathcal{P}_\infty$. Formally, such extreme points of $\mathcal{P}_\infty$ are characterized by the solution of finite linear programs of the form

$$\min_{\tilde{p}_\infty \in \mathcal{P}_\infty} \sum_{x \in X \setminus X_{n_X}} \tilde{p}_\infty(x)\phi(x). \tag{4.9}$$

The above problem also bounds the expectation of the observable $\phi \in \mathcal{C}(\cup_{i=1}^{n_X-1} X_i)$ from below. As such, it provides a natural way for rigorous error quantification. Specifically, for the observable $\phi(x) = \pm\mathbb{1}_{X_i}(x)$, the optimal value of (4.9) furnishes hard upper and lower bounds for the stationary probability mass of the state x_i.

The crux of Kuntz *et al.*'s approach lies in deriving an informative, yet guaranteed tail bound. To that end, they consider tail bounds from concentration inequalities and generalized moment bounds for "norm-like" functions.

Definition 4.3 (Norm-like function)**.** *A norm-like function $m : X \to \mathbb{R}$ satisfies*

(i) *m is non-negative on X*

(ii) *m has compact sublevel sets*

(iii) *the r-superlevel sets of m, i.e., $X_{\geq r} = \{x \in X : m(x) \geq r\}$, are nested: $r_1 \leq r_2 \implies X_{\geq r_2} \subset X_{\geq r_1}$, and the inclusion is strict for sufficiently large $r_2 - r_1$.*

A valid moment bound $c > 0$ for such a norm-like function, i.e.,

$$\sum_{x \in X} p_\infty(x) m(x) \leq c,$$

implies a tail bound for the r-superlevel sets of m via the concentration inequality

$$\sum_{x \in X_{\geq r}} p_\infty(x) \leq \sum_{x \in X_{\geq r}} \frac{m(x)}{r} p_\infty(x) \leq \frac{1}{r} \sum_{x \in X} m(x) p_\infty(x) \leq \frac{c}{r}. \tag{4.10}$$

A valid tail bound for the probability mass in X_{n_X} is thus obtained by choosing r small enough such that $X_{n_X} \subset X_{\geq r}$. A particularly appealing property of this construction is that the tail bound for X_{n_X} may be tightened arbitrarily by increasing the threshold r and enlarging the state space truncation correspondingly to ensure that $X_{n_X} \subset X_{\geq r}$. Based on this feature, Kuntz *et al.* [81] establish several convergence results in the limit of suitably growing state space truncations.

While this construction defers the problem of deriving a tail bound to that of deriving a moment bound for norm-like function, the latter task is readily addressed for polynomial norm-like functions, such as $m(x) = \sum_{i=1}^n x_i^\alpha$ with even α, by stationary moment bounding schemes [12–14]. This connection establishes a direct relation between Problem (4.9) and MSOS relaxations of (local-OM$_\infty$). Specifically, when the moment bound c is derived from (or implied by) a MSOS relaxation of (local-OM$_\infty$), then Problem (4.9) reduces to a relaxation of the MSOS relaxations of (local-OM$_\infty$). The following proposition formalizes this claim.

Proposition 4.2. *Let $X_1, \dots, X_{n_X}$ be a partition of the reachable set X such that $X_1 = \{x_1\}, \dots, X_{n_X-1} = \{x_{n_X-1}\}$ are singletons and $X_{n_X} = X \setminus \cup_{i=1}^{n_X-1} X_i$. Further, let m be a polynomial norm-like function such that $m(x) \geq r$ for all $x \in X_{n_X}$. Finally, let $\langle 1, \mu_k \rangle$, $k \in [n_X]$ and $\langle 1, \pi_{kj} \rangle$, $k \neq j \in [n_X]$ be the zeroth order moments of a* *feasible point of a MSOS relaxation of* (local-OM$_\infty$) *which implies the moment bound* $\sum_{k=1}^{n_X}\langle m, \mu_k\rangle \leq c$ *and explicitly incorporates positive semidefiniteness of the localizing matrix generated by the constraint* $m(x) - r \geq 0$ *on* X_{n_X} *(see Definition 2.21). Then,* $\tilde{p}(x_k) = \langle 1, \mu_k\rangle$, $k \in [n_X]$ *is feasible for* (4.9) *with tail bound* $\epsilon = c/r$.

Proof. First note that MSOS relaxations constrain all zeroth order moments of non-negative measures to be non-negative. Thus, $\langle 1, \mu_k\rangle \geq 0$ must hold for all $k \in [n_X]$. Further note that $\langle 1, \mu_k\rangle$ and $\langle 1, \sum_{j\neq k} \pi_{kj}\rangle$ coincide on the singleton partition elements (cf. Equation 4.5). Thus, for the constant test function $w \equiv 1$, the equality constraints in (local-OM$_\infty$) reduce to MSOS constraints for the partition elements centered on the states $x \in X^\circ$ and involve only the zeroth moments $\langle 1, \mu_1\rangle, \dots, \langle 1, \mu_{n_X-1}\rangle$. These constraints are further easily confirmed to be equivalent to the stationarity conditions (4.8) for the respective states (see the derivation in Appendix B.1). Similarly, summing the equality constraints of (local-OM$_\infty$) for the constant test function $w \equiv 1$ over all partition elements, implies that any feasible point of a MSOS relaxation of (local-OM$_\infty$) must satisfy

$$\sum_{k=1}^{n_X-1} \langle 1, \mu_k\rangle = 1 - \langle 1, \mu_{n_X}\rangle \tag{4.11}$$

To finally establish that also the tail bound

$$\sum_{k=1}^{n_X-1} \langle 1, \mu_k\rangle \leq 1 - \frac{c}{r}$$

must hold for any feasible point of the described MSOS relaxation, we note that positive semidefiniteness of the localizing matrix generated by the constraint $m(x) - r \geq 0$ implies that

$$\langle m - r, \mu_{n_X}\rangle = \langle m, \mu_{n_X}\rangle - r\langle 1, \mu_{n_X}\rangle \geq 0 \iff \langle 1, \mu_{n_X}\rangle \geq \frac{\langle m, \mu_{n_X}\rangle}{r}. \tag{4.12}$$

The tail bound thus follows by combining the asserted moment bound $\sum_{k=1}^{n_X}\langle 1, \mu_k\rangle \leq c$ with (4.11) and (4.12). □

Proposition 4.2 notably formalizes conditions under which all convergence guarantees of Kuntz *et al.*'s [100] bounding scheme carry over to MSOS relaxations generated via the local occupation measure framework.

4.6 Approximating stationary distributions: local occupation measures & the maximum entropy principle

The local occupation measure framework naturally extends to bounding the expectations of piecewise polynomial observables ϕ through MSOS relaxations of (local-OM) or (local-OM$_\infty$). To that end, one may simply choose a spatial partition of the state space (and semialgebraic overapproximation thereof) that coincides with the domain of the polynomial pieces defining ϕ. The standard MSOS hierarchy then generates without modification semidefinite relaxations of (local-OM) and (local-OM$_\infty$) which in turn furnish valid bounds. This observation is of particular practical interest as it enables bounding the probability mass of a given set of states. For instance, tractable bounding problems for the stationary probability mass of the subdomain X_i are readily construction from the infinite-dimensional linear program

$$\begin{aligned}
\inf_{\mu,\pi} \quad & \langle 1, \mu_i \rangle \\
\text{s.t.} \quad & \langle \mathcal{A}w, \mu_k \rangle + \sum_{j=1}^{n_X} \langle \mathcal{F}_k w, \pi_{jk} \rangle - \langle \mathcal{F}_j w, \pi_{kj} \rangle = 0, \quad \forall w \in \mathcal{C}(\bar{X}_k), \ \forall k \in P, \\
& \mu_k \in \mathcal{M}_+(X_k), \quad \forall k \in P, \\
& \pi_{jk} \in \mathcal{M}_+(N_{kj}), \quad \forall (j,k) \in \partial P.
\end{aligned}$$

which is a special case of (local-OM$_\infty$) for the piecewise constant observable $\phi(x) = \mathbb{1}_{X_i}(x)$. While a similar bounding problem has been considered in prior work by Dowdy and Barton [14], their proposal comes with additional limitations. On the one hand, Dowdy and Barton [14] only consider a bipartition of the state space into two subsets $\hat{X}_1 = X_i$ and $\hat{X}_2 = X \setminus X_i$. On the other hand, they do not account for exchange measures; instead, they impose the less restrictive global conservation constraint

$$\langle \mathcal{A}w, \hat{\mu}_1 \rangle + \langle \mathcal{A}w, \hat{\mu}_2 \rangle = 0, \quad \forall w \in \mathcal{C}(X),$$

on the stationary local occupation measures $\hat{\mu}_1$ and $\hat{\mu}_2$ associated with $\hat{X}_1$ and $\hat{X}_2$, respectively. Overall, Dowdy and Barton's [14] bounding problems are therefore weaker and less flexible when $X \setminus X_i$ does not admit tractable semialgebraic overapproximation.

Another advantage of the flexibility of (local-OM$_\infty$) is that it enables one-shot approximation of stationary measures through convex optimization and the maximum entropy principle. To that end, let us first consider systems with finite state space X. By drawing inspiration from the maximum entropy principle, one may approximate the stationary measure associated with a stochastic reaction system on a such finite state space as the optimal point of the following convex optimization problem.

$$\begin{aligned} S^* = \sup_{\mu,\pi} \quad & -\sum_{k=1}^{n_X} \langle 1, \mu_k \rangle \log \frac{\langle 1, \mu_k \rangle}{|X_k|} && \text{(uniform-S}_\infty\text{)} \\ \text{s.t.} \quad & \langle \mathcal{A}w, \mu_k \rangle + \sum_{j=1}^{n_X} \langle \mathcal{F}_k w, \pi_{jk} \rangle - \langle \mathcal{F}_j w, \pi_{kj} \rangle = 0, \quad \forall w \in \mathcal{C}(\bar{X}_k),\ \forall k \in P, \\ & \mu_k \in \mathcal{M}_+(X_k), \quad \forall k \in P, \\ & \pi_{jk} \in \mathcal{M}_+(N_{kj}), \quad \forall (j,k) \in \partial P, \end{aligned}$$

where $|X_k|$ denotes the cardinality of the partition element X_k. The connection to the maximum entropy principle is made explicit by the following proposition establishing that (uniform-S$_\infty$) bounds the entropy of the all stationary solutions of the CME from above.

Proposition 4.3. *Let X be finite and p_∞ be a stationary solution to the CME with entropy $S\left[p_\infty\right] = -\sum_{x\in X} p_\infty(x)\log p_\infty(x)$. Then, $S^* \geq S[p_\infty]$.*

Proof. Consider the local stationary occupation and exchange measures (μ, π) as defined in Equation (4.5). By construction (μ, π) is feasible for (uniform-S$_\infty$) and satisfies $\langle 1, \mu_k\rangle = \sum_{x\in X_k} p_\infty(x)$ for all $k \in P$. Further recall that the function $f(x) = -\sum_{i=1}^{N} x \log x$ attains its global maximum over the (scaled) $N-$simplex $\{x \in \mathbb{R}_+^N : \sum_{i=1}^{N} x_i = c\}$ at the uniform distribution $x = \frac{c}{N}[1, \dots, 1]^\top$. It therefore follows that the entropy contribution of partition element X_k is bounded above by

$$S_k = -\sum_{x\in X_k} p_\infty(x)\log p_\infty(x) \leq -\sum_{x\in X_k} \frac{\langle 1, \mu_k\rangle}{|X_k|}\log\frac{\langle 1, \mu_k\rangle}{|X_k|}$$

for all $k \in P$. Thus,

$$S\left[p_\infty\right] = \sum_{k=1}^{n_X} S_k \leq -\sum_{x\in X_k} \frac{\langle 1, \mu_k\rangle}{|X_k|}\log\frac{\langle 1, \mu_k\rangle}{|X_k|} \leq S^*,$$

where the second inequality follows from feasibility of (μ, π) for (uniform-S$_\infty$). □

Remark 4.2. *When X is finite,* (uniform-S$_\infty$) *is a finite, convex optimization problem as the occupation measures are fully characterized by the probability mass assigned to each state $x \in X$. Moreover, we may then obtain an exact entropy bound by choosing $n_X = |X|$ partition elements to coincide with the singleton states of X. The resultant problem, however, will become computationally intractable when the state space X is vast, as is typically the case for stochastic reaction systems. In that scenario, a coarser partitioning of X will still give rise to tractable MSOS relaxations of* (uniform-S$_\infty$)*. We show in Appendix B.2 that these relaxations preserve the upper bounding property of Problem* (uniform-S$_\infty$) *and admit reformulation as finite conic programs.*

Proposition 4.3 shows that for systems with finite state space X, (uniform-S$_\infty$) seeks a stationary measure μ that is consistent with the CME and whose "piecewise uniform" approximation

$$\tilde{\mu}(A) = \sum_{k=1}^{n_X} \sum_{x \in A \cap X} \frac{\langle 1, \mu_k \rangle}{|X_k|} \tag{4.13}$$

attains maximum entropy. This interpretation and the bounding property established in Proposition 4.3 is naturally inherited by MSOS relaxations of (uniform-S$_\infty$), giving rise to a computational method for computing maximum entropy approximations to stationary measures for stochastic reaction systems via convex optimization.

The proof of Proposition 4.3, however, also suggests that such a "piecewise uniform" approximation of the stationary measure and the associated entropy bound will deteriorate the coarser the partition of X is chosen. In particular, the uniform entropy bound is expected to become increasingly loose the coarser the partition is. While this is intuitive, it calls the utility of this an approximation scheme into question. After all, a key advantage of the local occupation measure framework is that we may partition vast state spaces rather coarsely and still obtain informative approximations (or bounds) for key statistics. In the context of approximating stationary measures this is of particular interest when the true stationary measure concentrates in a small subset of the state space. This subset may then be partitioned very finely while the remainder is covered by only a few partition elements. The entropy contribution from the uniform bound, however, is by construction disproportionately loose for such regions of large cardinality.

As pathologies arise from conservatively estimating the entropy contribution of regions of large cardinality with that of a uniform distribution, the following more flexible regularization strategy is natural: We may approximate the entropy contribution of a subdomain X_k simply by asserting the associated stationary occupation measure μ_k to resemble the shape of a given measure q. Concretely, we may approximate the associated entropy contribution $S[\mu_k] = \sum_{x \in X_k} \mu_k(x) \log \mu_k(x)$ as the entropy of q with its mass rescaled to $\langle 1, \mu_k \rangle$:

$$S[\mu_k] \approx -\sum_{x \in X} \langle 1, \mu_k \rangle q(x) \log \left(\langle 1, \mu_k \rangle q(x) \right) = -\langle 1, \mu_k \rangle \log \langle 1, \mu_k \rangle + \langle 1, \mu_k \rangle S[q].$$

When choosing q as the uniform measure on X_k, we recover the same entropy contributions as considered in (uniform-S_∞). By choosing q differently, however, we can not only leverage prior knowledge but also accommodate systems with infinite state spaces where the uniform entropy bound is uninformative. For instance, if we suspect X_k to only contain the exponential tails of the stationary distribution, we may choose q to be an exponential measure. The following regularized variant of Problem (regularized-S_∞) is the result:

$$\begin{aligned} \sup_{\mu,\pi} \quad & \sum_{k=1}^{n_X} \langle 1, \mu_k \rangle S[q_k] - \langle 1, \mu_k \rangle \log \langle 1, \mu_k \rangle \\ \text{s.t.} \quad & \langle \mathcal{A}w, \mu_k \rangle + \sum_{j=1}^{n_X} \langle \mathcal{F}_k w, \pi_{jk} \rangle - \langle \mathcal{F}_j w, \pi_{kj} \rangle = 0, \quad \forall w \in \mathcal{C}(\bar{X}_k), \ \forall k \in P, \\ & \mu_k \in \mathcal{M}_+(X_k), \quad \forall k \in P, \\ & \pi_{jk} \in \mathcal{M}_+(N_{kj}), \quad \forall (j,k) \in \partial P, \end{aligned} \qquad \text{(regularized-}\mathrm{S}_\infty\text{)}$$

where q_k denotes the probability measure which is assumed to resemble the measure μ_k up to rescaling of its mass.

Remark 4.3. *Problem* (regularized-S_∞) *is convex and admits reformulation as infinite dimensional linear program over the intersection of convex cones of non-negative Borel measures and exponential cones. Its MSOS relaxations are convex conic representable through positive semidefinite and exponential cone constraints. As such, the MSOS relaxations can be solved with off-the-shelf conic solvers such as Mosek [33]. A detailed derivation of the conic reformulation is provided in Appendix B.2.*

It is lastly worth emphasizing that the formulation (regularized-S_∞) does not require detailed specification of q_k but only its entropy $S[q_k]$. The latter is typically substantially less demanding and high-level intuition often suffices to specify it adequately. For instance, if μ_k is assumed to capture the tails of the stationary measure decaying at rate λ, then we shall choose $S[q_k] = 1 - \log \lambda$ which coincides with the entropy of an exponential distribution with parameter λ. Similarly, if we assume μ_k to resemble a Gaussian with variance σ^2, we do not need to explicitly specify its mean and simply choose $S[q_k] = \frac{1}{2}(1 + \log 2\pi\sigma^2)$.

4.6.1 Example: Schlögl's system

We illustrate the utility and advantages of Problem (regularized-S$_\infty$) for approximating stationary distributions of stochastic reaction systems with Schlögl's system [101]. Schlögl's system is a nonlinear birth-death process

$$2A \underset{k_2}{\overset{k_1}{\rightleftharpoons}} 3A, \qquad \emptyset \underset{k_4}{\overset{k_3}{\rightleftharpoons}} A$$

with reaction propensities

$$\begin{aligned}
a_1(x) &= k_1 x(x-1),\\
a_2(x) &= k_2 x(x-1)(x-2),\\
a_3(x) &= k_3,\\
a_4(x) &= k_4 x.
\end{aligned}$$

The stationary distribution of Schlögl's system can be determined analytically such that it provides a suitable test case. A recursive formula for its stationary distribution is given by

$$p_\infty(x) = \frac{a_1(x-1) + a_3(x-1)}{a_2(x) + a_4(x)} p_\infty(x-1),$$
$$\text{with } p_\infty(0) = {}_2F_2\left(-\frac{c_1+1}{2}, \frac{c_1-1}{2}; -\frac{c_2+1}{2}, \frac{c_2-1}{2}; \frac{k_1}{k_2}\right),$$

where ${}_2F_2$ denotes the generalized hypergeometric function with $c_1 = \sqrt{1 - 4k_3/k_1}$ and $c_2 = \sqrt{1 - 4k_4/k_2}$ [81]. Depending on the kinetic parameters $k_1, \ldots, k_4$, p_∞ is either uni- or bimodal. Here, we consider a bimodal case obtained for $k_1 = 0.15$, $k_2 = 1.5 \times 10^{-3}$, $k_3 = 20$, and $k_4 = 3.5$.

In order to emphasize the value of maximum entropy regularization when approximating stationary measures, we contrast approximations obtained from MSOS relaxations of two distinct local occupation measure problems: (regularized-S_∞) and (local-OM_∞) for the observable $\phi(x) = -x$. Both problems seek to approximate the stationary distribution from the same set of candidate distributions framed by necessary MSOS conditions. However, while the former does so by maximizing an entropy proxy, the latter returns the candidate distribution with maximum mean. We further compare both local occupation measure approaches against the independet baseline provided by the stationary FSP algorithm [18].

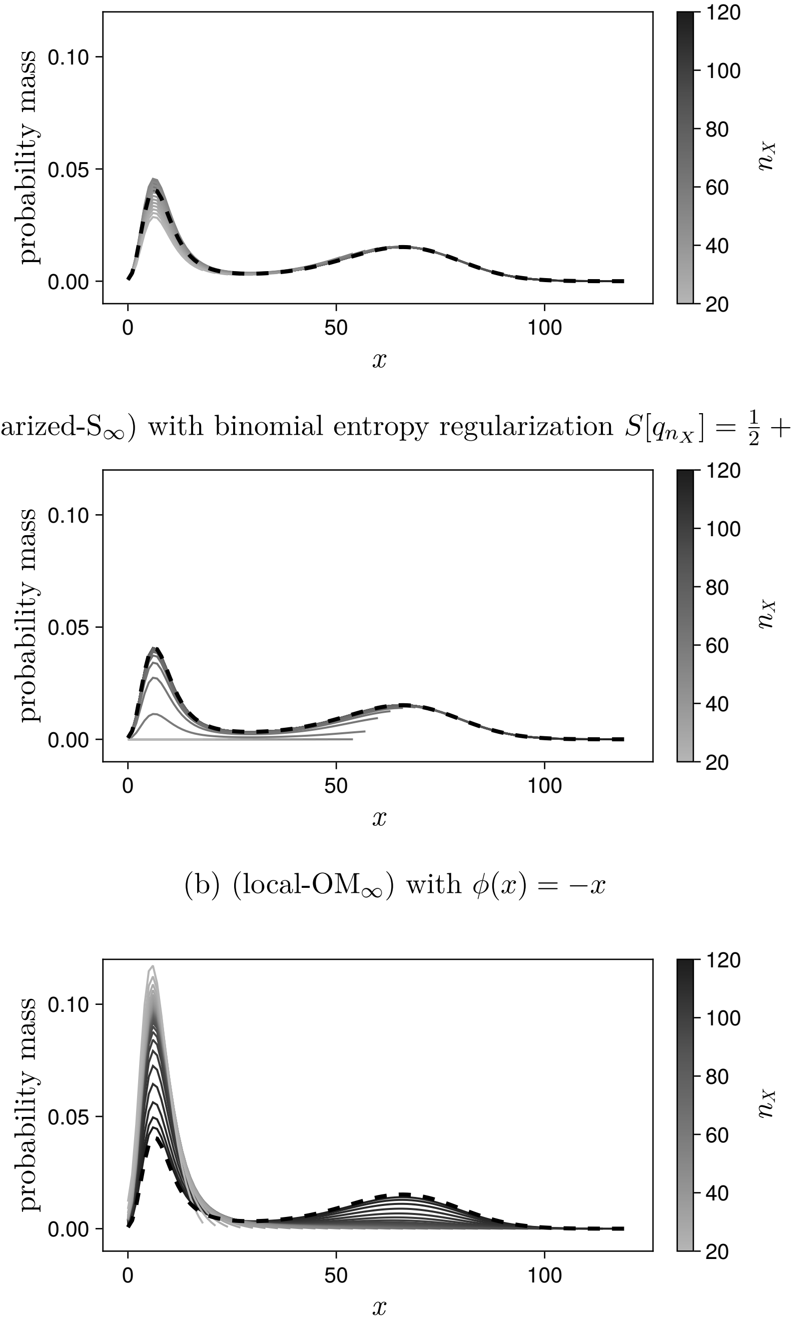
For all approximation schemes we utilize the following partition to overapproximate the state space $X = \mathbb{Z}_+$ of Schlögl's system:

$$X_k = \begin{cases} \{k-1\}, & k = 1, \dots, n_X - 1 \\ \{x \in \mathbb{R} : x \geq n_X\}, & k = n_X. \end{cases}$$

We refer to the states $\{0, \dots, n_X - 2\}$ as the state space truncation. For entropy regularization, we choose $S[q_k] = 0$ for the partition elements that are singletons.[3] For the entropy contribution of the remaining states lumped in the partition element X_{n_X}, we leverage the prior knowledge that Schlögl's system has a bimodal stationary distribution with approximately binomial tails. We thus expect that the probability mass concentrating in X_{n_X} is well appoimated by a binomial distribution. In line with this rationale, we choose the entropy regularization $S[q_{n_X}] = \frac{1}{2} + \frac{1}{2} \log \frac{\pi n}{2}$ which approximates the entropy of a binomial distribution with parameters n and $p = 0.5$. We choose $n = 100$ but show in Appendix B.3 that the results are rather insensitive to this choice.

Figure 4-3 shows approximations of the stationary distribution obtained by all three approaches for successively growing state space truncations. Notably, the local occupation measure framework allows for significantly more accurate approximations than the FSP algorithm for small state space truncations. This is a consequence of the increased flexibility of the local occupation measure framework which, in congruence with necessary MSOS conditions, allows probability mass to escape from the

[3] Note that this corresponds to choosing q_k as the Dirac measure at X_k and therefore recovers the exact contribution of μ_k to the entropy of the stationary measure proxy $\mu = \sum_{k=1}^{n_X} \mu_k$.



(a) (regularized-S_∞) with binomial entropy regularization $S[q_{n_X}] = \frac{1}{2} + \frac{1}{2}\log 50\pi$

(b) (local-OM_∞) with $\phi(x) = -x$

(c) stationary FSP algorithm [18] with designated state $x = 0$

Figure 4-3: Approximation of the stationary distribution of Schlögl's system for different approximation algorithms. Approximations are shown with solid lines on the state space truncation $\{0, \ldots, n_X - 1\}$. The true stationary distribution is indicated by the dashed line.

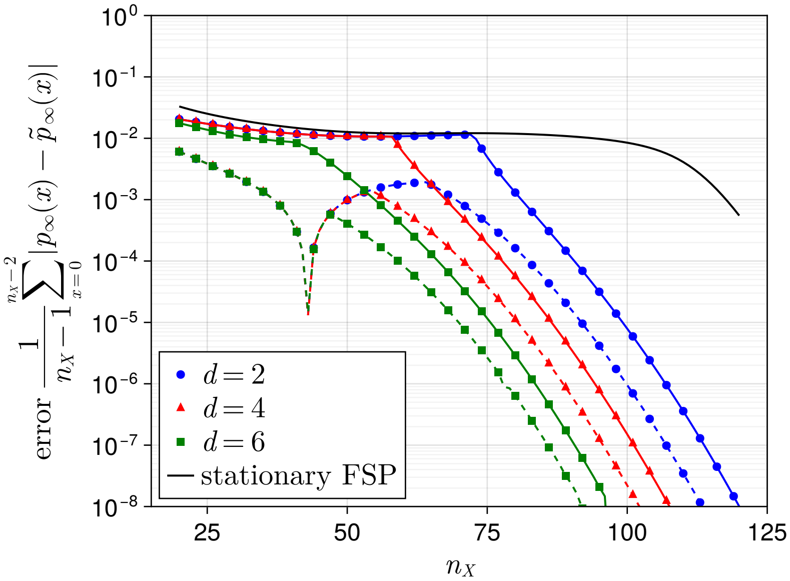


Figure 4-4: Errors in stationary measure approximations for Schlögl's system computed via degree-d MSOS relaxations of (regularized-S_∞) (colored dashed lines), (local-OM_∞) (colored solid lines), and the stationary FSP algorithm [18] (black solid line) for growing state space truncations. The entropy regularization for truncated region in (regularized-S_∞) is chosen as $S[q_{n_X}] = \frac{1}{2} + \frac{1}{2}\log 50\pi$. The objective function in (local-OM_∞) is chosen as $\phi(x) = -x$.

state space truncation. The FSP algorithm, in contrast, forces by construction all probability mass to concentrate in the state space truncation, rendering accurate approximation impossible when the true distribution assigns significant mass to states outside of the truncation.

Figure 4-4 further contrasts the approximation error attained by the FSP and both local occupation measure approaches. Remarkably, the MSOS relaxations of (regularized-S_∞) furnish significantly more accurate approximations than both other approaches. The improvement over FSP is particularly notable. Interestingly, the approximation error of the entropy-regularized local occupation measure approach does not decay monotonically and has a local minimum at a truncation that covers the first ~ 40 states. This truncation captures only the first mode of the true stationary distribution and omits the second. For this partition, the chosen binomial entropy regularization is thus particularly adequate as it aligns closely with the truncated mode of the true distribution.

4.7 Conclusion

The analysis of stochastic reaction systems relies widely on approximations that exploit unverifiable assumptions. Rigorous error control is therefore imperative to rule out erroneous conclusions. A host of recently proposed moment bounding schemes constitute a practical way to certify correctness. Here, we have unified and extended these bounding schemes under the framework of local occupation measures. We further establish direct connections of this framework to FSP algorithms and the truncation-based bounding scheme of Kuntz *et al.* [81]. Notably, the notion of localized occupation measures gives rise to tractable entropy-regularized bounding problems which are shown with an example to enable excellent approximation for stationary distributions. As such, the local occupation measure framework bridges the gap between (truncation-based) moment bounding schemes and moment closure approximations invoking the maxmimum entropy principle [11].

Part II

Quantifying the limits of quantum control

Chapter 5

A (very) brief primer on the mathematics of quantum mechanics

In this chapter, we introduce the quintessential mathematical concepts underpinning the modern description of quantum mechanics. We focus on the concepts most relevant for the task of quantum control. For a more thorough treatment of the subject, the reader is referred to [102–104] and the references therein.

5.1 Notation

In contrast to most material on the subject, we do not use Dirac notation. Instead, vectors (kets) are simply denoted by lower-case symbols and dual vectors (bras) are indicated with a superscript asterisk, i.e., the dual vector of ψ is denoted ψ^*. Beyond that, we rely on the following notational conventions throughout Part II of this thesis.

Linear algebra & analysis – The adjoint of a matrix A will be denoted by A^*. The commutator and anticommutator of two square matrices A and B will be denoted by $[A, B] = AB - BA$ and $\{A, B\} = AB + BA$, respectively. The notation $\langle \cdot\,, \cdot \rangle$ should not be confused with the Dirac notation commonly used in quantum physics but instead should be understood more broadly as bilinear form (duality bracket) between two (dual) vector spaces. We use $\mathcal{C}^{1,2}(A)$ to denote functions that are once, respectively twice, continuously differentiable with respect to their first, respectively second, argument on the domain A; when A is closed, differentiability shall be understood in the sense of Whitney [23].

Probability – For the sake of a light notation, we denote the classical expectation of a random variable x by $\mathbb{E}[x]$ and omit explicit reflection of the underlying probability measure as that will be clear from context throughout. We use δ_x to refer to the Dirac measure at the singleton $\{x\}$. Finally, we differentiate intrinsically stochastic processes from deterministic dynamical systems with our notation by indicating the time dependence of the former as subscript and of the latter as an argument.

Algebraic geometry – The set of polynomials with real coefficients in the variables x will be denoted by $\mathbb{R}[x]$; similarly, we refer to the restriction of $\mathbb{R}[x]$ to polynomials with degree at most d with $\mathbb{R}_d[x]$. The set of sum-of-squares polynomials will be denoted by $\Sigma^2[x]$. Whenever we refer to polynomials in $\mathbb{R}[\rho]$ where $\rho \in \mathbb{C}^{n \times n}$, we mean a polynomial with real coefficients jointly in the elements of $\mathrm{Re}(\rho)$ and $\mathrm{Im}(\rho)$. Lastly, we refer to vector- and matrix-valued functions as polynomials when all of their components are polynomials.

5.2 Hilbert space

A Hilbert space is a Banach space, i.e., a normed vector space in which all Cauchy sequences converge to a point in the vector space, endowed with an inner product structure [105]. In the context of quantum mechanics, Hilbert spaces form the home of quantum states. The inner product structure allows the comparison between two quantum states and quantification of their overlap (or similarity) simply through inner products. Although Hilbert spaces may generally be finite- or infinite-dimensional, we will consider only finite-dimensional Hilbert spaces here and thus not worry about the many complications that arise in the infinite-dimensional case.

5.3 Pure and mixed quantum states

The elements (vectors) ψ of an n-dimensional Hilbert space $\mathcal{H}_n$ may be considered the states of an n-level quantum system. That said, depending on how we choose to represent $\mathcal{H}_n$ not every element in $\mathcal{H}_n$ will correspond to a distinct quantum state. In the following and without loss of generality, we will simply choose $\mathcal{H}_n = \mathbb{C}^n$ (endowed with the usual inner product) and then consider all quantum states identical up to an arbitrary scaling (norm). To avoid ambiguities, we then insist that quantum states have unit norm. In this formalism, the set

$$B = \{\psi \in \mathbb{C}^n : \|\psi\|_2 = 1\},$$

describes all distinct configurations of an n-level quantum system. This characterization is particularly convenient since it admits a straightforward interpretation of a quantum state ψ as the encoding of a probability distribution over a set of basis states. More precisely, if we choose an orthonormal basis $e_1, \dots, e_n$ of $\mathbb{C}^n$, then any $\psi \in B$ admits a decomposition

$$\psi = \sum_{i=1}^{n} (e_i^* \psi) e_i$$

such that $\sum_{i=1}^{n} |e_i^* \psi|^2 = 1$. ψ can therefore be interpreted as encoding of a probability distribution that assigns probability $|e_i^* \psi|^2$ to the basis state e_i.

States $\psi \in B$ are referred to as *pure* quantum states. The description of quantum systems with pure states is useful and complete only if there is no uncertainty about the state of the system. If the state of the quantum system is not known with complete certainty, for example, due to noise or decoherence as a consequence of interaction with the environment, a strictly more general description in terms of density matrices is required. A density matrix provides an encoding of an entire ensemble of quantum states in the following sense: Suppose we lack complete knowledge of the precise state of a quantum system but instead know that the system occupies one of the finitely many distinct pure states $\psi_1, \dots, \psi_m \in B$, each with probability $p_1, \dots, p_m > 0$ such that $\sum_{i=1}^m p_i = 1$. Such an ensemble of quantum states is encoded by the density matrix

$$\rho = \sum_{i=1}^m p_i \psi_i \psi_i^*.$$

We say that the quantum state ρ encodes a *mixed* quantum state. Note that this is in line with the interpretation of each pure quantum state ψ_i as a distribution over basis states which lets us interpret ρ as a classical probabilistic mixture of these distributions with mixture weights $p_1, \ldots, p_m$. When expressed in a given basis, the diagonal entries in ρ are the average populations of the different basis states. The following theorem provides a complete characterization of density matrices.

Theorem 5.1 (Characterization of density matrices [102])**.** *A Hermitian matrix $\rho \in \mathbb{C}^{n\times n}$ is a density matrix if and only if it satisfies*

(i) $\operatorname{tr}(\rho) = 1$

(ii) $\psi^* \rho \psi \geq 0$ *for any* $\psi \in \mathbb{C}^n$

Proof. First consider a density matrix $\rho = \sum_{i=1}^m p_i \psi_i \psi_i^*$. Condition (i) is satisfied since

$$\operatorname{tr}(\rho) = \operatorname{tr}\left(\sum_{i=1}^m p_i \psi_i \psi_i^*\right) = \sum_{i=1}^m p_i (\psi_i^* \psi_i) = \sum_{i=1}^m p_i = 1,$$

where the second equality follows from the cyclic property of the trace. Furthermore, ρ satisfies condition (ii) since

$$\psi^* \rho \psi = \sum_{i=1}^m p_i |\psi_i^* \psi|^2 \geq 0.$$

Conversely, consider a Hermitian matrix ρ that satisfies conditions (i) and (ii). By the spectral theorem, ρ admits a decomposition

$$\rho = \sum_{i=1}^m p_i \psi_i \psi_i^*$$

with $m \leq n$ and $\psi_1, \ldots, \psi_m$ orthonormal. Conditions (i) and (ii) thus imply directly that

$$\operatorname{tr}(\rho) = \sum_{i=1}^{m} p_i \operatorname{tr}(\psi_i \psi_i^*) = \sum_{i=1}^{m} p_i = 1$$

and $p_1, \ldots, p_m \geq 0$, respectively. In other words, the spectral decomposition of ρ furnishes an ensemble of (orthogonal) quantum states $\psi_1, \ldots, \psi_m$ with probabilities $p_1, \ldots, p_m$. □

A corollary to Theorem 5.1 is the relation between the rank of the density matrix and purity of the state.

Corollary 5.1. *A density matrix ρ encodes a mixed state if and only if it has rank greater than one. Conversely, a density matrix encodes a pure state if and only if it has rank one.*

The following closed basic semialgebraic characterization of the set of density matrices encoding pure states will play an important role for the following chapters.

