Geometry of Optimization in Markov Decision Processes and Neural Network Based PDE Solvers

Der Fakultät für Mathematik und Informatik
der Universität Leipzig
eingereichte

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet

Mathematik

vorgelegt

von Johannes Christoph Müller, M.Sc. M.Sc.
geboren am 03.06.1995 in Augsburg (Deutschland)

Johannes Christoph Müller – Dissertation abstract

This thesis is divided into two parts dealing with the optimization problems in Markov decision processes (MDPs) and different neural network based numerical solvers for partial differential equations (PDEs).

In Part I we analyze the optimization problem arising in (partially observable) Markov decision processes using tools from algebraic statistics and information geometry, which can be viewed as neighboring fields of applied algebra and differential geometry, respectively. Here, we focus on infinite horizon problems and memoryless stochastic policies. Markov decision processes provide a mathematical framework for sequential decision making on which most current reinforcement learning algorithms are built. They formalize the task of optimally controlling the state of a system through appropriate actions. For fully observable problems, the action can be selected knowing the current state of the system. This case has been studied extensively and optimizing the action selection is known to be equivalent to solving a linear program over the (generalized) stationary distributions of the Markov decision process, which are also referred to as state-action frequencies.

In Chapter 3, we study partially observable problems where an action must be chosen based solely on an observation of the current state, which might not fully reveal the underlying state. We characterize the feasible state-action frequencies of partially observable Markov decision processes by polynomial inequalities. In particular, the optimization problem in partially observable MDPs is described as a polynomially constrained linear objective program that generalizes the (dual) linear programming formulation of fully observable problems. We use this to study the combinatorial and algebraic complexity of this optimization problem and to upper bound the number of critical points over the individual boundary components of the feasible set. Furthermore, we show that our polynomial programming formulation can be used to effectively solve partially observable MDPs using interior point methods, numerical algebraic techniques, and convex relaxations. Gradient-based methods, including variants of natural gradient methods, have gained tremendous attention in the theoretical reinforcement learning community, where they are commonly referred to as (natural) policy gradient methods.

In Chapter 4, we provide a unified treatment of a variety of natural policy gradient methods for fully observable problems by studying their state-action frequencies from the standpoint of information geometry. For a variety of NPGs and reward functions, we show that the trajectories in state-action space are solutions of gradient flows with respect to Hessian geometries, based on which we obtain global convergence guarantees and convergence rates. In particular, we show linear convergence for unregularized and regularized NPG flows with the metrics proposed by Kakade and Morimura and co-authors by observing that these arise from the Hessian geometries of conditional

entropy and entropy respectively. Further, we obtain sublinear convergence rates for Hessian geometries arising from other convex functions like log-barriers. We provide experimental evidence indicating that our predicted rates are essentially tight. Finally, we interpret the discrete-time NPG methods with regularized rewards as inexact Newton methods if the NPG is defined with respect to the Hessian geometry of the regularizer. This yields local quadratic convergence rates of these methods for step size equal to the inverse penalization strength, which recovers existing results as special cases.

Part II addresses neural network-based PDE solvers that have recently experienced a tremendous growth in popularity and attention in the scientific machine learning community. We focus on two approaches that represent the approximation of a solution of a PDE as the minimization over the parameters of a neural network: the deep Ritz method and physically informed neural networks.

In Chapter 5, we study theoretical properties of the boundary penalty for these methods and obtain a uniform convergence result for the deep Ritz method for a large class of potentially nonlinear problems. For linear PDEs, we estimate the error of the deep Ritz method in terms of the optimization error, the approximation capabilities of the neural network, and the strength of the penalty. This reveals a trade-off in the choice of the penalization strength, where too little penalization allows large boundary values and too strong penalization leads to a poor solution of the PDE inside the domain. For physics informed networks, we show that when working with neural networks that have zero boundary values also the second derivatives of the solution are approximated where otherwise only lower order derivatives are approximated.

In Chapter 6, we propose energy natural gradient descent, a natural gradient method with respect to second order information in the the function space, as an optimization algorithm for physics-informed neural networks and the deep Ritz method. We show that this method, which can be interpreted as a generalized Gauss-Newton method, mimics Newton's method in function space except for an orthogonal projection onto the tangent space of the model. We show that for a variety of PDEs, natural energy gradients converge rapidly and approximations to the solution of the PDE are several orders of magnitude more accurate than gradient descent, Adam and Newton's methods, even when these methods are given more computational time.

Authorship

This thesis contains results from various projects that have been completed within different collaborations. Section 2.3 is based on joint work with Guido Montúfar [211] where the rest of Chapter 2 is written solely by myself.

All results in Chapter 3 apart from Subsections 3.2.3, 3.3.3 and 3.4.2 are join work with Guido Montúfar and content of the publication [211] and the extended abstract [210]. Subsections 3.2.3, 3.3.3 and 3.4.2 contain results from a joint project with Mareike Dressler, Marina Garrote-López, Guido Montúfar and Kemal Rose with equal contribution; results are published in [105]. Subsection 3.2.4 is build on ideas developed with Guido Montúfar and Friedrich Wicke.

Chapter 4 is based on joint work with Guido Montúfar; results are published in [209].

Section 5.3 containts joint work with Patrick Dondl and Marius Zeinhofer with equal contributions; results are published in [102]. Section 5.4, 5.5 and Chapter 6 are based on joint work with Marius Zeinhofer with equal contribution. The results are published in [213, 214, 212].

Acknowledgements

First and foremost, I would like to thank my supervisors Nihat Ay and Guido Montúfar for sharing their geometric way of thinking with me, for trusting me and giving me room for unconventional projects, and most importantly, for assuring me of their full support and thus providing me with security. I would like to thank Guido for the countless hours he spent with me discussing all aspects of my projects and career and during which I learned so much.

I want to thank all people that I was fortunate with to collaborate with: Nihat Ay, Patrick Dondl, Mareike Dressler, Marina Garotte-López, Guido Montúfar, Jesse van Oostrum, Kemal Rose and Marius Zeinhofer. Further, I thank all other colleages that I have had the pleasure to work and discuss with, in particular Pradeep Banerjee, Benjamin Bowman, Pierre Bréchet, Katerina Papagiannouli, Hanna Tseran and Rishi Sonthalia, who hosted me so kindly in LA.

I want to thank the members of my thesis advisory committee who are Jürgen Jost, Max von Renesse and Jörg Lehnert. Further, I am grateful to Bernd Sturmfels for connecting me with various people and giving me valuable advice for my academic future.

I would like to express my sincere gratitude to all the staff of the institute that make it the great place it is and offering levels of support that I will probably only understand after leaving, a special shoutout going to Katharina Matschke.

I am grateful to Villigst for supporting my personal development in so many different ways and for providing a place where I have met so many amazing people.

I want to thank my flatmates for providing me with comfort during the times of repeated lockdowns as well as Yunus and Maria for being there and always having an open ear. I am thankful for all the people who provide me with a home in so many different places, especially Öhmchen, Rebecca, Lea, Thomas, Moritz, Jasper, Ana, Hannah, Johanna, the PA crew, and the Paddeln und Planschen group. I am grateful to Patiani for sharing my journey over the years. I want to thank my family and in particular my parents for their continuous and unbounded love and support.

Funding acknowledgement. During my PhD I was financially supported by the International Max Planck Research School *Mathematics in the Sciences*, the Evangelisches Studienwerk Villigst e.V. and received support by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 757983).

Contents

Chapter 1. Introduction 1

1.1 Notation and conventions 7

Part I. Geometry of Markov decision processes 11

Chapter 2. Background on Markov decision processes 12

2.1 State-action frequencies 19

2.2 The advantage function and Bellman optimality 23

2.3 Rational structure of the reward and an explicit line theorem 26

2.4 Solution methods for Markov decision processes 35

Chapter 3. State-action geometry of partially observable MDPs 44

3.1 The state-action polytope of fully observables systems 45

3.2 State-action geometry of partially observable systems 54

3.3 Number and location of critical points 69

3.4 Reward optimization in state-action space (ROSA) 83

Chapter 4. Geometry and convergence of natural policy gradient methods 94

4.1 Natural gradients 96

4.2 Natural policy gradient methods 101

4.3 Convergence of natural policy gradient flows 107

4.4 Locally quadratic convergence for regularized problems 128

4.5 Discussion and outlook 131

Part II. Neural network based PDE solvers 133

Chapter 5. Theoretical analysis of the boundary penalty method for neural network based PDE solvers 134

5.1 Presentation and discussion of the main results 137

5.2 Preliminaries regarding Sobolev spaces and neural networks 146

5.3 Proofs regarding uniform convergence for the deep Ritz method 150

5.4 Proofs of error estimates for the deep Ritz method 156

5.5 Proofs of implications of exact boundary values in residual minimization 167

Chapter 6. Energy natural gradients for neural network based PDE solvers 174

6.1 Energy natural gradients 176

6.2 Experiments 183

6.3 Conclusion and outlook 192

Bibliography 193

CHAPTER 1

Introduction

This thesis is divided into two parts: In Part I, we analyse the optimization problem encountered in (partially observable) Markov decision processes in state-action space using tools from algebraic statistics as well as information geometry, which can be viewed as adjacent fields of applied algebraic and differential geometry, respectively. Part II is concerned with the optimization problems encountered in neural network based approaches for solving partial differential equations. We describe the content as well as the relation between these topics in more detail below.

Part I: Geometry of Markov decision processes. Markov decision processes were established as a mathematical framework for sequential decision making in the late 1950s and early 1960s with Richard Bellman, Ronald A. Howard, David Blackwell and Cyrus Derman being among the early pioneers of this field [46, 45, 142, 55, 94, 54, 56, 93]. Since their introduction they have received considerable attention from the theoretical side. They also lie at the heart of many modern algorithms that led to important advancements in the field of robotics [232] and reinforcement learning [276] which was lately also used in the development of ChatGPT [227]. The task in a *Markov decision process (MDP)* is to control the state s_t of a system through suitable actions a_t in such a way such that the states $s_0, s_1, \dots$ evolve optimally in some notion. An action-selection strategy, which is commonly referred to as a *policy*, consists of the specification of the probability $\pi(a|s)$ to select action a when the system is currently in state s. Following a policy π yields random sequences $S_0, S_1, \dots$ and $A_0, A_1, \dots$ of states and actions. A common criterion for optimality is the infinite horizon reward, which in the mean reward formulation takes the form

$$R(\pi) = \mathbb{E}_\pi\left[\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1} r(S_t, A_t)\right], \tag{1.1}$$

where $r(s, a)$ describes how good the action a is when in state s and $\mathbb{E}_\pi$ denotes the expectation when the actions are selected according to the policy π. For the maximization of the reward function R over all policies, a variety of now classic methods have been developed, most notably value iteration, policy iteration and linear programming approaches, see Section 2.4. Gradient based methods, including variants of natural gradient methods for reward maximization, were pioneered in [277, 40, 39, 153, 206, 208] and have recently gained fast growing attention from the theoretical reinforcement learning community. In Chapter 4 we provide an overview of policy gradient methods and study them from a stand point of information geometry.

Karl Johan Åström extended the framework of MDPs to model for sequential decision problems with uncertainty about the state s, which led to the notion of *partially observable Markov decision processes (POMDPs)* [22]. Here, the action a_t is selected based on an observation o_t that is made from the state s_t. Intuitively, it is harder to make optimal decisions when there is uncertainty about the underlying state [180]. This is also reflected in the computational complexity of the optimization problem, which is NP-hard [228, 286] whereas fully observable MDPs can be solved in polynomial time [309]. In many applications the underlying state is not known exactly when selecting an action. For example a human or a robot has to act purely on its sensations or rather the history of sensations. In partially observable environments it is beneficial to base the decision on all previous observations, or a sufficient statistic thereof, as they might reveal more information about the current state of the system compared to the most recent observation. Various techniques for the optimization of such decision rules have been proposed, which are often based on the formulation of the underlying belief state MDP, see for example articles [272, 254] for surveys of these approaches. However, the storage of the entire sequence of observations would require infinite memory, which is infeasible to implement in practice and the optimization of history dependent policies is known to be undecidable for infinite horizons [186, 73]. Therefore, the maximization of the reward function in POMDPs within the smaller class of policies with finite or no memory[1] was posed as an important open problem at the 29th Conference on Learning Theory (COLT 2016) [25]. When optimizing memoryless policies in a POMDP the methods developed for believe state MDPs and MDPs in general are not applicable and very few approaches exist, see Section 2.4. Chapter 4 is dedicated to the study of exactly this optimization problem from the standpoint of algebraic statistics.

Throughout this thesis we analyse Markov decision processes via their state-action frequencies, which generalize the concept of stationary distributions of the underlying Markov process. For example, the mean reward $R(\pi)$ of a policy π can be computed according to

$$R(\pi) = \sum_{s,a} r(s,a)\eta^\pi(s,a), \tag{1.2}$$

where η^π is the stationary distribution over states and actions, referred to as the state-action frequency, which is given by

$$\eta^\pi(s,a) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{P}^\pi(S_t = s, A_t = a), \tag{1.3}$$

where $\mathbb{P}^\pi(S_t = s, A_t = a)$ denotes the probability that at time t the system is in state s and action a is selected when following the policy π. Since the reward of a policy is a linear function of the stationary distribution induced by the policy, the maximization of the reward R can be studied and solved over the stationary distributions. We refer to this ansatz as reward optimization in state-action space (or shortly ROSA) as the stationary distributions are probability distributions over pairs of states and actions. The objective in state-action space is linear and hence the complexity of the optimization problem is determined by the geometry of the set of stationary distributions. The geometry of this set was first studied in systematic fashion by Cyrus Derman who showed that for fully observable MDPs it is given by a polytope [93]. For partially observable problems a

[1] Note that finite memory can be augmented into the environment and hence often memoryless policies are studied, see for example [179, 231, 145]

decomposition of the set of feasible stationary distributions into infinitely many convex pieces with dimensions controlled by the degree of observability of the model has been given in [203], however, a more explicit description remained elusive.

In Chapter 2, we provide a self-contained introduction and overview over existing solution methods; in addition we perceive the reward function as a rational function in the entries of the policy and bound the degree of this rational function in terms of the observation mechanism where the degree essentially corresponds to the number of states that can lead to a particular observation, see Theorem 2.25. We use this to establish an explicit version of the line theorem due to [81], see Theorem 2.28, and to provide an algebraic proof for the existence of optimal policies, which are deterministic on all observations that identify the state, which has previously been shown in [203], see Theorem 2.30. In Section 2.4 we give an overview of classical solution methods of MDPs with emphasis on value iteration, policy iteration and linear programming. In particular, we bound the number of iterations required by value iteration and policy iteration to generate an optimal policy in terms of the minimum distance of the value function of suboptimal deterministic policies to the optimal value function, see Theorem 2.50 and Theorem 2.52, respectively. This bound depends on the geometry of the set of value functions rather than the size of the Markov decision process, as is the case with existing bounds, and cannot be derived from, nor does it imply, such bounds.

In Chapter 3, we study the geometry of the set of (generalized) stationary distribution, i.e., state-action frequencies, of a partially observable Markov decision process. In Section 3.2 we extend the classic result by Cyrus Derman [93] who characterized the stationary distributions of fully observable MDPs as a polytope and the analysis in [203] by providing a characterization based on polynomial inequalities of the feasible state-action frequencies inside this polytope. This characterization is formulated for general polynomially constrained policy models, see Theorem 3.18, where we derive explicit expressions of the defining polynomial inequalities under a rank assumption on the observation kernel, see Subsection 3.2.2. For deterministic observations we show that the feasible state-action frequencies are given by the intersection of a product of determinantal varieties of rank one matrices with the polytope of state-action frequencies, see Theorem 3.25. Further, we study the feasible state-action frequencies of multi-agent problems and provide explicit characterizations via polynomial conditions, see Subsection 3.2.4. Our description of feasible state-action frequencies yields a reformulation of the reward optimization problem as a linear objective problem over a polynomially constrained subset of the simplex, which can be regarded as a generalization of the linear programming formulation of MDPs. Using tools from applied algebraic geometry we describe the combinatorial and algebraic complexity of the optimization problem. We obtain upper bounds on the number of critical points of the reward function over the individual faces of the domain of the optimization problem, see Section 3.3. Further, we demonstrate that this reformulation of the reward optimization problem as a linear objective polynomially constrained program can be used to solve POMDPs effectively by the means of interior point methods, numerical algebraic approaches and convex relaxations, see Section 3.4. We find that using interior point methods in state-action space is fast and stable even for large discount factors where many optimizer relying on the policy suffer from instability. A benefit of numerical algebraic techniques and convex relaxation is that they are able to provide globally optimal solutions.

Chapter 4 is concerned with natural policy gradient methods for fully observable Markov decision processes where once again we choose to work in state-action space. We show that the dynamics of Kakade's NPG and Morimura's NPG solve a gradient flow with respect to the Hessian geometries of conditional entropic and entropic regularization of the reward (Sections 4.2.2 and 4.2.3 and Proposition 4.13). Leveraging results on gradient flows in Hessian geometries, we derive linear convergence rates for Kakade's and Morimura's NPG flow for the unregularized reward, which is a linear and hence not strictly concave function in state-action space, and also for regularized reward, see Theorems 4.26 and 4.27 and Corollaries 4.33 and 4.34. Further, for a class of NPG methods, which correspond to β-divergences and which generalize Morimura's NPG, we show sub-linear convergence in the unregularized case and linear convergence in the regularized case, see Theorem 4.27 and Corollary 4.34, respectively. For an overview of the convergence rates established in this work see Table 4.1 in Section 4.5. We complement our theoretical analysis with experimental evaluation, which indicates that the established linear and sub-linear rates for unregularized problems are essentially tight. For discrete-time gradient optimization, our ansatz in state-action space yields an interpretation of the regularized NPG method as an inexact Newton iteration if the step size is equal to the inverse regularization strength. This yields a relatively short proof for the local quadratic convergence of regularized NPG methods with Newton step sizes, see Theorem 4.36. This recovers as a special case the local quadratic convergence of Kakade's NPG under state-wise entropy regularization previously shown in [71].

Part II: Neural network based PDE solvers. The second part of this thesis is devoted to both the theoretical aspects of as well as development of a natural gradient method for neural network based numerical solvers for partial differential equations (PDEs). Here, we consider the so called *deep Ritz method (DRM)* as well as *physics informed neural networks (PINNs)* [110, 237]. For both methods, it is the goal to approximate the solution u^* of a PDE by a function u_θ computed by some neural network with parameters θ. Consider the Poisson equation

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{1.4}$$

where $\Omega \subseteq \mathbb{R}^d$ and $f\colon \Omega \longrightarrow \mathbb{R}$ is a square integrable function. The objective function used in the deep Ritz method for the optimization of the parameters θ of the neural networks is given by

$$L_{DRM}(\theta) = \frac{1}{2}\int_\Omega |\nabla u_\theta|^2 \mathrm{d}x - \int_\Omega f u_\theta \mathrm{d}x + \lambda \cdot \int_{\partial\Omega} u_\theta^2 \mathrm{d}s \tag{1.5}$$

for some $\lambda > 0$. PINNs work with the following objective function

$$L_{PINN}(\theta) = \int_\Omega |\Delta u_\theta + f|^2 \mathrm{d}x + \lambda \int_{\partial\Omega} u_\theta^2 \mathrm{d}s. \tag{1.6}$$

The deep Ritz method is inspired by the variational formulation of the PDE that characterizes the solution u^* of (1.4) as the unique minimizer of

$$E_{DRM}(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 \mathrm{d}x - \int_\Omega f u \mathrm{d}x \tag{1.7}$$

over all (sufficiently smooth) functions with zero boundary values. In contrast, PINNs use that u^* is uniquely characterized by $E_{PINN}(u) = 0$, where

$$E_{PINN}(u) = \frac{1}{2}\int_\Omega |\Delta u + f|^2 \mathrm{d}x + \lambda \int_{\partial\Omega} u^2 \tag{1.8}$$

for any $\lambda > 0$. In the last years these approaches have enjoyed tremendous attention as they offer the promise of performing well in the numerical approximation of high dimensional problems, which arise in optimal control, financial mathematics and quantum mechanics [129]. However, it has been documented that they often fail to produce highly accurate solutions even for simple problems as (stochastic) gradient descent methods saturate, which limits the adhoc applicability of these approaches for settings that require reliable solutions. This observation is what sparked our interest in this topic and therefore it is our main motivation to contribute to the theoretical understanding of the different factors influencing the error of these method as well as the development of efficient optimizers.

In Chapter 5, we study theoretical properties of these methods and obtain a uniform convergence result for the deep Ritz method for a large class of potentially nonlinear problems, see Theorem 5.1. For linear PDEs, we establish an error estimate, which informally reads as

$$\|u_\theta - u^*\|_{H^1(\Omega)} \lesssim \sqrt{\text{opt. error} + \lambda \cdot \text{appr. error}} + \lambda^{-1}, \tag{1.9}$$

where opt. error denotes the error of the optimization process of the parameters of the neural network and appr. error denotes the approximation error of the used network, see Theorem 5.3. This reveals the trade-off in the choice of the penalization strength λ where too little penalization allows large boundary values and too strong penalization leads to a poor solution of the PDE inside the domain. If f is r-times differentiable for this implies that for $n \in \mathbb{N}$, there is a ReLU network with $\mathcal{O}(\log_2^2(n^{(r+2)/d}) \cdot n)$ parameters such that if $\lambda_n \sim n^\sigma$ for $\sigma = \frac{2r+3}{2d}$ one has for any $\rho < \frac{2r+3}{4d}$ that

$$\|u_{\theta_n} - u_f\|_{H^1(\Omega)} \lesssim \sqrt{\text{opt. error} + n^{-2\rho}} + n^{-\rho} \quad \text{for all } \theta_n \in \Theta_n, \tag{1.10}$$

see Theorem 5.4. Note that the solution $u_f \in H^{r+2}(\Omega)$ can be approximated at a rate of $O(n^{-(r+1)/d})$, see Theorem 5.11, which is a faster rate than the rate $O(n^{-\rho})$ in the bound (1.10) for the deep Ritz method with successful training. For physics informed networks, we show that when working with neural networks that have zero boundary values also the second derivatives of the solution u^* are approximated where otherwise only lower order derivatives are approximated at the same rate, see Theorem 5.5 and Theorem 5.6, respectively.

Since the theoretical results ensure that successful optimization yields approximate solutions of the PDE we turn towards the design of efficient optimization schemes in Chapter 6. Here, we draw inspiration from natural policy gradient methods, which provide a locally quadratic convergence (for regularized reward). In the context of the deep Ritz method and PINNs this leads to a natural gradient induced by the Hessian geometry of the function space objective, which is often referred to as the energy. We call this method, which can also be interpreted as a generalized Gauß-Newton method, *energy natural gradient descent* and show it mimics Newton's method in function space except for an orthogonal projection onto the tangent space of the model, see Theorem 4.2. We demonstrate that for a variety of PDEs energy natural gradients converge fast and produce approximations to the solution of the PDE several orders of magnitude more accurate than gradient descent, Adam and Newton's method, see Section 6.2. As the low accuracy of these methods when directly optimized is regarded as a major challenge we believe that energy natural gradients can represent an important step in the development of neural network based PDE solvers.

Outlook and open questions. At the end of most sections and chapters we provide a conclusion and collect directions for future research. Here, we collect the most important directions for future research.