Proposition 5.1. *The set of pure quantum states encoded by a density matrix is described by $B = \{\rho \in \mathbb{C}^{n \times n} : \rho^* = \rho, \operatorname{tr}(\rho) = 1, \operatorname{tr}(\rho^2) = 1\}$.*

Proof. By Theorem 5.1 and Corollary 5.1, any density matrix ρ representing a pure state admits a factorization $\rho = \psi\psi^*$ where $\psi \in \mathbb{C}^n$ has unit norm. It follows that $\operatorname{tr}(\rho^2) = \operatorname{tr}(\psi(\psi^*\psi)\psi^*) = \operatorname{tr}(\rho) = 1$. □

This result further motivates an overapproximation of the set of density matrices with a simple inequality for the *purity* $\operatorname{tr}(\rho^2)$ of a quantum state.

Proposition 5.2. *If ρ is a density matrix, then its purity satisfies the inequality $\operatorname{tr}(\rho^2) \leq 1$. If ρ further encodes a mixed quantum state, the inequality is strict.*

Proof. Let ρ be a density matrix. By Theorem 5.1, ρ admits a decomposition

$$\rho = \sum_{i=1}^{m} p_i \psi_i \psi_i^*$$

with $p_1, \dots, p_m > 0$, $\sum_{i=1}^m p_i = 1$ and $\psi_1, \dots, \psi_m$ orthonormal. It follows from orthonormality of the ψ_i that

$$\operatorname{tr}\left(\rho^2\right) = \sum_{i=1}^m p_i^2 \le 1.$$

Finally note that, by strong convexity, the inequality above is tight if and only if $m = 1$. □

5.4 Quantum measurement

Quantum information science sets out to exploit the principles of quantum mechanics to derive protocols and build devices that can process data in a targeted manner. A key ingredient to achieve this goal is the ability to interact with a quantum system in two ways: targeted manipulation of the system state to process information and measurement of the system's observables to retrieve the result. Quantum control is concerned with the former kind of interaction, while quantum measurement deals with the latter.

The simplest kinds of measurements are projective measurements. Projective measurements are defined by observables which in turn are Hermitian matrices defined on the underlying Hilbert space. Consider such an observable, say $M \in \mathbb{C}^n$. By the spectral theorem, M admits a decomposition $M = \sum_{j=1}^m \lambda_j \Pi_j$ where $\Pi_1, \dots, \Pi_m$ are mutually orthogonal projectors onto the eigenspaces of M. When M has n distinct eigenvalues, each projector has rank one and can be represented as $\Pi_j = \phi_j \phi_j^*$, where ϕ_j is the eigenvector corresponding to the eigenvalue λ_j of M. A projective measurement generated by such an observable is called a von Neumann measurement. It has n distinct outcomes that coincide with the eigenvalues of M. When measuring a system in a pure quantum state ψ, the probability of measuring λ_j is

$$\mathbb{P}\left[\text{measuring } \lambda_j\right] = |\phi_j^* \psi|^2 = \psi^* \Pi_j \psi.$$

A natural extension of this definition to the case of measuring a mixed quantum state

$\rho = \sum_{j=1}^{m} p_j \psi_j \psi_j^*$ is obtained by interpreting the mixed quantum state as an ensemble of pure quantum states. We thus define

$$\begin{aligned}\mathbb{P}\left[\text{measuring } \lambda_j\right] &= \sum_{k=1}^{m} \mathbb{P}\left[\text{measuring } \lambda_j \mid \text{system is in state } \psi_k\right] p_k \\ &= \sum_{k=1}^{m} p_k(\psi_k^* \Pi_j \psi_k) = \sum_{k=1}^{m} p_k \mathrm{tr}\left(\Pi_j \psi_k \psi_k^*\right) = \mathrm{tr}\left(\Pi_j \rho\right).\end{aligned}$$

It is worth emphasizing that the above rules for projective quantum measurements need not be treated as definitions but instead can be derived from more fundamental primitives [103, Chapter 1.2].

So far we have laid out the rules for how likely projective measurement outcomes are. Lastly, we need to define what happens to the quantum state upon measurement. To that end, we need to distinguish two scenarios: After measurement, we can either look at the measurement outcome and use it or we can throw the measurement result away. In the case of a von Neumann measurement, if we use the measurement, the quantum state collapses to the eigenstate of the measured eigenvalue, say ϕ_j. Formally, if ψ_+ (or ρ_+ in density matrix form) denotes the quantum state post measurement, we are left with a pure state $\psi_+ = \phi_j$ (or $\rho_+ = \phi_j \phi_j^*$ in density matrix form). If we instead apply the measurement but ignore the outcome, we are left with the entire ensemble of all possible outcomes, i.e., the mixed quantum state

$$\rho_+ = \sum_{j=1}^{n} \mathrm{tr}\left(\Pi_j \rho\right) \Pi_j = \sum_{j=1}^{n} \Pi_j \rho \Pi_j.$$

In the more general case of a projective measurement, the eigenspaces of the observable can be degenerate. In those cases, the same construction as above applies except that a measurement need not lead to collapse to a pure quantum state coinciding with the measured eigenstate. If a degenerate eigenvalue is measured, the state collapses to a mixed state in the corresponding eigenspace

$$\rho_+ = \frac{\Pi_j \rho \Pi_j}{\mathrm{tr}\left(\Pi_j \rho\right)}.$$

(Note that this relation reduces readily to $\rho_+ = \phi_j \phi_j^*$ if the eigenspace is not degenerate and spanned by ϕ_j.)

The concept of projective measurements is further generalized by the notion of positive operator-valued measurements (POVMs). POVMs play an important role in describing and constructing continuous measurements of quantum systems and thus for quantum feedback control [103]. In contrast to projective measurements, POVMs are defined not by projectors onto the eigenspaces of an observable but by a collection of operators $M_1, \ldots, M_m \in \mathbb{C}^{n \times n}$ such that $\sum_{i=1}^m M_i^* M_i = I$. While the associated operators M_j need no longer be projectors, the rules for quantum measurement remain similar. The probability of measurement outcome j (now corresponding to the operator M_j as opposed to the projector onto the j^{th} eigenspace) conditioned on measuring a quantum system in state ρ is $\operatorname{tr}\left(M_j \rho M_j^*\right)$; the effect of the measurement on the quantum state is

$$\rho_+ = \frac{M_j\,\rho M_j^*}{\operatorname{tr}\left(M_j\,\rho M_j^*\right)}$$

when the measurement outcome is used and

$$\rho_+ = \sum_{j=1}^{m} M_j\,\rho M_j^*$$

when it is ignored. Note that these rules reduce to the previously introduced rules when the $M_1, \ldots, M_m$ are chosen to be projectors onto the eigenspaces of an observable.

5.5 The dynamics of closed quantum systems

The time evolution of the state of a closed quantum system with Hermitian Hamiltonian $H \in \mathbb{C}^{n \times n}$ is governed by the (time-dependent) Schrödinger equation

$$\frac{\mathrm{d}\psi}{\mathrm{d}t}(t) = -iH\psi(t). \tag{5.1}$$

(We have chosen units in which Planck's constant is unity.) The Hamiltonian encodes how the system components interact energetically. We emphasize that the dynamics (5.1) give rise to a unitary evolution, i.e., the state $\psi(t)$ will remain of unit norm if evolved according to Equation (5.1). To see this, note that the closed-form solution of Equation (5.1) for a given initial state ψ_0 is

$$\psi(t) = \exp\left(-itH\right)\psi_0.$$

Since H is Hermitian, $-itH$ is skew-symmetric, and $\exp\left(-itH\right)$ thus unitary.

The analog of the Schrödinger equation for the density matrix description of quantum systems is known as the (Liouville-)von Neumann Equation. It is obtained directly from the Schrödinger Equation by a simple calculation: Consider a density matrix $\rho = \sum_{j=1}^{m} p_j \psi_j \psi_j^*$ and note that

$$\begin{aligned}\frac{\mathrm{d}\rho}{\mathrm{d}t} &= \sum_{j=1}^{m} p_j \frac{\mathrm{d}(\psi_j\psi_j^*)}{\mathrm{d}t} \\ &= \sum_{j=1}^{m} p_j \left(\frac{\mathrm{d}\psi_j}{\mathrm{d}t}\psi_j^* + \psi_j \frac{\mathrm{d}\psi_j^*}{\mathrm{d}t}\right) \\ &= \sum_{j=1}^{m} p_j \left(-iH\psi_j\psi_j^* + i\psi_j\psi_j^* H\right) \\ &= -i\left[H, \rho\right], \qquad (5.2)\end{aligned}$$

where $[A, B] = AB - BA$ denotes the commutator and we used that H is Hermitian.

5.6 The dynamics of open quantum systems

The time evolution behavior of open quantum systems is more diverse than that of their closed counterparts. The dynamics of open quantum systems are not necessarily unitary and depend on the specific details of the interactions with the environment. If the interactions are measurements, it is further important to distinguish if the state is conditioned on the measurements or not. Here, we focus on two particular kinds of continuous measurements: direct photon counting and homodyne detection. Among continuous measurements, they are the most relevant for quantum information science and engineering [106–112], and also for the contributions of this thesis (see Chapter 7).

We follow [103, Chapter 4] to introduce the idea of continuous measurements and sketch the derivation of the governing equations for the state of a continuously measured quantum system. For a more formal treatment, the reader is referred to Belavkin [113].

5.6.1 Photon counting

The dynamics of a quantum system with Hamiltonian H subjected to photon counting measurements can be deduced by designing a suitable POVM as introduced in Section 5.4. For the sake of simplicity, let us consider a POVM with only two outcomes described by the measurement operators

$$M_1(h) = I - \left(\frac{1}{2}\sigma^*\sigma + iH\right) h \text{ and } M_2(h) = \sqrt{h}\sigma.$$

The so-called jump operator σ characterizes this measurement as it defines via M_2 the state the system jumps to upon emission of a photon. It is easily verified that

$$M_1(h)^* M_1(h) + M_2(h)^* M_2(h) = I + O(h^2)$$

so that M_1 and M_2 form indeed an asymptotically valid POVM as h approaches 0. If we now assume that the measurement takes time h and the state of the system is ρ_t at time t, we find that the evolution of the system state conditioned on repeated application of the POVM is characterized by the recursion

$$\rho_{t+h} = \frac{M_i(h)\rho_t M_i(h)^*}{\operatorname{tr}\left(M_i(h)\rho_t M_i(h)^*\right)} \text{ with probability } \operatorname{tr}\left(M_i(h)\rho_t M_i(h)^*\right).$$

Expanding the above expression to first order in h, we obtain

$$\rho_{t+h} = \begin{cases} \rho_t + \Big(-i[H,\rho_t] - \dfrac{1}{2}\{\sigma^*\sigma,\rho_t\} + \operatorname{tr}(\sigma\rho_t\sigma^*)\,\rho\Big)\,h, & \text{with prob. } 1 - \operatorname{tr}(\sigma\rho_t\sigma^*)\,h \\ \dfrac{\sigma\rho_t\sigma^*}{\operatorname{tr}(\sigma\rho_t\sigma^*)} & \text{with prob. } \operatorname{tr}(\sigma\rho_t\sigma^*)\,h \end{cases}$$

In the limit $h \to 0$ this reduces to a stochastic process driven by a Poisson counter $\mathrm{d}n_t$ with rate $\operatorname{tr}(\sigma\rho_t\sigma^*)$:

$$\mathrm{d}\rho_t = \Big(-i[H,\rho_t] - \frac{1}{2}\{\sigma^*\sigma,\rho_t\} + \operatorname{tr}(\sigma\rho_t\sigma^*)\,\rho_t\Big)\,\mathrm{d}t + \left(\frac{\sigma\rho_t\sigma^*}{\operatorname{tr}(\sigma\rho_t\sigma^*)} - \rho_t\right)\mathrm{d}n_t. \quad (5.3)$$

This is a form of a stochastic master equation which describes the random paths of the state of a quantum system when conditioned on the result of continuous photon counting measurements. If we apply continuous photon counting measurements but ignore the result, the system state evolves deterministically according to the mean $\bar{\rho}(t) = \mathbb{E}[\rho_t]$ over all possible paths of the above process. The resultant governing equation, known as the Lindblad master equation [114], reads

$$\frac{\mathrm{d}\bar{\rho}}{\mathrm{d}t}(t) = -i[H,\bar{\rho}(t)] + \sigma\bar{\rho}(t)\sigma^* - \frac{1}{2}\{\sigma^*\sigma,\bar{\rho}(t)\}.$$

5.6.2 Homodyne detection

For homodyne detection, the system's photon emissions are superimposed with the emissions of a strong local oscillator [103]. The advantage of this setup is that it allows the measurement of specific amplitude components or quadratures of the measured signal which has various practical advantages for quantum feedback control and other applications [115–117, for example].

A stochastic master equation analogous to that derived for photon counting in the previous section can be obtained from an analogous POVM as before. This time, however, the system Hamiltonian and jump operator must be adjusted to account for the superimposed oscillator field. For an oscillator with strength $\gamma \in \mathbb{R}$, the Hamiltonian and jump operator change according to $H \to H - \frac{i\gamma}{2}(\sigma - \sigma^*)$ and $\sigma \to$

$\sigma + \gamma I$ [103, Chapter 4.4]. Proceeding as before, we obtain a stochastic master equation of the form (5.3). It is easy to see from Equation (5.3) that the rate of photon detection is then proportional to γ^2 in the homodyne detection setup. In other words, the photon detections are dominated by the local oscillator. In the ideal limit of an infinitely strong oscillator, this dependence allows us to further simplify the stochastic master equation to a diffusion equation. In this limit, the statistics of the driving Poisson counter increments $\int_t^{t+h} \mathrm{d}n_t$ will be approximately Gaussian distributed with mean $\left(\gamma^2 + \gamma \operatorname{tr}\left((\sigma + \sigma^*)\rho\right)\right) h$ and variance $\gamma^2 h$ [118]. Combing this with Equation (5.3) and letting γ approach infinity while scaling $h \sim \gamma^{-3/2}$ finally yields

$$\mathrm{d}\rho_t = \left(-i[H, \rho_t] + \sigma\rho_t\sigma^* - \frac{1}{2}\{\sigma^*\sigma, \rho_t\}\right) \mathrm{d}t + \left(\sigma\rho_t + \rho\sigma^* - \operatorname{tr}\left(\rho_t\sigma + \rho_t\sigma^*\right)\rho_t\right) \mathrm{d}b_t,$$

where $\mathrm{d}b_t$ denote standard Brownian increments. Physically, the Brownian motion b_t is in one-to-one correspondence with a physically measured photon current induced by the superposition of the oscillator field and the system's photon emissions. The above equation can therefore be used to infer a quantum system's state from experimentally observed measurements.

Chapter 6

Quantifying the limits of open-loop quantum control

6.1 Introduction

The speed and accuracy with which quantum states can be controlled put a limit on the capabilities of quantum information technology. Quantum speed limits in the spirit of the celebrated Mandelstam-Tamm bound [19] or Margolus-Levitin theorem [20] establish that these limitations are non-trivial for problems of minimal time transition and, likewise, we ought to expect non-trivial limitations for other performance metrics in quantum control. Two natural questions therefore arise when designing quantum control protocols.

Question 1 (Certification)**.** *Is a projected performance goal for the operation of a given quantum device fundamentally unattainable?*

and the natural follow-up question

Question 2 (Quantification)**.** *What is the best attainable performance?*

In this Chapter, we apply the (local) occupation measure framework developed Part I of this thesis to devise a method to answer Question 1 definitively and Question 2 to arbitrary precision for a large class of open-loop quantum control problems.

For problems of minimal time transition, Question 1 can in some situations be answered by one of many quantum speed limits [19, 20, 119–122] providing hard lower bounds on the shortest possible time for the transition between quantum states [123]; see the review [124] for a detailed account. The performance bounds implied by such quantum speed limits, however, are often not informative in practice. They derive from fundamental trade-offs like energy-time uncertainty relations and leverage only coarse-grained system information, such as summary statistics of the system's energy. In practice, however, technological constraints like bounds on control drives and the detailed structure of the control fields and other energetic interactions impose more stringent performance barriers. With the exception of some recent work [125, 126] such information remains generally unaccounted for in the derivation of quantum speed limits, rendering them simply too conservative in many situations.

In principle, a rich set of tools from classical control theory like the Hamilton-Jacobi-Bellman equations or Pontryagin's maximum principle is available to characterize and identify globally optimal quantum control protocols [127–130]. While elegant when possible, these tools, however, remain limited to rather specific, usually two-level [130, 131], quantum systems without constraints on the system's state trajectory and the control drives. More complicated settings are often out of reach for this approach due to the difficulty of posing optimality conditions and establishing their sufficiency [46]. Practitioners instead resort in most cases to numerical local search techniques to design control protocols [132]; for example, Krotov's method [133–135], gradient ascent pulse engineering [136, 137], nonlinear programming via direct collocation [138], reinforcement learning [139], or the chopped random basis technique [140]. The empirical success of these techniques is often attributed to the well-known fact that the performance landscape generated by transition fidelities between given quantum states has no spurious local optima under the assumption of full and unconstraintred controllability [141]. In practice, however, these assumptions are typically unrealistic; control fields and drives are generally confined by technological constraints. Local search methods can in those cases, at best, ensure local optimality due to the typically non-convex nature of quantum control problems, and consequently are fundamentally incapable of providing affirmative answers to the Certification Question 1, let alone quantitative answers to Question 2.

To remove any doubt that a given control protocol is in fact (near-)optimal, its performance must match (or at least be close to) a guaranteed bound on the best attainable performance. Zhang *et al.* [21] recently proposed a systematic, convex optimization-based framework to compute such bounds for a large class of quantum optimal control problems. They recast the quantum optimal control problem as a quadratically constrained quadratic program by introducing a sufficient characterization of quantum evolution in terms of generalized probability conservation laws in the form of quadratic integral equations. Upon discretization and semidefinite relaxation of the quadratic constraints, a finite semidefinite program bounds the best attainable performance. In contrast to quantum speed limits, this framework incorporates constraints, fine-grained system information, accommodates a wider range of performance metrics for quantum control, and is shown to consistently furnish (more) informative bounds. That said, the discretization and semidefinite relaxation step required for its construction introduces an unknown level of error and conservatism. We show here that a deterministic analog of the (local) occupation measure framework introduced in Chapter 3 gives rise to a hierarchy of semidefinite bounding problems that do not suffer from the same limitations and that can under mild assumptions reliably answer Questions 1 and 2.

6.2 Open-loop quantum optimal control

The goal of open-loop quantum optimal control is to direct the dynamical evolution of a quantum system in a way that minimizes a given objective functional without relying on feedback information. The conceptually simplest and, due to its relevance for many quantum computing tasks, most commonly studied open-loop quantum optimal control problem is that of minimal time transition, also known as the quantum Brachistochrone problem [128, 129, 142–144]. Analogous to its classical counterpart, the objective of a quantum Brachistochrone problem is to drive a quantum system

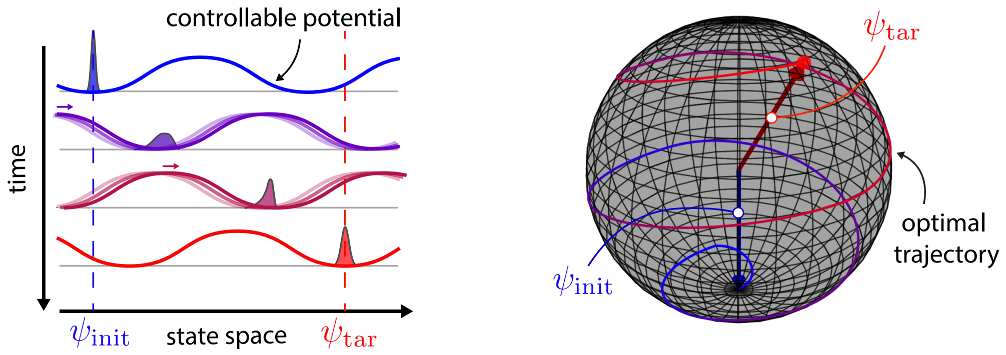


Figure 6-1: Illustration of quantum optimal control for minimal time transition.

from a given initial state ψ_{init} (for example the outcome of a projective measurement) to a given target state ψ_{tar} (for example an input state needed to implement a quantum information protocol) in the shortest possible time. As illustrated in Figure 6-1, such a targeted evolution of the system is facilitated by manipulation of its Hamiltonian. Designing protocols that realize or approximate an optimal Hamiltonian manipulation, typically through the application of external, time-dependent electromagnetic fields, is at the heart of any open-loop quantum optimal control problem. In the following, we introduce a mathematical formalism and model abstraction for this design problem.

We consider quantum systems with Hamiltonians of the form

$$H(u) = H_0 + \sum_{k=1}^{K} u_k H_k,$$

where H_0 is the Hamiltonian of the nominal, uncontrolled system and $H_1, \ldots, H_K$ are external control fields with real-valued drives $u = [u_1 \ \cdots \ u_K]$. We assume without loss of generality that the control fields are constant[1] and take the control drives and their variation over time to be the subject of design. Note that this is in line with most experimental setups where the structure of the control fields is dictated by available instrumentation (laser sources) but their intensity (power supply) can be varied. Following this intuition, we also assume that the control drives are confined

[1] If for a given problem the control fields are subject to design, it suffices in this formalism to let $H_1, \ldots, H_K$ be a basis of the feasible design space.

to an admissible set $U \subset \mathbb{R}^K$, representing technological constraints such as bounded power supply or safety restrictions. We further assume complete knowledge of the initial system state ψ_{init} and that the system is closed. Consequently, we treat the system state as pure and describe its dynamical evolution over the control horizon $[0, T]$ with the time-dependent Schrödinger equation

$$\begin{cases} \dfrac{\mathrm{d}\psi}{\mathrm{d}t}(t) = -iH(u(t))\psi(t), \quad t \in [0, T] \\ \psi(0) = \psi_{\text{init}}, \end{cases} . \tag{6.1}$$

Here, $\{u(t) \in U : t \in [0, T]\}$ denotes a feasible control protocol[2]; that is, explicitly time-dependent admissible control drives. We further consider restrictions on trajectories of the quantum system during and at the end of the control horizon T. Formally, we will impose that the quantum state evolves and terminates in feasible sets X and X_T, respectively. Restrictions of this form are common in practice where they encode constraints on maximum allowable leakage into undesirable states [145] or terminal accuracy [21].

Equipped with the model abstraction outlined above, the following optimization problem provides a mathematical formalism for open-loop quantum optimal control.

$$\begin{aligned} J^* := \inf_{u} \quad & \int_0^T \ell(t, \psi(t), u(t))\,\mathrm{d}t + m(\psi(T)) \\ \text{s.t.} \quad & \psi(t) \text{ satisfies (6.1)}, \\ & \psi(t) \in X \text{ on } [0, T], \\ & \psi(T) \in X_T, \\ & u(t) \in U \text{ on } [0, T]. \end{aligned} \tag{QOCP}$$

To quantify the limits of quantum control we seek lower bounds for the optimal value J^* of (QOCP), encoding the best attainable control performance as quantified by an accumulating stage cost ℓ and terminal cost m. In the following section, we list several common quantum control tasks which reduce to special cases of (QOCP).

[2] We assume throughout implicitly that a feasible control protocol is absolutely integrable.

6.3 Special cases

6.3.1 Quantum Brachistochrone problems

The quantum Brachistochrone problem seeks the minimal time transition between an initial quantum state ψ_{init} and a target quantum state ψ_{tar}. Although the control horizon T is fixed in (QOCP), one may simply rephrase the search for the minimal time transition as a sequence of feasibility problems of the form

$$\begin{aligned} \text{find} \quad & u \\ \text{s.t.} \quad & \psi(t) \text{ satisfies (6.1)}, \\ & \psi(T) = \psi_{\text{tar}}, \\ & u(t) \in U \text{ on } [0, T], \end{aligned}$$

The search for the minimal allowable control horizon T that renders the above problem feasible coincides with the search for the minimal time transition. The former can be done by successively bisecting the interval $[0, T]$.

We emphasize that the feasibility problem above is obtained as a special case of (QOCP) by choosing vanishing stage and terminal cost, and letting the terminal constraint set be the singleton $X_T = \{\psi_{\text{tar}}\}$. Another common variation of the quantum Brachistochrone problem is obtained when the terminal state is only required to be ϵ-close to the target state. This setting is also readily accommodated by choosing the terminal constraint set $X_T = \{\psi \in \mathbb{C}^n : |\psi^* \psi_{\text{tar}}|^2 \geq 1 - \epsilon\}$.

6.3.2 State preparation problems

A common quantum control problem closely related to the quantum Brachistochrone problem is that of state preparation. Here, the objective is to drive the system as close as possible to a given target state ψ_{tar} within a fixed afforded control horizon

$[0, T]$. This problem is obtained as a special case of (QOCP) with the following form

$$\begin{aligned} \inf_u \quad & 1 - |\psi(T)^* \psi_{\text{tar}}|^2 \\ \text{s.t.} \quad & \psi(t) \text{ satisfies (6.1)}, \\ & u(t) \in U \text{ on } [0, T]. \end{aligned}$$

Note that the choice of terminal cost $m(\psi(T)) = 1 - |\psi(T)^* \psi_{\text{tar}}|^2$ precisely quantifies the overlap between the terminal and target state and attains its minimum value of zero if and only if they coincide.

6.3.3 Minimum energy transition problems

The promise of faster and more accurate quantum devices has historically been driving the development and application of quantum optimal control [146]. In the wake of steadily growing uncertainty about the future of the global energy system [147–149], however, quantum optimal control is garnering increasing attention as a means to attenuate the energy requirements of emerging quantum technologies [150, 151]. A natural control problem arising in this context is that of minimum energy transitions; that is, the task of finding a transition protocol between a given initial and target quantum state within an afforded maximal time and with minimal required energy expenditure. Formally, this problem may be stated as

$$\begin{aligned} \inf_u \quad & \int_0^T E(u(t)) \, \mathrm{d}t \\ \text{s.t.} \quad & \psi(t) \text{ satisfies (6.1)}, \\ & \psi(T) = \psi_{\text{tar}}, \\ & u(t) \in U \text{ on } [0, T], \end{aligned}$$

which is easily identified as a special case of (QOCP). The accumulating cost $E(u)$ measures the energy expenditure associated with the control action u. Typical choices are $E(u) = \|H(u)\|_2$ [151] and $E(u) = \|u\|_2^2$ [152], which measure the maximum energy of the system and the power supplied through the external control fields, respectively.

6.4 A convex bounding approach

Even for convex cost functionals and control constraints, the open-loop quantum optimal control problem (QOCP) is generally a non-convex optimization problem due to the bilinear structure of the Schrödinger equation (6.1). Local optimization routines are therefore only guaranteed to converge to globally optimal control protocols under strong, typically unrealistic, and unverifiable assumptions [141]. To certify that a control protocol is near-optimal it is therefore necessary to compare its performance to a guaranteed (and informative) bound for the best attainable control performance. Using the moment-sum-of-squares hierarchy in combination with the (local) occupation measure framework, we next construct tractable optimization problems that furnish such bounds. To that end, however, we must adopt a reformulation of (QOCP) in terms of real-valued variables and impose additional restrictions on the nature of the functions and constraints that frame the problem.

6.4.1 Real reformulation

In order to draw on the tools from real algebraic geometry and the moment-sum-of-squares hierarchy as reviewed and extended in Chapters 2 and 3, respectively, we reformulate (QOCP) to involve only real variables. To that end, we represent the state vector ψ of the quantum system by its real and imaginary parts ψ_R and ψ_I, respectively. For simplicity, we abuse notation and from here onward implicitly identify a quantum state ψ with its real representation (ψ_R, ψ_I). In particular, we will not explicitly redefine sets and functions that were originally defined for complex quantum states but instead treat them as composed with the bijection

$$\mathbb{R}^{2n} \ni (\psi_R, \psi_I) \mapsto \psi_R + i\psi_I \in \mathbb{C}^n$$

between the real and complex representation of quantum states. The corresponding expressions are easily derived; for instance, the Schrödinger equation (6.1) takes in real coordinates the form

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\psi_R(t)\\ \psi_I(t)\end{bmatrix} = \begin{bmatrix} H_I(u(t)) & H_R(u(t))\\ -H_R(u(t)) & H_I(u(t))\end{bmatrix}\begin{bmatrix}\psi_R(t)\\ \psi_I(t)\end{bmatrix}, \quad t \in [0,T] \\ \begin{bmatrix}\psi_R(0)\\ \psi_I(0)\end{bmatrix} = \begin{bmatrix}\psi_{R,\text{init}}\\ \psi_{I,\text{init}}\end{bmatrix} \end{cases}, \tag{6.2}$$

where H_R and H_I refer to the real and imaginary part of the system Hamiltonian, respectively.

6.4.2 Assumptions on constraints and cost functionals

We make the following assumptions to ensure that the cost function and constraints in (QOCP) are described entirely by polynomials and closed basic semialgebraic sets, respectively. This property will be a key ingredient for constructing tractable convex lower bounding problems for (QOCP). Note that the following assumptions are based on the real representation of quantum states as discussed in the previous paragraph.

Assumption 6.1. *The sets U, X, and X_T in* (QOCP) *are compact closed basic semialgebraic sets. Accordingly, there exist finite collections of polynomials $\mathcal{U} \subset \mathbb{R}[u]$, and $\mathcal{X}, \mathcal{X}_T \subset \mathbb{R}[\psi]$ such that $U = \{u \in \mathbb{R}^K : p(u) \geq 0,\ \forall p \in \mathcal{U}\}$, $X = \{\psi \in \mathbb{R}^{2n} : p(\psi) \geq 0,\ \forall p \in \mathcal{X}\}$, and $X_T = \{\psi \in \mathbb{R}^{2n} : p(\psi) \geq 0,\ \forall p \in \mathcal{X}_T\}$ are compact. We refer to the polynomials in $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{X}_T$ as control, state, and terminal constraints, respectively.*

Assumption 6.2. *Let $B = \{\psi \in \mathbb{R}^{2n} : \|\psi\|_2^2 = 1\}$ denote the set of pure quantum states. The accumulating stage cost $\ell : [0,T] \times B \times U \to \mathbb{R}$ and terminal cost $m : B \to \mathbb{R}$ are polynomials jointly in all arguments.*

A few remarks are in order to put Assumptions 6.1 and 6.2 into perspective. First, it is worth noting that control constraints represent almost exclusively technological limitations. As such, they typically rule out unbounded control drives and describe a closed, and hence compact, set of admissible control actions. In fact, the characterization of technological limits is in practice rarely more complicated than simple box constraints which readily satisfy Assumption 6.1. Moreover, a review of common quantum control problems, as provided in Section 6.3, reveals that also state and terminal constraints are typically representable by polynomial inequalities of degree at most two. The state and terminal constraint sets X and X_T can further always be made compact as they may without loss of generality be intersected with the set of pure states. Assumption 6.1 can therefore be expected to hold in most practically relevant settings. Similarly, most cost functionals used in practice can be encoded by polynomials of degree at most two (cf. Section 6.3) which renders also Assumption 6.2 rather weak.

6.4.3 Open-loop quantum optimal control via infinite-dimensional linear programming

To construct convex bounding problems for (QOCP), we adopt the occupation measure approach for deterministic optimal control introduced by Lasserre *et al.* [46]. Although this method represents a strict special case of the framework outlined in Chapter 3, we revisit its core concepts and assumptions here to discuss them from the perspective of open-loop quantum control. To that end, we first introduce the notion of instantaneous and state-action occupation measures as an alternative description for the trajectories $\{(\psi(t), u(t)) : t \in [0,T]\}$ of a controlled quantum system and its dynamics.

The instantaneous occupation measure is defined as

$$\nu_T(C) = \mathbb{1}_C(\psi(T))$$

for any Borel subset of C of the set of pure quantum states B [46]. Here, $\mathbb{1}_C$ denotes the indicator of the set C, thus $\nu_T(C)$ simply indicates whether or not the quantum trajectory terminates in C. The associated duality bracket, i.e., the operation of averaging a continuous observable $w \in \mathcal{C}(B)$ with respect to ν_T, therefore reduces to

$$\langle w, \nu_T \rangle = \int_B w(\psi) \, \mathrm{d}\nu_T(\psi) = w(T, \psi(T)). \tag{6.3}$$

The state-action occupation measure is defined for any Borel subsets $C \subset [0, T]$, $D \subset B$, and $E \subset U$ such that

$$\xi(C \times D \times E) = \int_{C \cap [0,T]} \mathbb{1}_{D \times E}((\psi(t), u(t))) \, \mathrm{d}t.$$

Intuitively, $\xi(C \times C \times E)$ measures the time that the trajectory $\{(t, \psi(t), u(t)) : t \in [0, T]\}$ resides in $C \times D \times E$. Its duality bracket therefore reduces to the computation of time averages of continuous $w \in \mathcal{C}([0, T] \times B \times U)$ along this trajectory:

$$\langle w, \xi \rangle = \int_{X_T} w(\psi) \, \mathrm{d}\nu_T(\psi) = \int_0^T w(t, \psi(t), u(t)) \, \mathrm{d}t. \tag{6.4}$$

The support of the occupation measures encodes satisfaction of path and terminal constraints. This is immediately apparent from the definition of the instantaneous occupation measure since $\nu_T(X_T) = 1$ is equivalent to the condition that $\psi(T) \in X_T$. Similarly, the expected state-action occupation measure satisfies the condition

$$\xi([0, T] \times X \times U) = \int_{[0,T]} \mathbb{1}_{X \times U}((\psi(t), u(t))) \, \mathrm{d}t = T$$

if and only if $(\psi(t), u(t)) \in X \times U$ for almost every $t \in [0, T]$.

The key advantage of the occupation measure-based abstraction of quantum trajectories is that is gives rise to a linear description of quantum dynamics. Specifically, by the fundamental theorem of calculus, the instantaneous and state-action occupation measures describe the dynamics of a quantum system through the linear weak form condition

$$\langle w, \nu_T \rangle - w(0, \psi_{\text{init}}) = \langle \mathcal{A}w, \xi \rangle, \quad \forall w \in \mathcal{C}^1([0, T] \times X), \tag{6.5}$$

where $\mathcal{A}$ denotes the Liouville operator defined such that

$$\frac{\mathrm{d}w}{\mathrm{d}t}(t,\psi(t)) = \mathcal{A}w(t,\psi(t),u(t))$$

holds along the trajectories of the system. For controlled quantum systems governed by the (real) Schrödinger equation (6.2), $\mathcal{A}$ thus acts on smooth functions according to

$$\mathcal{A}w(t,\psi,u) = \frac{\partial w}{\partial t}(t,\psi) + \begin{bmatrix} \nabla_{\psi_R} w(t,\psi) \\ \nabla_{\psi_I} w(t,\psi) \end{bmatrix}^\top \begin{bmatrix} H_I(u) & H_R(u) \\ -H_R(u) & H_I(u) \end{bmatrix} \begin{bmatrix} \psi_R \\ \psi_I \end{bmatrix}. \tag{6.6}$$

It follows from these consideration that any feasible trajectory for (QOCP) generates non-negative occupation measures ν_T and ξ that satisfy Equation (6.5) and are supported only on X_T and $[0,T] \times X \times U$, respectively. Thus, the following linear program describes a valid lower bound for the optimal value of (QOCP):

$$\begin{aligned} J^*_{\mathrm{OM}} \coloneqq \inf_{\nu_T,\xi} \quad & \langle \ell, \xi\rangle + \langle m, \nu_T\rangle && \text{(OM-QOCP)} \\ \text{s.t.} \quad & \langle w, \nu_T\rangle - w(0,\psi_{\text{init}}) = \langle \mathcal{A}w, \xi\rangle, \quad \forall w \in \mathcal{C}^1([0,T]\times X), \\ & \nu_T \in \mathcal{M}_+(X_T), \\ & \xi \in \mathcal{M}_+([0,T]\times X\times U). \end{aligned}$$

Here, $\mathcal{M}_+(Y)$ denotes the cone of non-negative measures supported on Y. It can be shown by an analogous argument as for the stochastic case treated in Chapter 3 that the linear programming dual of (OM-QOCP),

$$\begin{aligned} J^*_{\mathrm{HJB}} \coloneqq \sup_{w} \quad & w(0,\psi_{\text{init}}) && \text{(HJB-QOCP)} \\ \text{s.t.} \quad & \mathcal{A}w + \ell \geq 0 \text{ on } [0,T]\times X\times U, \\ & m - w(T,\cdot) \geq 0 \text{ on } X_T, \\ & w \in \mathcal{C}^1([0,T]\times X), \end{aligned}$$

admits a useful interpretation as seeking the maximal smooth subsolution for the value function associated with (QOCP) (cf. Corollary 3.1). Any near-optimal point of (HJB-QOCP) can therefore be used to construct heuristic control protocols [47]. Moreover, the results of Lasserre *et al.* [46] and Vinter [59] establish easily verifiable conditions under which the linear programs (OM-QOCP) and (HJB-QOCP) characterize the true optimal control performance J^* defined by (QOCP).

Theorem 6.1. *Let Assumptions 6.1 and 6.2 be satisfied, the set of admissible control actions U be compact and convex, and the stage cost $\ell(t, \psi, \cdot)$ be convex on U for any $\psi \in X$ and $t \in [0, T]$. Further, assume that* (QOCP) *is feasible, i.e., there exists some control protocol $\{u(t) \in U : t \in [0, T]\}$ that induces a trajectory of the quantum state that satisfies all constraints of* (QOCP)*. Then, $J^* = J^*_{\text{OM}} = J^*_{\text{HJB}}$.*

Proof. This result is a special case of [46, Theorem 2.3]. To establish that the hypotheses of [46, Theorem 2.3] are satisfied, let $\psi \in X$ and $t \in [0, T]$ be arbitrary but fixed. Then, define the function

$$V(u) := \begin{bmatrix} H_I(u) & H_R(u) \\ -H_R(u) & H_I(u) \end{bmatrix} \begin{bmatrix} \psi_R \\ \psi_I \end{bmatrix}.$$

Clearly, the image of U under V is convex as V is an affine map [26]. It remains to show that the function

$$g(v) := \inf_{u \in U} \{\ell(t, \psi, u) : v = V(u)\}$$

is convex. To that end, consider v_1, v_2 such that $g(v_1)$ and $g(v_2)$ are finite. Since U is compact, and V and ℓ are continuous, the infimum in the definition of g is attained. Thus, there exist $u_1, u_2 \in U$ such that $g(v_1) = \ell(t, \psi, u_1)$ and $g(v_2) = \ell(t, \psi, u_2)$. Since further V is an affine function, it follows that $\alpha v_1 + (1 - \alpha)v_2 = V(\alpha u_1 + (1 - \alpha)u_2)$ holds for any $\alpha \in [0, 1]$. By convexity of U, definition of g, and convexity of ℓ in its last argument, we thus conclude that

$$g(\alpha v_1 + (1-\alpha)v_2) \leq \ell(t, \psi, \alpha u_1 + (1-\alpha)u_2) \leq \alpha g(v_1) + (1-\alpha)g(v_2).$$

□

We remark that the hypotheses of Theorem 6.1 hold for a broad class of quantum optimal control problems. In particular, the conclusion of Theorem 6.1 applies to all problems stated in Section 6.3 under the assumption box or polyhedral control constraints that imply bounded controls.

6.4.4 Tractable moment-sum-of-squares relaxations

The infinite-dimensional linear programs (OM-QOCP) and (HJB-QOCP) require further approximation to become computationally tractable. The moment-sum-of-squares hierarchy provides a systematic way to construct such approximations while preserving the lower-bounding property of (OM-QOCP) and (HJB-QOCP). On an intuitive level, the approximations are obtained by relaxing (OM-QOCP) to optimization over finite, truncated moment sequences of the measures ξ and ν_T, or conversely restricting (HJB-QOCP) to optimization over polynomials of bounded maximum degree. Upon these approximations, tractable and finite semidefinite programs are obtained by relaxing, respectively restricting, the non-negativity constraints in (OM-QOCP) and (HJB-QOCP): Non-negativity of the occupation measures is relaxed to necessary positive semidefiniteness constraints on moment and localizing matrices, while the non-negativity constraints in (HJB-QOCP) are restricted to sufficient sum-of-squares constraints; see Chapter 2 for more details on this construction. The approximation quality of the so-obtained restrictions and relaxations can further be improved by the partitioning approach introduced in Chapter 3. For the sake of brevity, however, we state only the traditional sum-of-squares restriction of

(HJB-QOCP) below.

$$
\begin{aligned}
J^*_{\mathrm{HJB},d} \coloneqq \inf_{w_d} \quad & w_d(0, \psi_{\mathrm{init}}) && \text{(HJB-QOCP}_d\text{)} \\
\text{s.t.} \quad & \mathcal{A} w_d + \ell \in Q_{d+1}\left(\mathcal{T} \cup \mathcal{X} \cup \mathcal{U}\right), \\
& m - w_d(T, \cdot) \in Q_d\left(\mathcal{X}_T\right), \\
& w_d \in \mathbb{R}_d[t, \psi].
\end{aligned}
$$

Here, $Q_d\left(\mathcal{S}\right)$ refers to the truncated quadratic module (cf. Definition 2.15) generated by a collection of polynomials $\mathcal{S} = \{p_1, \dots, p_n\}$. The constraint set $\mathcal{T} = \{t, T - t\}$ is chosen such that it frames the control horizon via $[0, T] = \{t : p(t) \geq 0, \ \forall p \in \mathcal{T}\}$. Note that (HJB-QOCP$_d$) is well-posed as the left-hand side of all constraints in (HJB-QOCP$_d$) are in fact polynomials due to Assumption 6.2 and the observation that the Liouville operator $\mathcal{A}$ maps polynomials to polynomials (cf. Equation (6.6)).

It is easily seen that $J^*_{\mathrm{HJB},d}$ forms by construction a sequence of monotonically improving lower bounds for J^* with increasing degree d. Further, we can establish mild and practically verifiable conditions under which these bounds converge to J^*_{HJB} (and via Theorem 6.1 also to J^*).

Theorem 6.2. *Suppose the control, state, and terminal constraints $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{X}_T$ as defined in Assumption 6.1 satisfy Putinar's condition (see Definition 2.16). Then, strong duality holds between* (OM-QOCP) *and* (HJB-QOCP) *and the optimal value $J^*_{\mathrm{HJB},d}$ converges from below to J^*_{HJB} as $d \to \infty$.*

Proof. The result follows as a special case of [75, Corollary 8] which proves strong duality as well as primal/dual convergence of the moment-sum-of-squares hierarchy for the class of generalized moment problems (cf. Problem (GMP)) with at most countably many moment constraints that satisfy Putinar's condition (cf. Definition 2.16) and explicitly bounded zeroth order moments. We show that (OM-QOCP) falls into this class.

To that end, we first recognize that, by the density of polynomials in continuous functions on compact sets, it suffices to impose condition (6.5) in (OM-QOCP) for all monomial test functions. Therefore (OM-QOCP) is equivalent to a generalized moment problem with countably many moment constraints. Next, we observe that $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{T}$ contain polynomials in distinct variables (u, ψ, and t). Thus, $\mathcal{T} \cup \mathcal{X} \cup \mathcal{U}$ satisfies Putinar's condition as $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{T}$ do. ($\mathcal{T}$ describes a bounded polyhedron and thus satisfies Putinar's condition [42].) Finally, we observe that for test functions $w(t, x) = 1$ and $w(t, x) = t$, the condition (6.5) implies that any feasible point of (OM-QOCP) must satisfy that $\langle 1, \nu_T \rangle = 1$ and $\langle 1, \xi \rangle = T$. The zeroth order moments of any feasible measures are therefore bounded. □

We finally remark that the hypotheses of Theorem 6.2 are broadly satisfied for quantum optimal control problems arising in practice. In particular, the state and terminal constraints can without loss of generality be augmented to satisfy Putinar's condition, simply by including the (redundant) constraint $1 - \|\psi\|_2^2 \geq 0$. Similarly, the control constraints typically describe a compact set. If the set is a polytope, for example an bounded interval box, it readily satisfies Putinar's condition. If it is more complicated but still compact, Putinar's condition can be enforced to hold by adding an additional redundant control constraint of the form $R - \|u\|_2^2 \geq 0$ with sufficiently large $R > 0$.

6.5 Extensions

6.5.1 Variable control horizon

Improving the processing speed of quantum devices is one of the key motivating applications of quantum control. Accordingly, many common quantum optimal control problems boil down to performing a specific manipulation of a quantum state in minimal time. We have discussed in Section 6.3 that these problems can be reformulated as a sequence of feasibility problems, each with a fixed control horizon. However, the bounding framework outlined above readily extends to dealing with problems that do not have a fixed horizon, allowing minimal time problems to be tackled directly. To that end, we consider the following generalization of (QOCP)

$$\begin{aligned}
\inf_{u,T} \quad & \int_0^T \ell(t,\psi(t),u(t))\,\mathrm{d}t + m(\psi(T)) \\
\text{s.t.} \quad & \psi(t) \text{ satisfies (6.1)}, \\
& \psi(t) \in X \text{ on } [0,T], \\
& \psi(T) \in X_T, \\
& u(t) \in U \text{ on } [0,T], \\
& T \in [0,\bar{T}].
\end{aligned}$$

Here, $\bar{T}$ denotes the maximal affordable control time. Minimal time problems are obtained from the above formulation when choosing $\ell \equiv 1$ and $m \equiv 0$. The infinite-dimensional linear programming counterpart of the above problem is given by the following primal-dual pair.

$$\begin{aligned}
\inf_{\nu_T,\xi} \quad & \langle \ell, \xi \rangle + \langle m, \nu_T \rangle \\
\text{s.t.} \quad & \langle w, \nu_T \rangle - w(0,\psi_{\text{init}}) = \langle \mathcal{A}w, \xi \rangle\,, \quad \forall w \in \mathcal{C}^1([0,\bar{T}] \times X), \\
& \nu_T \in \mathcal{M}_+([0,\bar{T}] \times X_T), \\
& \xi \in \mathcal{M}_+([0,\bar{T}] \times X \times U).
\end{aligned}$$

$$\begin{aligned}
\sup_{w} \quad & w(0,\psi_{\text{init}}) \\
\text{s.t.} \quad & \mathcal{A}w + \ell \geq 0 \text{ on } [0,\bar{T}] \times X \times U, \\
& m - w \geq 0 \text{ on } [0,\bar{T}] \times X_T, \\
& w \in \mathcal{C}^1([0,T] \times X).
\end{aligned}$$

As before, the decision variables ν_T and ξ admit interpretation as an occupation measure pair, albeit with slightly modified definition. To account for the variable control horizon T, the instantaneous occupation measure ν_T is now given by

$$\nu_T(A \times B) = \mathbb{1}_{A \times B}((T,\psi(T)))$$

for any Borel sets $A \subset [0, \bar{T}]$ and $B \subset X_T$. The dual interpretation of the linear programming approach as finding the maximal smooth HJB subsolution also remains valid. Furthermore, by analogous arguments as for the fixed control horizon case, the above pair of primal-dual linear programs admit tractable moment-sum-of-squares approximations under Assumptions 6.1 and 6.2.

6.5.2 Mixed quantum states

The formulation of the open-loop quantum optimal control problem (QOCP) as well as the associated convex bounding problems considered so far admit straightforward extension to handling mixed quantum states. For mixed quantum states, represented in terms of a density matrix, the open-loop quantum optimal control problem reads

$$\begin{aligned}
\inf_{u} \quad & \int_0^T \ell(t, \rho(t), u(t)) \, \mathrm{d}t + m(\rho(T)) \\
\text{s.t.} \quad & \begin{cases} \dfrac{\mathrm{d}\rho}{\mathrm{d}t}(t) = -i[H(u(t)), \rho(t)], \ t \in [0, T] \\ \rho(0) = \rho_{\text{init}} \end{cases}, \\
& \rho(t) \in X \text{ on } [0, T], \\
& \rho(T) \in X_T, \\
& u(t) \in U.
\end{aligned}$$

Following an analogous construction as in Section 6.4, this problem may again be conservatively approximated by a primal-dual pair of infinite-dimensional linear programs analogous to (OM-QOCP) and (HJB-QOCP), which under similar conditions as Assumptions 6.2 and 6.1 further admit tractable moment-sum-of-squares approximations. It must be noted, however, that in particular the representation of state and terminal constraints as closed basic semialgebraic sets is more complicated in the mixed state regime. A detailed discussion of this issue presented in Chapter 7 in the context of closed-loop control of open quantum systems where it bears greater relevance.

6.5.3 Gate design

Quantum gates are the elementary building blocks for programs in the circuit model of quantum computation [102]. As such, their design is essential for a practical realization of general purpose quantum computers. Mathematically, a quantum gate represents a unitary transformation, and its application to a quantum state is physically realized by the unitary evolution of the state under a time-dependent Hamiltonian according to the Schrödinger equation (6.1). The design of control protocols that induce this evolution in minimal time is therefore essential for the development of fast quantum computing hardware and hence a key application of open-loop quantum optimal control. To apply the bounding method outlined in Section 6.4 to this problem, we first note that the evolution of a unitary matrix $V(t)$ under a controlled Hamiltonian is described by

$$\begin{cases} \frac{\mathrm{d}V}{\mathrm{d}t}(t) = -iH(u(t))V(t), \quad t \in [0,T] \\ V(0) = V_{\text{init}} \end{cases}. \tag{6.7}$$

For the sake of gate design, V_{init} is typically taken as the identity and one seeks a control protocol that terminates in a target unitary matrix V_{tar} at the end of the control horizon T. Applying such a control protocol to a system with the same controlled Hamiltonian, which initially resides in state ψ, then amounts to the computation of $V_{\text{tar}}\psi$.

In the language of optimal control, the gate design problem for a target unitary $V_{\text{tar}} \in \mathbb{C}^{n\times n}$ can be stated as follows.

$$\begin{aligned} \inf_{u} \quad & 1 - \frac{|\mathrm{tr}\,(V_{\text{tar}}^* V(T))\,|^2}{n^2} \\ \text{s.t.} \quad & V(t) \text{ satisfies (6.7)}, \\ & u(t) \in U \text{ on } [0,T]. \end{aligned}$$

Here, $|\mathrm{tr}\,(V_{\text{tar}}^* V(T))\,|/n$ is the gate fidelity which assumes its maximal value of unity if and only if $V(T) = V_{\text{tar}}$.

The proposed bounding approach and all of its guarantees extend readily to the gate design problem as it can be reduced to a state preparation problem discussed in Section 6.3. To that end, simply consider an auxiliary n^2-level quantum system with initial and target states given by

$$\hat{\psi}_{\text{init}} = \frac{1}{n} \begin{bmatrix} V_{\text{init}} e_1 \\ \vdots \\ V_{\text{init}} e_n \end{bmatrix} \text{ and } \hat{\psi}_{\text{tar}} = \frac{1}{n} \begin{bmatrix} V_{\text{tar}} e_1 \\ \vdots \\ V_{\text{tar}} e_n \end{bmatrix},$$

and block diagonal Hamiltonian

$$\hat{H}(u) = \begin{bmatrix} H(u) & & \\ & \ddots & \\ & & H(u) \end{bmatrix}.$$

Clearly, through vectorization and rescaling, the state of this auxiliary system is in one-to-one correspondence with the unitary operation representing the gate. Similarly, it is easily verified that Schrödinger equation with this auxiliary system is equivalent to Equation (6.7) and that the gate fidelity coincides with the state fidelity.

6.6 Examples

6.6.1 State preparation and minimum time transitions

We demonstrate the efficacy of the proposed sum-of-squares-based bounding method by mapping out the performance boundary for minimal time transition problems under control and leakage constraints. Specifically, we compute upper bounds for the maximum attainable fidelity given a finite control horizon. The resultant performance boundary in turn sheds light on the minimal transition time.

Throughout, we compare the proposed sum-of-squares bounds with known quantum speed limits as well as bounds furnished by the recent semidefinite programming

(SDP)-based bounding method of Zhang *et al.* [21]. As representative quantum speed limits, we compare against a practical variant of the Mandelstam-Tamm [19] and Margolus-Levitin [122] bound. To that end, we recall the Mandelstam-Tamm bound:

$$T_{\min} \geq \frac{\arccos |\psi_{\text{tar}}^* \psi_{\text{init}}|}{\Delta E},$$

where ΔE refers to the time-averaged uncertainty of the system's energy

$$\Delta E = \frac{1}{T_{\min}} \int_0^{T_{\min}} \sqrt{\psi(t)^* H(u(t))^2 \psi(t) - \left(\psi(t)^* H(u(t)) \psi(t)\right)^2} \, \mathrm{d}t.$$

Clearly, this quantity and hence also the Mandelstam-Tamm bound cannot be computed without prior knowledge of the path of the minimal time transition. A practically computable and still guaranteed variant of the Mandelstam-Tamm bound is therefore proposed in [21]. This variant relies on conservatively bounding the instantaneous energy variance in terms of the magnitude-wise largest eigenvalue of the system Hamiltonian $\lambda_{\max}[H(u)]$:

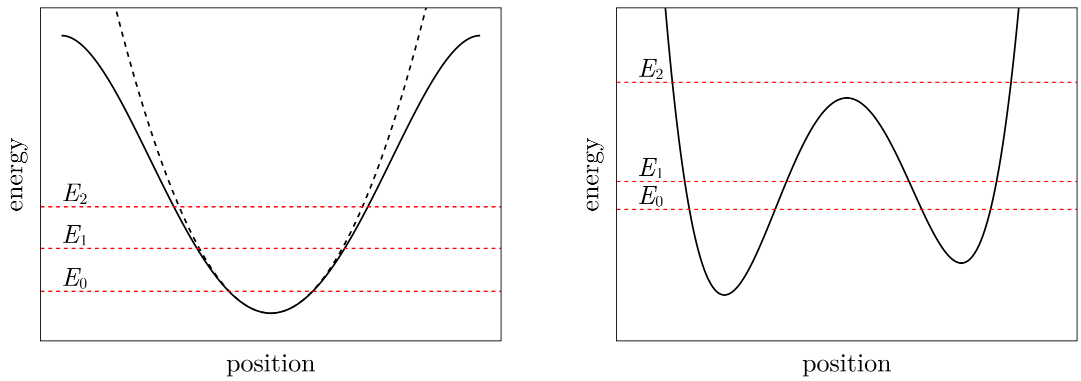
$$\psi(t)^* H(u(t))^2 \psi(t) - \left(\psi(t)^* H(u(t)) \psi(t)\right)^2 \leq \psi(t)^* H(u(t))^2 \psi(t) \leq \sup_{u \in U} \lambda_{\max}\left[H(u)\right]^2.$$

When composed with the Mandelstam-Tamm bound, this conservative estimate yields the valid lower bound

$$T_{\min} \geq \frac{\arccos |\psi_{\text{tar}}^* \psi_{\text{init}}|}{\sup_{u \in U} \lambda_{\max}\left[H(u)\right]}$$

for the minimal required transition time. Furthermore, this bound is also a conservative but practical variant of the Margolus-Levitin bound [153]

$$T_{\min} \geq \frac{\arccos |\psi_{\text{tar}}^* \psi_{\text{init}}|}{\bar{E}}$$



(a) Approximately harmonic potential (b) Asymmetric double-well potential [154]

Figure 6-2: Potential energy surfaces of the three-level quantum system models summarized in Table 6.1.

as $\lambda_{\max}[H(u)]$ also bounds the average energy

$$\bar{E} = \frac{1}{T_{\min}} \int_0^{T_{\min}} |\psi_{\text{init}}^* H(u(t)) \psi_{\text{init}}| \, \mathrm{d}t$$

from above.

For our examples, we follow Zhang *et al.* [21] and consider two three-level quantum systems with distinct features: a system in an asymmetric double-well potential [154] and a typical transmon qubit model given by an approximately harmonic potential with nearest level couplings [155]. The potential energy surfaces of both systems are illustrated in Figure 6-2 and their Hamiltonians are summarized in Table 6.1. The control goal is to drive the systems from their ground state to the first excited state in minimal time. To map out the entire performance boundary, however, we reformulate the minimal time transition problem as a sequence of state preparation problems seeking to maximize the terminal probability of residing in the first excited state, $P_2(T) = |e_2^* \psi(T)|^2$, for a range of control horizons T.

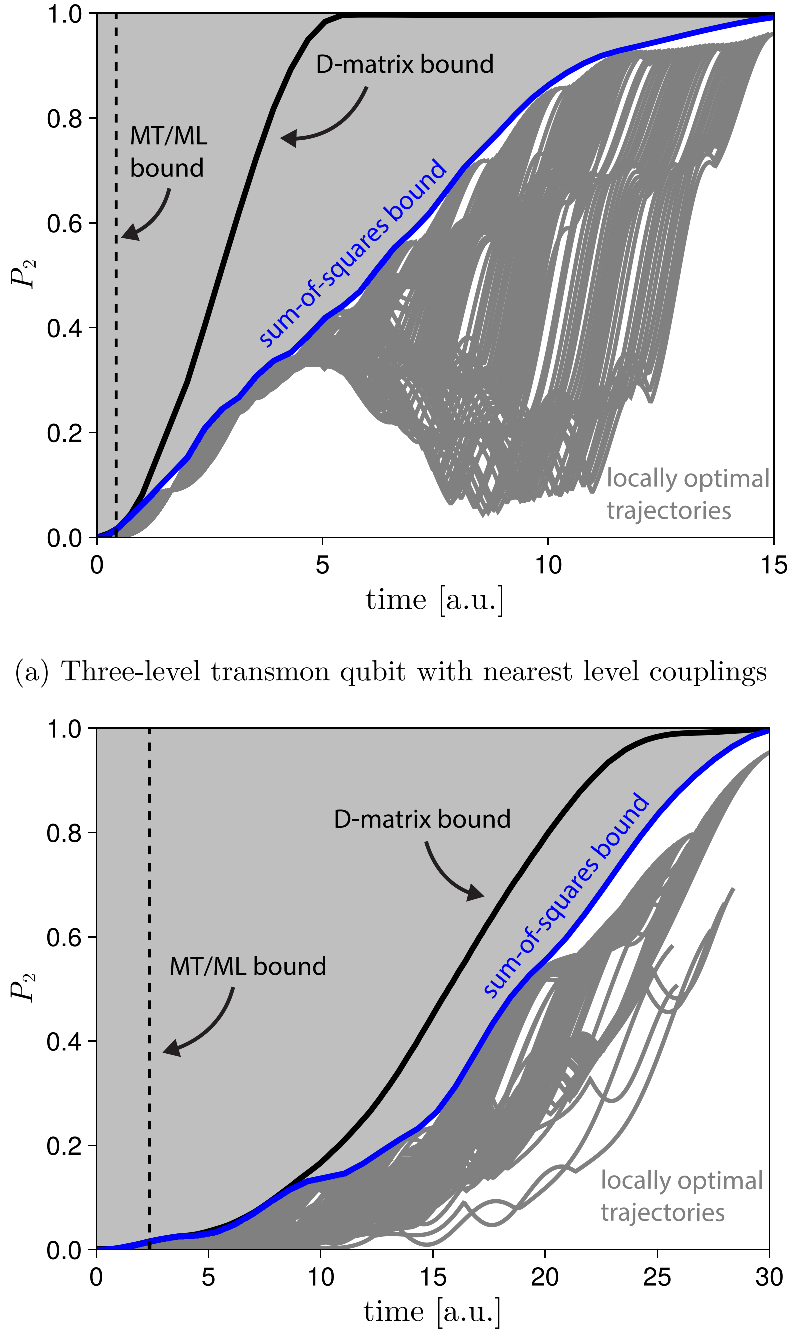
First, we consider the unconstrained case; that is, we let X and X_T coincide with the set of pure quantum states. For this case, Figure 6-3 compares the performance limits established by the sum-of-squares approach proposed here, the D-matrix SDP bound of Zhang *et al.* [21], and the described practical variant of the Mandelstam-Tamm/Margolus-Levitin bound. The performance limits are further contextual-

Table 6.1: Controlled Hamiltonians $H(u) = H_0 + uH_1$ for a three-level transmon qubit and asymmetric double-well model.

	Drift Hamiltonian H_0	**Control field** H_1	**Admissible controls** U
Approx. harmonic potential [21]	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1.9 & 0 \\ 0 & 0 & 3.7 \end{bmatrix}$	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & -\sqrt{2} \\ 0 & -\sqrt{2} & 0 \end{bmatrix}$	$[-0.15, 0.15]$
Double-well potential [154]	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.1568 & 0 \\ 0 & 0 & 0.7022 \end{bmatrix}$	$\begin{bmatrix} -2.5676 & 0.3921 & 0.6382 \\ 0.3921 & 2.3242 & -0.7037 \\ 0.6382 & -0.7037 & -0.5988 \end{bmatrix}$	$[-0.3, 0.3]$

ized by locally optimal trajectories obtained via gradient ascent pulse engineering (GRAPE) [136]. For both quantum systems, it is evident that the Mandelstam-Tamm/Margolus-Levitin bound is exceedingly conservative and not suitable to gauge performance limits, let alone to make meaningful conclusions about optimality of a given control protocol. The D-matrix bound provides a notably better quantification of the performance limits, in particular at early times. From the practical perspective, however, it is the case of long times over which a transition with near unit fidelities can be achieved that bears the most relevance. The D-matrix bound deteriorates in this regime. The sum-of-squares bounds proposed here, in contrast, remain within a few percent of the maximum attainable fidelity over the entire time horizon. As such, they provide meaningful certificates of near-global optimality for the locally optimal trajectories.

Next, we consider the minimum time transition problem under additional maximal allowable leakage constraints. While allowing the quantum system to occupy higher energy levels during transition between quantum states can enable faster state preparation, it can also be the source of additional errors that must be detected and corrected [145]. Designing transition protocols that balance transition time and leakage into undesirable states thus plays a vital role in realizing fast but fault-tolerant quantum computation. In the considered example, we incorporate such considerations by a seeking transition path subject to a maximal allowable leakage threshold $P_{3,\max}$ for the second excited state; formally, we constrain the quantum state to remain in



(a) Three-level transmon qubit with nearest level couplings

(b) Three-level system with double-well potential

Figure 6-3: Bounds for the maximum attainable fidelity for the preparation of the first excited state. The gray lines correspond to locally optimal trajectories computed via GRAPE [136]. The shaded area is fundamentally unattainable. The dashed black line marks the lower bound for the minimal transition time implied by the Mandelstam-Tamm [19] and Margolus-Levitin [20] bounds. The solid black and blue lines mark the performance boundary implied by the D-matrix bound [21] and the proposed sum-of-squares bound, respectively.

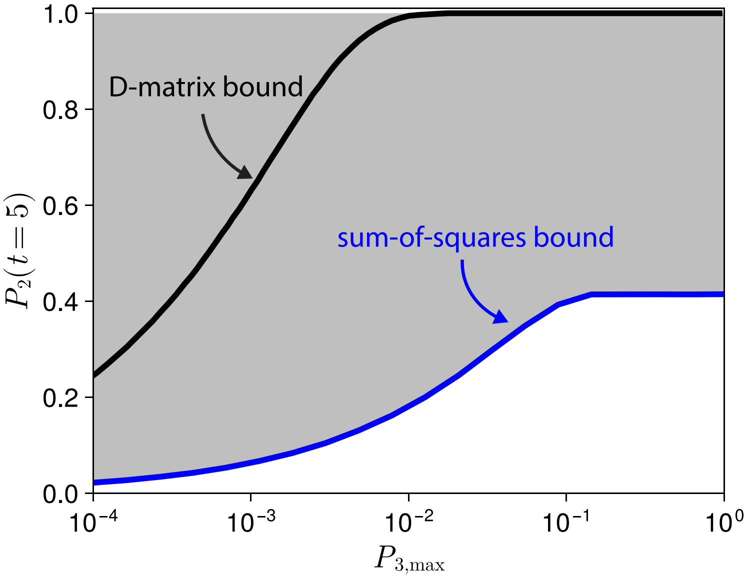


Figure 6-4: Upper bound for the attainable fidelity of preparing the first excited state of a three-level transmon qubit system (see Table 6.1) under the leakage constraint $|e_3^*\psi(t)|^2 \leq P_{3,\max}$ for the second excited state. The gray-shaded area is fundamentally not attainable. The black and blue solid lines indicate the performance boundary implied by the D-matrix bound [21] and the proposed sum-of-squares bound.

the set

$$X = \{\psi \in B : |e_3^*\psi|^2 \leq P_{3,\max}\}$$

during the transition. Due to its common use for modeling quantum computation applications, we focus here on the transmon qubit model from before (cf. Table 6.1).

Figure 6-4 compares the D-matrix [21] and sum-of-squares bounds for the maximal achievable fidelity of the first excited state at time $T = 5$ for a range of leakage budgets. The sum-of-squares bound outperforms the D-matrix bound again notably.

6.6.2 Entanglement generation

A maximally entangled pair of qubits is a common prerequisite for quantum information protocols [102]. Entanglement generation is consequently an important application for quantum control. To showcase that the sum-of-squares bounding approach outlined in this chapter applies to entanglement creation problems as well, we consider a simple two-qubit quantum system and map out its performance boundary for minimum time entanglement generation. To that end, we compute lower bounds for the optimal value of (QOCP) with terminal cost

$$m(\psi) = -4|\psi_1\psi_4 - \psi_2\psi_3|^2$$

measuring entanglement by the (negative) squared concurrence [102]. The Hamiltonian of the considered two-qubit system is given by

$$H(u) = -\frac{1}{2}\left(\sigma_z \otimes \sigma_z + I \otimes \sigma_z + \sigma_z \otimes I\right) - \frac{u}{2}\left(I \otimes \sigma_x + \sigma_x \otimes I\right), \quad u \in [-1, 1],$$

where $\otimes$ denotes the Kronecker product, and σ_x and σ_z the Pauli matrices

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ and } \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

We assume the system resides initially in a separable ground state, i.e., $\psi_{\text{init}} = e_1 \otimes e_1$.

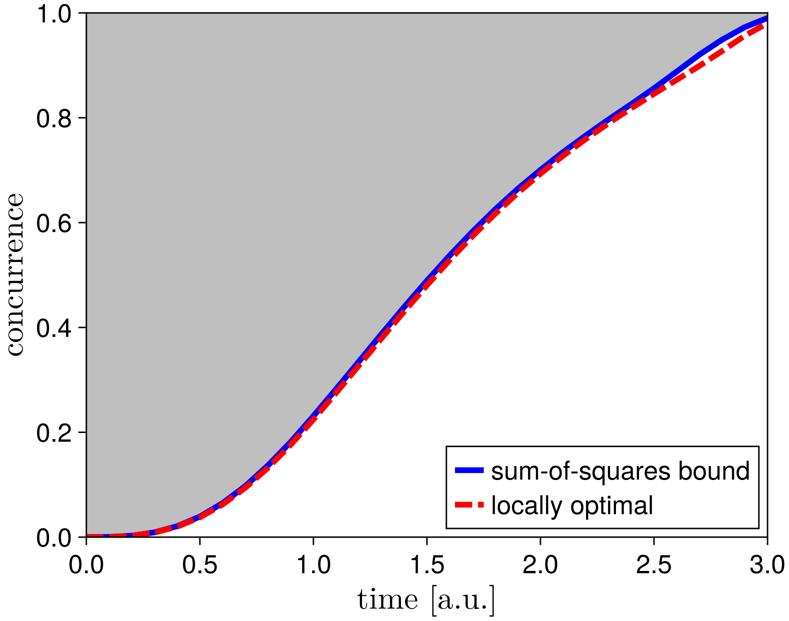


Figure 6-5: Performance boundary for minimal time entanglement generation in a two-qubit system. The gray-shaded area is fundamentally unattainable.

Figure 6-5 compares the computed performance boundary for maximal concurrence generation over time with that attained by a locally optimal control protocol (identified via GRAPE [136]). The performance boundary is quantified accurately, thus certifying near-optimality of the control protocol obtained via local search.

6.6.3 Gate design

Motivated by its relevance for quantum computing applications, we finally consider a gate design problem as discussed in Section 6.5. Following an example studied by Zhang *et al.* [21], we consider the problem of designing a single-qubit Hadamard gate,

$$V_{\text{tar}} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

a common building block for essentially all quantum circuits [102]. The controlled Hamiltonian to implement this gate is assumed to be of the form

$$H(u) = \omega \sigma_z + u \sigma_x,$$

where σ_x and σ_z again denote the Pauli matrices. The drift parameter is chosen as $\omega = 0.0784$ according to [154].

We quantify the best attainable gate fidelity over time. Figure 6-6 shows the performance boundary as quantified by the proposed bounding method. The boundary is again quantitatively accurate and can certify near-optimality of a given locally optimal control protocol to high precision over the entire fidelity range.

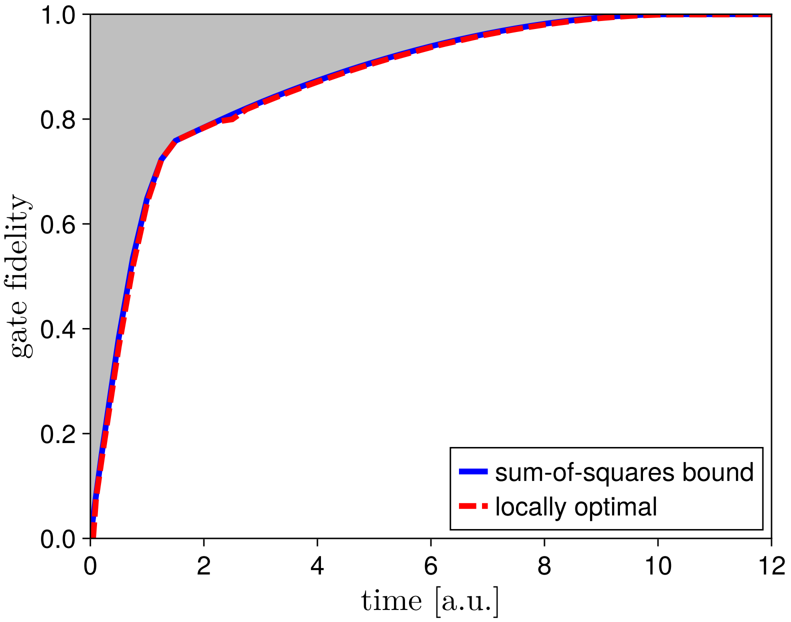


Figure 6-6: Performance boundary for Hadamard gate design problem. The gray-shaded area is fundamentally unattainable.

Chapter 7

Quantifying the limits of closed-loop quantum control

The content of this chapter has been accepted for publication in the *IEEE Transactions on Automatic Control*.

7.1 Introduction

Feedback control of devices at the quantum scale holds significant potential for current and future applications in the field of quantum information science [146, 156]. The nonlinear and stochastic nature of quantum systems under continuous observation, however, complicates the quest for an effective deployment of feedback control in practice [103]. While the characterization of optimal quantum feedback controllers through the dynamic programming principle has a longstanding history, dating back to the work of Belavkin [113] in 1988, designing such controllers by solving the quantum analog of the nonlinear Hamilton-Jacobi-Bellman (HJB) equation remains, barring a few simplified situations [157–159], an elusive challenge. It is instead common practice to rely on heuristics, often rooted in reinforcement learning, gradient-based optimization, or expert intuition, to design quantum feedback controllers in applications [132, 139, 160]. And although such heuristically derived control policies are frequently found to perform remarkably well, their degree of suboptimality essentially always remains unquantified, leaving uncertainty about whether any observed performance limitations are fundamental or simply due to a suboptimal controller design. In this section, we construct a method to bring clarity to such situations by computing informative bounds on the best attainable performance for a wide range of quantum feedback control problems. These bounds may serve as certificates of optimality, witnesses of fundamental limitations, or performance targets and, as such, complement controller design heuristics.