In Chapter 3, we describe the state-action frequencies of a POMDP as a polynomially constrained subset within the probability simplex connecting POMDPs to algebraic statistics. We use this description to describe the combinatorial and algebraic complexity of the optimization problem and to solve POMDPs using interior point methods and numerical algebriac techniques. During our work a number of natural questions arose where we list the ones we consider most important here. First, a study for finite memory policies would nicely complement our results and could provide guidelines for the design of memory. While the set of value functions of fully observable problems has been studied and used to design optimal representation, an analysis of the value functions associated to a POMPD remains elusive but would nicely complement our results. Our bounds on the number of critical points could be improved by studying the polar degrees of products of determinantal varities. Where we have described the feasible state-action frequencies of multi-agent MDPs, the optimization problem arising in multi-agent problems has not been studied conclusively. In particular, studying the number of critical points could yield insight into the role of the degree of decentralization for the algebraic complexity of the problem. Further, our analysis of POMDPs could be extended to cover information theoretic objectives like (conditional) entropy and mutual information that play an important role in regularized MDPs and unsupervised pre-training. In our experiments, we observed that the sequence of convex relaxations given by the moment-SOS hierarchy provided exact global solutions in the first order relaxation. We believe that this observation deserves a closer theoretical analysis. Our description of the state-action frequencies of POMDPs with deterministic observations via products of varieties of rank one matrices could provide a starting point for a Riemannian optimization technique for POMDPs

In Chapter 4, we study the geometry of a variety of natural policy gradient methods for fully observable problems, describe the evolution of their state-action frequencies by gradient flows with respect to Riemannian geometries on the state-action polytope and obtain global convergence guarantees for regularized and unregularized problems. Although, this offers a general framework that recovers known results as special cases, there are a couple of future directions that deserve further attention. First, our linear convergence guarantees for regularized problems degrade when the regularization strength decreases, however our experiments indicate that the actual convergence remains linear. This gap could be filled with an improved theoretical analysis. Our experiments indicate that various NPG methods suffer from plateaus, which are induced by the Riemannian geometry on the state-action polytope. The design of methods that reduce the influence of these plateaus could have a great impact in the field of reinforcement learning where policy gradient methods are currently among the most popular approaches. Where we have studied convergence behavior under the assumption of exact gradient evaluations it would be interesting to characterize the number of samples required to estimate the respective notions of natural policy gradients. Finally, a better understanding of the convergence of (natural) policy gradient methods for partially observable problems remains elusive.

Our theoretical analysis of the boundary penalty method in the context of the deep Ritz method and physics informed networks in Chapter 5 depends on the boundary values required to approximate a function where we show that ReLU networks can approximate functions with exact zero boundary values. Hence, a systematic study of the approximation rates of ReLU networks with zero boundary values and required boundary values. Based on our theoretical analysis of the deep Ritz method we suggest a coupling of the penalization strength with the approximation capabilities of the respective network. This suggestion remains to be complemented with an empirical study of different penalization strategies. Where our results describe the error made by the deep Ritz method and physics informed neural networks in terms of the optimization error a conclusive analysis of the convergence behavior of different optimizers remains open.

In Chapter 6 we propose energy natural gradient descent (E-NGD), a natural gradient method with respect to a Hessian-induced Riemannian metric as an optimization algorithm for physics-informed neural networks (PINNs) and the deep Ritz method and demonstrate its ability to produce highly accurate approximations of the solution of the PDE. Important steps in the pursue of efficient neural network based PDE solvers that can be applied at an industrial scale include the following. An efficient implementation of energy natural gradients – possibly in matrix-free fashion – would vastly improve the applicability of physics informed neural networks to large scale and industrial problems. Since the convergence of energy natural gradient descent is sensitive to the initialization we believe that it is important to gain a better understand the behavior of different initialization schemes. We observed Levenberg-Marquardt type modifications of energy natural gradient to reduce the plateaus of the energy natural gradient. Where our choices seemed to work well in practice a systematic procedure for the choice would improve the applicability of energy natural gradients.

1.1 Notation and conventions

Throughout the thesis we highlight new notation by typesetting it in blue color and italic and recall our notation where appropriate. We denote the real numbers by $\mathbb{R}$ and the non-negative and positive real numbers by $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{>0}$, respectively. For two sets $\mathcal{X}$ and $\mathcal{Y}$ we denote the set of mappings from $\mathcal{X}$ to $\mathcal{Y}$ by $\mathcal{Y}^{\mathcal{X}}$, in particular, we denote the free vector space over $\mathcal{X}$ by

$$\mathbb{R}^{\mathcal{X}} = \{(v_x)_{x\in\mathcal{X}} : v_x \in \mathbb{R} \text{ for all } x \in \mathcal{X}\}.$$

When $\mathcal{X}$ is finite we denote the *Euclidean inner product* of two vectors $v, w \in \mathbb{R}^{\mathcal{X}}$ by $\langle v, w\rangle_{\mathcal{X}} = \sum_{x\in\mathcal{X}} v_x w_x$ and write $\|v\|_2$ for its *Euclidean norm* given by $\sqrt{\langle v, v\rangle_{\mathcal{X}}}$, where we sometimes omit the subscript. More generally, for $p \in [1, \infty)$ the *p-norm* of $v \in \mathbb{R}^{\mathcal{X}}$ is given by $\|v\|_p := \sqrt[p]{\sum_{x\in\mathcal{X}} |v_x|^p}$ and the *∞-norm* is given by $\|v\|_\infty := \max_{x\in\mathcal{X}} |v_x|$. When H is a vector space with a scalar product $\langle\cdot,\cdot\rangle$, i.e., a positive definite symmetric bilinear form, we denote the *orthogonal complement* of a subset $S \subseteq H$ by $A^\perp := \{v \in H : \langle v, w\rangle = 0 \text{ for all } w \in S\}$. For a subset $A \subseteq \mathcal{X}$ of a topological space we denote its closure, interior and boundary by $\overline{A}$, $\operatorname{int}(A)$ and ∂A. For a subset $A \subseteq \mathcal{X}$ of a metric space $(\mathcal{X}, d)$ and a point $x \in \mathcal{X}$ we denote the *distance* of x to A by $\operatorname{dist}(x, A) := \inf_{y\in A} d(x, y)$. Note that $\operatorname{dist}(\cdot, A)$ is Lipschitz continuous with constant one.

For a matrix $A \in \mathbb{R}^{n\times m}$, we denote its transpose by $A^\top$ and a pseudoinverse by $A^+ \in \mathbb{R}^{m\times n}$. Note that if A^+ is the Moore-Penrose inverse then AA^+ is the orthogonal (Euclidean) projection onto the range $\operatorname{range}(A) = \{Ax : x \in \mathbb{R}^m\}$ and A^+A is the orthogonal (Euclidean) projection onto the kernel $\ker(A) = \{x \in \mathbb{R}^m : Ax = 0\}$. We denote the set of symmetric and positive definite matrices by $\mathbb{S}^{sym}_{>0}$.

For functions f, g we write $f(t) = O(g(t))$ for $t \to t_0$ if there is a constant $c > 0$ such that $f(t) \le cg(t)$ for $t \to t_0$, where we allow $t_0 = +\infty$.

We denote the simplex of probability distributions over a finite set $\mathcal{X}$ by

$$\Delta_{\mathcal{X}} := \left\{\mu \in \mathbb{R}^{\mathcal{X}}_{\ge 0} : \sum_{x\in\mathcal{X}} \mu_x = 1\right\}.$$

We refer to $\Delta_{\mathcal{X}}$ as the *probability simplex* and for $\mu \in \Delta_{\mathcal{X}}$ we sometimes write $\mu(x) = \mu_x$ for the mass at $x \in \mathcal{X}$. The probability simplex is the convex hull of the *Dirac measures* $\{\delta_x\}_{x\in\mathcal{X}}$ that place the entire mass on individual $x \in \mathcal{X}$ and hence correspond to the unit vectors, i.e., $\delta_x(y) = \delta_{xy} = 1$ if and only if $x = y$. A *Markov kernel* from a finite set $\mathcal{X}$ to another finite set $\mathcal{Y}$ is a mapping $Q \in \Delta^{\mathcal{X}}_{\mathcal{Y}}$ i.e., it corresponds to a stochastic mapping $Q\colon \mathcal{X} \to \Delta_{\mathcal{Y}}$. An element $Q \in \Delta^{\mathcal{X}}_{\mathcal{Y}}$ is a $|\mathcal{X}| \times |\mathcal{Y}|$ row stochastic matrix with entries $Q_{xy} = Q(y|x)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and can be interpreted as conditional probability distributions. We call the set $\Delta^{\mathcal{Y}}_{\mathcal{X}}$ the *conditional probability polytope* as it is indeed polytope given as a Cartesian product of probability simplices. Given $Q^{(1)} \in \Delta^{\mathcal{X}}_{\mathcal{Y}}$ and $Q^{(2)} \in \Delta^{\mathcal{Y}}_{\mathcal{Z}}$ their composition into a kernel $Q^{(2)} \circ Q^{(1)} \in \Delta^{\mathcal{X}}_{\mathcal{Z}}$ from $\mathcal{X}$ to $\mathcal{Z}$ is given by

$$(Q^{(2)} \circ Q^{(1)})(z|x) := \sum_{y\in\mathcal{Y}} Q^{(2)}(z|y)Q^{(1)}(y|x) \quad \text{for all } x \in \mathcal{X}, z \in \mathcal{Z}.$$

Given $p \in \Delta_{\mathcal{X}}$ and $Q \in \Delta^{\mathcal{X}}_{\mathcal{Y}}$ we denote their composition into a joint probability distribution by $p * Q \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ given by

$$(p * Q)(x, y) := p(x)Q(y|x).$$

The *support* of a vector $v \in \mathbb{R}^{\mathcal{X}}$ is the set $\operatorname{supp}(v) = \{x \in \mathcal{X} \colon v_x \neq 0\}$.

For a vector $\mu \in \mathbb{R}^{\mathcal{X}}_{\ge 0}$ we denote its *Shannon entropy* by

$$H(\mu) := -\sum_x \mu(x)\log(\mu(x)),$$

with the usual convention that $0\log(0) := 0$. For $\mu \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}_{\ge 0}$ we denote the X-marginal by $\mu_X \in \mathbb{R}^{\mathcal{X}}_{\ge 0}$, where $\mu_X(x) := \sum_y \mu(x, y)$. Further, we denote the conditional entropy of μ

conditioned on X by

$$H(\mu|\mu_X) \coloneqq -\sum_{x,y} \mu(x,y) \log \frac{\mu(x,y)}{\mu_X(x)} = H(\mu) - H(\mu_X). \tag{1.11}$$

For any strictly convex function $\phi\colon \Omega \to \mathbb{R}$ defined on a convex subset $\Omega \subseteq \mathbb{R}^d$, the associated *Bregman divergence* $D_\phi\colon \Omega \times \Omega \to \mathbb{R}$ is given by

$$D_\phi(x,y) \coloneqq \phi(x) - \phi(y) - \langle \nabla\phi(y), x-y \rangle.$$

The Bregman divergence of the Shannon entropy is given by the *Kullback–Leibler divergence* or shortly *KL-divergence* that we denote by

$$D_{KL}(\mu,\nu) = \sum_{x\in\mathcal{X}} \mu_x \log\left(\frac{\mu_x}{\nu_x}\right) \quad \text{for } \mu,\nu \in \Delta_{\mathcal{X}},$$

which is well defined whenever $\operatorname{supp}(\nu) \subseteq \operatorname{supp}(\mu)$.

Given two smooth manifolds $\mathcal{M}$ and $\mathcal{N}$ and a smooth function $f\colon \mathcal{M} \to \mathcal{N}$, we denote the differential of f at $p \in \mathcal{M}$ by $df_p\colon T_p\mathcal{M} \to T_{f(p)}\mathcal{N}$. We denote the gradient of a smooth function $f\colon \mathcal{M} \to \mathbb{R}$ defined on a Riemannian manifold $(\mathcal{M}, g)$ by $\nabla^g f\colon \mathcal{M} \to T\mathcal{M}$ and denote the values of the vector field by $\nabla^g f(p) \in T_p\mathcal{M}$ for $p \in \mathcal{M}$. When the Riemannian metric is unambiguous we drop the superscript. In the Euclidean case, we write $Df(p)$ for the Jacobian matrix with entries $Df(p)_{ij} = \partial_j f_i(p)$. For a univariate differentiable function we write $D^2f(p)$ or $\nabla^2 f(p)$ for the *Hessian matrix* with entries $Df(p)_{ij} = \nabla^2 f(p)_{ij} = \partial_i\partial_j f(p)$.