The key insight to enable our construction is that quantum filtering theory reduces the problem of optimal feedback control of continuously observed quantum systems to the optimal control of structured jump-diffusion processes. Therefore, by combining quantum filtering theory with the (local) occupation measure framework described in Chapter 3, we devise a method for computing informative bounds on the best attainable feedback control performance for a rich class of quantum systems via the moment-sum-of-squares hierarchy. We establish conditions under which the bounds converge to the true optimal control performance and discuss extensions of this method to account for measurement imperfections and decoherence due to non-observed decay channels.

The remainder of this Chapter is structured as follows. In Section 7.2, we define the class of quantum feedback control problems under consideration and discuss key assumptions. Our main contribution, the construction and convergence analysis of a hierarchy of increasingly tight convex bounding problems for the best attainable quantum feedback control performance, is presented in Section 7.3. Section 7.4 discusses a heuristic for the construction of near optimal control policies as well as practically relevant extensions of this hierarchy to account for infinite horizon problems, decoherence due to unobserved decay channels, and measurement imperfections. Finally, we demonstrate the practical utility of the proposed bounding method with a qubit control example in Section 7.5.

7.2 Quantum stochastic optimal control

We consider quantum systems with a Hermitian Hamiltonian of the form

$$H(u) = H_0 + \sum_{k=1}^{K} u_k H_k,$$

where H_0 denotes the drift Hamiltonian of the system, and $H_1, \dots, H_K$ are control fields with tunable drives $u = [u_1 \ \cdots \ u_K]$. The control drives are assumed to be confined to an admissible set $U \subset \mathbb{R}^K$. To enable feedback control, we consider systems subjected to a continuous measurement process ζ_t. Conditioned on this measurement process, the density matrix ρ_t encoding the state of such a system follows stochastic dynamics described by the Quantum Filtering Equation [161]

$$\mathrm{d}\rho_t = \mathcal{L}(u_t)\rho_t \, \mathrm{d}t + \mathcal{G}\rho_t \, \mathrm{d}\zeta_t. \tag{QFE}$$

Note that, in contrast to Chapter 7, we here deliberately use the more general density matrix formalism as potential measurement imperfections or entirely unobserved decay channels lead to classical probabilistic mixtures of quantum states [103]. For systems of the described structure, the action of the Lindbladian $\mathcal{L}(u)$ is given by

$$\mathcal{L}(u)\rho = -i[H(u), \rho] + \sum_{l=1}^{L} \left(\sigma_l \rho \sigma_l^* - \frac{1}{2}\{\sigma_l^* \sigma_l, \rho\} \right),$$

where the jump operators σ_l characterize the interaction between the quantum system and its environment due to observation. We focus on systems that are subjected to a combination of homodyne detection and photon counting measurements. For notational convenience, we partition the index set of measurements $\{1, \dots, L\}$ into sets HD and PC, covering the homodyne detection and photon counting measurements, respectively. For sake of conceptual simplicity, we further assume that there are no measurement inefficiencies and that all measurement channels are observed, i.e., $\mathrm{HD} \cup \mathrm{PC} = \{1, \dots, L\}$; we comment in Section 7.4 on the necessary modifications in the presence of unobserved decay channels or imperfect detection.

The innovation operator $\mathcal{G}$ decomposes under these assumptions into two separate contributions associated with homodyne detection and photon counting measurements, respectively:

$$\mathcal{G}\rho_t \,\mathrm{d}\zeta_t = \sum_{l\in\mathrm{HD}} \mathcal{G}_l\rho_t \,\mathrm{d}b_t^l + \sum_{l\in\mathrm{PC}} \mathcal{G}_l\rho_t \,\mathrm{d}n_t^l - \mathcal{L}_l\rho_t \,\mathrm{d}t.$$

Homodyne detection causes diffusive innovations described by standard Gaussian increments $\mathrm{d}b_t^l$ which are in one-to-one correspondence with a measured homodyne current [103]. The associated innovation operator acts according to

$$\mathcal{G}_l\rho_t \,\mathrm{d}b_t^l = \left(\sigma_l\rho_t + \rho_t\sigma_l^* - \mathrm{tr}\left(\sigma_l\rho_t + \rho_t\sigma_l^*\right)\rho_t\right)\mathrm{d}b_t^l.$$

Photon counting, in contrast, causes a deterministic drift

$$\mathcal{L}_l\rho_t \,\mathrm{d}t = \left(\sigma_l\rho_t\sigma_l^* - \mathrm{tr}\left(\sigma_l\rho_t\sigma_l^*\right)\rho_t\right)\mathrm{d}t.$$

and leads to discrete innovations upon detection of photon emissions as described by Poisson counters n_t^l which fire at rate $\lambda_l(\rho) = \mathrm{tr}\left(\sigma_l\rho\sigma_l^*\right)$ [103]. Conditioned on the measurement of an emitted photon, the state of the system jumps to $h_l(\rho) = \sigma_l\rho\sigma_l^*/\mathrm{tr}\left(\sigma_l\rho\sigma_l^*\right)$. The associated innovation operator accordingly acts as

$$\mathcal{G}_l\rho_t \,\mathrm{d}n_t^l = \left(h_l(\rho_t) - \rho_t\right)\mathrm{d}n_t^l.$$

As it will be relevant throughout, it is worth noting here that the dynamics described by (QFE) inherently preserve purity of the (conditioned) quantum state due to the assumed complete and lossless observation of all measurement channels.

Lemma 7.1. *The set of pure quantum states*

$$B = \left\{\rho \in \mathbb{C}^{n\times n} : \rho^* = \rho, \mathrm{tr}\left(\rho\right) = \mathrm{tr}\left(\rho^2\right) = 1\right\}$$

is invariant under the dynamics (QFE).

Table 7.1: Common quantum feedback control problems as special cases of (QSOCP).

Control task	Stage cost $\ell(\rho_t, u_t)$	Terminal cost $m(\rho_T)$
State preparation	0	$1 - \psi_{\text{tar}}^* \rho_T \psi_{\text{tar}}$
State stabilization	$1 - \psi_{\text{tar}}^* \rho_t \psi_{\text{tar}}$	0
State purification	$\text{tr}\,(\rho_t^2)$	0
Entanglement creation[1]	0	$1 - \text{tr}\,(\tilde{\rho}_T^2)$

[1] $\tilde{\rho}_T$ denotes the reduced density matrix for a given subsystem [162]

Proof. Applying Itô's lemma to (QFE) shows that

$$\mathrm{d}\left[\text{tr}\,(\rho_t)\right] = \mathrm{d}\left[\text{tr}\left(\rho_t^2\right)\right] = 0$$

if $\rho_0 \in B$. Moreover, the right-hand side of (QFE) maps Hermitian matrices into Hermitian matrices. □

Given the described abstraction of feedback-controlled quantum systems under continuous observation, we now turn to the task of computing guaranteed lower bounds on the best attainable feedback control performance as characterized by the quantum stochastic optimal control problem

$$\begin{aligned} J^* = \inf_{u_t} \quad & \mathbb{E}\left[\int_0^T \ell(\rho_t, u_t)\,\mathrm{d}t + m(\rho_T)\right] \\ \text{s.t.} \quad & \rho_t \text{ satisfies (QFE) on } [0,T] \text{ with } \rho_0 \sim \nu_0, \\ & u_t \in U \text{ is non-anticipative on } [0,T]. \end{aligned} \tag{QSOCP}$$

Here, ℓ and m frame the control task by encoding the control performance in terms of an accumulating stage and terminal cost, respectively. Several common quantum feedback control tasks alongside the associated cost functions are listed in Table 7.1, showcasing the versatility of the formulation (QSOCP).

To ensure the well-posedness and tractability of the bounding problems for (QSOCP) as will be derived in the following section, we finally make some assumptions on the nature of the initial state of the quantum system, the representation of control constraints and cost functions, as well as the structure of photon counting measurements.

Assumption 7.1. *The initial distribution ν_0 of the quantum state satisfies* $\operatorname{supp}\nu_0 \subset B$*, i.e., the initial state is guaranteed to be pure albeit potentially uncertain.*

Assumption 7.2. *The set of admissible control actions U is a compact and basic closed semialgebraic set, i.e., there exist polynomials $\mathcal{U} = \{q_1, \dots, q_r\}$ such that $U = \{u \in \mathbb{R}^K : q(u) \geq 0, \forall q \in \mathcal{U}\}$ is compact. We refer to $\mathcal{U}$ as the control constraints.*

Assumption 7.3. *The cost functions ℓ and m are polynomials.*

Assumption 7.4. *The jump operators σ_l with $l \in \mathrm{PC}$ are such that $h_l(\rho)$ is a polynomial of degree at most one.*

Assumption 7.1 ensures that the quantum state, conditioned on the measurements, remains pure as per Lemma 7.1 and thus confined to a basic closed semialgebraic set. Assumptions 7.2 – 7.4 guarantee further that the dynamics (QFE) and control problem (QSOCP) are described entirely by polynomials and thus establish the basis for application of the semialgebraic and moment-sum-of-squares to derive tractable bounding problems.

It is worth emphasizing that, while Assumptions 7.1 – 7.3 are extremely mild and may even be relaxed (see Section 7.4), Assumption 7.4 is more limiting. Notwithstanding, it remains consistent with many practically relevant photon counting measurement setups; for example, measurements associated with unitary jump operators or measurements that cause a jump to the same quantum state independent of the state the photon emission occurred in.

7.3 A convex bounding approach

To construct computable bounds on the optimal value of (QSOCP), we follow the construction of Chapter 3. For sake of brevity, we focus here exclusively on the dual perspective of constructing (polynomial) subsolutions to the Hamilton-Jacobi-Bellman equations. To that end, we draw on the dynamic programming heuristic, which asserts that the value function associated with (QSOCP), i.e., the minimal cost-to-go

$$\begin{aligned} V(t,\rho) = \inf_{u_s} \quad & \mathbb{E}\left[\int_t^T \ell(\rho_s,u_s)\,ds + m(\rho_T)\right] \\ \text{s.t.} \quad & \rho_s \text{ satisfies (QFE) on } [t,T] \text{ with } \rho_t \sim \delta_\rho, \\ & u_s \in U \text{ is non-anticipative on } [t,T], \end{aligned} \tag{7.1}$$

satisfies the Hamilton-Jacobi-Bellman (HJB) equation [73]:

$$\begin{cases} \inf\limits_{u\in U} \mathcal{A}V(\cdot,\cdot,u) + \ell(\cdot,u) = 0 \text{ on } [0,T) \times B, \\ V(T,\cdot) = m \text{ on } B. \end{cases}$$

Here, $\mathcal{A}$ denotes to the infinitesimal generator [73] associated with the process (QFE); the action of $\mathcal{A}$ on a smooth observable $w \in \mathcal{C}^{1,2}([0,T]\times B)$ is given by

$$\begin{aligned} \mathcal{A}w(t,\rho,u) =& \frac{\partial w}{\partial t}(t,\rho) + \langle \tilde{\mathcal{L}}(u)\rho, \nabla_\rho w(t,\rho)\rangle \\ &+ \frac{1}{2}\sum_{l\in\mathrm{HD}} \langle \mathcal{G}_l\rho, \nabla_\rho^2 w(t,\rho)\,\mathcal{G}_l\rho\rangle \\ &+ \sum_{l\in\mathrm{PC}} \lambda_l(\rho)\left(w(t,h_l(\rho)) - w(t,\rho)\right), \end{aligned} \tag{7.2}$$

where $\tilde{\mathcal{L}}(u) = \mathcal{L}(u) - \sum_{l\in PC}\mathcal{L}_l$ is the effective drift operator associated with the control action u. Note that, due to Assumption 7.1, it suffices to solve the HJB equation on $[0,T]\times B$ as B is invariant under (QFE) as per Lemma 7.1 and thus constitutes the effective state space of the quantum system.

While the HJB equation is a nonlinear partial differential equation which is extremely difficult to solve even for low-dimensional systems, we can cast the search for a smooth HJB subsolution as a convex, albeit infinite-dimensional, optimization problem:

$$\begin{aligned} \sup_{w \in \mathcal{C}^{1,2}([0,T]\times B)} \quad & \int_B w(0,\cdot)\,\mathrm{d}\nu_0 && \text{(sub-HJB)} \\ \text{s.t.} \quad & \mathcal{A}w + \ell \geq 0 \text{ on } [0,T] \times B \times U, \\ & m - w(T,\cdot) \geq 0 \text{ on } B. \end{aligned}$$

Lemma 7.2. *Any feasible point w of* (sub-HJB) *underestimates the value function* (7.1) *on* $[0,T] \times B$ *and so* $\int_B w(0,\cdot)\,\mathrm{d}\nu_0$ *underestimates the best attainable control performance* J^*.

Proof. At $t = T$, feasibility of w implies that $w(T,\cdot) \leq V(T,\cdot) = m$ on B. Now consider any time $0 \leq t < T$, any state $\rho \in B$, and any feedback controller $\{u_s\}_{s\in[t,T]}$ admissible on $[t,T]$. By feasibility of w, it follows that for $\rho_t \sim \delta_\rho$,

$$\mathbb{E}\left[\int_t^T \ell(\rho_s, u_s)\,ds + m(\rho_T)\right] \geq \mathbb{E}\left[\int_t^T -\mathcal{A}w(s,\rho_s,u_s)\,ds + w(T,\rho_T)\right] = w(t,\rho),$$

where we used Dynkin's formula [74] in the last step. Finally, taking the infimum of the left-hand side over all admissible controllers establishes that $V(t,\rho) \geq w(t,\rho)$. □

The infinite-dimensional nature of (sub-HJB) renders its immediate practical value rather limited. We therefore proceed by constructing tractable finite dimensional restrictions of (sub-HJB) using the moment-sum-of-squares hierarchy. To that end, we restrict the optimization to polynomials of fixed maximum degree d instead of arbitrary smooth functions and further strengthen the non-negativity constraints in (sub-HJB) to sufficient sum-of-squares constraints. For this restriction to be well-posed and tractable, we must ensure that the left-hand side of the non-negativity constraints in (sub-HJB) are polynomials and that non-negativity is imposed on closed basic semialgebraic sets. The latter is guaranteed by Assumption 7.2 and the fact that B is basic closed semialgebraic. The former follows from Assumptions 7.3, 7.4, and the following result.

Lemma 7.3. *Under Assumption 7.4, the infinitesimal generator $\mathcal{A}$ [cf. Equation (7.2)] maps polynomials to polynomials.*

Proof. Let w be a polynomial. Then, $\frac{\partial w}{\partial t}, \nabla_\rho w$, and $\nabla^2_\rho w$ are componentwise polynomials as polynomials are closed under differentiation. Further note that $\tilde{\mathcal{L}}(u)\rho$, $\mathcal{G}_l\rho$, $\lambda_l(\rho)$, and by Assumption 7.4 also $h_l(\rho)$, are componentwise polynomials. Since polynomials are also closed under addition, multiplication, and composition, it thus follows that $\langle \tilde{\mathcal{L}}(u)\rho, \nabla_\rho w(t,\rho)\rangle$, $\langle \mathcal{G}_l\rho, \nabla^2_\rho w(t,\rho)\, \mathcal{G}_l\rho\rangle$, and $\lambda_l(\rho)\,(w(t,h_l(\rho)) - w(t,\rho))$ are polynomials and therefore so is $\mathcal{A}w$. □

The resultant sum-of-squares restriction of (sub-HJB) reads

$$\begin{aligned} J_d^* = \sup_{w_d \in \mathbb{R}_d[t,\rho]} \quad & \int_B w_d(0,\cdot)\,\mathrm{d}\nu_0 \\ \text{s.t.} \quad & \mathcal{A}w_d + \ell \in Q_{d+2}\left(\mathcal{T} \cup \mathcal{B} \cup \mathcal{U}\right), \\ & m - w_d(T,\cdot) \in Q_d\left(\mathcal{B}\right), \end{aligned} \tag{sos-HJB$_d$}$$

where we use $Q_d\left(\mathcal{S}\right)$ to refer to the bounded-degree quadratic modulus associated with a set of polynomials $\mathcal{S} = \{s_1, \dots, s_p\}$ (see Definition 2.15).

The set of control constraints $\mathcal{U}$ is defined as in Assumption 7.2. The sets $\mathcal{T}$ and $\mathcal{B}$ similarly denote collections of polynomial constraints whose non-negativity describe the sets $[0,T]$ and B, respectively. There is not a unique choice for these constraints, so we make the following choice which endows the sequence of restrictions (sos-HJB$_d$) with favorable convergence guarantees as shown later.

Assumption 7.5. *For the construction of* (sos-HJB$_d$) *we choose $\mathcal{T} = \{t, T - t\}$ so that $[0,T] = \{t \in \mathbb{R} : p(t) \geq 0,\ \forall p \in \mathcal{T}\}$. Moreover, to keep the computational burden associated with solving* (sos-HJB$_d$) *at a minimum, we explicitly eliminate the symmetry and unit trace constraints in B and represent density matrices only in terms of the remaining degrees of freedom, i.e.,* $\mathrm{Re}(\rho_{ii})$, *for $1 \leq i < n$, and ρ_{ij}, for $1 \leq i < j \leq n$. The set of such reduced density matrix representations will be denoted by D. In the following, we abuse notation and refer to reduced representations simply* *by $\rho \in D$. In these reduced coordinates, the set of pure density matrices B is given by a single polynomial equality constraint*

$$\operatorname{tr}\left(\rho^2\right) = \left(1 - \sum_{i=1}^{n-1} \operatorname{Re}(\rho_{ii})\right)^2 + \sum_{i=1}^{n-1} \operatorname{Re}(\rho_{ii})^2 + 2 \sum_{1 \le i < j \le n} |\rho_{ij}|^2 = 1. \tag{7.3}$$

Accordingly, we let $\mathcal{B} = \{1 - \operatorname{tr}(\rho^2), \operatorname{tr}(\rho^2) - 1\}$ so that $B = \{\rho \in D : p(\rho) \ge 0,\ \forall p \in \mathcal{B}\}$.

The hierarchical structure of Problem (sos-HJB$_d$) described in the following corollary is desirable from a practical point of view as it allows to trade off more computation for tighter bounds.

Corollary 7.1. *Any feasible point w_d of Problem* (sos-HJB$_d$) *underestimates the value function* (7.1) *on $[0,T] \times B$ and as such $\int_B w_d(0,\cdot)\,\mathrm{d}\nu_0$ underestimates J^*. Moreover, the optimal values J_d^* form a monotonically increasing sequence.*

Proof. Any feasible point of Problem (sos-HJB$_d$) is also feasible for (sub-HJB) so underestimates the value function by Lemma 7.2. Since $Q_d(\mathcal{S}) \subset Q_{d+1}(\mathcal{S})$ for any set of polynomials $\mathcal{S}$, it follows that (sos-HJB$_{d+1}$) is a relaxation of (sos-HJB$_d$) and hence $J_{d+1}^* \ge J_d^*$. □

A natural question that arises from Corollary 7.1 is if the bound J_d^* approaches the true optimal value J^* as the degree d is increased. In the following, we make a first step toward analyzing this convergence question. Specifically, we prove convergence whenever (QSOCP) admits a smooth value function and the control constraints satisfy Putinar's condition (see Definition 2.16).

First, we observe that the polynomials that frame Problem (sos-HJB$_d$) naturally satisfy Putinar's condition as long as the control constraints do.

Lemma 7.4. *The set $\mathcal{B}$ as defined in Assumption 7.5 satisfies Putinar's condition. If further the set of control constraints $\mathcal{U}$ satisfies Putinar's condition, then so does the set $\mathcal{T} \cup \mathcal{B} \cup \mathcal{U}$.*

Proof. From the description in Assumption 7.5, it is easily verified that $\mathcal{B}$ satisfies Putinar's condition since Equation (7.3) yields for $\rho \in D$ that

$$1 - \sum_{i=1}^{n-1} \mathrm{Re}(\rho_{ii})^2 - \sum_{1 \leq i < j \leq n} |\rho_{ij}|^2 = \left(1 - \sum_{i=1}^{n-1} \mathrm{Re}(\rho_{ii})\right)^2 + \sum_{1 \leq j < j \leq n} |\rho_{ij}|^2 + 1 - \mathrm{tr}\left(\rho^2\right).$$

The right-hand side of the relation above is clearly an element of $Q_2(\mathcal{B})$ as the first two terms are sums of squares and the last term $1 - \mathrm{tr}(\rho^2)$ is an element of $\mathcal{B}$. Further, $\mathcal{T}$ satisfies Putinar's condition as it is a set of degree one polynomials defining a bounded polyhedron [42]. Finally note that $a \in Q_d(\mathcal{T})$, $b \in Q_d(\mathcal{B})$, $c \in Q_d(\mathcal{U})$ implies that $a + b + c \in Q_d(\mathcal{T} \cup \mathcal{B} \cup \mathcal{U})$ as $\mathcal{T}$, $\mathcal{B}$, and $\mathcal{U}$ are comprised of polynomials in distinct variables. The conclusion follows. □

With this in hand, the convergence of the bounds furnished by (sos-HJB$_d$) can be established by application of Putinar's Positivstellensatz [40] according to the following theorem.

Theorem 7.1. *If the value function* (7.1) *is* $\mathcal{C}^{1,2}([0,T] \times B)$ *and the set of control constraints* $\mathcal{U}$ *satisfies Putinar's condition, then* $J_d^* \uparrow J^*$.

Proof. Let $\epsilon > 0$ and recall that on a compact set any continuously differentiable function and its (partial) derivatives can be approximated uniformly by a polynomial and its derivatives [23]. Therefore, there exists a polynomial w such that

$$\|V - w\|_\infty, \|\mathcal{A}V - \mathcal{A}w\|_\infty < \epsilon,$$

where $\|\cdot\|_\infty$ refers to the sup norm on the respective the domains, $[0,T] \times B$ and $[0,T] \times B \times U$. Under the assumed smoothness of the value function V, it is well-known that V satisfies the HJB equation (see e.g. [73, Thm. 3.1]) and thus in particular it holds that

$$\mathcal{A}V + \ell \geq 0 \text{ on } [0,T] \times B \times U,$$
$$m - V(T,\cdot) \geq 0 \text{ on } B.$$

Now consider $\hat{w} = w + 2\epsilon(t - T - 1)$ and note that, by construction, $\mathcal{A}\hat{w} = \mathcal{A}w + 2\epsilon$ and $\hat{w}(T, \cdot) = w(T, \cdot) - 2\epsilon$. It follows that

$$\begin{aligned} &\mathcal{A}\hat{w} + \ell \geq \mathcal{A}V + \ell + \epsilon > 0 \text{ on } [0, T] \times B \times U, \\ &m - \hat{w}(T, \cdot) \geq m - V(T, \cdot) + \epsilon > 0 \text{ on } B. \end{aligned}$$

By Lemma 7.4, Putinar's Positivstellensatz [40, Lemma 4.1] therefore guarantees for sufficiently large d that $\mathcal{A}\hat{w} + \ell \in Q_{d+2}(\mathcal{T} \cup \mathcal{B} \cup \mathcal{U})$ and likewise $m - \hat{w}(T, \cdot) \in Q_d(\mathcal{B})$ such that $\hat{w}$ is feasible for (sos-HJB$_d$). The result follows by noting that

$$\begin{aligned} J^* - J_d^* &\leq \int_B |V(0, \cdot) - \hat{w}(0, \cdot)| \, \mathrm{d}\nu_0 \\ &\leq \max_{\rho \in B} |V(0, \rho) - w(0, \rho)| + |2\epsilon(T + 1)| \\ &< (2T + 3)\epsilon. \end{aligned}$$

□

Remark 7.1. *It should be emphasized that the assumption that* (QSOCP) *admits a smooth value function is by no means weak and, even if satisfied, generally not easily verified. Theorem 7.1 is only a first step toward establishing a formal basis for our empirical observation that the bounds in fact often do appear tight. Related work [5, 6, 46] suggests that the conditions under which convergence can be guaranteed may be substantially relaxed.*

We conclude this section with a few remarks about the practicality of the derived bounding problems. First, the local occupation measure framework proposed in Chapter 3 provides straightforward generalizations of the bounding problems constructed here. And although these generalizations add little conceptual depth from the perspective of quantum control, they boost the practicality of the general approach. By considering piecewise polynomial value function underapproximators, the local occupation measure framework provides more fine-grained control over the construction of tighter and numerically better conditioned bounding problems. Second, the sum-of-squares bounding problems (sos-HJB$_d$) (as well as their generalizations derived from the local occupation measure framework) are equivalent to finite semidefinite programs (SDP) [30, 44]. As such, they are readily solved by a range of powerful off-the-shelf available solvers [33–35, 163–165]. Finally, these SDPs can further be automatically constructed from symbolic representations of the underlying sum-of-squares programs by openly available optimization modeling tools [45, 77, 78, 166].

7.4 Extensions

7.4.1 Infinite horizon problems

While we detailed our analysis for the finite horizon problem (QSOCP), one can construct analogous bounding problems for (discounted) infinite horizon problems. Specifically, it suffices to note that for control objectives of the form

$$\mathbb{E}\left[\int_0^\infty e^{-\gamma t}\ell(\rho_t, u_t)\,\mathrm{d}t\right]$$

with discount rate $\gamma > 0$ a global value function underestimator $e^{-\gamma t}w$ is obtained from any smooth function $w \in \mathcal{C}^{1,2}([0,\infty) \times B)$ that satisfies

$$\mathcal{A}w - \gamma w + \ell \geq 0 \text{ on } [0,\infty) \times B \times U.$$

This conclusion follows by analogous arguments as in the proof of Lemma 7.2 after noting that

$$\mathcal{A}(e^{-\gamma t}w) = e^{-\gamma t}\left(\mathcal{A}w - \gamma w\right).$$

It is thus straightforward to adapt the hierarchy of the bounding problems (sos-HJB$_d$) to furnish valid bounds on the best attainable control performance for discounted infinite horizon problems.

7.4.2 Decoherence and mixed initial states

The relaxation of Assumption 7.1 to mixed initial quantum states or the consideration of unobserved decay channels is possible at the expense of introducing additional conservatism. For initially mixed quantum states or unobserved decay channels, the purity of the quantum state can no longer be guaranteed. (sub-HJB), however, still characterizes valid bounds when the constraints are enforced on the set of all mixed quantum states

$$\bar{B} = \left\{\rho \in \mathbb{C}^{n\times n} : \rho = \rho^*, \operatorname{tr}(\rho) = 1, \psi^* \rho \psi \geq 0 \ \forall \psi \in \mathbb{C}^n \right\}.$$

As $\bar{B}$ can be defined by a finite set of polynomial inequality constraints, the resulting problem in principle also admits valid sum-of-squares restrictions akin to (sos-HJB$_d$). However, these restrictions are typically deemed impractical. Representing the positivity requirement for density matrices in terms of polynomial inequalities necessitates enforcing the non-negativity of all its principal minors, as dictated by Sylvester's criterion. The large number and high degree of the associated polynomials render the corresponding sum-of-squares restrictions computationally cumbersome. A more practical approach is to impose the constraints in (sub-HJB) instead on a simpler closed basic semialgebraic overapproximation of $\bar{B}$; for example,

$$\tilde{B} = \left\{\rho \in \mathbb{C}^{n\times n} : \rho^* = \rho, \operatorname{tr}(\rho) = 1, \operatorname{tr}\left(\rho^2\right) \leq 1 \right\}.$$

This modification potentially introduces additional conservatism as $\tilde{B}$ is a strict superset of $\bar{B}$ but leads to practical sum-of-squares restrictions. Moreover, the restriction $\tilde{B}$ can be refined flexibly by adding additional constraints of the form $\psi^* \rho \psi \geq 0$ for any number of fixed state vectors $\psi \in \mathbb{C}^n$.

7.4.3 Imperfect measurements

So far we have assumed lossless or perfect homodyne and photon counting measurements. In practice, however, various factors lead to imperfect detection [103]. While in most such cases the bounds furnished by (sos-HJB$_d$) will remain valid due to the simple fact that additional losses typically lead to more stringent performance limitations, it is often of interest to quantify explicitly the limitations induced by measurement imperfections. The presented bounding method extends naturally to this task. To that end, it is necessary to account for measurement inefficiencies in (QFE). For homodyne measurements with efficiency $\eta \in [0,1]$, the innovation operator in (QFE) acts according to

$$\mathcal{G}_l \rho = \eta(\sigma_l \rho_t + \rho \sigma_l^* - \operatorname{tr}(\sigma_l \rho_t + \rho \sigma_l^*)\, \rho),$$

and for inefficient photon detection, the drift operator $\mathcal{L}_l$ must be modified to

$$\mathcal{L}_l \rho = \eta \left(\sigma_l \rho \sigma_l^* - \operatorname{tr}\left(\sigma_l \rho \sigma_l^*\right) \rho_t\right)$$

alongside the arrival rate of the driving Poisson counter which decays to $\lambda_l(\rho) = \eta \operatorname{tr}(\sigma_l \rho \sigma_l^*)$ [103, Section 4.8]. It is easily observed that under these modifications the conclusion of Lemma 7.3 remains valid. Given imperfect measurements ($\eta < 1$), however, the conclusion of Lemma 7.1 no longer holds and initially pure quantum states no longer remain pure as they evolve under the dynamics (QFE). As a consequence, the set of pure quantum states B in (sub-HJB) must be replaced by a basic semialgebraic overapproximation of the set of mixed states as discussed in Section 7.4.2.

7.4.4 Extraction of heuristic controllers

Bounds computed via (sos-HJB$_d$) may be used to verify the near-optimality of any given control policy. As such, the proposed bounding method complements heuristic approaches for the design of control policies. The solution of (sos-HJB$_d$), however, can also be used to inform controller design directly. At the optimal point of (sos-HJB$_d$), the optimization variable w_d approximates by construction the best possible polynomial underapproximator of the value function. Thus, it is reasonable to treat w_d as a proxy for the value function [47, 53] and construct a heuristic controller by greedily descending on w_d along the trajectory, i.e.,

$$u_t^*(\rho) \in \underset{u \in U}{\arg\min} \; \mathcal{A} w_d(t, \rho, u) + \ell(\rho, u). \tag{7.4}$$

The above requires minimization of a polynomial over the set of admissible control actions U, which is only expected to be tractable in the case of one or few control inputs. Otherwise, we argue that the inherently heuristic nature of this construction may justify the use of fast heuristics to find local or approximate minimizers instead, for example by relying on recent advances in machine learning [167, 168].

7.5 Example

We finally demonstrate the utility of the proposed bounding method for the problem of stabilizing the state of a qubit in a cavity [139]. Figure 7-1 shows a schematic of the associated control loop.

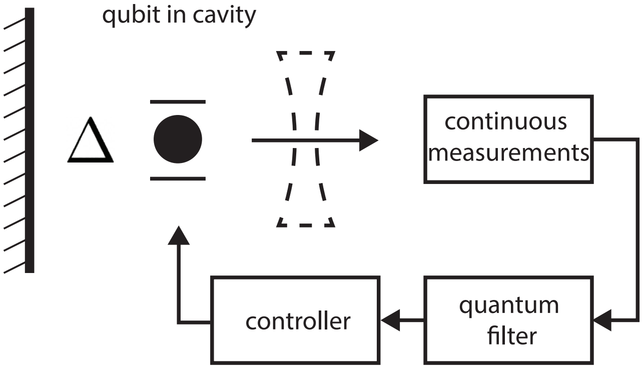


Figure 7-1: Closed-loop controlled qubit in a cavity subjected to continuous measurements.

The Hamiltonian of the qubit is given by

$$H(u) = \frac{\Delta}{2}\sigma_z + \frac{\Omega}{2}u\sigma_x,$$

where

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ and } \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

denote the Pauli matrices. To enable feedback, we assume that the qubit is subjected to continuous measurements associated with the jump operator

$$\sigma = \kappa \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Note that such a measurement conforms with Assumption 7.4. The parameters are chosen as $\Delta = \Omega = 5$ and $\kappa = 1$; the set of admissible control actions is $U = [-1, 1]$. In the following, we consider a realization of the measurements through homodyne detection and photon counting setups and contrast the two.

The objective of the control problem is to stabilize the excited state $\psi_{\text{tar}} = [1\ 0]^*$ with minimal expected infidelity (viz. maximum expected fidelity)

$$\mathbb{E}\left[\int_0^T 1 - \psi_{\text{tar}}^* \rho_t \psi_{\text{tar}} \,\mathrm{d}t\right].$$

The qubit is assumed to reside initially in its ground state $\psi_0 = [0\ 1]^*$ and the distribution of the initial state in density matrix form is hence given by $\nu_0 = \delta_{\psi_0 \psi_0^*}$.

For the implementation of the presented bounding method, we relied on the optimization ecosystem in Julia. Specifically, we used the packages `MarkovBounds.jl` [45] and `SumOfSquares.jl` [77] to assemble the bounding problems and pass the resultant SDPs via the `MathOptInterface` [78] to Mosek v10 [33]. All computations were performed on a MacBook M1 Pro with 16GB unified memory.

Table 7.2 summarizes upper bounds for the maximal average fidelity attainable with both measurement setups as obtained by solving the bounding problems (sos-HJB$_d$) for increasing degree d. The bounds are clearly non-trivial and suggest to be informative even for moderate degrees. To emphasize this point, we further constructed heuristic controllers for both measurement setups from the optimal solution

Table 7.2: Performance bounds for feedback controlled qubit in a cavity subjected to homodyne detection and photon counting measurements.

Homodyne detection

Degree d	Fidelity bound	Computational time [s]
2	0.8502	0.008
4	0.8111	0.078
6	0.7973	0.64
8	0.7893	5.0
10	0.7856	27.9

Best known fidelity: 0.7750

Photon counting

Degree d	Fidelity bound	Computational time [s]
2	0.9602	0.0043
4	0.7497	0.031
6	0.7153	0.180
8	0.6902	1.67
10	0.6798	14.9

Best known fidelity: 0.6547

of (sos-HJB$_4$) as described in Section 7.4.4. Their empirical performance serves as an achievable lower bound for the best attainable mean fidelity. Figure 7-2 shows the mean fidelity and noise level attained by both controllers, alongside a visualization of the associated control policy as a function of the polarisations of the quantum state. The controllers achieve mean fidelities of 77.50 % and 65.47 % (ensemble averages over 10,000 sample trajectories) for the homodyne detection and photon counting setup, respectively. Against the backdrop of the computed bounds, the controllers are thus certifiably near-optimal, showcasing the practical utility of the proposed bounding method.

An interesting spillover of this example is that, barring (highly unlikely) major statistical errors in the estimates of the fidelity attained by the heuristic controllers, this case study constitutes a computational proof that under the assumed circumstances a homodyne detection setup allows for strictly and significantly greater average mean

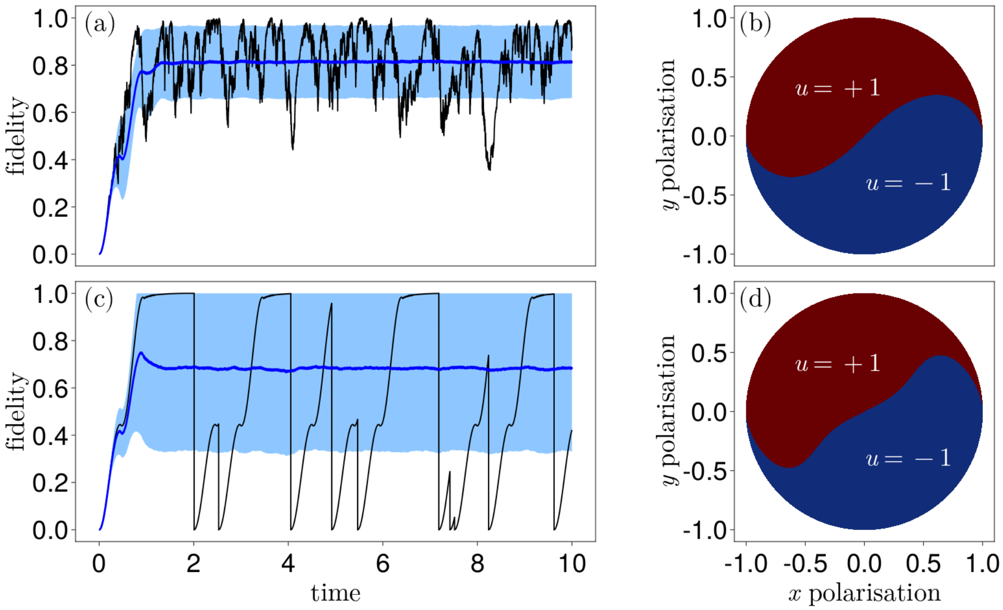


Figure 7-2: Fidelity of the closed-loop controlled qubit alongside a visualization of the heuristic controller at $t = 5$ in the x-y plane of the Bloch sphere for homodyne (a,b) and photon counting measurements (c,d). Mean trace and standard deviation band are shown in blue. A representative sample path is shown in black.

fidelity than photon counting. This demonstrates that the proposed bounding method may indeed provide relevant insights for the design of quantum devices.

7.6 Conclusion

We have showed in Part II of this thesis that viewing controlled quantum systems through the lens of occupation measures and combining this perspective with moment-sum-of-squares techniques yields a powerful approach for revealing the limits of quantum control at large. For the case of open-loop quantum control, we have established mild conditions under which this approach is capable of mapping out the performance boundary for various common open-loop quantum control tasks to arbitrary precision. By combining it further with quantum filtering theory, we have presented the first ever technique for bounding the best attainable feedback control performance for a broad class of quantum systems subjected to continuous measurements. Convergence of the bound sequence to the true optimal feedback control performance was proved under technical conditions.

Our theoretical considerations are complemented by demonstrations of the proposed bounding methods on a range of quantum systems and control tasks, both open- and closed-loop. In all cases, we have found that the established bounds are tight or nearly so. Moreover, for open-loop controlled quantum systems, the obtained bounds were found to provide distinct improvements over performance limits implied by quantum speed limits such as the celebrated Mandelstam-Tamm [19] and Margolus-Levitin [20] bounds as well as other more recent algorithmic proposals [21].

Between the established theoretical guarantees and the empirically good performance, we argue that the proposed bounding methods can have relevant implications for the design of the next generation of quantum devices. On the one hand, they provide access to heuristic controllers alongside performance bounds which may guide controller design or certify the optimality of a given control policy. On the other hand, the bounds may serve as witnesses of fundamental limitations and so inform design decisions at an early stage.

Part III

Dynamical low-rank approximation

Chapter 8

Robust dynamical low-rank approximation for nonlinear matrix differential equations

8.1 Introduction

We consider the task of approximating the solution of intractably large matrix-valued initial value problems (IVPs),

$$\begin{cases} \dot{X}(t) = F(X(t), t), \quad t \in [0, T] \\ X(0) = X_0 \in \mathbb{R}^{n\times m} \end{cases}, \tag{8.1}$$

with a nonlinear vector field $F : \mathbb{R}^{n\times m} \times [0, T] \to \mathbb{R}^{n\times m}$. Many tasks in computational and engineering science involve, boil down to, are well-approximated by, or can be recast as such problems; for instance, the solution of high-dimensional ordinary (ODE) and partial differential equation (PDE) models [169–173], forward propagation of uncertainties through such models [174–178], filtering and smoothing problems in high-dimensions [179, 180], the training of machine learning models [181, 182], compression of matrix-valued data streams [183], and sensitivity analysis [184, 185] to list only a few.

Dynamical low-rank approximation (DLRA) [186] offers a systematic framework for computing a rank-r approximation for the solution of the IVP (8.1). To that end, DLRA seeks the solution of the auxiliary problem

$$\begin{cases} \dot{\hat{X}}(t) = \Pi_{\mathcal{T}_{\mathcal{M}_r}(\hat{X}(t))} F(\hat{X}(t), t), \quad t \in [0, T] \\ \hat{X}(0) = \Pi_{\mathcal{M}_r} X_0, \end{cases} \tag{8.2}$$

where $\mathcal{M}_r$ refers to the manifold of rank-r $n \times m$ matrices, $\mathcal{T}_{\mathcal{M}_r}(Y)$ to its tangent space at the point $Y \in \mathcal{M}_r$, and $\Pi_{\mathcal{X}}$ to orthogonal projection onto the manifold $\mathcal{X}$. Intuitively, Equation (8.2) greedily minimizes the error accumulation rate incurred by approximating the solution of the IVP (8.1) in the low-rank manifold – an approach also known as the Dirac-Frenkel variational principle [187, 188]. By construction, the solution $\hat{X}$ of the IVP (8.2) can therefore be viewed as a rank-r approximation to the full solution X of the original IVP (8.1).

Due to the reduced complexity of representing and processing low-rank matrices, DLRA often yields tractable approximations when solving the full problem is out of computational reach. The greatest potential for recovering tractability is realized when the vector field F (and its projection onto the tangent bundle of the low-rank manifold) can be evaluated more efficiently by exploiting low-rank structure in its argument. A common setting where this applies is when F leads to a structurally bounded rank growth; that is, when for all t and $X = USV^\top \in \mathcal{M}_r$, $F(X, t)$ admits an explicit rank-$r_F \ll m, n$ factorization that can be computed efficiently from the individual factors of X, in particular without computing full ambient representations of the matrices X and $F(X, t)$. In practice, this situation is encountered for example when F is a polynomial of moderate degree. Accordingly, DLRA has been widely demonstrated to enable dramatic speedups for the approximate solution of IVPs of the form (8.1) where F is linear, bilinear or quadratic [169, 175, 176, 178]. When F does not lead to structurally bounded rank growth or the rank growth is too steep, the computational advantages of DLRA, beyond a reduction in memory footprint for storing and processing the approximate solution, are less clear. Moreover, even when

F admits an explicit and efficiently computable low-rank factorization, it typically remains a highly intrusive, error-prone process to exploit this structure in numerical DLRA schemes.

For the approximate propagation of parametric uncertainties through nonlinear PDEs via DLRA, Naderi and Babaee [22] have recently proposed a remedy to the mentioned complications arising from nonlinear vector fields. In broad strokes, they combine sparse approximation and interpolatory projectors [189–192] to devise a DLRA scheme that approximates the evaluation of the vector field F in each integration step from a small subset of strategically selected rows and columns. The resultant approximation and its projection onto the tangent bundle of the low-rank manifold can then be performed efficiently without implementation of tailored, problem-specific routines, even in the presence of nonlinearities.

In this chapter, we put the on-the-fly sparse approximation heuristic of Naderi and Babaee [22] on a more general footing, showing that it enables with minor modifications the efficient DLRA for matrix-valued IVPs of the form (8.1) with local vector fields. We show further that this heuristic composes naturally with a rich set of robust geometric integration routines for DLRA, yielding distinctly improved robustness properties in the presence of small singular values of the low-rank approximation. Lastly, we present `LowRankIntegrators.jl` – a performant, yet high-level package for DLRA in the Julia programming language [193]. As part of a rich feature set, `LowRankIntegrators.jl` notably exploits the composable nature of on-the-fly sparse approximation and robust DLRA integrators to enable efficient DLRA for generic IVPs of the form (8.1) with minimal intrusion. The minimal required user input are row-, column- and element-wise evaluation oracles for the vector field F.

8.2 Notation & terminology

Throughout this chapter we rely on the following notational conventions, terminology, and blanket assumptions.

Basic linear algebra – We assume that all matrices and vectors are real, and unless specified otherwise, matrices will be assumed to be of dimension $n \times m$. Accordingly, we default to considering approximation in the manifold of real $n \times m$ rank-r matrices, which we denote by $\mathcal{M}_r = \{\hat{X} \in \mathbb{R}^{n \times m} : \operatorname{rank} \hat{X} = r\}$. We loosely refer to $\mathcal{M}_r$ and elements $\hat{X} \in \mathcal{M}_r$ as the low-rank manifold and low-rank matrices, respectively. The tangent space of $\mathcal{M}_r$ at a point $\hat{X} \in \mathcal{M}_r$ will be denoted by $\mathcal{T}_{\mathcal{M}_r}(\hat{X})$ and the tangent bundle by $T_{\mathcal{M}_r} = \{(\hat{X}, V) : \hat{X} \in \mathcal{M}_r, V \in \mathcal{T}_{\mathcal{M}_r}(\hat{X})\}$. The real Stiefel manifold of (semi-)orthogonal $n \times k$ matrices will be denoted by $\mathcal{V}_{n,k} = \{U \in \mathbb{R}^{n \times k} : U^\top U = I\}$, where I refers to the identity matrix of appropriate dimension. Lastly, the euclidean projection onto a set A will be denoted by Π_A.

Low-rank representations – For considerations of computational efficiency, it is relevant to distinguish between different representations of low-rank matrices. Throughout, we refer to the full representation of a low-rank matrix $\hat{X} \in \mathcal{M}_r$ in terms of its $n \times m$ individual entries as the ambient or full matrix representation. When $\hat{X}$ is instead represented in terms of the factors $U \in \mathcal{V}_{n,r}$, $V \in \mathcal{V}_{m,r}$, and $S \in \mathbb{R}^{r \times r}$, such that $\hat{X} = USV^\top$ we refer to it as an SVD-like factorization. Similarly, when the factors are unstructured, i.e., $U \in \mathbb{R}^{n \times r}$, $V \in \mathbb{R}^{m \times r}$ and $S \in \mathbb{R}^{r \times r}$, we simply refer to a low-rank factorization of $\hat{X}$ without further qualification.

8.3 Numerical methods for dynamical low-rank approximation

Numerical schemes for DLRA fall broadly into one of two categories. The first category of methods relies on solving the IVP (8.2) in intrinsic coordinates of the low-rank manifold with standard ODE integrators [174, 186, 194]. To that end, $\hat{X}(t)$ is replaced by a structurally fixed low-rank factorization and ordinary evolution equations for the individual factors are derived from the Dirac-Frenkel variational principle. For instance, when expressing $\hat{X}(t)$ with an SVD-like factorization $\hat{X}(t) = U(t)S(t)V(t)^\top$ with time-dependent factors $U(t) \in \mathcal{V}_{n,r}$, $V(t) \in \mathcal{V}_{m,r}$, and $S(t) \in \mathbb{R}^{r \times r}$ and imposing the gauge conditions $\dot{U}(t)^\top U(t) = 0$ and $\dot{V}(t)^\top V(t) = 0$ to preserve semiorthogonality of $U(t)$ and $V(t)$ over time, it is easily verified that the IVP (8.2) is equivalent to the ordinary IVP [186].

$$\begin{cases} \begin{bmatrix} \dot{U}(t) \\ \dot{S}(t) \\ \dot{V}(t) \end{bmatrix} = \begin{bmatrix} (I - U(t)U(t)^\top)F(\hat{X}(t),t)V(t)S^{-1}(t) \\ U(t)^\top F(\hat{X}(t),t)V(t) \\ (I - V(t)V(t)^\top)F(\hat{X}(t),t)^\top U(t)S^{-\top}(t) \end{bmatrix}, \quad t \in [0,T] \\ \hat{X}(0) = U(0)S(0)V(0)^\top = \Pi_{\mathcal{M}_r} X_0 \text{ such that } U(0) \in \mathcal{V}_{n,r}, V(0) \in \mathcal{V}_{m,r} \end{cases} \tag{8.3}$$

Analogous IVPs are obtained for structurally different factorizations such as $\hat{X} = UZ^\top$ with $U \in \mathcal{V}_{n,r}$ and $Z \in \mathbb{R}^{m \times r}$ [174], or when a weighted inner product is used to characterize orthogonality and projections [174, 194]. The resultant IVPs can in principle be solved with standard ODE integrators due to the intrinsic parameterization of the low-rank manifold. While conceptually straightforward, this approach is recognized as coming with various practical limitations. Most notably, it leads to stringent stepsize restrictions in the presence of small singular values of $\hat{X}$ [195] – a situation that is closely tied to achieving an accurate approximation of the true solution and so can rarely be avoided in practice. The stepsize restrictions originate from the stiffness inherent to the IVP (8.2) in the presence of small singular values: the Lipschitz constant of the vector-field F is amplified by the inverse of the smallest non-zero singular value of $\hat{X}$ [186, Lemma 4.2].

Remarkably, there still exists a class of time-stepping methods for the IVP (8.2) with accuracy guarantees that are independent of the magnitude of the smallest singular value. We refer to these methods here loosely as robust geometric integrators. Methods from this class avoid a fixed parameterization of the low-rank manifold. Instead, they take a fundamentally geometric approach to discretize the IVP (8.2), alternating between applying explicit time-stepping routines in the ambient space and retracting back to the low-rank manifold [196, 197]. Figure 8-1 shows an illustration of the explicit Euler method,

$$\hat{X}_{k+1} = R_{\hat{X}_k}\left(h\Pi_{\mathcal{T}_{\mathcal{M}_r}(\hat{X}_k)}F(\hat{X}_k,t_k)\right), \tag{8.4}$$

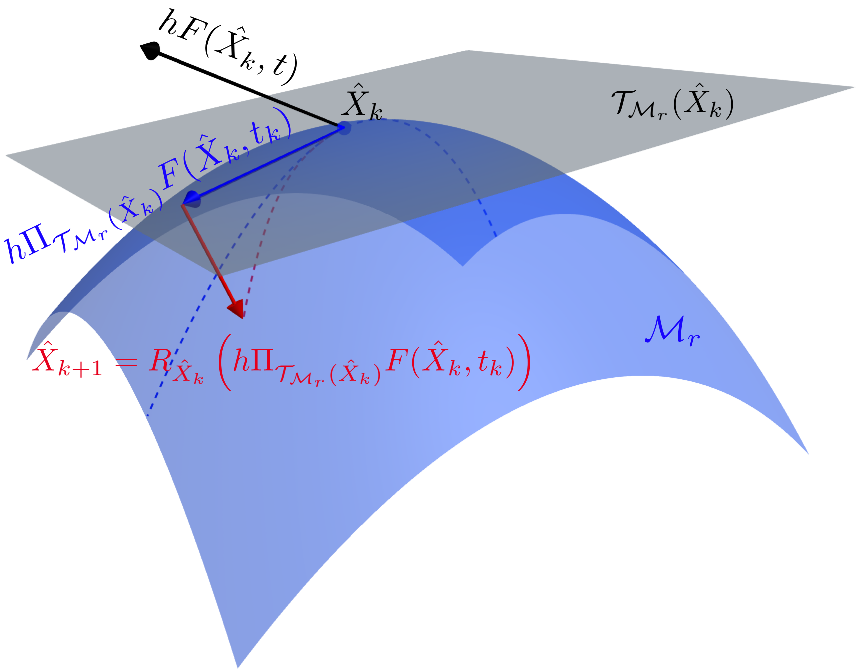

$hF(\hat{X}_k, t)$
$\hat{X}_k$
$\mathcal{T}_{\mathcal{M}_r}(\hat{X}_k)$
$h\Pi_{\mathcal{T}_{\mathcal{M}_r}(\hat{X}_k)}F(\hat{X}_k, t_k)$
$\hat{X}_{k+1} = R_{\hat{X}_k}\left(h\Pi_{\mathcal{T}_{\mathcal{M}_r}(\hat{X}_k)}F(\hat{X}_k, t_k)\right)$
$\mathcal{M}_r$

Figure 8-1: Projected Euler step for integration along the low-rank manifold.

in this framework. Here, the map $T_{\mathcal{M}_r} \ni (Y, V) \mapsto R_Y(V) \in \mathcal{M}_r$ refers to a retraction to the low-rank manifold $\mathcal{M}_r$. Formally, a retraction R_Y is defined so that for any $Y \in \mathcal{M}_r$ and $V \in \mathcal{T}_{\mathcal{M}_r}(Y)$, the function $\gamma(t) = R_Y(tV)$ is (for sufficiently small t) a well-defined, smooth curve in $\mathcal{M}_r$ satisfying $\gamma(0) = Y$ and $\dot{\gamma}(0) = V$ [198]. These conditions ensure that the discretization (8.4) is consistent with (8.2) when passing to the limit $h \to 0$ [199, Theorem 1]. A plethora of retractions for the low-rank manifold exist [177, 197, 200, 201, for example] and each generates a consistent integrator by Equation (8.4) [197, 199]. Furthermore, many computationally tractable and practically common retractions are *extended*; that is, they map not only elements from the tangent bundle of the low-rank manifold consistently to smooth curves in the low-rank manifold but any sufficiently small element from the ambient space. The most commonly used extended retraction for the low-rank manifold is the projective retraction

$$\bar{R}_Y^{\perp}(V) = \operatorname*{arg\,min}_{Z \in \mathcal{M}_r} \|Y + V - Z\|_F.$$

(We use equality above because the minimizer is unique for sufficiently small V [200, Lemma 3.1].) More generally, an extended retraction $\bar{R}_Y : \mathbb{R}^{m\times n} \to \mathcal{M}_r$ satisfies that $\gamma(t) = \bar{R}_Y(tV)$ is a smooth curve in $\mathcal{M}_r$ with $\gamma(0) = Y$ and $\dot{\gamma}(0) = \Pi_{\mathcal{T}_{\mathcal{M}_r}(Y)}V$ for any $Y \in \mathcal{M}_r$ and $V \in \mathbb{R}^{n\times m}$ [202]. In other words, the recursion

$$X_{k+1} = \bar{R}_{\hat{X}_k}(hF(\hat{X}_k, t_k)) \tag{8.5}$$

is also a consistent discretization for the IVP (8.2). Recursion (8.5) generalizes readily to higher-order explicit projected Runge-Kutta (PRK) schemes [196]; see Algorithm 1 for the skeleton of practical PRK schemes.

Algorithm 1 Explicit s-step projected Runge-Kutta integrator [196]

```
Input: Current state $\hat{X}_0 \in \mathcal{M}_r$, extended low-rank retraction $\bar{R}$, current time $t$,
time step $h$, Butcher tableau $(a, b, c)$.
Compute
    $Y_1 = \hat{X}_0$
    $K_1 = \Pi_{\mathcal{T}_{\mathcal{M}_r}(Y_1)}F(Y_1, t)$
for $i = 2, \ldots, s$ do
    Compute
        $Y_i = \bar{R}_{\hat{X}_0}\left(h\sum_{j=1}^{i-1} a_{ij}K_j\right)$
        $K_i = \Pi_{\mathcal{T}_{\mathcal{M}_r}(Y_i)}F(Y_i, t + c_i h)$
end for
Compute
    $\hat{X}_1 = \bar{R}_{\hat{X}_0}\left(h\sum_{i=1}^{s} b_i K_i\right)$
Output: Approximation of evolved state $\hat{X}_1 \in \mathcal{M}_r$
```

8.4 Computational cost of dynamical low-rank approximation

It is clear that numerically approximating the low-rank solution of the IVP (8.2) yields a reduced memory footprint relative to computing and storing the solution of its higher-dimensional counterpart (8.1). In this section, we discuss conditions under which DLRA also promises a asymptotically lower computational load as the side dimensions of the matrix grow. To that end, we analyze the cost associated with a projected Euler step (8.4) in two scenarios: First, we consider the case when $F(\hat{X}, t)$ has no exploitable structure and must be evaluated element-wise from an ambient $n \times m$ matrix representation of $\hat{X} \in \mathcal{M}_r$. Second, we assume that F leads to structurally bounded rank growth; that is, $F(\hat{X}, t)$ admits an explicit rank-r_F (with $r_F \ll m, n$) factorization that, given a low-rank factorization of the input $\hat{X} \in \mathcal{M}_r$, can be evaluated in $O(r_F(m + n))$ operations. Throughout, we assume that $\hat{X}$ is available as an SVD-like rank-r factorization $\hat{X} = USV^\top$. This is in line with most numerical DLRA schemes. The analysis is presented for the projected Euler method (8.4) for sake of simplicity. It is readily extended to more complicated geometric time-stepping routines and yields identical conclusions.

The projected Euler step (8.4) involves four distinct computational substeps: evaluation of the vector field $F(\hat{X}, t)$ and its tangent space projection, an update step in the ambient space, and retraction to the low-rank manifold. When F has no notable structure, it must be evaluated element-wise from an ambient matrix representation of $\hat{X}$, rendering the first step most expensive. Assuming that every element of F can be evaluated in constant time[1], $O(mn)$ and $O(mnr)$ operations are required for the evaluation of $F(\hat{X}, t)$ and the computation of the ambient representation of $\hat{X}$ from its low-rank factorization $USV^\top$, respectively. Since further the result of the evaluation of $F(\hat{X}, t)$ is an unstructured matrix $Z \in \mathbb{R}^{n \times m}$, its tangent space projection [186,

[1] While this assumption is not always valid, it generally holds for vector fields generated by PDEs or more broadly vector fields with local structure.

Lemma 4.1]

$$
\begin{aligned}
\Pi_{\mathcal{T}_{\mathcal{M}_r}(\hat{X})} Z &= UU^\top Z + ZVV^\top - UU^\top ZVV^\top \\
&= \begin{bmatrix} U & ZV \end{bmatrix} \begin{bmatrix} I & -U^\top ZV \\ 0 & I \end{bmatrix} \begin{bmatrix} Z^\top U & V \end{bmatrix}^\top
\end{aligned} \tag{8.6}
$$

requires additional $O(mnr)$ operations. The expression (8.6) for the tangent space projection underlines further that all tangent space elements have at most rank $2r$. It follows that the update step in the ambient space as well as the retraction to the low-rank manifold can be computed relatively cheaply as long as $r \ll n, m$; using the truncated SVD for example, computing the projective retraction requires only $O(r^2(m+n))$ operations [203]. In contrast, a classical Euler step for solving the original IVP (8.1) requires $O(mn)$ operations. Leaving memory requirements aside, the computational cost for DLRA (at fixed rank r) scales therefore asymptotically equivalently with the side dimensions of the matrix as computing the full solution. The vector field evaluation remains the driving cost. The situation is different, however, when $F(\hat{X}, t)$ admits an explicit rank-r_F factorization that can be evaluated in $O(r_F(n+m))$ operations from the low-rank factorization of $\hat{X}$. Not only does this lead to asymptotically reduced costs for evaluating $F(\hat{X}, t)$ but also yields a low-rank factorization as a result, which can be projected into the tangent space at a reduced cost. In light of Equation (8.6), the cost for the tangent space projection of a rank-r_F factorization of $F(\hat{X}, t)$ is $O(\max\{r_F r, r^2\}(n+m))$ operations. In this situation, the solution of the DLRA problem (8.2) is overall asymptotically cheaper ($O(\max\{r_F r, r^2\}(n+m))$ operations) than solving the original IVP (8.1) ($O(mn)$ operations).

8.5 On-the-fly sparse approximation

We have established in the previous section that DLRA is only expected to yield computational advantages beyond reduced memory requirements when the vector field F leads to structurally bounded rank growth. While this applies in some important cases, most notably when F is a polynomial of low degree and low-rank constant term[2], it does not hold in the presence of many commonly encountered nonlinearities; for example, element-wise exponential, trigonometric or logarithmic functions. Furthermore, computational routines to take advantage of structurally bounded rank growth in DLRA are, even if they exist, often not readily available. Instead, they must be implemented in an intrusive, often tedious and error-prone process. At the same time, a simple empirical observation inspires hope that the computational advantages of DLRA can be leveraged more easily and extended to more general vector fields: Even if the vector field F involves nonlinearities that do not lead to structurally bounded rank growth, it often admits a (locally) excellent low-rank approximation. This empirical observation has been leveraged with great success in the context of model order reduction [204]. In particular, variants of the empirical interpolation method (EIM) [205] and its discrete analog (DEIM) [189, 206] are utilized widely to determine structural low-rank approximation for nonlinear vector fields in the context of projective model order reduction [204]. The key advantage of DEIM-induced low-rank factorizations over other approximation schemes for this purpose is that the factorizations are constructed by combining sparse interpolation and projection. Put differently, the individual factors of the vector field approximation are computed from only a few strategically selected elements of the vector field. If the vector field has local structure, the computation of a full ambient representation of the input can then be avoided, rendering evaluation of the vector field approximation at low-rank inputs efficient. Bringing this approach to DLRA, however, comes with unique challenges. Specifically, DEIM schemes typically require an offline stage where input-output samples of the vector field are collected to determine which elements should best be interpolated and how. For typical applications of DLRA, however, such an offline stage is considered intractable. To circumvent this limitation in these situations, Naderi and Babaee [22] proposed a DLRA scheme for nonlinear stochastic PDEs that builds and adapts sparse approximations of the vector field without

[2] We say F is a polynomial when F is obtained from the composition of finitely many matrix additions as well as matrix and Hadamard products.

offline stage. Instead, they combine discrete interpolation with recursive on-the-fly adaptation of the interpolation indices to approximate the evolution equations for a low-rank factorization akin to Equation (8.3). Here, we extend this construction to a parameterization-independent on-the-fly sparse approximation heuristic that applies for DLRA of generic matrix-valued IVPs (8.1) with local vector field F. The heuristic composes readily with geometric DLRA schemes which is shown in Section 8.8 to be numerically favorable. The proposed heuristic involves executing the following three-step procedure at every iterate $\hat{X}_k$ of a numerical DLRA scheme.

Heuristic 1 (On-the-fly sparse approximation)**.** *Given a current low-rank iterate $\hat{X}_k$, time step t_k, and a desired rank r_F, construct a sparse rank-r_F approximator of F via the following steps:*

1. *Identify $U_F \in \mathbb{R}^{n\times r_F}$ and $V_F \in \mathbb{R}^{m\times r_F}$ so that the range of U_F and V_F approximate the range and co-range of $F(\hat{X}_k, t_k)$, respectively.*

2. *Based on U_F and V_F, identify $O(r_F)$ row and column indices that are best suited to interpolate the remaining rows and columns of $F(\hat{X}_k, t_k)$. Let R_k and C_k be submatrices of the identity such that left multiplication with $R_k^\top$ and right multiplication with C_k extract the respective rows and columns.*

3. *Compute oblique projectors $P_k = U_F(R_k^\top U_F)^\dagger R_k^\top$ and $Q_k = V_F(C_k^\top V_F)^\dagger C_k^\top$, and approximate F by $P_k F Q_k^\top$.*[3]

Step 3 of the above heuristic is easily executed once step 1 and 2 are completed. Similarly, once step 1 has been performed, step 2 is readily addressed by a host of DEIM procedures [189–191, 206, 207] or CUR decomposition methods based on leverage scores [208, 209]. The crux of Heuristic 1 lies in its first step. Given an iterate $\hat{X}_k$, it is not obvious how the range and co-range of $F(\hat{X}_k, t_k)$ should be approximated when the full evaluation of $F(\hat{X}_k, t_k)$ is exceedingly expensive. A straightforward recursive solution to this problem is to use the sparse approximator obtained from the previous iteration. Concretely, one may simply compute U_F and V_F as the leading

[3] $Y^\dagger$ denotes the Moore-Penrose inverse of Y.

left and right singular vectors of $F(\hat{X}_k, t_k)Q_{k-1}^\top$ and $P_{k-1}F(\hat{X}_k, t)$, respectively. This is the essence of the proposal by Naderi and Babaee [22]. We will discuss several tractable alternatives for performing step 1 in Section 8.6.

We emphasize that the sparse approximator $P_k F Q_k^\top$ furnished by Heuristic 1 admits by construction an explicit rank-r_F factorization that can be evaluated from $O(r_F^2)$ individual entries of F. When composed with a numerical DLRA scheme, Heuristic 1 therefore recovers the favorable scaling properties of DLRA. A tractable approximation for the projected Euler method (8.4) for nonlinear vector fields is for instance

$$\hat{X}_{k+1} = R_{\hat{X}_k}\left(h\Pi_{\mathcal{T}_{\mathcal{M}_r}(\hat{X}_k)} P_k F(\hat{X}_k, t_k) Q_k^\top\right). \tag{8.7}$$

It follows from the discussions in Section 8.4 that the approximate DLRA recursion (8.7) achieves favorable scaling properties, even in the presence of nonlinearities in the vector field F. Moreover, implementation of this recursion requires minimal intrusion as it relies exclusively on simple element-, row- and column-wise evaluation oracles for the vector field F.

An alternative, but equally natural approximation is obtained by taking advantage of the splitting of the tangent space projection into subprojections as given in Equation (8.6). The subprojections onto the range and co-range of $\hat{X}_k = USV^\top$ with (semi-)orthogonal factors U and V can be approximated from the rows and columns selected in Heuristic 1:

$$\begin{aligned}\Pi_{\mathcal{T}_{\mathcal{M}_r}(\hat{X}_k)} F(\hat{X}_k, t) \approx& \\ &UU^\top P_k F(\hat{X}_k, t) + F(\hat{X}_k, t_k) Q_k^\top VV^\top - UU^\top P_k F(\hat{X}_k, t_k) Q_k^\top VV^\top.\end{aligned}$$

Note in particular that this construction preserves the favorable scaling properties: the application of only one of the oblique projectors P_k and Q_k to F yields explicit rank-r_F factorizations approximating F which require the evaluation of only $O(r_F)$ of its rows and columns, respectively. Heuristic 1 therefore composes readily in a natural way with many numerical DLRA schemes that exploit the splitting of the tangent space projection explicitly, such as projector-splitting integrator [210] or basis-update and Galerkin integrators [171, 211].

8.6 On-the-fly approximation of range and co-range

The main complication of DLRA with on-the-fly sparse approximation is computing informative rank-r_F approximations of the range and co-range of the vector field F in each integration step. To ultimately best approximate F in a least squared error sense, the most natural approach to this problem is computing the r_F leading left and right singular vectors of $F(\hat{X}_k, t_k)$ at a given low-rank iterate $\hat{X}_k$. In the following, we discuss two tractable routines to approximate this computation.

Recursive range and co-range approximation – The continuous and causal evolution of the state $\hat{X}(t)$ according to the ODE (8.2) implies under suitable regularity conditions on F that also the leading left and right singular values of $F(\hat{X}(t), t)$ evolve continuously (or at least nearly so) [212]. This suggests a recursive approach that incrementally updates approximations of $\hat{X}(t)$ alongside range and co-range approximations of $F(\hat{X}(t), t)$. Given a numerical DLRA scheme that recursively updates approximations $\hat{X}_k$ from the previous iterate $\hat{X}_{k-1}$, a natural update rule for the range and co-range is described in Algorithm 2. This update rule is indeed tractable as $F(\hat{X}_k, t_k)Q_{k-1}^\top$ and $P_{k-1}F(\hat{X}_k, t_k)$ are explicit rank-r_F factorization which are computed by evaluating $O(r_F)$ columns and rows of F explicitly. The computation of

Algorithm 2 Recursive range and co-range approximation

Input: Current iterate $\hat{X}_k$, time t_k, oblique projectors P_{k-1} and Q_{k-1} from previous time step, desired rank r_F.
Compute rank-r_F truncated SVDs (in parallel)

$$\begin{aligned} U_F S_C V_C^\top &= \mathrm{SVD}_{r_F}\left(F(\hat{X}_k, t_k)Q_{k-1}^\top\right) \\ V_F S_R U_R^\top &= \mathrm{SVD}_{r_F}\left(F(\hat{X}_k, t_k)^\top P_{k-1}^\top\right) \end{aligned}$$

Output: Approximations U_F and V_F for the r_F leading singular vectors $F(\hat{X}_k, t_k)$.

the truncated SVDs in Algorithm 2 therefore requires only $O(r_F^2(n+m))$ operations. Note further that it is critical to update the range from $F(\hat{X}_k, t_k)Q_{k-1}^\top$ and the co-range from $P_{k-1}F(\hat{X}_k, t_k)$ and not vice versa. Otherwise, the computed range and co-range would be subsets of the range of P_{k-1} and Q_{k-1}, respectively, and hence remain unchanged between iterations. Lastly, this recursive process must be initialized. A natural initialization is obtained from the truncated SVD of $F(\hat{X}_0, 0)$. When this is intractable, we discuss a clustering approach in the next subsection that may be applied instead.

Naderi and Babaee [22] propose a similar update rule to Algorithm 2 in the context of DLRA for stochastic PDEs:

1. Compute U_F as the leading left singular vectors of $F(\hat{X}_k, t_k)C_{k-1}$, where C_{k-1} is the column selector matrix identified in the previous time step.

2. Based on U_F compute a row selector matrix R_k.

3. Compute V_F as the leading right singular vectors of $R_k^\top F(\hat{X}_k, t_k)$.

4. Based on V_F compute a column selector matrix C_k.

There are several notable distinctions between this procedure and embedding Algorithm 2 in Heuristic 1. First, this procedure intertwines the range estimation and index selection step of Heuristic 1. As such, the range and co-range approximation can no longer be parallelized. Second, it introduces an ordering into the range and co-range estimation step. This ordering appears natural in the problem-specific context of [22], but arbitrary in the generic setting treated here. Lastly, the range and co-range approximations from the above procedure are obtained from singular vectors of the row and column-wise evaluation of the vector field directly. Algorithm 2 in contrast computes the singular vectors of a low-rank approximation of F generated by the oblique projectors determined in a previous time step. While both approaches yield equivalent range approximations when the oblique projectors have full rank, the approach in Algorithm 2 more closely mimics the computation of the leading singular vectors of F. As such it is expected to return more accurate approximations for the leading singular vectors. Chaturantabut and Sorensen [189] argue that this property is advantageous when the classical DEIM routine [189, Algorithm 1] is used to identify suitable approximation indices in Step 2 of Heuristic 1. Moreover, as a useful by-product Algorithm 2 produces an estimate of the singular values of $F(\hat{X}_k, t_k)$. This information enables error estimates for the sparse approximation of F [206] and thus can inform adaptation of the approximation rank r_F as discussed by Naderi and Babaee [22] who compute this information in a separate step.

Cluster-based range and co-range approximation – In many scientific applications of DLRA, the state X in the IVP (8.1) corresponds to a discretization of a continuous spatio-temporal field; for example a temperature, concentration, or pressure field. The vector field F then typically derives from a partial-differential operator governing the field's dynamics. In this setting, the rows and columns of $F(X(t), t)$ typically inherit patterns of "smoothness" and "similarity" from the field $X(t)$. Given a low-rank approximation $\hat{X}(t)$ for the field, it is therefore a reasonable strategy to approximate the range of $F(\hat{X}(t), t)$ by identifying rows and columns in $\hat{X}(t)$ that loosely speaking are mutually as dissimilar as possible. The intuition is that such rows and columns capture the distinct spatial features of the physical field and that the corresponding rows and columns of $F(\hat{X}(t), t)$ therefore collectively span the co-range and range of the dynamics in close approximation.