A closed basic semialgebraic set is a set described by finitely many polynomial inequalities lets say $S = \{x \in \mathbb{R}^n : p_i(x) \geq 0 \text{ for } i = 1,\dots,k\}$ for some polynomials p_i. We call $F \subseteq S$ a *face* or *boundary component* of S if it is of the form $F = \{x \in S : p_i(x) = 0 \text{ for } i \in I\}$ for some index set $I \subseteq \{1,\dots,k\}$. We denote the set of faces of S by $\mathcal{F}(S)$, which is a partially ordered set (or shortly *poset*) with the partial order of inclusion. We endow the poset $\mathcal{F}(S)$ with the *join* and *meet* operation

$$F \wedge G \coloneqq F \cap G \quad \text{and} \quad F \vee G \coloneqq \bigcap_{\substack{H\in\mathcal{F}(S)\\ F,G\subseteq H}} H.$$

These turn $\mathcal{F}(S)$ into a *lattice* as the join and meet satisfy the absorption laws $F \vee (F \wedge G) = F$ and $F \wedge (F \vee G) = F$ for all $F, G \in \mathcal{F}$. A *morphism* between two lattices $\mathcal{F}$ and $\mathcal{G}$ is a mapping $\varphi\colon \mathcal{F} \to \mathcal{G}$ that respects the join and the meet, i.e., such that $\varphi(F \wedge G) = \varphi(F) \wedge \varphi(G)$ and $\varphi(F \vee G) = \varphi(F) \vee \varphi(G)$ for all $F, G \in \mathcal{F}$. A lattice *isomorphism* is a bijective lattice morphism where the inverse is also a morphism. We say that two basic semialgebraic sets with isomorphic face lattice are *combinatorially equivalent*.

We denote the space of functions on $\Omega \subseteq \mathbb{R}^d$ that are integrable in p-th power by $L^p(\Omega)$, where we assume that $p \in [1,\infty)$. Endowed with

$$\|u\|_{L^p(\Omega)}^p \coloneqq \int_\Omega |u|^p \mathrm{d}x$$

this is a Banach space, i.e., a complete normed space. If u is a multivariate function with values in $\mathbb{R}^m$ we interpret $|\cdot|$ as the Euclidean norm. We denote the subspace of $L^p(\Omega)$

of functions with weak derivatives up to order k in $L^p(\Omega)$ by $W^{k,p}(\Omega)$, which is a Banach space with the norm

$$\|u\|^p_{W^{k,p}(\Omega)} := \sum_{l=0}^{k} \|D^l u\|^p_{L^p(\Omega)}.$$

This space is called a *Sobolev space* and we denote its dual space, i.e., the space consisting of all bounded and linear functionals on $W^{k,p}(\Omega)$ by $W^{k,p}(\Omega)^*$. The closure of all compactly supported smooth functions $C_c^\infty(\Omega)$ in $W^{k,p}(\Omega)$ is denoted by $W_0^{k,p}(\Omega)$. If Ω has a Lipschitz continuous boundary the operator that restricts a Lipschitz continuous function onto the closure $\overline{\Omega}$ to the boundary admits a linear and bounded extension $\mathrm{tr}\colon W^{1,p}(\Omega) \longrightarrow L^p(\partial\Omega)$, which we call the *trace operator*. Its kernel coincides with $W_0^{1,p}(\Omega)$. We write $\|u\|_{L^p(\partial\Omega)}$ whenever we mean $\|\mathrm{tr}(u)\|_{L^p(\partial\Omega)}$ and for $p = 2$ we write $H^k_{(0)}(\Omega)$ instead of $W^{k,2}_{(0)}(\Omega)$.

Part I

Geometry of Markov decision processes

CHAPTER 2

Background on Markov decision processes

Markov decision processes (MDPs) constitute an important mathematical framework for sequential decision making and originated in the late 1950s and early 1960s [46, 45, 142, 55, 54, 56, 93]. We work with the more general concept of partially observable MDPs, which were introduced by Karl Johan Åström [22] and allows to incorporate uncertainty about the state of the Markov process. In this chapter we provide a self contained introduction to the fundamental concepts in the theory of Markov decision processes and restrict ourselves to time discrete infinite horizon problems with finite state, action and observation space and to large extend to memoryless stochastic policies. In particular, we introduce visitation frequencies, which are the central object of interest in this dissertation. For more comprehensive introduction to the field of Markov decision processes including a collection of historical references we refer to [298, 235, 236, 116, 135, 136].

We begin by giving the definition of a partially observable Markov decision process.

Definition 2.1 (Partially observable Markov decision process). A *partially observable Markov decision process* or shortly *POMDP* is a tuple $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$, where

- $\mathcal{S}$, $\mathcal{O}$ and $\mathcal{A}$ are finite sets called the *state, observation* and *action space* respectively,
- $\alpha \in \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$ is a Markov kernel, which we call *transition mechanism*,
- $\beta \in \Delta_{\mathcal{O}}^{\mathcal{S}}$ is a Markov kernel, which we call *observation mechanism* and
- $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, which we call *instantaneous reward vector*.

We call the system *fully observable* if the supports of $\{\beta(\cdot|s)\}_{s\in\mathcal{S}}$ are disjoint subsets of $\mathcal{O}$, in which case the POMDP simplifies to an MDP, i.e., a POMDP with $\mathcal{O} = \mathcal{S}$ and $\beta = \mathrm{id}_{\mathcal{S}}$. We denote the cardinatlities of the sets $\mathcal{S}$, $\mathcal{O}$ and $\mathcal{A}$ by $n_{\mathcal{S}}$, $n_{\mathcal{O}}$ and $n_{\mathcal{A}}$ respectively.

Throughout this part of the thesis, we will always assume that $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$ denotes a POMDP and we will not always repeat this when stating theorems or other results.

Example 2.2 (Crying baby). To illustrate the abstract concept of partially observable Markov decision processes we consider the problem of optimally feeding a baby based where we have to base our decision of feeding it or not on whether it is crying. For this we model the baby to be in either of the two states $\mathcal{S} = \{s_1, s_2\} = \{$"hungry", "not hungry"$\}$. There are two possible observations $\mathcal{O} = \{o_1, o_2\} = \{$"crying", "not crying"$\}$ and two possible actions $\mathcal{A} = \{a_1, a_2\} = \{$"feed", "don't feed"$\}$. For the transition of the states of the baby we make the following assumptions: After being fed, the baby is never hungry, i.e., $\alpha(s_2|s_1, a_1) = \alpha(s_2|s_2, a_1) = 1$. When hungry and not being fed, the baby remains hungry, i.e., $\alpha(s_1|s_1, a_2) = 1$. When not hungry and not being fed, the baby will be hungry with probability 10% at the next time step, i.e., $\alpha(s_1|s_2, a_2) = 0.1$. See also Figure 2.1 for a visualization of the transition mechanism. Further, we assume that the baby always cries when it is hungry, i.e., $\beta(o_1|s_1) = 1$, and cries with probability 50% when it is not hungry[1],

[1] Of course this only meant as a didactic example and not an attempt to realistically model a baby.

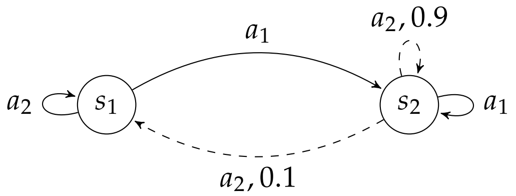

Figure 2.1. Transition model for the crying baby example.

i.e., $\beta(o_1|s_2) = 0.5$, and hence obtain the following observation kernel

$$\beta = \begin{array}{c c} & \begin{array}{cc} o_1 & o_2 \end{array} \\ \begin{array}{c} s_1 \\ s_2 \end{array} & \begin{pmatrix} 1 & 0 \\ 0.5 & 0.5 \end{pmatrix} \end{array} \in \Delta_{\mathcal{O}}^{\mathcal{S}}.$$

Finally, for the instantaneous reward we assume that we obtain no reward when we feed the baby when it is hungry or when we don't feed it when it is not hungry, we assume that we obtain a negative reward of -1 when we feed the baby when it is not hungry and a negative reward of -10 when we don't feed the baby when it is hungry. Overall, we obtain the reward vector

$$r = \begin{array}{c c} & \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{array}{c} s_1 \\ s_2 \end{array} & \begin{pmatrix} 0 & -10 \\ -1 & 0 \end{pmatrix} \end{array} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}.$$

We will revisit this example throughout the thesis.

The goal in Markov decision processes is to optimally control the state s_t of the system through suitable actions a_t. In a partially observable setting only partial information about the true state s_t is revealed through the observation o_t made from s_t. The rule how to choose the next action a_t is usually referred to as a *policy* or *action selection mechanism* and typically denoted by π. The following classes of policies are common in Markov decision processes and reinforcement learning.

Memoryless stochastic policies. A *memoryless policy*, *reactive policy* or *Markov policy* selects the action a_t at time t purely based on the observation of the most recent observation o_t. In general it can do so in a stochastic way in which case it is modelled by a Markov kernel $\pi_t \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ from the observation space $\mathcal{O}$ to the action space $\mathcal{A}$. The sequence $\pi = (\pi_0, \pi_1, \dots)$ is referred to as the policy. The stochastic process of the states $S_0, S_1, \dots$ arising through actions selected according to a policy π satisfy the Markov property

$$\mathbb{P}(S_{t+1} = s_{t+1}|S_0 = s_0, \dots, S_t = s_t) = \mathbb{P}(S_{t+1} = s_{t+1}|S_t = s_t) \tag{2.1}$$

justifying the term Markov policies. When the action selection π_t at is independent of t we call the policy *stationary* and associate it with a single Markov kernel $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$. We refer to the family $\Delta_{\mathcal{A}}^{\mathcal{O}}$ of memoryless stochastic policies as the *policy polytope*. As a product of simplices it is indeed a polytope, which in a general context is sometimes referred to as the *conditional probability polytope*. The optimization of memoryless stochastic stationary policies in POMDPs is known to be NP-hard in general [286].

Memory based policies and internal state controller. A more general class of policies is described by policies with *finite memory*. Here, the next action a_t is based on a finite horizon window

$$h_t^{(k)} = (o_l, , a_l, o_{l+1}, a_{l+1}, \dots, o_t) \in \mathcal{H}_t^{(k)} := (\mathcal{O} \times \mathcal{A})^{\max(t-k,0)} \times \mathcal{O}$$

for some horizon length $k \in \mathbb{N}$ and $l = \max(t-k, 0)$. Mathematically finite memory policies can be modelled as a sequence $\pi = (\pi_0, \pi_1, \dots)$ of Markov kernels

$$\pi_t \in \Delta_{\mathcal{A}}^{\mathcal{H}_t^{(k)}}.$$

Note that the stochastic process of the states $S_0, S_1, \dots$ arising from a memory based policy do not in general satisfy the Markov property (2.1) but rather

$$\mathbb{P}(S_{t+1} = s_{t+1} | S_0 = s_0, \dots, S_t = s_t) = \mathbb{P}(S_{t+1} = s_{t+1} | S_{\max(t-k,0)} = s_{\max(t-k,0)}, \dots, S_t = s_t)$$

A class of policies, which allows for more general memories than sliding windows is given by the family of *finite state controllers*. Here, the policy or controller is based on an internal state $i \in \mathcal{I}$, where $\mathcal{I}$ is a finite set. More precisely a policy is a sequence $\pi = (\pi_0, \pi_1, \dots)$ of Markov kernels $\pi_t \in \Delta_{\mathcal{A}}^{\mathcal{I}}$. In addition there evolution of the internal states are described by Markov kernels $\psi_t \in \Delta_{\mathcal{I}}^{\mathcal{O}\times\mathcal{I}}$ where $\psi(i_{t+1}|o_{t+1}, i_t)$ describes the probability that the internal state transitions from i_t to i_{t+1} if the observation o_t is made. At time $t = 0$ the kernel ψ_0 only depends on the observation o_0, i.e., $\psi_0 \in \Delta_{\mathcal{I}}^{\mathcal{O}}$. One can either fix the internal state transitions or interpret them as flexible and a second search variable – next to the policy – in the optimization of Markov decision processes. It is easy to see that the framework incorporates memoryless stochastic policies, where the internal state i_t simply agrees with the observation o_t and also policies with finite memory where the internal state i_t agrees with the history $h_t^{(k)}$.

History and belief state policies. In theory one can allow for infinite memory and consider policies selecting actions based on the full history

$$h_t = (o_0, a_0, o_1, a_1, \dots, o_t) \in \mathcal{H}_t := (\mathcal{O} \times \mathcal{A})^t \times \mathcal{O}.$$

Here, a policy corresponds to a sequence $\pi = (\pi_0, \pi_1, \dots)$, where $\pi_t \in \Delta_{\mathcal{A}}^{\mathcal{H}_t}$.

Instead of working with the full history one often works with the corresponding *belief state MDP* $(\mathcal{B}, \mathcal{A}, \omega, r, \gamma)$. The state space of this MDP is given by the set $\mathcal{B} = \Delta_{\mathcal{S}}$ of probability distributions b over the state space of the original POMDP, which are commonly referred to as *beliefs*. The action space remains the same. In slight abuse of notation the reward of a belief $b \in \Delta_{\mathcal{S}}$ and an action $a \in \mathcal{A}$ is given by

$$r(b, s) := \sum_{s\in\mathcal{S}} b(s) r(s, a).$$

The transition of the belief state b to a new belief state b' under action $a \in \mathcal{A}$ is given by

$$b' = b'(b, a) := \sum_{o\in\mathcal{O}} b'(s'|b, a, o) \sum_s b(s)\alpha(s'|s, a),$$

where the belief update when taking action $a \in \mathcal{A}$ and observing $o \in \mathcal{O}$ afterwards is given by Bayes rule

$$b'(s'|b, a, o) \propto \sum_s b(s)\alpha(s'|s, a)\beta(o|s').$$

The belief serves as a sufficient statistics [38] and solving the belief state MDP is equivalent to computing an optimal history dependent policy $\pi = (\pi_0, \pi_1, \dots)$, which is known to be PSPACE-complete for finite horizons [228] and undecidable for infinite horizons [186, 73].

Comparison between the policy classes. It is a classic result that every fully observable MDP admits a deterministic stationary memoryless policy that is optimal under all (possiblty) history dependent policies [93, 236, 276]. We provide a proof for the reduction from history based policies to stationary memoryless policies for fully observable problems in Theorem 3.16 and the existence of deterministic optimal policies in Theorem 2.23.

For partially observable problems memoryless neither of these two results hold. Indeed, memoryless policies sometimes perform strictly worse compared to history dependent policies and deterministic memoryless policies perform in general worse compared to memoryless stochastic policies [267]. Although they are more restrictive than policies with memory, memoryless policies are attractive as they are easier to optimize and are versatile enough for certain applications [278, 182, 299, 157]. Further, finite memory policies and finite state controllers with fixed internal state transition can be cast as memoryless policies by augmenting the state space with the space of histories of length k or the space of internal states respectively [179, 231, 145]. The mathematical equivalence of finite state controllers and memoryless policies – with different underlying POMDPs – motivates us to restrict our analysis to memoryless stochastic policies.

State-action Markov process. Let us now introduce the Markov processes induced by the policies of a POMDP. A policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ gives rise to transition kernels $P_\pi \in \Delta_{\mathcal{S}\times\mathcal{A}}^{\mathcal{S}\times\mathcal{A}}$ and $p_\pi \in \Delta_{\mathcal{S}}^{\mathcal{S}}$ by

$$P_\pi(s', a'|s, a) := \alpha(s'|s, a)(\pi \circ \beta)(a'|s') \tag{2.2}$$

and

$$p_\pi(s'|s) := \sum_{a\in\mathcal{A}} (\pi \circ \beta)(a|s)\alpha(s'|s, a) \tag{2.3}$$

Definition 2.3 (State-action Markov process). For any initial state distribution $\mu \in \Delta_{\mathcal{S}}$, a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ defines a Markov process on $\mathcal{S}\times\mathcal{A}$ with transition kernel P_π and we induce the associated probability measure on $(\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$ by $\mathbb{P}^{\pi,\mu}$, $\mathbb{P}^{\pi}(\cdot|S_0 \sim \mu)$ or $\mathbb{P}^{\pi}(\cdot|S_0 \sim s_0)$ if $\mu = \delta_{s_0}$. In general, for any policy π, which is not necessarily memoryless or stationary we denote the resulting stochastic process on $\mathcal{S} \times \mathcal{A}$ by $\mathbb{P}^{\pi,\mu}$. Note that this process is in general not Markovian.

Notation for MDPs. With $\mu \in \Delta_{\mathcal{S}}$ we denote the initial state distribution. Following a policy π yields a series of states and actions, which we denote by $s_0, a_0, s_1, \dots$. In contrast to the specific states and actions visited in one trajectory we denote the random variables at the different times by $S_0, A_0, S_1, \dots$. We denote the expectation with respect to the probability measure $\mathbb{P}^{\pi,\mu}$ on $(\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$ by $\mathbb{E}_{\pi,\mu}$ or $\mathbb{E}_\pi[\cdot|S_0 \sim \mu]$ or $\mathbb{E}_\pi[\cdot|S_0 = s]$ when the initial distribution is the Dirac $\mu = \delta_s$ concentrated at $s \in \mathcal{S}$. Further, when considering the expectation of the Markov process with transition kernel P_π with an initial distribution over states and actions $\nu \in \Delta_{\mathcal{S}\times\mathcal{A}}$ we denote the expectation by $\mathbb{E}_\pi[\cdot|(S_0, A_0) \sim \nu]$ and $\mathbb{P}^{\pi}(\cdot|(S_0, A_0) \sim \nu)$ for the corresponding measure on $(\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$. If the initial distribution is deterministic, i.e., $\nu = \delta_{(s,a)}$, then we write $\mathbb{E}_\pi[\cdot|S_0 = s, A_0 = a]$ and $\mathbb{P}^\pi(\cdot|S_0 = a, A_0 = a)$ for the associated measure on $(\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$. For POMDPs we sometimes denote state based policies by $\tau \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ to distinguish them from observation based policies $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$.

The reward function. Every policy leads to a stochastic process over states and actions and we require a criterion scoring these processes and thus providing an objective for policy optimization. Here, we work with the infinite horizon accumulated reward.

Definition 2.4 (Infinite horizon reward). Consider a *discount factor* $\gamma \in [0, 1]$. We define

$$R(\pi) = R_\gamma^\mu(\pi) := \begin{cases} (1-\gamma) \cdot \mathbb{E}_{\pi,\mu}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)\right] & \text{if } \gamma \in [0,1) \\ \mathbb{E}_{\pi,\mu}\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} r(S_t, A_t)\right] & \text{if } \gamma = 1, \end{cases}$$

called the *(infinite horizon) expected discounted reward* for $\gamma \in [0, 1)$ and the *(infinite horizon) expected mean reward* for $\gamma = 1$. Further, the case $\gamma = 0$ is sometimes referred to as the *bandit case* and the expected reward is also called the *return* in the literature [276].

Although not changing the resulting optimization problem we choose to work the factor $(1-\gamma)$ for the discounted reward, which is not commonly done in the literature. The reason for this is that the factor normalizes the weights $\sum_{t=0}^{\infty} \gamma^t = (1-\gamma)^{-1}$ and thus the discounted reward can be interpreted as a weighted average of the rewards of the individual time steps. This allows us to develop a common frame work for discounted and mean reward optimization in state-action space later on. Another effect of working with the normalized discounted reward is that we have $R_\gamma \to R_1$ for $\gamma \to 1$, see Proposition 2.13. We only use the sub- and superscripts when we want to highlight the dependence on the discount factor and initial distribution. The expected and mean reward exist without any assumptions and are continuous with respect to γ as we show in Proposition 2.13.

The reward optimization problem (ROP). In Markov decision processes it is the goal to maximize the reward over all policies. In our case, we study the optimization problem

$$\text{(ROP)} \qquad \text{maximize } R(\pi) \quad \text{subject to } \pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$$

of maximizing the infinite horizon reward over all memoryless stochastic policies.

Example 2.5 (Crying baby example continued). In order to illustrate the concept of the infinite horizon reward we compute the reward for the deterministic policies in the crying baby Example 2.2. For this we fix an initial distribution $\mu \in \Delta_{\mathcal{S}}$ and a discount factor $\gamma \in [0, 1)$ where the computations for $\gamma = 1$ can be carried out analogously. With two observations and two actions there are four deterministic policies corresponding to the four vertices of the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{O}} \cong [0, 1]^2$. We begin by computing the reward of the policy π_1, where we choose to always feed the baby no matter whether it is crying or not. If the baby is hungry at the beginning, which happens with probability μ_{s_1} then a reward of 0 is obtained at time 0. After this the baby is never hungry but always fed and hence we obtain a reward of -1 at every future time. If the baby is not hungry at the beginning, which happens with probability μ_{s_2} then we obtain a reward of -1 at every time step. Hence, the reward is given by

$$R(\pi_1) = -\mu_{s_1}(1-\gamma)\sum_{t=1}^{\infty}\gamma^t - \mu_{s_2}(1-\gamma)\sum_{t=0}^{\infty}\gamma^t = -\gamma\mu_{s_1} - \mu_{s_2}.$$

Let us now consider the policy π_2 of never feeding the baby. When the baby is hungry at time $t = 0$ it stays hungry and we collect a reward of -10 at every time step and hence a reward of $-10 \cdot (1-\gamma)\sum_{t=0}^{\infty}\gamma^t = -10$. If the baby is not hungry at time $t = 0$ and is hungry for the first time at time $t > 0$ then we collect a reward of $-10 \cdot \gamma^t$, which happens with probability $0.1 \cdot 0.9^{t-1}$. Hence, when the baby is not hungry at time $t = 0$ we collect an overall reward of $\sum_{t=0}^{\infty}(-10)\cdot\gamma^t\cdot 0.1\cdot 0.9^{t-1} = -\frac{\gamma}{1-0.9\gamma}$. Overall, for an initial distribution $\mu \in \Delta_{\mathcal{S}}$ we obtain an overall reward of

$$R(\pi_2) = -10\mu_{s_1} - \frac{\gamma\mu_{s_2}}{1-0.9\gamma}.$$

Let us now consider the deterministic policy π_3, which is the most intuitive one of feeding the baby when it is crying and not feeding it when it is not crying. When the baby is hungry at the beginning it will cry and we will feed it and obtain a reward of 0. After one time step the baby will not be hungry and hence from the on we receive the same reward – but discounted with a factor of γ – as in the case where the baby was not hungry in the first place. When the baby is not hungry at the beginning there is a 50% chance that it will cry and we feed it in which case we receive a reward of -1, which we normalize by $1-\gamma$, and the baby will remain not hungry and hence after time 0 we obtain the same reward but discounted with γ. When the baby is not crying we don't feed it in which case we receive a reward of 0 and the baby will be not hungry with probability 90% and hungry with probability 10% at the next time step. When we denote the reward obtained when the baby is hungry at $t = 0$ by V_{s_1} and the reward obtained when the baby is not hungry by V_{s_2} our considerations lead to the following system of linear equations:

$$\begin{aligned} V_{s_1} &= \gamma V_{s_2} \\ V_{s_2} &= 0.5(-(1-\gamma) + \gamma V_{s_2}) + 0.5(0.9\gamma V_{s_2} + 0.1\gamma V_{s_1}). \end{aligned} \tag{2.4}$$

The unique solution to this system is given by

$$V_{s_2} = \frac{-0.5(1-\gamma)}{1-0.95\gamma-05\gamma^2} = -\frac{10}{\gamma+20} \quad \text{and } V_{s_1} = -\frac{10\gamma}{\gamma+20}.$$

Overall, we obtain the reward according to

$$R(\pi_3) = \mu_{s_1}V_{s_1} + \mu_{s_2}V_{s_2} = -\frac{10(\gamma\mu_{s_1}+\mu_{s_2})}{\gamma+20}.$$

For the fourth deterministic policy π_4 where we feed the baby when it is not crying and don't feed it when it is crying we can compute the reward in similar fashion and obtain a reward of

$$R(\pi_4) = -10\mu_{s_1} - \frac{10\mu_{s_2}}{20-19\gamma}.$$

We have seen that it is convenient to consider the reward obtained when starting in a given state. This leads us to the important concept of value functions where the recurrence relationship (2.4) is formalized in the Bellman equations, see Theorem 2.9. Further, we see in this example that the reward function is a rational function of the discounted factor.