The identification of a diverse set of rows and columns in $\hat{X}$ is naturally posed as a clustering problem. Among the wide range of clustering problem formulations and algorithms available [213, 214], K-means clustering stands out as a suitable choice for this application. On the one hand, K-means clustering algorithms can be warm-started [214] which is important when clustering is performed in every iteration of a numerical DLRA scheme as per Heuristic 1. On the other hand, low-rank structure can be exploited to reduce the associated computational load further. For the identification of representative columns of $\hat{X}$, the K-means clustering problem seeks clusters $S_1, \ldots, S_{r_F}$ (index sets referencing collections of columns) that approximate the solution to the following combinatorial optimization problem:

$$\begin{aligned}
\min_{S_1,\ldots,S_{r_F}} \quad & \sum_{i=1}^{r_F} \sum_{k\in S_i} \|\hat{X}e_k - \mu_i\|_F^2 \\
\text{s.t.} \quad & \mu_i = \frac{1}{|S_i|} \sum_{k\in S_i} \hat{X}e_k, \quad i \in [r_F], \\
& S_i \cap S_j = \emptyset, \quad 1 \le i \ne j \le r_F, \\
& \cup_{i=1}^{r_F} S_i = [m],
\end{aligned}$$

where $[r]$ denotes the index range $\{1,\ldots,r\}$. When $\hat{X}$ is available as a rank-r factorization $UZ^\top$ with $U \in \mathcal{V}_{n,r}$,[4] the invariance of the Euclidean norm under orthogonal transformations can be exploited to solve the clustering problem instead in r-dimensional space:

$$\begin{aligned}
\min_{S_1,\ldots,S_{r_F}} \quad & \sum_{i=1}^{r_F} \sum_{k\in S_i} \|Z^\top e_k - \mu_i\|_F^2 \\
\text{s.t.} \quad & \mu_i = \frac{1}{|S_i|} \sum_{k\in S_i} Z^\top e_k, \quad i \in [r_F], \\
& S_i \cap S_j = \emptyset, \quad 1 \le i \ne j \le r_F, \\
& \cup_{i=1}^{r_F} S_i = [m].
\end{aligned}$$

The identification of representative rows of $\hat{X}$ via K-means clustering proceeds analogously and can be performed in parallel.

Once the clusters of rows and columns of $\hat{X}$ are identified, there are different conceivable ways to approximate the range and co-range of $F(\hat{X},t)$. On the one hand, one or few representative rows and columns can be sampled from each cluster. On the other hand, when F represents a discretized partial-differential operator, it is often feasible to evaluate F at the cluster means directly. In either case, the range and co-range of $F(\hat{X},t)$ can be approximated by the singular vectors of the submatrices generated by evaluating F at the identified rows and columns.

[4] Such a factorization is readily computed from the SVD-like factorization $\hat{X} = USV^\top$ used internally by numerical DLRA routines.

8.7 LowRankIntegrators.jl – dynamical low-rank approximation in Julia

Numerical DLRA schemes are designed to exploit low-rank structures in every operation they are composed of. While this gives rise to a distinctly improved memory footprint and scaling properties, it also comes at the cost of a notably more complicated computational implementation. It is no easy feat to realize the promised computational benefits of such an algorithmic design, especially not while maintaining an easy-to-use, yet general and extensible code base. As a consequence, it remains common practice to tailor implementations of numerical DLRA schemes to a given application exploiting the structure of the field F wherever possible. Once a new application arises, this process is repeated, starting from low-level primitives [215]. While such a workflow has been shown to deliver on the promises of DLRA in several applications [172, 173, 216, 217], it renders DLRA rather inaccessible non-experts. Moreover, this approach is cumbersome when developing algorithms or in prototyping situations where the problem at hand is subject to frequent change. With this perspective, it is natural to wonder how much performance must really be sacrificed by a generic implementation of numerical DLRA schemes with a high-level interface. `LowRankIntegrators.jl`[5] provides such an implementation. It integrates sparse on-the-fly approximation, robust geometric DLRA time-stepping routines, and ancillary techniques for the manipulation of low-rank matrix factorizations to enable DLRA for generic matrix-valued IVPs with minimal intrusion. Performance losses due to its problem-agnostic implementation are mitigated by leveraging the features of the Julia programming language [193], in particular specialization on parametric types. A particular emphasis is placed on on-the-fly sparse approximation. The user is only required to supply element-, row- and column-wise evaluation oracles for the vector field F and choose a sparse approximation scheme. Not only does this extend applicability of DLRA to problems where routines for an efficient evaluation and tangent space projection of the vector field at low-rank inputs are not read-

[5] https://github.com/fholtorf/LowRankIntegrators.jl

ily available, but it also enables a workflow that closely mimics that of solving the IVP (8.1) with one of many widely adopted tools for solving ordinary IVPs [218, for example]. As such, `LowRankIntegrators.jl` complements software efforts, such as `Ensign`[6] [215], aimed at streamlining the low-level implementation of tailored, problem-specific DLRA schemes.

To achieve a performant, yet problem-agnostic implementation of numerical DLRA schemes, `LowRankIntegrators.jl` relies on Julia's automatic specialization on parametric types. On the one hand, automatic type specialization enables a unified and high-level user interface which accepts DLRA problems in a generic form but still allows problem-specific structure to be exploited during the solve step. Most importantly, the solve routine specializes on how the vector field shall be evaluated and projected: Is the vector field or a component of it linear? Shall nonlinear components be approximated via on-the-fly sparse approximation, or did the user provide a function that maps low-rank inputs to an explicit low-rank factorization of the vector field in an efficient way? Can these functions be evaluated in-place to avoid memory allocations or are they inherently allocating? Information that answers these and more questions is encoded by the problem type and the solve routines take advantage of them. On the other hand, `LowRankIntegrators.jl` leverages Julia's type system to implement a rich and extensible library of DLRA integrators. The implementation exploits the common structure of geometric DLRA integrators as outlined in Section 8.3 and in greater detail in [197]; for instance, `LowRankIntegrators.jl` implements the generic projected Runge-Kutta scheme of [196] which may be composed with a range of retractions, including user-defined ones, and Butcher tableaus to generate a rich set of consistent integrators. Moreover, integrators that involve the solution of ordinary IVPs as substeps [171, 210, 211, for example] can be composed with any of the routines in `DifferentialEquations.jl` [218] for these substeps. A list of integration and ancillary routines available in `LowRankIntegrators.jl` at the time of writing this thesis is given in Table 8.1.

Lastly, the numerical DLRA schemes implemented in `LowRankIntegrators.jl`

[6] https://github.com/leinkemmer/Ensign

Table 8.1: DLRA routines and ancillary utilities implemented in `LowRankIntegrators.jl`

Time-stepping routines
Projected Euler (arbitrary retraction) [199]
Projected Runge-Kutta (arbitrary retraction & explicit Butcher tableau) [196]
(Rank-adaptive) BUG/unconventional algorithm [211, 219]
Lie-Trotter projector splitting algorithm [210]
Strang projector splitting algorithm [195]
Retractions
SVD/projective retraction
KSL/Lie-Trotter retraction [202]
KLS retraction [197, 219]
orthographic retraction [200]
Sparse approximation
DEIM index selection [189]
QDEIM index selection [190]
LDEIM index selection [191]
recursive range approximation [22]
K-means clustering approximation

depend inherently on efficient routines for processing and manipulating low-rank matrix factorizations. As illustrated in Figure 8-3, this is facilitated by the Julia package `LowRankArithmetic.jl`[7]. `LowRankArithmetic.jl` leverages Julia's type system by introducing custom types for common low-rank factorization formats alongside specialized methods for common operations on them. A concise overview of the supported low-rank formats and operations is outlined in Table 8.2. Furthermore, `LowRankArithmetic.jl` facilitates the propagation of low-rank factorizations through generic Julia functions composed of the supported operations without the need for custom implementations. As exemplified for the case of matrix addition below, this feature hinges on the property that the supported operations preserve the structure of the low-rank factorization formats and lead to at most bounded rank growth.

Example 8.1 (Addition of low-rank factorizations)**.** *Consider two real low-rank factorizations $X_i = U_iS_iV_i^\top$, with $U_i \in \mathbb{R}^{n\times r_i}$, $S_i \in \mathbb{R}^{r_i\times r_i}$ and $V_i \in \mathbb{R}^{m\times r_i}$ for $i = 1, 2$.*

[7] https://github.com/fholtorf/LowRankArithmetic.jl

Their sum admits the structurally identical factorization

$$X_1 + X_2 = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^\top .$$

The addition of low-rank matrices represented as factor triples (U_i, S_i, V_i) *can therefore be reduced to suitable concatenation of the factors. This is computationally favorable when* $r_1 + r_2 \ll n, m$*.*

Similar, albeit more complicated, rules apply for the remaining operations listed in Table 8.2. This perspective underlines that `LowRankArithmetic.jl` can streamline the process of exploiting low-rank structure beyond internal manipulations of low-rank factorizations in numerical DLRA schemes. Most notably, it enables automation of the otherwise cumbersome but crucial step of implementing custom routines for efficient evaluation of the vector field at inputs in low-rank format. Via `LowRankArithmetic.jl` a generic Julia function for the vector field may be automatically specialized on such inputs. This is demonstrated in Figure 8-2 for a vector field of the form

$$F(X) = LX - X \odot (GX) \tag{8.8}$$

as obtained from discretization of Burgers' equation with an uncertain initial condition; see [175] for details.[8] Figure 8-2 shows the computational advantage realized by automatic specialization of the generic implementation of the vector field (8.8) in Julia

```julia
function F(X, p)
    L, G = p
    return L*X - X .* (G*X)
end
```

as it is evaluated on low-rank matrix representations of `LowRankArithmetic.jl`'s SVD-like factorization type. This specialization intrinsically capitalizes on the fact

[8] The matrices L and G represent second-order finite difference approximations of the Laplacian and gradient on an equidistant one-dimensional grid.

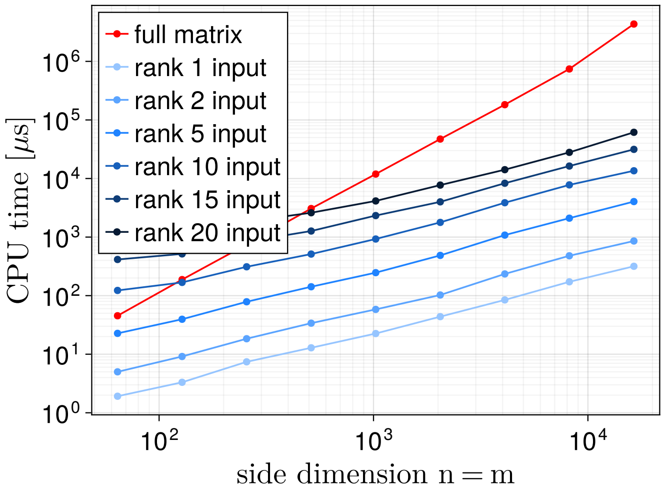


Figure 8-2: Computational cost of propagating an SVD-like low-rank factorization through the bilinear vector field (8.8) via `LowRankArithmetic.jl`.

that the vector field (8.8) maps matrices of rank r to matrices of rank at most $r + r^2$ and returns a low-rank factorization in place of a dense matrix. The computational savings therefore increase as the side dimensions of the input grow relative to its rank.

8.8 Examples

8.8.1 Test equation

We first demonstrate the value of composing on-the-fly sparse approximation with robust geometric integrators. The additional error induced by sparse approximation, even if small, is found to amplify the stiffness of evolution equations of a fixed parameterization of the low-rank manifold given in Equation (8.3). As a consequence, even more stringent step size restrictions are imposed on explicit integrators for such approximate evolution equations in the presence of small singular values. To demonstrate this effect, we consider the test problem from Kieri *et al.* [195], seeking low-rank approximation of the time-dependent matrix

$$A(t) = \exp\left(tW_1\right) D(t) \exp\left(tW_2\right),$$

Table 8.2: Low-rank factorization formats and operations supported by `LowRankArithmetic.jl`.

Low-rank factorization formats
SVD-like representation ($X = USV^\top$)
Two factor representation ($X = UZ^\top$)
Arithmetic operations
Matrix addition
Matrix multiplication
Hadamard products
Element-wise integer powers
Utilities
Concatenation
Slicing & indexing
Adjoints
QR
SVD (rounding)
Orthonormalization of factors (by SVD, QR, Gram-Schmidt, gradient flow [178], or second-moment matching [220])

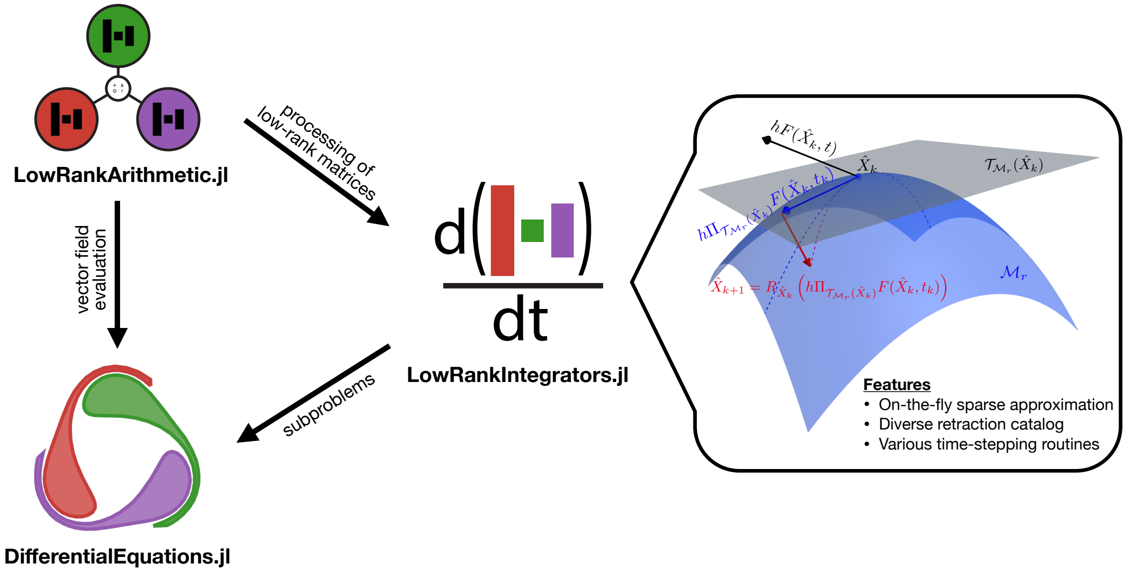
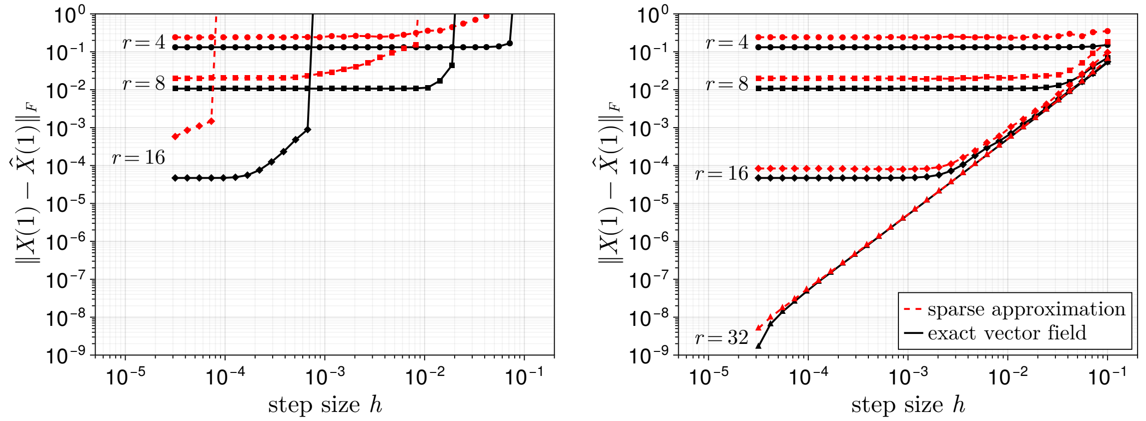

LowRankArithmetic.jl
processing of low-rank matrices
vector field evaluation
$\frac{d(\ldots)}{dt}$
LowRankIntegrators.jl
subproblems
DifferentialEquations.jl
$hF(\hat{X}_k, t)$
$\hat{X}_k$
$\mathcal{T}_{\mathcal{M}_r}(\hat{X}_k)$
$h\Pi_{\mathcal{T}_{\mathcal{M}_r}(\hat{X}_k)}F(\hat{X}_k, t_k)$
$\hat{X}_{k+1} = R_{\hat{X}_k}\left(h\Pi_{\mathcal{T}_{\mathcal{M}_r}(\hat{X}_k)}F(\hat{X}_k, t_k)\right)$
$\mathcal{M}_r$
Features
• On-the-fly sparse approximation
• Diverse retraction catalog
• Various time-stepping routines

Figure 8-3: Overview of dynamical low-rank approximation ecosystem in Julia.



(a) RK4 on factors

(b) Strang projector-splitting algorithm

Figure 8-4: Comparison between the accuracy of robust geometric integrators for DLRA with on-the-fly sparse approximation and standard integrators for evolution equations of low-rank factors. The number of approximation points is chosen as $r_F = 2r$.

where $D(t) = \exp(t)\text{diag}(1/2, 1/4, \ldots, 1/2^n)$ and $W_1, W_2 \in \mathbb{R}^{n \times n}$ skew-symmetric matrices with independent normal entries. The vector field in the DLRA IVP (8.2) is thus given by $F(X,t) = \dot{A}(t)$. We consider the case of $n = 100$ and a final time $T = 1$.

Figure 8-4a compares the final approximation error obtained by applying a standard integrator (the 4-stage Runge-Kutta method from [218]) to the evolution equations of an SVD-like low-rank approximation (8.3) with and without on-the-fly sparse approximation of the vector field, for different step sizes, and approximation ranks. We observe that the step size restriction due to small singular values is notably exacerbated by sparse approximation of the vector field. In contrast, Figure 8-4b illustrates that the robust Strang projector splitting integrator [195, 210] remains stable across the entire range of step sizes. Remarkably, the robustness persists when the vector field is sparsely approximated which is in stark contrast to the evolution equation approach. The computational cost of both numerical schemes compared in Figure 8-4 is dominated by matrix multiplications of identical cost and is thus similar [219].

8.8.2 Solar wind prediction under uncertainty

The quantification of uncertainty in predictions for the dynamics of the solar-terrestrial system remains an open challenge in numerical space weather prediction [221]. The large scale of the involved models and nonlinearity of the governing equations pose significant complications for this task. DLRA with on-the-fly sparse approximation promises to alleviate both issues at the cost of a small approximation error. In the following, we demonstrate that these promises can indeed be kept. To that end, we consider the propagation of uncertainty through the heliospheric upwind extrapolation (HUX) model [222, 223], describing the propagation of solar wind streams in the ecliptic plane. Leaving physical details aside, the HUX model boils down to the following nonlinear PDE:

$$\begin{cases} \frac{\partial}{\partial r} v(r,\phi,\omega) = \Omega \frac{\partial}{\partial \phi} \log v(r,\phi,\omega), & \phi \in [0, 2\pi], r \in [0.14\,\text{AU}, 1\,\text{AU}] \\ v(r,0,\omega) = v(r,2\pi,\omega), & r \in [0.14\,\text{AU}, 1\,\text{AU}] \\ v(r_0,\phi) = v_0(\phi,\omega), & \phi \in [0, 2\pi] \end{cases}, \tag{8.9}$$

where $v(r,\phi,\omega)$ denotes the radial velocity of solar wind at distance r from the sun's center, angle ϕ in the ecliptic plane, and realization of uncertainty ω. Ω denotes the sun's rotational frequency. For a detailed derivation, we refer the reader to [222].

The largest source of uncertainty in space weather models stems from uncertain initial and boundary conditions [221]. We therefore assume that the initial solar wind profile at the inner heliosphere is uncertain and admits a Karhunen-Loève expansion,

$$v_0(\phi,\omega) = \mathbb{E}[v_0(\phi,\omega)] + \sum_{i=1}^{10} \sqrt{\lambda_i}\, \xi_i(\phi)\, \omega_i,$$

where ξ_i and λ_i are the leading eigenfunction and eigenvalue pairs of the periodic kernel [224]

$$k(\phi,\phi') = \sigma^2 \exp \frac{\sin^2 \pi|\phi - \phi'|}{L^2}.$$

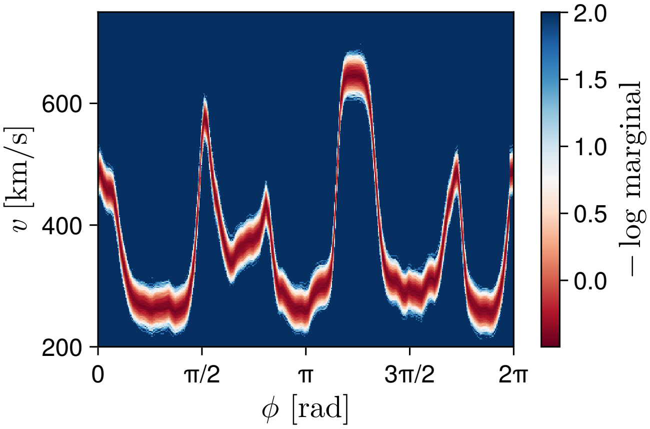


Figure 8-5: Marginal density of initial solar wind velocity profile in the inner heliosphere ($r = 0.14\,\mathrm{AU}$).

We chose a velocity and length scale of $\sigma = 20\,\mathrm{km\,s^{-1}}$ and $L = 0.2$, respectively. The mean velocity profile is given as the solution of a high-fidelity magnetohydrodynamic model simulated under the conditions of the Carrington rotation 2068 [223]. The marginal density of the corresponding initial velocity profile is shown in Figure 8-5.

In order to turn this problem into a form amenable to DLRA, we discretize the angular domain into $n = 512$ uniformly spaced intervals and draw $m = 4096$ independent samples for ω from a standard multivariate normal distribution. A sample approximation of the stochastic solution of the HUX model (8.9) is then encoded in the matrix-valued state $X_{i,j}(r) = v(r, \phi_i, \omega_j)$. Upon first-order accurate upwind discretization of the angular gradient operator as proposed in [223], the associated vector field is given by

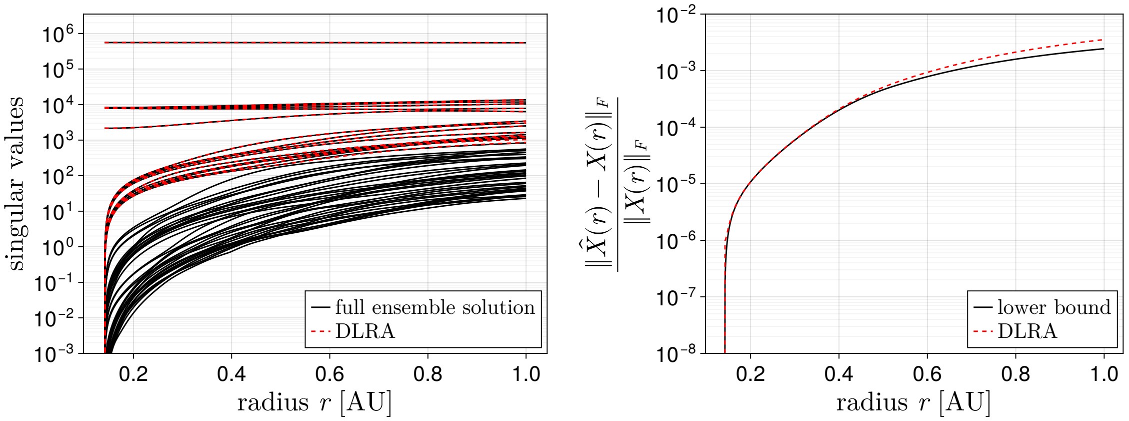
$$\frac{\mathrm{d}X_{i,j}}{\mathrm{d}r}(r) = F_{i,j}(X(r), r) = \begin{cases} \Omega \dfrac{\log X_{i+1,j}(r) - \log X_{i,j}(r)}{\Delta\phi}, & 1 \leq i < N \\ \Omega \dfrac{\log X_{1,j}(r) - \log X_{N,j}(r)}{\Delta\phi}, & i = N \end{cases} .$$

We note that without further approximation or transformation, this vector field does not admit an efficient projection onto the tangent bundle of the low-rank manifold. Thus, on-the-fly sparse approximation is necessary to leverage the scaling benefits of DLRA.

In the following, we compare a rank-15 DLRA solution to the full solution of the HUX model for the entire ensemble of initial conditions. The DLRA solution was computed using the 4$^{\text{th}}$ order projected Runge-Kutta method of Kieri and Vandereycken [196] with KSL retraction [197] as implemented in `LowRankIntegrators.jl`. The vector field was sparsely approximated from $r_F = 30$ rows and column indices which were updated after every integration step via Heuristic 1 with recursive range and co-range estimates from Algorithm 2. The approximation indices were selected with the classical DEIM procedure [189, Algorithm 1]. Computation of the rank-15 DLRA solution took $\sim 30\,\text{s}$ on a MacBook M1 Pro with 16 GB unified memory. Despite the modest size of the problem, this corresponds to four-fold speed-up over computing the full ensemble solution with the canonical fourth order Runge-Kutta method from `DifferentialEquations.jl`. Figure 8-6 shows that the DLRA solution accurately tracks the leading singular values of the full ensemble solution from the inner heliosphere to the earth's orbit and thus achieves near-optimal compression. As further illustrated in Figure 8-7, the recursive updating strategy for the selection of approximation indices aligns with common intuition. The selected approximation indices cluster around shocks in the velocity profile and track them as they propagate to earth's radius. Figure 8-8 finally contrasts the predictions of the rank-15 DLRA and full ensemble solution for the velocity profile at the earth's orbit (1 AU). As indicated by the accurate tracking of the leading singular values, the solutions are qualitatively and quantitatively accurate, differing by no more than 0.5% on average. Moreover, characteristics of the stochastic solution such as non-Gaussian marginals are preserved by the low-rank approximation, rendering it suitable to quantify prediction uncertainties of the HUX model originating from an uncertain initial condition.

8.8.3 2-D nonlinear heat equation

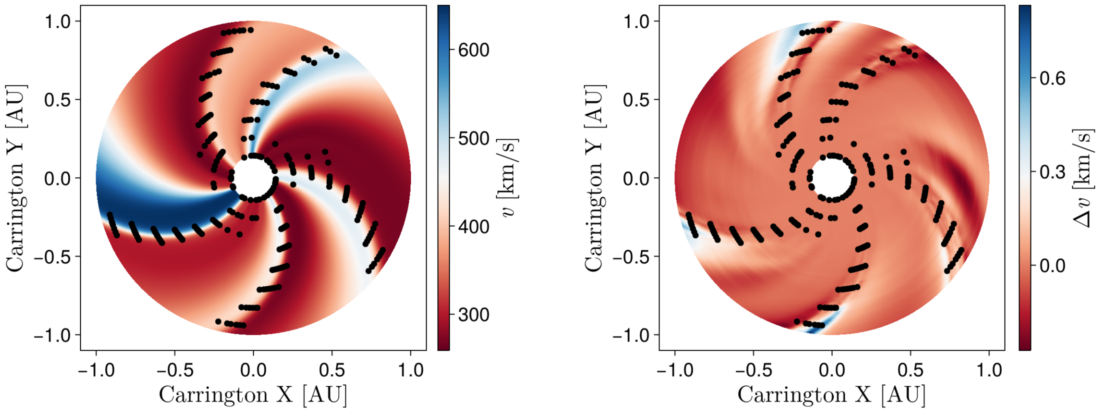
With this last example, we demonstrate the favorable scaling properties attained by composing on-the-fly sparse approximation with robust DLRA routines. To that end, we consider a heat conduction problem on a square with Dirichlet boundary



(a) leading singular values

(b) relative error

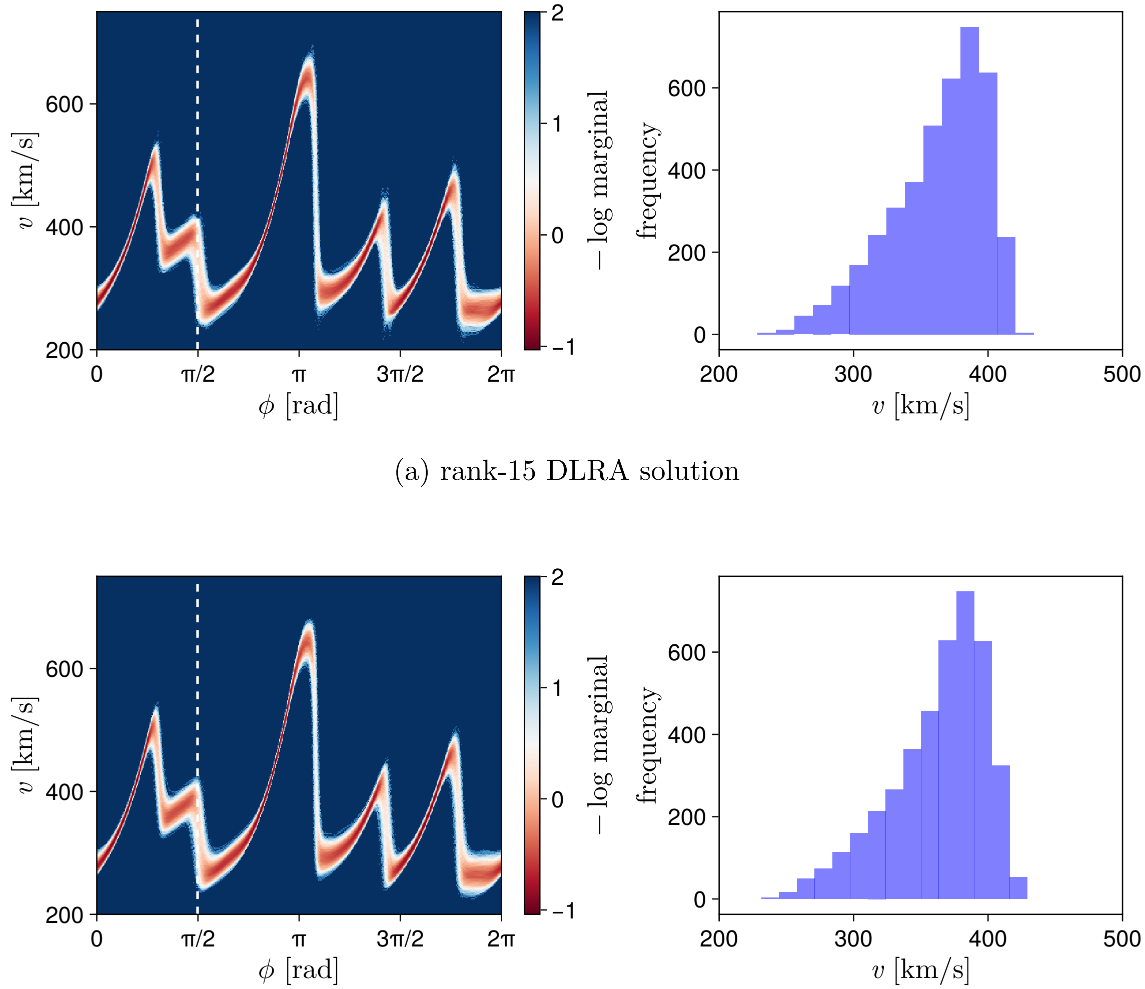
Figure 8-6: Near-optimal compression of the stochastic solution of the HUX model. The best attainable relative error (lower bound) is computed via the rank-15 truncated SVD of the full ensemble solution.



(a) mean

(b) absolute error

Figure 8-7: Mean solar wind velocity profile prediction of rank-15 DLRA solution with on-the-fly sparse approximation. Approximation points points are shown in black.



(a) rank-15 DLRA solution

(b) full ensemble solution

Figure 8-8: Predictions of solar wind velocity profile at the earth's orbit ($r = 1\,\mathrm{AU}$). Left: negative logarithm of marginal density on the entire orbit. Right: histogram of velocity predictions at $\phi = \pi/2$.

conditions and a nonlinear, time-dependent heat source. The source term consists of a volumetric heating flux with an extended Arrhenius rate expression and a balancing cooling term. The corresponding (dimensionless) PDE reads

$$\begin{cases} \dot{T} + \alpha \Delta T = T^{\beta} \exp\left(-\frac{\gamma}{T}\right) Q - \delta(T - T_a), & t \in [0,1],\ (x,y) \in (0,1)^2 \\ T(t,x,y) = T_a, & t \in [0,1],\ (x,y) \in \{0,1\}^2 \, , \\ T(0,x,y) = T_a, & (x,y) \in [0,1]^2 \end{cases} \tag{8.10}$$

where Δ denotes the Laplacian operator $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$. The spatially and temporally varying heat flux Q is given by

$$Q(t,x,y) = 10 \exp\left(-\frac{(x - \bar{x}(t))^2 + (y - \bar{y}(t))^2}{2L^2}\right) \text{ with } \begin{bmatrix} \bar{x}(t) \\ \bar{y}(t) \end{bmatrix} = \begin{bmatrix} 0.1 + 0.8t \sin^2 \frac{3\pi}{2} t \\ 0.1 + 0.8t^2 \end{bmatrix}.$$

The remaining parameters are chosen as $\alpha = 1 \times 10^{-4}$, $\beta = 0.3$, $\gamma = 0.5$, $\delta = 8$, $L = 0.1$, and $T_a = 0.5$. The large cooling rate δ leads to a locally sharply peaked temperature profile ideally suited for low-rank approximation.

To apply DLRA to the heat equation (8.10), the spatially continuous temperature field is approximated by a time-dependent $n \times n$ matrix X such that $X_{i,j}(t) \approx T(t, x_i, y_j)$ on a uniform grid $\{(x_i, y_i) = (i,j)\Delta l : 1 \leq i, j \leq n\}$ of width $\Delta l = 1/(n+1)$. The vector field encoding the dynamics of X is obtained by discretizing the Laplacian operator with a second-order centered difference stencil. For the numerical solution of the DLRA problem, we use the basis update and Galerkin time-stepping scheme [171] from `LowRankIntegrators.jl` with Euler substeps (also known as the KLS integrator [197]). For a tractable approximation of the nonlinear vector field, the integrator is composed with on-the-fly sparse approximation. The approximation indices are determined with the classical DEIM procedure [189, Algorithm 1]. The range and co-range approximations are updated recursively via Algorithm 2. For a rank-r DLRA approximation, we use a rank-$2r$ sparse approximation of the vector field constructed from $2r$ rows and columns. Figure 8-9 shows snapshots of a rank-5 approximation to the solution of Equation (8.10) as computed with the described numerical scheme. As illustrated, the rows and columns identified for sparse approximation concentrate on the temperature peak induced by the heat source and follow closely as it travels through the domain. Moreover, the width of the approximation grid increases as the temperature profile becomes more diffuse.

Finally, Figure 8-10 shows the cost-accuracy trade-off offered by the described DLRA scheme. The composition of on-the-fly sparse approximation with robust geometric DLRA integrators yields the desired linear scaling of the computational cost with respect to the number of grid points n along one dimension of the domain. This is in stark contrast to the quadratically scaling cost incurred by solving the full equation and enables substantial computational savings for sufficiently fine grids. At the same time, DLRA yields highly accurate low-rank approximations to the true solution. Relative to the full solution, the rank-10 DLRA solution for instance accumulates less than 0.01% error over the simulation horizon.

All computational experiments were conducted on MIT supercloud's Intel Xeon Platinum 8260 processor with 48 cores and 187.5 GB RAM [225].