In Section 2.3 we obtain a general expression of the infinite horizon discounted reward as a rational function of the entries of the policy as well as the discounted factor and μ, β, α. Finally, we note that our expressions extend to the mean reward case $\gamma = 1$ in which case we obtain $R(\pi_1) = -1, R(\pi_2) = R(\pi_4) = -10$ and $R(\pi_3) = -\frac{10}{21}$. Where with policy π_1 we end up feeding the baby all the time and with policies π_2 and π_4 we end up never feeding the hungry baby, with the policy π_3 we adapt our choice in a meaningful way to the crying of the baby. We will see later however, that we can obtain a higher reward deciding randomly whether to feed the baby when we hear it crying, see Example 2.32.

Value functions and Bellman's equation. Value functions encode the reward that is obtained when starting in a given state. They play an important role for the design of various solution algorithms for Markov decision and their theoretical analysis. In order to keep our introduction to Markov decision processes short we restrict our attention in this chapter to discounted value functions and refer [187] for the suitable generalizations for the mean reward case.

Definition 2.6 (Value function). For a policy $\pi \in \Delta_{\mathcal{A}}^{O}$ and $\gamma \in [0, 1)$ we define the *(state) value function* $V^\pi = V^\pi_\gamma \in \mathbb{R}^{\mathcal{S}}$ via

$$V^\pi(s) := R_\gamma^{\delta_s}(\pi) = (1-\gamma) \cdot \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) | S_0 = s\right] \quad \text{for } s \in \mathcal{S}. \tag{2.5}$$

Note that $R(\pi) = \sum_{s\in\mathcal{S}} \mu(s) V^\pi(s) = \langle \mu, V^\pi \rangle_{\mathcal{S}}$ for any policy $\pi \in \Delta_{\mathcal{A}}^{O}$.

Definition 2.7 (State-action value function). For a policy $\pi \in \Delta_{\mathcal{A}}^{O}$ and $\gamma \in [0, 1)$ we define the *state-action value function* or *Q-value function* $Q^\pi = Q^\pi_\gamma \in \mathbb{R}^{\mathcal{S}\in\mathcal{A}}$ via

$$Q^\pi(s, a) := (1-\gamma) \cdot \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) | S_0 = s, A_0 = a\right] \quad \text{for } s \in \mathcal{S}, a \in \mathcal{A}. \tag{2.6}$$

The state and state-action value functions are closely related by the following two formulas, which are elementary to verify

$$V^\pi(s) = \sum_{a\in\mathcal{A}} (\pi \circ \beta)(a|s) Q^\pi(s, a) \quad \text{and} \tag{2.7}$$

$$Q^\pi(s, a) = (1-\gamma) r(s, a) + \gamma \sum_{s'\in\mathcal{S}} \alpha(s'|s, a) V^\pi(s'). \tag{2.8}$$

Further, for both value functions determine the reward in the following way

$$R(\pi) = \sum_{s\in\mathcal{S}} \mu(s) V^\pi(s) = \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \mu(s) (\pi \circ \beta)(a|s) Q^\pi(s, a), \tag{2.9}$$

where we again leave the proof to the reader.

Definition 2.8 (One step reward). For $\pi \in \Delta_{\mathcal{A}}^{O}$ we define the *one step reward* $r_\pi \in \mathbb{R}^{\mathcal{S}}$ as

$$r_\pi(s) := \sum_{a\in\mathcal{A}} r(s, a) (\pi \circ \beta)(a|s).$$

The classic algorithm of value iteration and also Q-learning are based on the following characterization of the value functions as fixed points of contracting operators.

Theorem 2.9 (Bellman equation). *For $\gamma \in [0,1)$ the value functions are uniquely determined by the Bellman equations*

$$V^\pi = \gamma p_\pi V^\pi + (1-\gamma) r_\pi \quad \textit{and} \tag{2.10}$$

$$Q^\pi = \gamma P_\pi Q^\pi + (1-\gamma) r. \tag{2.11}$$

Proof. Using the dominated convergence theorem and the Neumann series we compute

$$\begin{aligned} Q^\pi(s,a) &= (1-\gamma) \sum_{t=0}^\infty \gamma^t \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \mathbb{P}^\pi(S_t = s', A_t = a' | S_0 = s, A_0 = a) r(s', a') \\ &= (1-\gamma) \sum_{t=0}^\infty \gamma^t P_\pi^t r = (1-\gamma)(I - \gamma P_\pi)^{-1} r. \end{aligned}$$

Multiplication with $(I - \gamma P_\pi)$ and rearranging yields the Bellman equation for Q^π. Note that (2.11) uniquely determines Q^π as $I - \gamma P_\pi$ is invertible.

For the state value function (2.10) can be shown with an analogue computation or deduced from (2.11) using the relation (2.7). □

The Bellman equations characterize the value functions as the solutions of a system of linear equations. Since the solution of such a system is a rational operation by Cramer's rule we obtain the following result expressing the reward function as a rational function.

Proposition 2.10 (Reward as a rational function). *Consider $\gamma \in [0,1)$. It holds that*

$$R_\gamma^\mu(\pi) = (1-\gamma) \cdot \frac{\det(I - \gamma p_\pi + r_\pi \mu^\top)}{\det(I - \gamma p_\pi)} - 1 + \gamma. \tag{2.12}$$

In particular, the entries of the value function are given by

$$V^\pi(s) = (1-\gamma) \cdot \frac{\det(I - \gamma p_\pi)_s^{\delta_s}}{\det(I - \gamma p_\pi)} = (1-\gamma) \cdot \frac{\det(I - \gamma p_\pi + r_\pi \delta_s^\top)}{\det(I - \gamma p_\pi)} - 1 + \gamma, \tag{2.13}$$

where $(I - \gamma p_\pi)_s^{\delta_s}$ denotes the matrix that is obtained by replacing the s-th column of $I - \gamma p_\pi$ with the unit vector $\delta_s \in \mathbb{R}^{\mathcal{S}}$.

Proof. By the matrix determinant lemma [99, Lemma 1.1] it holds for an invertible matrix $A \in \mathbb{R}^{d \times d}$ and vectors $v, w \in \mathbb{R}^d$ that $\det(A + uv^\top) = (1 + v^\top A^{-1} u) \det(A)$. Applying this to the reward function $R_\gamma^\mu(\pi) = \mu^\top V^\pi = \mu^\top (1-\gamma)(I - \gamma p_\pi)^{-1} r_\pi$ yields (2.12). The expression for the value function follows from Cramer's rule and by setting $\mu = \delta_s$ respectively. □

2.1 State-action frequencies

The reward arising from a policy is determined by the time the Markov process spends at the individual states while selection specific actions. Hence, instead of optimizing the policy directly one could try to optimize the time spent at favorable states doing favorable actions. The time spent at these are encoded by the so called state-action frequencies, which turn out to be generalizations of stationary distribution incorporating the discount factor. They have appeared in linear programming formulations of Markov decision processes [189, 86, 80, 139] and have been studied systematically under this name by Cyrus Derman [93] who showed that they form a polytope. Apart from their use in linear programming approaches these frequencies appear naturally in the analysis of other solution algorithms like policy gradient methods [2].

By the dominated convergence theorem the discounted reward satisfies

$$\begin{aligned} R(\pi) &= (1-\gamma)\cdot\mathbb{E}_{\pi,\mu}\left[\sum_{t=0}^{\infty}\gamma^t r(S_t,A_t)\right] = (1-\gamma)\sum_{t=0}^{\infty}\mathbb{E}_{\pi,\mu}\left[r(S_t,A_t)\right] \\ &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\sum_{s\in\mathcal{S},a\in\mathcal{A}}\mathbb{P}^{\pi,\mu}(S_t=s,A_t=a)r(s,a). \end{aligned} \tag{2.14}$$

Hence, the reward is determined by weighted averages of the visitation probabilities $\mathbb{P}^{\pi,\mu}(S_t=s,A_t=a)$ of the state-action Markov process at time $t\in\mathbb{N}$. This motivates the definition of the state-action frequencies, which can be interpreted as a measure of how much time the state-action process spends at the different states and actions.

Definition 2.11 (State-action frequencies). For a policy π we define the *(discounted) state-action frequency* or *(discounted) state-action distribution* $\eta^\pi\in\Delta_{\mathcal{S}\times\mathcal{A}}$ by

$$\eta^\pi(s,a)=\eta^\pi_\gamma(s,a):=\begin{cases}(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}^{\pi,\mu}(S_t=s,A_t=a) & \text{if } \gamma\in[0,1)\\ \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{P}^{\pi,\mu}(S_t=s,A_t=a) & \text{if } \gamma=1.\end{cases} \tag{2.15}$$

We denote the set of all state-action frequencies in the fully and in the partially observable cases by

$$\mathcal{N}:=\left\{\eta^\pi:\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}\right\}\subseteq\Delta_{\mathcal{S}\times\mathcal{A}}\quad\text{and}\quad\mathcal{N}^\beta:=\left\{\eta^\pi:\pi\in\Delta_{\mathcal{A}}^{\mathcal{O}}\right\}\subseteq\mathcal{N}\subseteq\Delta_{\mathcal{S}\times\mathcal{A}}.$$

The name *state-action frequency* dates back to the seminal work [93]; other common names of state action frequencies include *(discounted) visitation/occupancy measure*, or *on-policy distribution* [276].

The reward optimization problem in state-action space. An analogue computation to (2.14) holds for the mean reward case and hence

$$R(\pi)=\langle r,\eta^\pi\rangle_{\mathcal{S}\times\mathcal{A}}\quad\text{for all } \pi\in\Delta_{\mathcal{A}}^{\mathcal{O}}. \tag{2.16}$$

Hence, the reward optimization problem (ROP) has the same optimal value as the *reward optimization problem in state-action space*

$$\text{maximize } \langle r,\eta\rangle_{\mathcal{S}\times\mathcal{A}}\quad\text{subject to } \eta\in\mathcal{N}^\beta. \tag{ROP-SA}$$

Indeed, we see later that these two problems are equivalent in the sense that a solution π^* of (ROP) can efficiently be computed from an optimizer η^* of (ROP-SA) through conditioning, see Lemma 3.2. Note that these two optimization problems are rather different, where (ROP) is a linearly constrained problem with non-linear objective and (ROP-SA) is a linear objective problem with a possibly non-linear constraint set. The complexity of the problem lies in the feasible region $\mathcal{N}^\beta$ and Chapter 3 studies to the geometry of $\mathcal{N}^\beta$.

Analogously to state-action frequencies we can introduce state frequencies, which play an important role as they appear in the iteration complexity of gradient based methods for reward optimization. Further, they appear naturally when computing a policy corresponding to a state-action distribution.