8.9 Conclusion

DLRA has enabled the computation of accurate low-rank approximations to otherwise intractable matrix-valued IVPs in applications across various domains. These success stories have traditionally been achieved on the back of low-level implementations of numerical DLRA schemes tailored to structured, typically linear, bi-linear, or at most quadratic vector fields. In this chapter, we have presented methods and computational tools that extend the applicability of DLRA to general nonlinear vector fields with local structure while alleviating the need for tailored low-level implementations. From a methodological perspective, we have shown that a simple modification of Naderi and Babaee's on-the-fly sparse approximation heuristic [22] enables DLRA for generic nonlinear matrix-valued IVPs with vector fields and composes readily with a host of geometric DLRA schemes. The composition with geometric DLRA schemes was demonstrated to endow the heuristic with notably improved

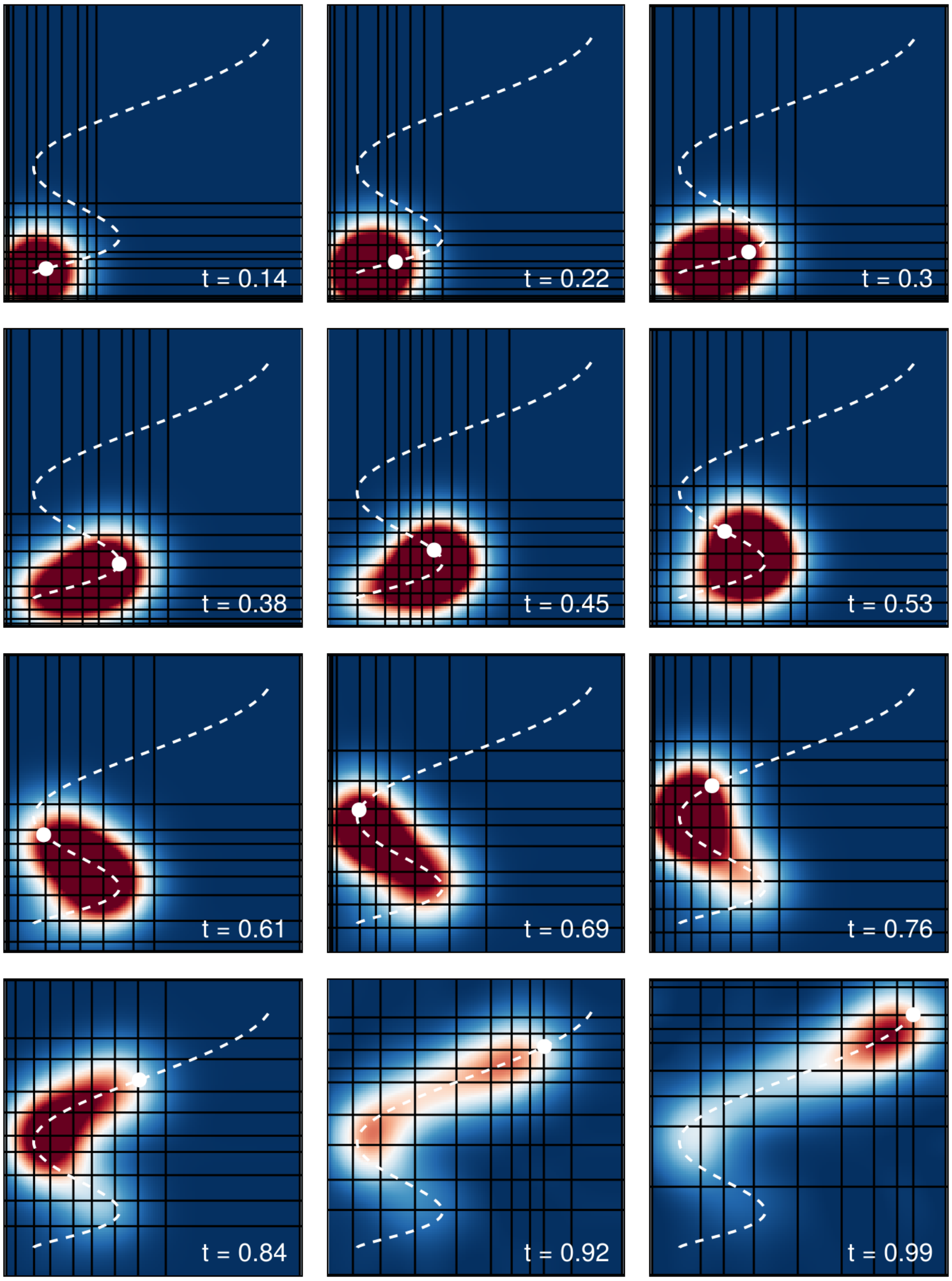
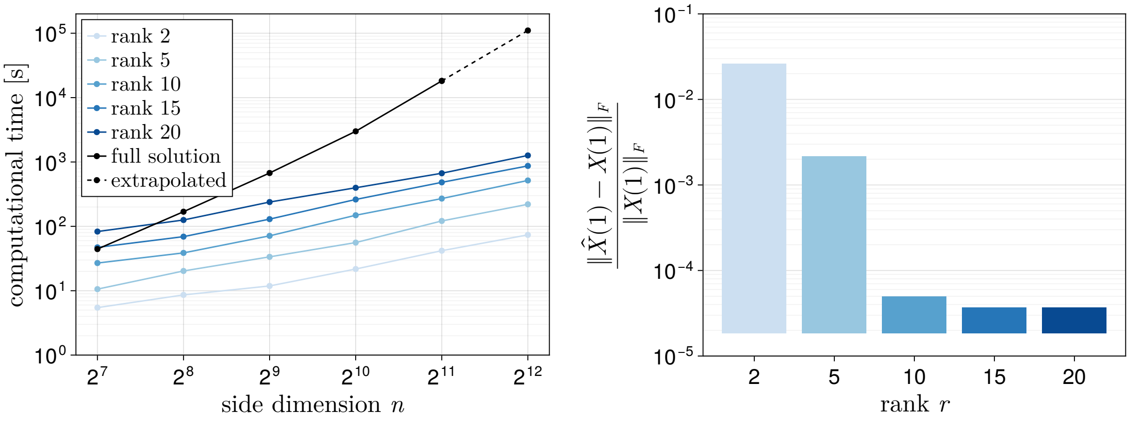


Figure 8-9: Snapshots of a rank-5 DLRA solution of the heat equation (8.10). The on-the-fly adapted sparse approximation grid is indicated in black. Temperatures are shown on the scale between 0.5 (dark blue) and 1.0 (dark red). The center and path of the source is shown with a white marker and white dashed line, respectively.



(a) computational cost

(b) accuracy ($n = 1024$)

Figure 8-10: Trade-off between cost and accuracy for DLRA of the nonlinear heat equation (8.10) with increasing grid resolution and rank.

numerical robustness properties. From the tooling perspective, we have developed `LowRankIntegrators.jl`, a high-level software package for DLRA in the Julia programming language. `LowRankIntegrators.jl` leverages Julia's type system to automatically specialize numerical DLRA schemes on problem-specific information. As such, it enables performant DLRA with minimal intrusion from high-level input specifications.

Chapter 9

Concluding Remarks

There is no such thing as a free lunch. What renders Markovian stochastic processes and control problems so fascinating is broadly what makes them difficult to study. Uncertain, intrinsically noisy dynamics, even if governed by conceptually simple principles, can give rise to rich and complex behaviors. While this endows stochastic processes with the capacity of modeling phenomena of remarkable complexity and so motivates myriads of established and nascent use cases in science and engineering, it in turn also complicates quantitative analyses in various ways. In this thesis, we have presented new mathematical techniques alongside computational tools for bounding the statistics of controlled jump-diffusion processes and dynamical low-rank approximation. These contributions address two kinds of such complications: certification and scale. As demonstrated by quantifying the limits of quantum control, our bounding techniques may produce witnesses of fundamental limitations, certificates of optimality or robustness, and performance targets for challenging Markov control problems. The developed tools for dynamical low-rank approximation in turn offer a way forward in analyzing otherwise exceedingly large matrix-valued dynamical systems as commonly encountered in the study of stochastic partial differential equations by extracting and tracking only their dominant features.

Our contributions line up with many efforts aimed at unlocking the full potential of stochastic Markov process models to support scientific and engineering activities at large. One of the major challenges of this vision is bringing methodological advances to practitioners and their applications. Only when new methods are tested on real, meaningful problems can we as a scientific community hope to establish a positive feedback loop where those developing methods cater to the problems faced by practitioners using them. A glaring obstacle in this pursuit is that techniques for analyzing stochastic processes tend to build on a complicated web of different mathematical and computational domains and the connections between them. As such, new methods are difficult to deploy. It appears natural to take advantage of the emergence of performant, high-level programming languages to overcome this obstacle. Programming languages like Julia allow us to hide sophisticated and complicated methods behind approachable interfaces at an all time low effort and without compromising on performance. We hope to lead by example in making the methods developed in this thesis available in this manner.

Bibliography

[1] A. Einstein, “Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen,” *Annalen der Physik*, vol. 4, pp. 549–560, 1905.

[2] M. v. Smoluchowski, “Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen,” *Annalen der Phys.*, vol. 21, pp. 756–780, 1906.

[3] P. Langevin, “Sur la théorie du mouvement brownien,” *Compt. Rendus*, vol. 146, pp. 530–533, 1908.

[4] A. Genthon, “The concept of velocity in the history of brownian motion: From physics to mathematics and back,” *The European Physical Journal H*, vol. 45, no. 1, pp. 49–105, 2020.

[5] W. H. Fleming and D. Vermes, “Convex duality approach to the optimal control of diffusions,” *SIAM Journal on Control and Optimization*, vol. 27, no. 5, pp. 1136–1155, 1989.

[6] A. G. Bhatt and V. S. Borkar, “Occupation Measures for Controlled Markov Processes: Characterization and Optimality,” *The Annals of Probability*, vol. 24, no. 3, pp. 1531–1562, 1996.

[7] D. T. Gillespie, “Approximate accelerated stochastic simulation of chemically reacting systems,” *Journal of Chemical Physics*, vol. 115, no. 4, pp. 1716–1733, 2001.

[8] A. Ale, P. Kirk, and M. P. H. Stumpf, “A general moment expansion method for stochastic kinetic models,” *The Journal of Chemical Physics*, vol. 138, no. 17, p. 174 101, 2013.

[9] M. J. Keeling, “Multiplicative moments and measures of persistence in ecology,” *Journal of Theoretical Biology*, vol. 205, no. 2, pp. 269–281, 2000.

[10] I. Nåsell, “An extension of the moment closure method,” *Theoretical Population Biology*, vol. 64, no. 2, pp. 233–239, 2003.

[11] P. Smadbeck and Y. N. Kaznessis, “A closure scheme for chemical master equations,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 35, pp. 14 261–14 265, 2013.

[12] K. R. Ghusinga, C. A. Vargas-Garcia, A. Lamperski, and A. Singh, “Exact lower and upper bounds on stationary moments in stochastic biochemical systems,” *Physical Biology*, vol. 14, no. 4, 04LT01, 2017.

[13] Y. Sakurai and Y. Hori, “A convex approach to steady state moment analysis for stochastic chemical reactions,” in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, IEEE, 2017, pp. 1206–1211.

[14] G. R. Dowdy and P. I. Barton, “Bounds on stochastic chemical kinetic systems at steady state,” *The Journal of Chemical Physics*, vol. 148, no. 8, p. 84 106, 2018.

[15] M. Backenköhler, L. Bortolussi, and V. Wolf, “Bounding First Passage Times in Chemical Reaction Networks,” in *International Conference on Computational Methods in Systems Biology*, Springer, 2019, pp. 379–382.

[16] F. Holtorf and P. I. Barton, “Tighter bounds on transient moments of stochastic chemical systems,” *Journal of Optimization Theory and Applications*, vol. 200, no. 1, pp. 104–149, 2024.

[17] B. Munsky and M. Khammash, “The finite state projection algorithm for the solution of the chemical master equation,” *The Journal of Chemical Physics*, vol. 124, no. 4, p. 44 104, 2006.

[18] A. Gupta, J. Mikelson, and M. Khammash, “A finite state projection algorithm for the stationary solution of the chemical master equation,” *The Journal of Chemical Physics*, vol. 147, no. 15, 2017.

[19] L. Mandelstam and I. Tamm, “The uncertainty relation between energy and time in nonrelativistic quantum mechanics,” *J. Phys.(USSR)*, vol. 9, p. 249, 1945.

[20] N. Margolus and L. B. Levitin, “The maximum speed of dynamical evolution,” *Physica D: Nonlinear Phenomena*, vol. 120, no. 1-2, pp. 188–195, 1998.

[21] H. Zhang, Z. Kuang, S. Puri, and O. D. Miller, “Conservation-law-based global bounds to quantum optimal control,” *Physical Review Letters*, vol. 127, no. 11, p. 110 506, 2021.

[22] M. H. Naderi and H. Babaee, “Adaptive sparse interpolation for accelerating nonlinear stochastic reduced-order modeling with time-dependent bases,” *Computer Methods in Applied Mechanics and Engineering*, vol. 405, p. 115 813, 2023.

[23] H. Whitney, “Analytic extensions of differentiable functions defined in closed sets,” *Hassler Whitney Collected Papers*, pp. 228–254, 1992.

[24] R. Durrett, *Probability: Theory and Examples*. Cambridge University Press, 2019, vol. 5.

[25] E. Kowalski, “Measure and Integral,” ETH Zürich, Tech. Rep., 2011.

[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[27] G. Blekherman, P. A. Parrilo, and R. R. Thomas, *Semidefinite optimization and convex algebraic geometry*. SIAM, 2012.

[28] R. Freund, “Introduction to Semidefinite Programming (SDP),” Massachusetts Institute of Technology, Tech. Rep., 2004, pp. 1–54.

[29] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory.* SIAM, 1994.

[30] P. A. Parrilo, “Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization,” Ph.D. dissertation, California Institute of Technology, 2000, p. 117.

[31] M. Goemens and F. Rendl, “Combinatorial Optimization,” in *Handbook of Semidefinite Programming: Theory, Algorithms and Applications*, H. Wolkowicz, R. Saigal, and L. Vandenberghe, Eds., Springer, 2000, pp. 343–360.

[32] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming.* SIAM, 1994.

[33] E. Andersen and K. Andersen, “The MOSEK interior point optimizer for linear programming: an implementation of the homogeneous algorithm,” in *High Performance Optimization*, Springer, 2000, pp. 197–232.

[34] K.-C. Toh, M. J. Todd, and R. H. Tütüncü, “SDPT3—a MATLAB software package for semidefinite programming, version 1.3,” *Optimization Methods and Software*, vol. 11, no. 1-4, pp. 545–581, 1999.

[35] J. F. Sturm, “Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones,” *Optimization Methods and Software*, vol. 11, no. 1-4, pp. 625–653, 1999.

[36] A. Shapiro, “On duality theory of conic linear problems,” *Nonconvex Optimization and its Applications*, vol. 57, pp. 135–155, 2001.

[37] P. Nash and E. J. Anderson, *Linear programming in infinite-dimensional spaces: theory and applications.* Wiley, 1987.

[38] D. Hilbert, “Ueber die Darstellung definiter Formen als Summe von Formenquadraten,” *Mathematische Annalen*, vol. 32, no. 3, pp. 342–350, 1888.

[39] G. Strang, *Introduction to Linear Algebra*, 5th ed. 2016.

[40] M. Putinar, “Positive polynomials on compact semi-algebraic sets,” *Indiana University Mathematics Journal*, vol. 42, no. 3, pp. 969–984, 1993.

[41] M. Marshall, “Positive polynomials and sums of squares,” in American Mathematical Soc., 2008.

[42] J. B. Lasserre, *Moments, Positive Polynomials and Their Applications.* World Scientific, 2010, vol. 1.

[43] M. Putinar, “Jean Bernard Lasserre: Moments, Positive Polynomials and Their Applications,” *Foundations of Computational Mathematics*, vol. 11, no. 4, pp. 489–497, 2011.

[44] J. B. Lasserre, “Global Optimization with Polynomials and the Problem of Moments,” *SIAM Journal on Optimization*, vol. 11, no. 3, pp. 796–817, 2001.

[45] F. Holtorf, A. Edelman, and C. Rackauckas, "Stochastic Optimal Control via Local Occupation Measures," *arXiv:2211.15652v2*, 2024.

[46] J. B. Lasserre, D. Henrion, C. Prieur, and E. Trélat, "Nonlinear Optimal Control via Occupation Measures and LMI-Relaxations," *SIAM Journal on Control and Optimization*, vol. 47, no. 4, pp. 1643–1666, 2008.

[47] D. Henrion, J. B. Lasserre, and C. Savorgnan, "Nonlinear optimal control synthesis via occupation measures," in *2008 47th IEEE Conference on Decision and Control*, IEEE, 2008, pp. 4749–4754.

[48] V. Gaitsgory and M. Quincampoix, "Linear programming approach to deterministic infinite horizon optimal control problems with discounting," *SIAM Journal on Control and Optimization*, vol. 48, no. 4, pp. 2480–2512, 2009.

[49] D. Henrion, M. Korda, M. Kružík, and R. Rios-Zertuche, "Occupation measure relaxations in variational problems: The role of convexity," *arXiv:2303.02434*, 2023.

[50] M. Korda, D. Henrion, and J. B. Lasserre, "Moments and convex optimization for analysis and control of nonlinear PDEs," in *Handbook of Numerical Analysis*, vol. 23, Elsevier, 2022, pp. 339–366.

[51] D. Henrion, M. Infusino, S. Kuhlmann, and V. Vinnikov, "Infinite-dimensional moment-SOS hierarchy for nonlinear partial differential equations," *arXiv:2305.18768*, 2023.

[52] O. Hernández-Lerma, J. B. Lasserre, O. Hernández-Lerma, and J. B. Lasserre, "The linear programming approach," *Further Topics on Discrete-Time Markov Control Processes*, pp. 203–249, 1999.

[53] C. Savorgnan, J. B. Lasserre, and M. Diehl, "Discrete-time stochastic optimal control via occupation measures and moment relaxations," in *2009 48th IEEE Conference on Decision and Control*, IEEE, 2009, pp. 519–524.

[54] K. Helmes and R. Stockbridge, "Numerical comparison of controls and verification of optimality for stochastic control problems," *Journal of Optimization Theory and Applications*, vol. 106, pp. 107–127, 2000.

[55] M. J. Cho and R. H. Stockbridge, "Linear programming formulation for optimal stopping problems," *SIAM Journal on Control and Optimization*, vol. 40, no. 6, pp. 1965–1982, 2002.

[56] A. Lamperski, K. R. Ghusinga, and A. Singh, "Analysis and control of stochastic systems using semidefinite programming over moments," *IEEE Transactions on Automatic Control*, vol. 64, no. 4, pp. 1726–1731, 2018.

[57] D. Henrion, M. Junca, and M. Velasco, "Moment-SOS hierarchy and exit time of stochastic processes," *arXiv:2101.06009*, 2021.

[58] F. Holtorf, F. Schäfer, J. Arnold, C. Rackauckas, and A. Edelman, "Performance bounds for quantum control," *arXiv:2304.03366*, 2023.

[59] R. Vinter, "Convex duality and nonlinear optimal control," *SIAM Journal on Control and Optimization*, vol. 31, no. 2, pp. 518–538, 1993.

[60] T. G. Kurtz and R. H. Stockbridge, "Existence of Markov Controls and Characterization of Optimal Markov Controls," *SIAM Journal on Control and Optimization*, vol. 36, no. 2, pp. 609–653, 1998.

[61] S. Shin, V. M. Zavala, and M. Anitescu, "Decentralized Schemes with Overlap for Solving Graph-Structured Optimization Problems," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 3, pp. 1225–1236, Sep. 2020.

[62] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, 2010.

[63] V. Cibulka, M. Korda, and T. Haniš, "Spatio-temporal decomposition of sum-of-squares programs for the region of attraction and reachability," *IEEE Control Systems Letters*, vol. 6, pp. 812–817, 2021.

[64] C. Riener, T. Theobald, L. J. Andrén, and J. B. Lasserre, "Exploiting symmetries in SDP-relaxations for polynomial optimization," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 122–141, 2013.

[65] N. Augier, D. Henrion, M. Korda, and V. Magron, "Symmetry reduction and recovery of trajectories of optimal control problems via measure relaxations," *arXiv:2307.03787*, 2023.

[66] C. Schlosser and M. Korda, "Sparse moment-sum-of-squares relaxations for nonlinear dynamical systems with guaranteed convergence," pp. 1–34, 2020.

[67] J. Wang, C. Schlosser, M. Korda, and V. Magron, "Exploiting term sparsity in moment-sos hierarchy for dynamical systems," *IEEE Transactions on Automatic Control*, 2023.

[68] Y. Zheng, G. Fantuzzi, and A. Papachristodoulou, "Sparse sum-of-squares (SOS) optimization: A bridge between DSOS/SDSOS and SOS optimization for sparse polynomials," in *2019 American Control Conference (ACC)*, IEEE, 2019, pp. 5513–5518.

[69] A. A. Ahmadi and A. Majumdar, "DSOS and SDSOS optimization: LP and SOCP-based alternatives to sum of squares optimization," in *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, IEEE, 2014, pp. 1–5.

[70] A. A. Ahmadi, S. Dash, and G. Hall, "Optimization over structured subsets of positive semidefinite matrices via column generation," *Discrete Optimization*, vol. 24, pp. 129–151, 2017.

[71] A. A. Ahmadi and G. Hall, "On the construction of converging hierarchies for polynomial optimization based on certificates of global positivity," *Mathematics of Operations Research*, vol. 44, no. 4, pp. 1192–1207, 2019.

[72] A. A. Ahmadi and A. Majumdar, “DSOS and SDSOS Optimization: More Tractable Alternatives to Sum of Squares and Semidefinite Optimization,” *SIAM Journal on Applied Algebra and Geometry*, vol. 3, no. 2, pp. 193–230, 2019.

[73] B. Øksendal and A. Sulem, *Applied Stochastic Control of Jump Diffusions*, 2nd ed. Springer, 2007.

[74] B. Oksendal, *Stochastic Differential Equations*, 5th ed. New York: Springer, 2003.

[75] M. Tacchi, “Convergence of Lasserre’s hierarchy: the general case,” *Optimization Letters*, vol. 16, no. 3, pp. 1015–1033, 2022.

[76] S. Jiang, B. Natura, and O. Weinstein, “A faster interior-point method for sum-of-squares optimization,” *Algorithmica*, pp. 1–42, 2023.

[77] T. Weisser, B. Legat, C. Coey, L. Kapelevich, and J. P. Vielma, “Polynomial and moment optimization in Julia and JuMP,” in *JuliaCon*, 2019.

[78] B. Legat, O. Dowson, J. D. Garcia, and M. Lubin, “MathOptInterface: a data structure for mathematical optimization problems,” *INFORMS Journal on Computing*, vol. 34, no. 2, pp. 672–689, 2022.

[79] D. T. Gillespie, “A rigorous derivation of the chemical master equation,” *Physica A: Statistical Mechanics and its Applications*, vol. 188, no. 1-3, pp. 404–425, 1992.

[80] L. Breuer, *From Markov jump processes to spatial queues.* Springer Science & Business Media, 2003.

[81] J. Kuntz, P. Thomas, G.-B. Stan, and M. Barahona, “Bounding the stationary distributions of the chemical master equation via mathematical programming,” *The Journal of Chemical Physics*, vol. 151, no. 3, p. 34 109, 2019.

[82] D. Schnoerr, G. Sanguinetti, and R. Grima, “Approximation and inference methods for stochastic biochemical kinetics—a tutorial review,” *Journal of Physics A: Mathematical and Theoretical*, vol. 50, no. 9, p. 093 001, 2017.

[83] A. Arkin, J. Ross, and H. H. McAdams, “Stochastic kinetic analysis of developmental pathway bifurcation in phage $λ$-infected Escherichia coli cells,” *Genetics*, vol. 149, no. 4, pp. 1633–1648, 1998.

[84] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, “Stochastic gene expression in a single cell,” *Science*, vol. 297, no. 5584, pp. 1183–1186, 2002.

[85] Q. Liu and Y. Jia, “Fluctuations-induced switch in the gene transcriptional regulatory system,” *Physical Review E*, vol. 70, no. 4, p. 41 907, 2004.

[86] M. N. Artyomov, J. Das, M. Kardar, and A. K. Chakraborty, “Purely stochastic binary decisions in cell signaling models without underlying deterministic bistabilities,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 48, pp. 18 958–18 963, 2007.

[87] A. Eldar and M. B. Elowitz, "Functional roles for noise in genetic circuits," *Nature*, vol. 467, no. 7312, pp. 167–173, 2010.

[88] D. T. Gillespie, "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions," *Journal of Computational Physics*, vol. 22, no. 4, pp. 403–434, 1976.

[89] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[90] Y. Sakurai and Y. Hori, "Interval analysis of worst-case stationary moments for stochastic chemical reactions with uncertain parameters," *Automatica*, vol. 146, p. 110 647, 2022.

[91] D. Schnoerr, G. Sanguinetti, and R. Grima, "Comparison of different moment-closure approximations for stochastic chemical kinetics," *The Journal of Chemical Physics*, vol. 143, no. 18, p. 185 101, 2015.

[92] D. Schnoerr, G. Sanguinetti, and R. Grima, "Validity conditions for moment closure approximations in stochastic chemical kinetics," *The Journal of Chemical Physics*, vol. 141, no. 8, p. 84 103, 2014.

[93] R. Grima, "A study of the accuracy of moment-closure approximations for stochastic chemical kinetics," *The Journal of Chemical Physics*, vol. 136, no. 15, p. 154 105, 2012.

[94] G. R. Dowdy and P. I. Barton, "Dynamic bounds on stochastic chemical kinetic systems using semidefinite programming," *The Journal of Chemical Physics*, vol. 149, no. 7, p. 74 103, 2018.

[95] Y. Sakurai and Y. Hori, "Bounding transient moments of stochastic chemical reactions," *IEEE Control Systems Letters*, vol. 3, no. 2, pp. 290–295, 2019.

[96] D. Del Vecchio, A. J. Dy, and Y. Qian, "Control theory meets synthetic biology," *Journal of The Royal Society Interface*, vol. 13, no. 120, p. 20 160 380, 2016.

[97] S. I. Resnick, *Adventures in stochastic processes.* Springer Science & Business Media, 1992.

[98] G. R. Dowdy, "Using semidefinite programming to bound distributions in chemical engineering systems," Ph.D. dissertation, Massachusetts Institute of Technology, 2019.

[99] A. Gupta, C. Briat, and M. Khammash, "A scalable computational framework for establishing long-term behavior of stochastic reaction networks," *PLoS Computational Biology*, vol. 10, no. 6, e1003669, 2014.

[100] J. Kuntz, M. Ottobre, G.-B. Stan, and M. Barahona, "Bounding stationary averages of polynomial diffusions via semidefinite programming," *SIAM Journal on Scientific Computing*, vol. 38, no. 6, A3891–A3920, 2016.

[101] F. Schlögl, "Chemical reaction models for non-equilibrium phase transitions," *Zeitschrift für Physik*, vol. 253, no. 2, pp. 147–161, 1972.

[102] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information.* Cambridge university press, 2010.

[103] H. M. Wiseman and G. J. Milburn, *Quantum measurement and control.* Cambridge university press, 2009.

[104] K. Jacobs and D. A. Steck, "A straightforward introduction to continuous quantum measurement," *Contemporary Physics*, vol. 47, no. 5, pp. 279–303, 2006.

[105] W. Rudin, *Functional Analysis*, 2nd ed. McGraw-Hill, 1991.

[106] P. Bushev *et al.*, "Shot-noise-limited monitoring and phase locking of the motion of a single trapped ion," *Physical Review Letters*, vol. 110, no. 13, p. 133 602, 2013.

[107] J. Combes and K. Jacobs, "Rapid state reduction of quantum systems using feedback control," *Physical Review Letters*, vol. 96, no. 1, p. 10 504, 2006.

[108] C. D'Helon and M. James, "Stability, gain, and robustness in quantum feedback networks," *Physical Review A*, vol. 73, no. 5, p. 53 803, 2006.

[109] A. C. Doherty and K. Jacobs, "Feedback control of quantum systems using continuous state estimation," *Physical Review A*, vol. 60, no. 4, p. 2700, 1999.

[110] H. Wiseman and A. Doherty, "Optimal unravellings for feedback control in linear quantum systems," *Physical Review Letters*, vol. 94, no. 7, p. 70 405, 2005.

[111] K. Reuer *et al.*, "Realizing a deep reinforcement learning agent for real-time quantum feedback," *Nature Communications*, vol. 14, no. 1, p. 7138, 2023.

[112] R. Vijay *et al.*, "Stabilizing Rabi oscillations in a superconducting qubit using quantum feedback," *Nature*, vol. 490, no. 7418, pp. 77–80, 2012.

[113] V. Belavkin, "Nondemolition stochastic calculus in Fock space and nonlinear filtering and control in quantum systems," in *Proceedings XXIV Karpacz winter school, Stochastic methods in mathematics and physics*, World Scientific Singapore, 1988, pp. 310–324.

[114] G. Lindblad, "On the generators of quantum dynamical semigroups," *Communications in Mathematical Physics*, vol. 48, pp. 119–130, 1976.

[115] H. Yuen and J. Shapiro, "Optical communication with two-photon coherent states–Part I: Quantum-state propagation and quantum-noise," *IEEE Transactions on Information Theory*, vol. 24, no. 6, pp. 657–668, 1978.

[116] J. Shapiro, H. Yuen, and A. Mata, "Optical communication with two-photon coherent states–Part II: Photoemissive detection and structured receiver performance," *IEEE Transactions on Information Theory*, vol. 25, no. 2, pp. 179–192, 1979.

[117] H. Yuen and J. Shapiro, "Optical communication with two-photon coherent states–Part III: Quantum measurements realizable with photoemissive detectors," *IEEE Transactions on Information Theory*, vol. 26, no. 1, pp. 78–92, 1980.

[118] H. M. Wiseman and G. J. Milburn, "Quantum theory of field-quadrature measurements," *Physical review A*, vol. 47, no. 1, p. 642, 1993.

[119] V. Giovannetti, S. Lloyd, and L. Maccone, "Quantum limits to dynamical evolution," *Physical Review A*, vol. 67, no. 5, p. 052 109, 2003.

[120] P. M. Poggi, F. C. Lombardo, and D. Wisniacki, "Quantum speed limit and optimal evolution time in a two-level system," *Europhysics Letters*, vol. 104, no. 4, p. 40 005, 2013.

[121] A. del Campo, I. L. Egusquiza, M. B. Plenio, and S. F. Huelga, "Quantum speed limits in open system dynamics," *Physical Review Letters*, vol. 110, no. 5, p. 050 403, 2013.

[122] S. Deffner and E. Lutz, "Quantum speed limit for non-Markovian dynamics," *Physical Review Letters*, vol. 111, no. 1, p. 010 402, 2013.

[123] T. Caneva *et al.*, "Optimal control at the quantum speed limit," *Physical Review Letters*, vol. 103, no. 24, p. 240 501, 2009.

[124] S. Deffner and S. Campbell, "Quantum speed limits: from Heisenberg's uncertainty principle to optimal quantum control," *Journal of Physics A: Mathematical and Theoretical*, vol. 50, no. 45, p. 453 001, 2017.

[125] C. Arenz, B. Russell, D. Burgarth, and H. Rabitz, "The roles of drift and control field constraints upon quantum control speed limits," *New Journal of Physics*, vol. 19, no. 10, p. 103 015, 2017.

[126] J. Lee, C. Arenz, H. Rabitz, and B. Russell, "Dependence of the quantum speed limit on system size and control complexity," *New Journal of Physics*, vol. 20, no. 6, p. 63 002, 2018.

[127] N. Khaneja, R. Brockett, and S. J. Glaser, "Time optimal control in spin systems," *Physical Review A*, vol. 63, no. 3, p. 032 308, 2001.

[128] A. Carlini, A. Hosoya, T. Koike, and Y. Okudaira, "Time-optimal quantum evolution," *Physical Review Letters*, vol. 96, no. 6, p. 060 503, 2006.

[129] A. Carlini, A. Hosoya, T. Koike, and Y. Okudaira, "Time-optimal unitary operations," *Physical Review A*, vol. 75, no. 4, p. 042 308, 2007.

[130] G. C. Hegerfeldt, "Driving at the quantum speed limit: Optimal control of a two-level system," *Physical Review Letters*, vol. 111, no. 26, p. 260 501, 2013.

[131] G. C. Hegerfeldt, "High-speed driving of a two-level system," *Physical Review A*, vol. 90, no. 3, p. 032 110, 2014.

[132] M. Abdelhafez, D. I. Schuster, and J. Koch, "Gradient-based optimal control of open quantum systems using quantum trajectories and automatic differentiation," *Physical Review A*, vol. 99, no. 5, p. 052 327, 2019.

[133] V. F. Krotov and I. Feldman, "An iterative method for solving optimal control problems," *Engineering Cybernetics*, vol. 21, pp. 123–130, 1983.

[134] D. M. Reich, M. Ndong, and C. P. Koch, "Monotonically convergent optimization in quantum control using Krotov's method," *The Journal of Chemical Physics*, vol. 136, no. 10, 2012.

[135] M. Goerz *et al.*, "Krotov: A Python implementation of Krotov's method for quantum optimal control," *SciPost Physics*, vol. 7, no. 6, p. 080, 2019.

[136] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrüggen, and S. J. Glaser, "Optimal control of coupled spin dynamics: design of NMR pulse sequences by gradient ascent algorithms," *Journal of Magnetic Resonance*, vol. 172, no. 2, pp. 296–305, 2005.

[137] P. de Fouquieres, S. G. Schirmer, S. J. Glaser, and I. Kuprov, "Second order gradient ascent pulse engineering," *Journal of Magnetic Resonance*, vol. 212, no. 2, pp. 412–417, 2011.

[138] A. Trowbridge, A. Bhardwaj, K. He, D. I. Schuster, and Z. Manchester, "Direct collocation for quantum optimal control," in *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)*, IEEE, vol. 1, 2023, pp. 1278–1285.

[139] F. Schäfer, P. Sekatski, M. Koppenhöfer, C. Bruder, and M. Kloc, "Control of stochastic quantum dynamics by differentiable programming," *Machine Learning: Science and Technology*, vol. 2, no. 3, p. 035 004, 2021.

[140] T. Caneva, T. Calarco, and S. Montangero, "Chopped random-basis quantum optimization," *Physical Review A*, vol. 84, no. 2, p. 022 326, 2011.

[141] H. A. Rabitz, M. M. Hsieh, and C. M. Rosenthal, "Quantum optimally controlled transition landscapes," *Science*, vol. 303, no. 5666, pp. 1998–2001, 2004.

[142] A. Borras, C. Zander, A. Plastino, M. Casas, and A. Plastino, "Entanglement and the quantum brachistochrone problem," *Europhysics Letters*, vol. 81, no. 3, p. 30 007, 2007.

[143] A. Frydryszak and V. Tkachuk, "Quantum brachistochrone problem for a spin-1 system in a magnetic field," *Physical Review A*, vol. 77, no. 1, p. 014 103, 2008.

[144] X. Wang, M. Allegra, K. Jacobs, S. Lloyd, C. Lupo, and M. Mohseni, "Quantum brachistochrone curves as geodesics: Obtaining accurate minimum-time protocols for the control of quantum systems," *Physical Review Letters*, vol. 114, no. 17, p. 170 501, 2015.

[145] C. J. Wood and J. M. Gambetta, "Quantification and characterization of leakage errors," *Physical Review A*, vol. 97, no. 3, p. 032 306, 2018.

[146] S. J. Glaser *et al.*, "Training Schrödinger's cat: Quantum optimal control," *The European Physical Journal D*, vol. 69, no. 12, pp. 1–24, 2015.

[147] S. G. Yalew *et al.*, “Impacts of climate change on energy systems in global and regional scenarios,” *Nature Energy*, vol. 5, no. 10, pp. 794–802, 2020.

[148] A. T. Hoang *et al.*, “Impacts of COVID-19 pandemic on the global energy system and the shift progress to renewable energy: Opportunities, challenges, and policy implications,” *Energy Policy*, vol. 154, p. 112 322, 2021.