Definition 2.12 (State frequencies). For a policy π we define the *(discounted) state frequency* or *(discounted) state distribution* $\rho^\pi \in \Delta_{\mathcal{S}}$ by

$$\rho^\pi(s) = \rho^\pi_\gamma(s) := \begin{cases} (1-\gamma)\displaystyle\sum_{t=0}^{\infty} \gamma^t \mathbb{P}^{\pi,\mu}(S_t = s) & \text{if } \gamma \in [0,1) \\ \displaystyle\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{P}^{\pi,\mu}(S_t = s) & \text{if } \gamma = 1. \end{cases} \tag{2.17}$$

It is elementary to check that $\rho^\pi(s) = \sum_{a\in\mathcal{A}} \eta^\pi(s,a)$, i.e., that ρ^π is the state-marginal of the joint probability disitrbution η^π.

Existence and continuity of reward and frequencies. Before we continue we convince ourselves that the discounted frequencies always exist even without requiring the ergodicity of the system. In particular this implies the well-definedness of the infinite horizon reward function. For this we follow the reasoning in [142]. In order to show that the expected state-action frequencies exist without any assumptions, we recall that for a (row or column) stochastic matrix P, the *Cesàro mean* is defined by

$$P^* := \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} P^t$$

and exists without any assumptions. Further, P^* is the projection onto the subspace of stationary distribution [103]. For $\gamma \in [0,1)$ the matrix

$$P^*_\gamma := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t P^t = (1-\gamma)(I-\gamma P)^{-1}$$

is known as the *Abel mean* of P, where we used the Neumann series. By the Tauberian theorem, it holds that $P^*_\gamma \to P^*$ for $\gamma \nearrow 1$, see for example [120, 144].

Proposition 2.13. *The frequencies $\eta^\pi_\gamma, \rho^\pi_\gamma$ and the reward $R_\gamma(\pi)$ exist for every $\pi \in \Delta^{\mathcal{O}}_{\mathcal{A}}$ and $\mu \in \Delta_{\mathcal{S}}$ and are continuous in $\gamma \in [0,1]$ for fixed π and μ.*

Proof. The existence of the state-action frequencies as well as the continuity with respect to the discount parameter follows directly from the general theory since

$$\eta^\pi_\gamma = (P^T_\pi)^*_\gamma(\mu * (\pi \circ \beta))$$

for $\gamma \in [0,1)$ and $\eta^\pi_1 = (P^T_\pi)^*(\mu * (\pi \circ \beta))$. With an analogue argument, the statement follows for the state frequencies and for the reward. □

For $\gamma = 1$ we work under the following standard assumption in the (PO)MDP literature.

Assumption 2.14 (Uniqueness of stationary distributions). If $\gamma = 1$, we assume that for any policy $\pi \in \Delta^{\mathcal{O}}_{\mathcal{A}}$ there exists a unique stationary distribution $\eta \in \Delta_{\mathcal{S}\times\mathcal{A}}$ of P_π.

Note that this assumption is weaker than ergodicity and is satisfied whenever the Markov chain with transition kernel P^π is irreducible and aperiodic for every policy π, e.g., when the transition kernel satisfies $\alpha > 0$. The following theorem shows in particular that for any initial distribution μ, the infinite time horizon state-action frequency $\eta_1^{\pi,\mu}$ is the unique stationary distribution of P_π. For $\gamma \in [0,1)$ the ergodicity Assumption 2.14 is not required, since the discounted stationary distributions are always unique since $I - \gamma P_\pi$ is invertible because the spectral norm of P_π is one. The result can be interpreted as a characterization similar to the Bellman equation for the value functions.

Theorem 2.15 (Characterization via discounted stationarity). *Fix $\gamma \in [0,1]$ and $\pi \in \Delta_{\mathcal{A}}^{O}$. Then the state and state-action frequency $\rho^\pi \in \Delta_{\mathcal{S}}$ and $\eta^\pi \in \Delta_{\mathcal{S}\times\mathcal{A}}$ satisfy*

$$\rho^\pi = \gamma p_\pi^T \rho^\pi + (1-\gamma)\mu \quad \textit{and} \tag{2.18}$$

$$\eta^\pi = \gamma P_\pi^T \eta^\pi + (1-\gamma)(\mu * (\pi \circ \beta)). \tag{2.19}$$

Further, let Assumption 2.14 hold then ρ^π and η^π are the unique elements in $\Delta_{\mathcal{S}}$ and $\Delta_{\mathcal{S}\times\mathcal{A}}$ satisfying (2.18) *and* (2.19) *respectively.*

Proof. By the general theory of Cesàro means, $(P_\pi^T)^*$ projects onto the space of stationary distributions and hence the $\eta^\pi = (P_\pi^T)^*(\mu * (\pi \circ \beta))$ is stationary; this can also be verified through explicit computation. Hence, by Assumption 2.14, η_1^π is the unique stationary distribution. For $\gamma \in [0,1)$ we have

$$\eta_\gamma^\pi = (P_\pi^T)_\gamma^*(\mu * (\pi \circ \beta)) = (I - \gamma P_\pi^T)^{-1}(\mu * (\pi \circ \beta)),$$

which yields the claim. For the state distributions $\rho_\gamma^{\pi,\mu}$ the claim follows analogously or by marginalization. □

Properties of frequencies. Since $\mathbb{P}^{\pi,\mu}(S_t = s, A_t = a) = \mathbb{P}^{\pi,\mu}(S_t = s)(\pi \circ \beta)(a|s)$ it holds that $\eta^\pi(s,a) = (\pi\circ\beta)(a|s)\rho^\pi(s)$ for all policies π and $s \in \mathcal{S}, a \in \mathcal{A}$. Our main motivation for studying state-action frequencies was the identity $R(\pi) = \langle r, \eta^\pi\rangle_{\mathcal{S}\times\mathcal{A}}$, which also implies the identity

$$R(\pi) = \langle r, \eta^\pi\rangle_{\mathcal{S}\times\mathcal{A}} \sum_{s\in\mathcal{S}} \rho^\pi(s) \sum_{a\in\mathcal{A}} (\pi \circ \beta)(a|s) r(s,a) = \langle r_\pi, \rho^\pi\rangle_{\mathcal{S}}. \tag{2.20}$$

Where the frequencies can be used to compute the reward of a policy we can also interpret frequencies as rewards allowing us to transfer many statements for infinite horizon rewards to frequencies. Indeed, if we choose $r(s',a') := \delta_{ss'}\delta_{aa'}$ then $R(\pi) = \eta^\pi(s,a)$ and similarly if $r(s',a') := \delta_{ss'}$ then $R(\pi) = \rho^\pi(s)$. With this interpretation Proposition 2.10 implies

$$\eta^\pi(s,a) = (1-\gamma)\cdot\frac{\det(I - \gamma p_\pi + (\pi\circ\beta)(a|s)\delta_s\mu^\top)}{\det(I-\gamma p_\pi)} - 1 + \gamma \tag{2.21}$$

and

$$\rho^\pi(s) = (1-\gamma)\cdot\frac{\det(I - \gamma p_\pi + \delta_s\mu^\top)}{\det(I-\gamma p_\pi)} - 1 + \gamma. \tag{2.22}$$

In addition, the following consequences of Cramer's rule are useful.

Proposition 2.16. *Consider $\gamma \in [0,1)$ and consider $\pi \in \Delta_{\mathcal{A}}^{O}$. It holds that*

$$\eta_{\gamma}^{\pi,\mu}(s,a) = (1-\gamma) \cdot \frac{(\pi \circ \beta)(a|s)\det(I-\gamma p_{\pi}^{T})_{s}^{\mu}}{\det(I-\gamma p_{\pi})} \tag{2.23}$$

and

$$\rho_{\gamma}^{\pi,\mu}(s) = (1-\gamma) \cdot \frac{\det(I-\gamma p_{\pi}^{T})_{s}^{\mu}}{\det(I-\gamma p_{\pi})}. \tag{2.24}$$

where $(I-\gamma p_{\pi})_{s}^{\mu}$ denotes the matrix that is obtained by replacing the s-th column of $I-\gamma p_{\pi}$ with the initial distribution $\mu \in \mathbb{R}^{\mathcal{S}}$.

Proof. Applying Cramer's rule to $\rho^{\pi} = (1-\gamma)(I-\gamma p_{\pi}^{T})^{-1}\mu$ and $\eta^{\pi}(s,a) = (\pi \circ \beta)(a|s)\rho^{\pi}(s)$ yields the result. □

2.2 The advantage function and Bellman optimality

Here, we review the fundamental principle of Bellman optimality for fully observable MDPs. This implies the existence of deterministic optimal policies and lies at the heart of the two classic algorithms of value and policy iteration. Further, the advantage function and the performance difference lemma are powerful tools in Markov decision processes.

Definition 2.17 (Advantage function). For a policy $\pi \in \Delta_{\mathcal{A}}^{O}$ we define the *advantage function* as $A^{\pi}(s,a) := Q^{\pi}(s,a) - V^{\pi}(s)$ for $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

The value $A^{\pi}(s,a)$ of the advantage function encodes how much better it is to fix the first action to be a when starting in state s compared to choosing it according to π given that all future actions will be selected according to π. The concept and name advantage function dates back to [30][2] and yields an elegant expression of the difference of the reward of two policies in terms of a state-action frequency and an advantage function.

Lemma 2.18 (Performance difference, [152]). *Let $\gamma \in [0,1)$ and consider $\pi, \pi' \in \Delta_{\mathcal{A}}^{O}$. It holds that*

$$R(\pi) - R(\pi') = \frac{\langle \eta^{\pi}, A^{\pi'} \rangle_{\mathcal{S}\times\mathcal{A}}}{1-\gamma}.$$

Proof. We compute

$$\langle \eta^{\pi}, A^{\pi'} \rangle_{\mathcal{S}\times\mathcal{A}} = \langle \eta^{\pi}, Q^{\pi'} \rangle_{\mathcal{S}\times\mathcal{A}} - \langle \eta^{\pi}, V^{\pi'} \rangle_{\mathcal{S}\times\mathcal{A}}.$$

The second term equals

$$\langle \eta^{\pi}, V^{\pi'} \rangle_{\mathcal{S}\times\mathcal{A}} = \sum_{s,a} \rho^{\pi}(s)\pi(a|s)V^{\pi'}(s) = \langle \rho^{\pi}, V^{\pi'} \rangle_{\mathcal{S}}.$$

[2] Here, an advantage function A^{*} assuming optimal actions after the first action was considered.

Using (2.8) and the discounted stationarity (2.18) the first term equals

$$\begin{aligned}\sum_{s\in\mathcal{S},a\in\mathcal{A}} \eta^\pi(s,a)Q^{\pi'}(s,a) &= \sum_{s\in\mathcal{S},a\in\mathcal{A}} \eta^\pi(s,a)\left((1-\gamma)r(s,a)+\gamma\sum_{s'\in\mathcal{S}}\alpha(s'|s,a)V^{\pi'}(s')\right)\\ &= (1-\gamma)R(\pi)+\gamma\sum_{s'\in\mathcal{S}}V^{\pi'}(s')\sum_{s\in\mathcal{S},a\in\mathcal{A}}\rho^\pi(s)\pi(a|s)\alpha(s'|s,a)\\ &= (1-\gamma)R(\pi)+\langle\gamma p_\pi^T\rho^\pi, V^{\pi'}\rangle_{\mathcal{S}}\\ &= (1-\gamma)R(\pi)+\langle\rho^\pi, V^{\pi'}\rangle_{\mathcal{S}}-(1-\gamma)\langle\mu, V^{\pi'}\rangle_{\mathcal{S}}\\ &= (1-\gamma)R(\pi)+\langle\rho^\pi, V^{\pi'}\rangle_{\mathcal{S}}-(1-\gamma)R(\pi').\end{aligned}$$

Combining the computations we obtain

$$\langle\eta^\pi, A^{\pi'}\rangle_{\mathcal{S}\times\mathcal{A}} = (1-\gamma)R(\pi)-(1-\gamma)R(\pi').$$

□

Theorem 2.19 (Bellman optimality criterion). *Consider a fully observable Markov decision process $(\mathcal{S},\mathcal{A},\alpha,r)$ and a discount factor $\gamma\in[0,1)$ and $\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}$. Then the following statements are equivalent:*

(*i*) *It holds that $V^\pi \geq V^{\pi'}$ for all $\pi'\in\Delta_{\mathcal{A}}^{\mathcal{S}}$.*

(*ii*) *It holds that $R_\gamma^\mu(\pi)\geq R_\gamma^\mu(\pi')$ for all $\pi'\in\Delta_{\mathcal{A}}^{\mathcal{S}}$ and $\mu\in\Delta_{\mathcal{S}}$.*

(*iii*) *For some $\mu\in\operatorname{int}(\Delta_{\mathcal{S}})$ it holds that $R_\gamma^\mu(\pi)\geq R_\gamma^\mu(\pi')$ for all $\pi'\in\Delta_{\mathcal{A}}^{\mathcal{S}}$.*

(*iv*) *It holds that $A^\pi(s,a)\leq 0$ for all $s\in\mathcal{S}, a\in\mathcal{A}$.*

(*v*) *It holds that*

$$V^\pi(s) = \max_{a\in\mathcal{A}} Q^\pi(s,a) \quad \textit{for all } s\in\mathcal{S}. \tag{2.25}$$

Proof. We first prove that (*i*) implies (*ii*). This follows as

$$R_\gamma^\mu(\pi) = \mu^\top V^\pi \geq \mu^\top V^{\pi'} = R_\gamma^\mu(\pi')$$

for any $\pi'\in\Delta_{\mathcal{A}}^{\mathcal{S}}$ and $\mu\in\Delta_{\mathcal{S}}$.