[149] R. Lowe and P. Drummond, “Solar, wind and logistic substitution in global energy supply to 2050–barriers and implications,” *Renewable and Sustainable Energy Reviews*, vol. 153, p. 111 720, 2022.

[150] A. Auffeves, “Quantum technologies need a quantum energy initiative,” *PRX Quantum*, vol. 3, no. 2, p. 20 101, 2022.

[151] M. Aifer and S. Deffner, “From quantum speed limits to energy-efficient quantum gates,” *New Journal of Physics*, vol. 24, no. 5, p. 55 002, 2022.

[152] K. Kobzar, T. E. Skinner, N. Khaneja, S. J. Glaser, and B. Luy, “Exploring the limits of broadband excitation and inversion: II. Rf-power optimized pulses,” *Journal of Magnetic Resonance*, vol. 194, no. 1, pp. 58–66, 2008.

[153] S. Deffner and E. Lutz, “Energy–time uncertainty relation for driven quantum systems,” *Journal of Physics A: Mathematical and Theoretical*, vol. 46, no. 33, p. 335 302, 2013.

[154] J. Werschnik and E. Gross, “Quantum optimal control theory,” *Journal of Physics B: Atomic, Molecular and Optical Physics*, vol. 40, no. 18, R175, 2007.

[155] F. Motzoi, J. M. Gambetta, P. Rebentrost, and F. K. Wilhelm, “Simple pulses for elimination of leakage in weakly nonlinear qubits,” *Physical Review Letters*, vol. 103, no. 11, p. 110 501, 2009.

[156] C. P. Koch *et al.*, “Quantum optimal control in quantum technologies. Strategic report on current status, visions and goals for research in Europe,” *EPJ Quantum Technology*, vol. 9, no. 1, p. 19, 2022.

[157] L. Bouten, S. Edwards, and V. Belavkin, “Bellman equations for optimal feedback control of qubit states,” *Journal of Physics B: Atomic, Molecular and Optical Physics*, vol. 38, no. 3, p. 151, 2005.

[158] V. P. Belavkin, A. Negretti, and K. Mølmer, “Dynamical programming of continuously observed quantum systems,” *Physical Review A*, vol. 79, no. 2, p. 022 123, 2009.

[159] V. P. Belavkin, “Nondemolition measurements, nonlinear filtering and dynamic programming of quantum stochastic processes,” in *Modeling and Control of Systems: in Engineering, Quantum Mechanics, Economics and Biosciences Proceedings of the Bellman Continuum Workshop 1988, June 13–14, Sophia Antipolis, France*, Springer, 2006, pp. 245–265.

[160] R. Porotti, A. Essig, B. Huard, and F. Marquardt, “Deep reinforcement learning for quantum state preparation with weak nonlinear measurements,” *Quantum*, vol. 6, p. 747, 2022.

[161] V. P. Belavkin, "Quantum stochastic calculus and quantum nonlinear filtering," *Journal of Multivariate Analysis*, vol. 42, no. 2, pp. 171–201, 1992.

[162] V. S. Bhaskara and P. K. Panigrahi, "Generalized concurrence measure for faithful quantification of multiparticle pure state entanglement using Lagrange's identity and wedge product," *Quantum Information Processing*, vol. 16, no. 5, p. 118, 2017.

[163] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd, "Conic Optimization via Operator Splitting and Homogeneous Self-Dual Embedding," *Journal of Optimization Theory and Applications*, vol. 169, no. 3, pp. 1042–1068, 2016.

[164] M. Garstka, M. Cannon, and P. Goulart, "COSMO: A conic operator splitting method for convex conic problems," *Journal of Optimization Theory and Applications*, vol. 190, no. 3, pp. 779–810, 2021.

[165] C. Coey, L. Kapelevich, and J. P. Vielma, "Solving natural conic formulations with Hypatia.jl," *INFORMS Journal on Computing*, vol. 34, no. 5, pp. 2686–2699, 2022.

[166] I. Dunning, J. Huchette, and M. Lubin, "JuMP: A Modeling Language for Mathematical Optimization," *SIAM Review*, vol. 59, no. 2, pp. 295–320, 2017.

[167] R. Deits, T. Koolen, and R. Tedrake, "LVIS: Learning from value function intervals for contact-aware robot controllers," in *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2019-May, IEEE, May 2019, pp. 7762–7768.

[168] D. Bertsimas and B. Stellato, "The voice of optimization," *Machine Learning*, vol. 110, pp. 249–277, 2021.

[169] J. Kusch and P. Stammer, "A robust collision source method for rank adaptive dynamical low-rank approximation in radiation therapy," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 57, no. 2, pp. 865–891, 2023.

[170] T. Jahnke and W. Huisinga, "A dynamical low-rank approach to the chemical master equation," *Bulletin of mathematical biology*, vol. 70, pp. 2283–2302, 2008.

[171] G. Ceruti, M. Frank, and J. Kusch, "Dynamical low-rank approximation for Marshak waves," *Karlsruhe Institute of Technology, CRC*, vol. 1173, 2022.

[172] L. Einkemmer, A. Ostermann, and C. Piazzola, "A low-rank projector-splitting integrator for the Vlasov–Maxwell equations with divergence correction," *Journal of Computational Physics*, vol. 403, p. 109 063, 2020.

[173] L. Einkemmer, J. Mangott, and M. Prugger, "A low-rank complexity reduction algorithm for the high-dimensional kinetic chemical master equation," *Journal of Computational Physics*, p. 112 827, 2024.

[174] T. P. Sapsis and P. F. Lermusiaux, "Dynamically orthogonal field equations for continuous stochastic dynamical systems," *Physica D: Nonlinear Phenomena*, vol. 238, no. 23-24, pp. 2347–2360, 2009.

[175] J. Kusch, G. Ceruti, L. Einkemmer, and M. Frank, "Dynamical Low Rank Approximation for Burgers' Equation with Uncertainty," *International Journal for Uncertainty Quantification*, vol. 12, no. 5, 2022.

[176] T. Sapsis, M. Ueckermann, and P. F. Lermusiaux, "Global analysis of Navier–Stokes and Boussinesq stochastic flows using dynamical orthogonality," *Journal of fluid mechanics*, vol. 734, pp. 83–113, 2013.

[177] A. A. S. Charous, "High-order retractions for reduced-order modeling and uncertainty quantification," Ph.D. dissertation, Massachusetts Institute of Technology, 2021.

[178] F. Feppon and P. F. Lermusiaux, "Dynamically orthogonal numerical schemes for efficient stochastic advection and Lagrangian transport," *SIAM Review*, vol. 60, no. 3, pp. 595–625, 2018.

[179] E. Vidlicková, "Dynamical low rank approximation for uncertainty quantification of time-dependent problems," Ph.D. dissertation, EPFL, 2022.

[180] J. Schmidt, P. Hennig, J. Nick, and F. Tronarp, "The rank-reduced Kalman filter: Approximate dynamical-low-rank filtering in high dimensions," *Advances in Neural Information Processing Systems*, vol. 36, pp. 61 364–61 376, 2023.

[181] S. Schotthöfer, E. Zangrando, J. Kusch, G. Ceruti, and F. Tudisco, "Low-rank lottery tickets: Finding efficient low-rank neural networks via matrix differential equations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 051–20 063, 2022.

[182] E. Zangrando, S. Schotthöfer, G. Ceruti, J. Kusch, and F. Tudisco, "Rank-adaptive spectral pruning of convolutional layers during training," *arXiv:2305.19059*, 2023.

[183] A. Nonnenmacher and C. Lubich, "Dynamical low-rank approximation: Applications and numerical experiments," *Mathematics and Computers in Simulation*, vol. 79, no. 4, pp. 1346–1357, 2008.

[184] A. Blanchard and T. P. Sapsis, "Analytical description of optimally time-dependent modes for reduced-order modeling of transient instabilities," *SIAM Journal on Applied Dynamical Systems*, vol. 18, no. 2, pp. 1143–1162, 2019.

[185] M. Donello, M. H. Carpenter, and H. Babaee, "Computing sensitivities in evolutionary systems: A real-time reduced order modeling strategy," *SIAM Journal on Scientific Computing*, vol. 44, no. 1, A128–A149, 2022.

[186] O. Koch and C. Lubich, "Dynamical low-rank approximation," *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 2, pp. 434–454, 2007.

[187] P. A. Dirac, "Note on exchange phenomena in the Thomas atom," in *Mathematical proceedings of the Cambridge philosophical society*, Cambridge University Press, vol. 26, 1930, pp. 376–385.

[188] J. Frenkel, "Wave mechanics," *Oxford, Calendron Press*, 1934.

[189] S. Chaturantabut and D. C. Sorensen, "Discrete empirical interpolation for nonlinear model reduction," in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, IEEE, 2009, pp. 4316–4321.

[190] Z. Drmac and S. Gugercin, "A new selection operator for the discrete empirical interpolation method—improved a priori error bound and extensions," *SIAM Journal on Scientific Computing*, vol. 38, no. 2, A631–A648, 2016.

[191] P. Y. Gidisu and M. E. Hochstenbach, "A hybrid DEIM and leverage scores based method for CUR index selection," in *European Consortium for Mathematics in Industry*, Springer, 2021, pp. 147–153.

[192] P. Astrid, S. Weiland, K. Willcox, and T. Backx, "Missing point estimation in models described by proper orthogonal decomposition," *IEEE Transactions on Automatic Control*, vol. 53, no. 10, pp. 2237–2251, 2008.

[193] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, 2017.

[194] M. Cheng, T. Y. Hou, and Z. Zhang, "A dynamically bi-orthogonal method for time-dependent stochastic partial differential equations I: Derivation and algorithms," *Journal of Computational Physics*, vol. 242, pp. 843–868, 2013.

[195] E. Kieri, C. Lubich, and H. Walach, "Discretized dynamical low-rank approximation in the presence of small singular values," *SIAM Journal on Numerical Analysis*, vol. 54, no. 2, pp. 1020–1038, 2016.

[196] E. Kieri and B. Vandereycken, "Projection methods for dynamical low-rank approximation of high-dimensional problems," *Computational Methods in Applied Mathematics*, vol. 19, no. 1, pp. 73–92, 2019.

[197] A. Séguin, G. Ceruti, and D. Kressner, "From low-rank retractions to dynamical low-rank approximation and back," *arXiv:2309.06125*, 2023.

[198] E. Hairer, C. Lubich, and G. Wanner, "Structure-preserving algorithms for ordinary differential equations," *Geometric numerical integration*, vol. 31, 2006.

[199] M. Shub, "Some remarks on dynamical systems and numerical analysis," *Proc. VII ELAM.(L. Lara-Carrero and J. Lewowicz, eds.), Equinoccio, U. Simón Bolıvar, Caracas*, pp. 69–92, 1986.

[200] P.-A. Absil and J. Malick, "Projection-like retractions on matrix manifolds," *SIAM Journal on Optimization*, vol. 22, no. 1, pp. 135–158, 2012.

[201] F. Feppon and P. F. Lermusiaux, "A geometric approach to dynamical model order reduction," *SIAM Journal on Matrix Analysis and Applications*, vol. 39, no. 1, pp. 510–538, 2018.

[202] P.-A. Absil and I. V. Oseledets, "Low-rank retractions: A survey and new results," *Computational Optimization and Applications*, vol. 62, no. 1, pp. 5–29, 2015.

[203] L. N. Trefethen and D. Bau, *Numerical linear algebra*. SIAM, 2022.

[204] P. Benner, S. Gugercin, and K. Willcox, “A survey of projection-based model reduction methods for parametric dynamical systems,” *SIAM Review*, vol. 57, no. 4, pp. 483–531, 2015.

[205] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera, “An ‘empirical interpolation’method: application to efficient reduced-basis discretization of partial differential equations,” *Comptes Rendus Mathematique*, vol. 339, no. 9, pp. 667–672, 2004.

[206] D. C. Sorensen and M. Embree, “A DEIM induced CUR factorization,” *SIAM Journal on Scientific Computing*, vol. 38, no. 3, A1454–A1482, 2016.

[207] B. Peherstorfer, Z. Drmac, and S. Gugercin, “Stability of discrete empirical interpolation and gappy proper orthogonal decomposition with randomized and deterministic sampling points,” *SIAM Journal on Scientific Computing*, vol. 42, no. 5, A2837–A2864, 2020.

[208] M. W. Mahoney and P. Drineas, “CUR matrix decompositions for improved data analysis,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 3, pp. 697–702, 2009.

[209] C. Boutsidis and D. P. Woodruff, “Optimal CUR matrix decompositions,” in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, 2014, pp. 353–362.

[210] C. Lubich and I. V. Oseledets, “A projector-splitting integrator for dynamical low-rank approximation,” *BIT Numerical Mathematics*, vol. 54, no. 1, pp. 171–188, 2014.

[211] G. Ceruti, J. Kusch, and C. Lubich, “A rank-adaptive robust integrator for dynamical low-rank approximation,” *BIT Numerical Mathematics*, vol. 62, no. 4, pp. 1149–1174, 2022.

[212] K. Wright, “Differential equations for the analytic singular value decomposition of a matrix,” *Numerische Mathematik*, vol. 63, pp. 283–295, 1992.

[213] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: A review,” *ACM Computing Surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[214] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4.

[215] F. Cassini and L. Einkemmer, “Efficient 6D Vlasov simulation using the dynamical low-rank framework Ensign,” *Computer Physics Communications*, vol. 280, p. 108 489, 2022.

[216] J. Coughlin, J. Hu, and U. Shumlak, “Robust and conservative dynamical low-rank methods for the vlasov equation via a novel macro-micro decomposition,” *arXiv:2311.09425*, 2023.

[217] C. Patwardhan, M. Frank, and J. Kusch, “Asymptotic-preserving and energy stable dynamical low-rank approximation for thermal radiative transfer equations,” *arXiv:2402.16746*, 2024.

[218] C. Rackauckas and Q. Nie, "Differentialequations. jl–A Performant and Feature-Rich Ecosystem for Solving Differential Equations in Julia," *Journal of open research software*, vol. 5, no. 1, pp. 15–15, 2017.

[219] G. Ceruti and C. Lubich, "An unconventional robust integrator for dynamical low-rank approximation," *BIT Numerical Mathematics*, vol. 62, no. 1, pp. 23–44, 2022.

[220] J. Lin and P. F. Lermusiaux, "Minimum-correction second-moment matching: Theory, algorithms and applications," *Numerische Mathematik*, vol. 147, pp. 611–650, 2021.

[221] S. K. Morley, "Challenges and opportunities in magnetospheric space weather prediction," *Space Weather*, vol. 18, no. 3, e2018SW002108, 2020.

[222] P. Riley and R. Lionello, "Mapping solar wind streams from the Sun to 1 AU: A comparison of techniques," *Solar Physics*, vol. 270, pp. 575–592, 2011.

[223] P. Riley and O. Issan, "Using a heliospheric upwinding extrapolation technique to magnetically connect different regions of the heliosphere," *Frontiers in Physics*, vol. 9, p. 679 497, 2021.

[224] D. J. MacKay *et al.*, "Introduction to Gaussian processes," *NATO ASI series F computer and systems sciences*, vol. 168, pp. 133–166, 1998.

[225] A. Reuther *et al.*, "Interactive supercomputing on 40,000 cores for machine learning and data analysis," in *2018 IEEE High Performance extreme Computing Conference (HPEC)*, IEEE, 2018, pp. 1–6.

[226] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes.* Elsevier, 2014.

Appendix A

Stochastic optimal control via local occupation measures

A.1 Proof of Corollary 1

Proof. Fix $z \in X$ and $t \in [t_{n_T-1}, T]$. Now consider an admissible control process u_s such that all paths of the controlled process (s, x_s, u_s) lie in $[t, T] \times X \times U$ with $x_t \sim \delta_z$. Further define $\tau_0 = t$ and τ_i for $i \geq 1$ to be the minimum between T and the time point at which the process crosses for the i^{th} time from one subdomain of the partition $X_1, \ldots, X_{n_X}$ to another. By construction, the process is confined to some (random) subdomain X_k in the interval $[\tau_i, \tau_{i+1}]$. Since $w_{n_T,k}$ is sufficiently smooth on $[\tau_i, \tau_{i+1}] \times X_k$, Ito's lemma applies and yields that

$$\begin{aligned} w_{n_T,k}(\tau_{i+1}, x_{\tau_{i+1}}) = w_{n_T,k}(\tau_i, x_{\tau_i}) + \int_{\tau_i}^{\tau_{i+1}} \mathcal{A} w_{n_T,k}(s, x_s, u_s) \, \mathrm{d}s \\ + \int_{\tau_i}^{\tau_{i+1}} \nabla_x w_{n_T,k}(s, x_s)^\top g(x_s, u_s) \, \mathrm{d}b_s. \end{aligned}$$

Now note that by Constraint (3.6),

$$\int_{\tau_i}^{\tau_{i+1}} \mathcal{A} w_{n_T,k}(s, x_s, u_s) \, \mathrm{d}s \geq -\int_{\tau_i}^{\tau_{i+1}} \ell(x_s, u_s) \, \mathrm{d}s.$$

Further note that

$$\mathbb{E}_{\delta_z}\left[\int_{\tau_i}^{\tau_{i+1}} \nabla_x w_{n_T,k}(s,x_s)^\top g(x_s,u_s)\,\mathrm{d}b_s\right] = 0$$

as the integrand is square-integrable by Assumption 3.2 and $\tau_i \leq \tau_{i+1}$ are stopping times with respect to the natural filtration [226, Chapter 2, Proposition 1.1]. Thus, after taking expectations, we obtain

$$\mathbb{E}_{\delta_z}\left[w_{n_T,k}(\tau_i, x_{\tau_i})\right] \leq \mathbb{E}_{\delta_z}\left[\int_{\tau_i}^{\tau_{i+1}} \ell(x_s,u_s)\,\mathrm{d}s + w_{n_T,k}(\tau_{i+1}, x_{\tau_{i+1}})\right].$$

Moreover, continuity holds at any crossing between any distinct subdomains X_k and X_j due to Constraint (3.8) such that

$$\mathbb{E}_{\delta_z}\left[w(\tau_i, x_{\tau_i})\right] = \mathbb{E}_{\delta_z}\left[w_{n_T,k}(\tau_i, x_{\tau_i})\right] = \mathbb{E}_{\delta_z}\left[w_{n_T,j}(\tau_i, x_{\tau_i})\right],$$

when the process crosses from X_k to X_j at τ_i. Now using that $\mathbb{E}_{\delta_z}\left[w(\tau_0, x_{\tau_0})\right] = w(t,z)$, we obtain by summing over the time intervals $[\tau_0,\tau_1],\dots,[\tau_N,\tau_{N+1}]$ that

$$w(t,z) \leq \mathbb{E}_{\delta_z}\left[\int_t^{\tau_{N+1}} \ell(x_s,u_s)\,\mathrm{d}s + w(\tau_{N+1}, x_{\tau_{N+1}})\,\mathrm{d}s\right].$$

Letting $N \to \infty$, it follows that

$$w(t,z) \leq \mathbb{E}_{\delta_z}\left[\int_t^{T} \ell(x_s,u_s)\,\mathrm{d}s + w(T, x_T)\right]$$

as $\tau_N \to T$ almost surely. Finally using that $w(T,x) \leq \phi(x)$ on X due to Constraint (3.9) and the fact that all results hold for any admissible control policy, we obtain the desired result $w(t,z) \leq V(t,z)$.

It remains to show that w preserves the lower bounding property across the boundaries introduced by discretization of the time domain. To that end, note that by an analogous argument as before, we have for any $t \in [t_{i-1}, t_i)$ that

$$w(t,z) \leq \mathbb{E}_{\delta_z}\left[\int_t^{t_i} \ell(x_s, u_s)\,\mathrm{d}s + \lim_{s \nearrow t_i} w(s, x_s)\right].$$

Since Constraint (3.7) implies that $\lim_{s \nearrow t_i} w(s,x) \leq w(t_i, x)$ on X, it finally follows by induction that $w(t,z) \leq V(t,z)$ for any $t \in [0,T]$ and $z \in X$. □

Appendix B

Analysis of stochastic reaction systems via local occupation measures

B.1 Direct derivation of local occupation and exchange measures for stochastic reaction systems

We have shown in the main text that suitably defined localized occupation and exchange measures lead to an infinite-dimensional linear program that characterizes the expectations of observables of stochastic reaction systems. Here, we approach the same construction from a different perspective. We show that the localized notions of occupation and exchange measures arise naturally from the CME when considering the time evolution of certain “local” observables.

Let us first recall the setup of Chapter 4: $X_1, \dots, X_{n_X}$ denotes a partition of the state space of the system X, i.e., $X_i \cap X_j = \emptyset$ if $i \neq j$ and $\cup_{i=1}^{n_X} X_i = X$. Combined with a grid of time points $0 = t_0 < t_1 < \cdots < t_{n_T}$, the sets $[t_{i-1}, t_i] \times X_k$ form a partition of the spatio-temporal domain $[0, T] \times X$. Now consider a smooth observable $w \in \mathcal{C}^{1,0}([0,T] \times X)$ and its restriction $\hat{w}(t,x) = \mathbb{1}_{X_k}(x)w(t,x)$ to X_k. By the fundamental theorem of calculus, the evolution of the expectation of $\hat{w}$ between the times t_{i-1} and t_i is given by

$$\sum_{x \in X_k} w(t_i, x)p(t_i, x) - w(t_{i-1}, x)p(t_{i-1}, x) = \int_{[t_{i-1}, t_i]} \sum_{x \in X_k} p(t,x)\frac{\partial w}{\partial t}(t,x) + w(t,x)\frac{\partial p}{\partial t}(t,x)\,\mathrm{d}t.$$

It follows from the CME that

$$\sum_{x \in X_k} w(t,x)\frac{\partial p}{\partial t}(t,x) = \sum_{x \in X_k} w(t,x) \sum_{r=1}^{n_R} a_r(x - \gamma_r)p(t, x - \gamma_r) - a_r(x)p(t,x).$$

This can not quite be identified as the expectation (or Lebesgue integral) of a suitable measures supported on X_k since the inner sum includes terms with respect to states that do lie in X_k. In order to separate these states, we reorder swap the order of summation and shift the summation index by the stoichiometric coefficients γ_r:

$$\sum_{x \in X_k} w(t,x)\frac{\partial p}{\partial t}(t,x) = \sum_{r=1}^{n_R} \sum_{z + \gamma_r \in X_k} w(z + \gamma_r, t)a_r(z)p(t,z) - \sum_{x \in X_k} w(t,x)a_r(x)p(t,x).$$

Now note that the inner summation range of the first sum may be decomposed into three components

$$\{x \in X : x + \gamma_r \in X_k\} = \left(X_k \cup X_k^{r,\mathrm{in}}\right) \setminus X_k^{r,\mathrm{out}}$$

where $X_k^{r,\mathrm{in}}$ and $X_k^{r,\mathrm{out}}$ denote the states outside of X_k that transition into X_k and conversely the states in X_k that leave X_k via reaction r with non-zero rate:

$$X_k^{r,\mathrm{in}} = \{x \in X : x \notin X_k, x + \gamma_r \in X_k, a_r(x) > 0\},$$
$$X_k^{r,\mathrm{out}} = \{x \in X_k : x + \gamma_r \notin X_k, a_r(x) > 0\}.$$

After reordering terms accordingly, we obtain

$$\sum_{x\in X_k} w(t,x)\frac{\partial p}{\partial t}(t,x) = \sum_{x\in X_k}\sum_{r=1}^{n_R} a_r(x)(w(t,x+\gamma_r)-w(t,x))p(t,x) +$$
$$\sum_{r=1}^{n_R}\left(\sum_{x\in X_k^{r,\text{in}}} a_r(x)w(t,x+\gamma_r)p(t,x) - \sum_{x\in X_k^{r,\text{out}}} a_r(x)w(t,x+\gamma_r)p(t,x)\right).$$

In this form, the right-hand side is readily interpreted as the sum of Lebesgue integrals with respect to non-negative measures. To tie the last two terms finally to the state space partition, we recall the notion of “relative neighborhoods” from the main text: the “relative neighborhood” of X_k in X_j denoted by N_{kj} comprises all states in X_j that transition into X_k with non-zero rate; formally,

$$N_{kj} = \{x\in X_j : R_k(x)\neq\emptyset\} \text{ where } R_k(x) = \{r : x+\gamma_r\in X_k \text{ and } a_r(x)>0\}.$$

This allows us to concisely express the states that lead to transition into and out of X_k in a way that is directly tied to the partition:

$$\cup_{r=1}^{n_R} X_k^{r,\text{in}} = \left(\cup_{j=1}^{n_X} N_{kj}\right)\setminus N_{kk} = \text{ all states that enter } X_k \text{ with non-zero rate.}$$
$$\cup_{r=1}^{n_R} X_k^{r,\text{out}} = \left(\cup_{j=1}^{n_X} N_{jk}\right)\setminus N_{kk} = \text{ all states that leave } X_k \text{ with non-zero rate.}$$

As a consequence, we can concisely express

$$\sum_{r=1}^{n_R}\sum_{x\in X_k^{r,\text{in}}} w(t,x+\gamma_r)a_r(x)p(t,x) = \sum_{j\neq k}\sum_{x\in N_{kj}}\sum_{r\in R_k(x)} a_r(x)w(t,x+\gamma_r)p(t,x),$$
$$\sum_{r=1}^{n_R}\sum_{x\in X_k^{r,\text{out}}} a_r(x)w(t,x+\gamma_r)p(t,x) = \sum_{j\neq k}\sum_{x\in N_{jk}}\sum_{r\in R_j(x)} a_r(x)w(t,x+\gamma_r)p(t,x).$$

Overall, it follows from the definition of the infinitesimal generator (cf. Equation

(4.2)) that

$$\begin{aligned}\sum_{x\in X_k} w(t_i,x)p(t_i,x) - \sum_{x\in X_k} w(t_{i-1},x)p(t_{i-1},t) = \\ \int_{[t_{i-1},t_i]} \sum_{x\in X_k} \mathcal{A}w(t,x)p(t,x)\,\mathrm{d}t \\ + \sum_{j\neq k}\int_{[t_{i-1},t_i]} \sum_{x\in N_{kj}} \sum_{r\in R_k(x)} a_r(x)w(t,x+\gamma_r)p(t,x)\,\mathrm{d}t \\ - \sum_{j\neq k}\int_{[t_{i-1},t_i]} \sum_{x\in N_{jk}} \sum_{r\in R_j(x)} a_r(x)w(t,x+\gamma_r)p(t,x)\,\mathrm{d}t.\end{aligned}$$

Clearly, every term above may be interpreted as the Lebesgue integral of a certain observable with respect to a local occupation or exchange measure as defined in the main text.

B.2 Conic reformulation of maximum entropy problems

Recall the maximum entropy problem (regularized-S$_\infty$) introduced in Chapter 4, Section 4.6:

$$\begin{aligned}\sup_{\mu,\pi} \quad & \sum_{k=1}^{n_X} \langle 1,\mu_k\rangle S[q_k] - \langle 1,\mu_k\rangle \log\langle 1,\mu_k\rangle && \text{(max-S}_\infty\text{)}\\ \text{s.t.} \quad & \langle \mathcal{A}w,\mu_k\rangle + \sum_{j=1}^{n_X} \langle \mathcal{F}_k w,\pi_{jk}\rangle - \langle \mathcal{F}_j w,\pi_{kj}\rangle = 0, \quad \forall w\in\mathcal{C}(\bar{X}_k),\ \forall k\in P,\\ & \mu_k\in\mathcal{M}_+(X_k), \quad \forall k\in P,\\ & \pi_{jk}\in\mathcal{M}_+(N_{kj}), \quad \forall (j,k)\in\partial P.\end{aligned}$$

Further recall the definition of the exponential cone.

Definition B.1 (Exponential cone)**.** *The exponential cone is defined as*

$$K_{\exp} = \{x\in\mathbb{R}^3 : x_1 \geq x_2 \exp\frac{x_3}{x_2},\ x_2>0\} \cup \{x\in\mathbb{R}^3 : x_1\geq 0, x_2=0, x_3\leq 0\}.$$

In order to reformulate (regularized-S$_\infty$) as a conic program, we consider the hypograph reformulation of the nonlinear terms in the objective function,

$$
\begin{aligned}
\sup_{\mu,\pi,s} \quad & \sum_{k=1}^{n_X} \langle 1, \mu_k\rangle S[q_k] + s_k \\
\text{s.t.} \quad & \langle \mathcal{A}w, \mu_k\rangle + \sum_{j=1}^{n_X} \langle \mathcal{F}_k w, \pi_{jk}\rangle - \langle \mathcal{F}_j w, \pi_{kj}\rangle = 0, \quad \forall w \in \mathcal{C}(\bar{X}_k),\ \forall k \in P, \\
& \mu_k \in \mathcal{M}_+(X_k), \quad \forall k \in P, \\
& \pi_{jk} \in \mathcal{M}_+(N_{kj}), \quad \forall (j,k) \in \partial P, \\
& s_k \leq -\langle 1, \mu_k\rangle \log\langle 1, \mu_k\rangle, \quad \forall k \in P.
\end{aligned}
$$

and pose the hypograph constraints as exponential cone constraints. To that end, note first that for $\langle 1, \mu_k\rangle > 0$, we have

$$
s_k \leq -\langle 1, \mu_k\rangle \log\langle 1, \mu_k\rangle \iff \langle 1, \mu_k\rangle \exp \frac{s_k}{\langle 1, \mu_k\rangle} \leq 1 \iff (1, \langle 1, \mu_k\rangle, s_k) \in K_{\text{exp}}.
$$

Further note that in the limit $\langle 1, \mu_k\rangle \to 0$, we also have that $\langle 1, \mu_k\rangle \log\langle 1, \mu_k\rangle \to 0$ so that by closedness of K_{exp} the conclusion

$$
s_k \leq -\langle 1, \mu_k\rangle \log \frac{\langle 1, \mu_k\rangle}{|X_k|} \iff (1, \langle 1, \mu_k\rangle, s_k) \in K_{\text{exp}}
$$

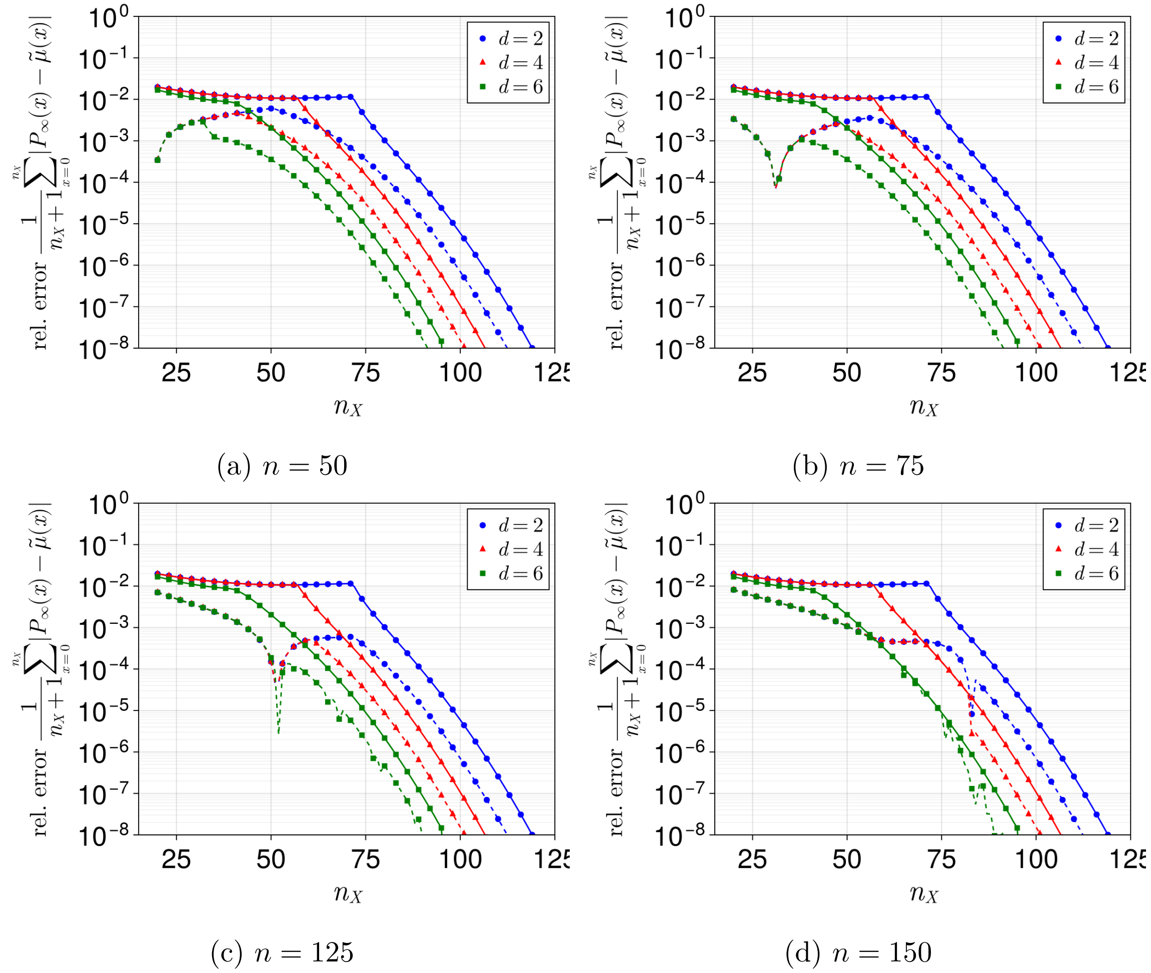
remains valid if $\langle 1, \mu_k\rangle = 0$. It follows that (regularized-S$_\infty$) is equivalent to the infinite dimensional linear program

$$
\begin{aligned}
\sup_{\mu,\pi,s} \quad & \sum_{k=1}^{n_X} \langle 1, \mu_k\rangle S[q_k] + s_k \\
\text{s.t.} \quad & \langle \mathcal{A}w, \mu_k\rangle + \sum_{j=1}^{n_X} \langle \mathcal{F}_k w, \pi_{jk}\rangle - \langle \mathcal{F}_j w, \pi_{kj}\rangle\,, \quad \forall w \in \mathcal{C}(\bar{X}_k),\ \forall k \in P, \\
& \mu_k \in \mathcal{M}_+(X_k), \quad \forall k \in P, \\
& \pi_{jk} \in \mathcal{M}_+(N_{kj}), \quad \forall (j,k) \in \partial P, \\
& (1, \langle 1, \mu_k\rangle, s_k) \in K_{\text{exp}}, \quad \forall k \in P.
\end{aligned}
$$

Upon moment sum-of-squares relaxation of this problem, the exponential cone constraints are preserved as they involve only the zeroth order moments of the measures μ_k.

B.3 Approximations to the stationary distribution of Schlögl's system

The results presented in Chapter 4, Section 4.6.1 are presented for entropy regularization $S[q_{n_X}] = \frac{1}{2} + \frac{1}{2}\log\frac{\pi n}{2}$ with $n = 100$. Here we present results for $n = 50, 75, 125, 150$ to showcase that the advantage persists for a range of regularizations.



(a) $n = 50$

(b) $n = 75$

(c) $n = 125$

(d) $n = 150$

Figure B-1: Errors in stationary measure approximations for Schlögl's system computed via degree-d moment-sum-of-squares relaxations of (regularized-S$_\infty$) (dashed lines) and (local-OM$_\infty$) (solid lines) for different partitions. The entropy regularization for truncated region in (regularized-S$_\infty$) is chosen as $S[q_{n_X}] = \frac{1}{2} + \frac{1}{2}\log\frac{\pi n}{2}$. The objective function in (local-OM$_\infty$) is chosen as $\phi(x) = -x$.