It is clear that (*ii*) implies (*iii*).

Now we show that (*iii*) implies (*iv*). For this we assume that

$$0 < A^\pi(s_0,a_0) = Q^\pi(s_0,a_0) - V^\pi(s_0)$$

for some $s_0\in\mathcal{S}, a_0\in\mathcal{A}$. Let $\pi'\in\Delta_{\mathcal{A}}^{\mathcal{S}}$ denote a greedy improvement of $\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}$, i.e., let $\pi'\colon\mathcal{S}\to\mathcal{A}$ be a deterministic policy satisfying

$$Q^\pi(s,\pi'(s)) = \max_{a\in\mathcal{A}} Q^\pi(s,a) \geq V^\pi(s) \quad \text{for all } s\in\mathcal{S},$$

which implies $A^\pi(s,\pi'(s))\geq 0$. By the performance difference lemma we have

$$(1-\gamma)(R_\gamma^\mu(\pi')-R_\gamma^\mu(\pi)) = \sum_{s\in\mathcal{S},a\in\mathcal{A}}\rho^{\pi'}(s)\pi'(a|s)A^\pi(s,a) = \sum_{s\in\mathcal{S}}\rho^{\pi'}(s)A^\pi(s,\pi'(s)) > 0$$

since $\rho_\gamma^{\pi',\mu}(s_0)\geq(1-\gamma)\mu(s_0)>0$ and $A^\pi(s_0,\pi'(s_0))\geq A^\pi(s_0,a_0)>0$.

Let us now proof the equivalence of (*iv*) and (*v*). Let (*iv*) hold, i.e., $0 \leq Q^\pi(s,a) - V^\pi(s)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. Then clearly $V^\pi(s) \geq \max_{a\in\mathcal{A}} Q^\pi(s,a)$ for all $s \in \mathcal{S}$. On the other hand by (2.7) it holds

$$V^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) Q^\pi(s,a) \leq \max_{a\in\mathcal{A}} Q^\pi(s,a). \tag{2.26}$$

Let now (*v*) hold, then we have $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s) \leq \max_{a\in\mathcal{A}} Q^\pi(s,a) - V^\pi(s) = 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}$.

It remains to show that (*iv*) implies (*i*). For this we assume that $A^\pi(s,a) \leq 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. For $\pi' \in \Delta_\mathcal{A}^\mathcal{S}$ the performance difference lemma implies that

$$R(\pi') - R(\pi) = (1-\gamma)^{-1} \langle \eta^{\pi'}, A^\pi \rangle_\mathcal{S} \leq 0$$

for an arbitrary initial distribution $\mu \in \Delta_\mathcal{S}$, which shows $V^\pi \geq V^{\pi'}$. □

Definition 2.20 (Bellman optimality). We call a policy $\pi^* \in \Delta_\mathcal{A}^\mathcal{S}$ satisfying the equivalent statements (*i*)-(*v*) from Theorem 2.19 a *Bellman optimal* policy. For any Bellman optimal policy π^* it holds that

$$V^{\pi^*}(s) = \max_{\pi\in\Delta_\mathcal{A}^\mathcal{S}} V^\pi(s) \quad \text{and } Q^{\pi^*}(s,a) = \max_{\pi\in\Delta_\mathcal{A}^\mathcal{S}} Q^\pi(s,a) \quad \text{for } s \in \mathcal{S}, a \in \mathcal{A}$$

and we write V^* and Q^* for V^{π^*} and Q^{π^*}.

The Bellman optimality criterion states that a policy is optimal if and only if the value of every state is the best possible value over all actions. This is precisely the case if

$$\sum_{a\in\mathcal{A}} \pi(a|s) Q^\pi(s,a) = \max_{a'\in\mathcal{A}} Q^\pi(s,a') \quad \text{for all } s \in \mathcal{S}, \tag{2.27}$$

where we used equation (2.25) using (2.7). Hence, a policy is Bellman optimal if and only if it selects only actions with positive probability that maximize the Q value function. Often, the optimality criterion (2.25) is stated as

$$V^\pi(s) = \max_{a\in\mathcal{A}} (1-\gamma) r(s,a) + \gamma \sum_{s'\in\mathcal{S}} \alpha(s'|s,a) V^\pi(s') \quad \text{for all } s \in \mathcal{S}, \tag{2.28}$$

which is due to (2.8). This formulation lies at the core of the value iteration algorithm as we will see later.

Definition 2.21 (Greedy policies). We call a policy $\pi \in \Delta_\mathcal{A}^\mathcal{S}$

(*i*) *greedy with respect to* $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ if

$$\sum_{a\in\mathcal{A}} \pi(a|s) Q(s,a) = \max_{a\in\mathcal{A}} Q(s,a) \quad \text{for all } s \in \mathcal{S}, \tag{2.29}$$

(*ii*) *greedy with respect to* π' if it is greedy with respect to $Q^{\pi'}$ and

(*iii*) *greedy with respect to* $V \in \mathbb{R}^\mathcal{S}$ if it is greedy with respect to $Q^V \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ defined by

$$Q^V(s,a) := (1-\gamma) r(s,a) + \gamma \sum_{s'\in\mathcal{S}} \alpha(s'|s,a) V(s').$$

With this terminology Theorem 2.19 states that a policy is Bellman optimal if and only if it is greedy with respect to itself. The following result implies the existence of deterministic Bellman optimal policies in and is useful in the analysis of policy iteration.

Lemma 2.22 (Policy improvement). *Let $\pi' \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ be greedy with respect to $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$. Then it holds that $R(\pi') \geq R(\pi)$ for all initial distribution $\mu \in \Delta_{\mathcal{S}}$. Further, if $R(\pi') = R(\pi)$ for some strictly positive initial distribution $\mu \in \operatorname{int}(\Delta_{\mathcal{S}})$ or equivalently $V^{\pi'} = V^{\pi}$ then π (and therefore π') is Bellman optimal.*

Proof. By the performance difference Lemma 2.18 it holds that

$$
\begin{aligned}
(1-\gamma)(R(\pi') - R(\pi)) &= \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \eta^{\pi'}(s,a) A^{\pi}(s,a) \\
&= \sum_{s\in\mathcal{S}} \rho^{\pi'}(s) \left(\sum_{a\in\mathcal{A}} \pi'(a|s) Q^{\pi}(s,a) - V^{\pi}(s) \right) \qquad (2.30)\\
&= \sum_{s\in\mathcal{S}} \rho^{\pi'}(s) \left(\max_{a\in\mathcal{A}} Q^{\pi}(s,a) - V^{\pi}(s) \right) \geq 0,
\end{aligned}
$$

where we used (2.26) in the last step. Assume now that $R(\pi') - R(\pi) = 0$ for some initial distribution $\mu \in \operatorname{int}(\Delta_{\mathcal{S}})$ then $\rho^{\pi'}(s) \geq (1-\gamma)\mu(s) > 0$. Now (2.30) implies that $V^{\pi}(s) = \max_{a\in\mathcal{A}} Q^{\pi}(s,a)$ for all $s \in \mathcal{S}$ and thus π is Bellman optimal. □

Theorem 2.23 (Existence of deterministic optimal policies in MDPs). *Consider a fully observable Markov decision process $(\mathcal{S}, \mathcal{A}, \alpha, r)$ and consider a discount factor $\gamma \in [0,1)$. Then there is a deterministic Bellman optimal policy $\pi^* \in \Delta_{\mathcal{A}}^{\mathcal{S}}$.*

Proof. Fix a positive initial distribution $\mu \in \operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$. The discounted reward $R\colon \Delta_{\mathcal{A}}^{\mathcal{S}} \to \mathbb{R}$ is a continuous function over a bounded set and hence admits a maximizer $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, which is Bellman optimal by Theorem 2.19. Now any deterministic greedy policy π^* induced by π is a deterministic optimal policy by Lemma 2.22. □

s We will see in Example 2.32 that in a POMDP the optimal memoryless policy might be required to be stochastic and might depend on the initial distribution.

2.3 Rational structure of the reward and an explicit line theorem

The reward optimization problem (ROP) is a linearly constrained problem and hence the complexity of this problem depends on the objective function. It is known that the reward function is non concave [50] and in the mean reward case it was shown to be a rational function of degree at most $|\mathcal{S}|$ [123]. We have seen in Proposition 2.10 that the reward function for discounted problems obtains an explicit expression as the fraction of two determinantal polynomials. We use this to establish an interpretable connection between the rational degree of the reward function and the observations available from the Markov decision process. We postpone the proofs to the later subsections and only discuss the consequences of the results here.

Definition 2.24 (Degree of a rational function). We say that a function $f\colon \Omega \to \mathbb{R}^m$ is a *rational function* if it admits a representation of the form $f_i = p_i/q_i$ for polynomials p_i and q_i. We say that f is *of degree at most k* if the polynomials p_i and q_i have degree at most k. Finally, we say that $f\colon \Omega \to \mathbb{R}^m$ is a rational function with *common denominator* if it admits a representation of the form $f_i = p_i/q$ for polynomials p_i and q.

We use the expression (2.12) of the reward function R as a rational function to bound its rational degreein terms of the observation kernel β. The result follows from the general result regarding the degree of determinantal polynomials in Proposition 2.37. In fact, this proposition can be used to establish a tighter bound , where we choose the looser bound as it offers a functional interpretation.

Theorem 2.25 (Uniqueness of stationary distributions). *Consider a POMDP $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$ and $\gamma \in [0,1)$. Let $\pi_0 \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ and $O \subseteq \mathcal{O}$ be a subset of observations and let*

$$\Pi_O := \left\{\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}} : \pi(\cdot|o) = \pi_0(\cdot|o) \text{ for all } o \in \mathcal{O} \setminus O\right\} \subseteq \Delta_{\mathcal{A}}^{\mathcal{O}}$$

denote all policies that agree with π_0 on all observations $o \in \mathcal{O} \setminus O$. The rational degree of $R|_{\Pi_O}$ is upper bounded by

$$\deg(R|_{\Pi_O}) \leq \left|\left\{s \in \mathcal{S} : \beta(o|s) > 0 \text{ for some } o \in O\right\}\right|. \tag{2.31}$$

Remark 2.26 (Degree of the value function and the frequencies). Where Theorem 2.25 is formulated for the reward function the bound (2.31) also holds for the value functions V^π as well as for the state and state-action frequencies η^π and ρ^π. Note that by (2.13), (2.21) and (2.22) these are rational functions with common denominator. In particular, this implies that the bound (2.31) also holds for $Q^\pi(s,a) = \sum_{s' \in \mathcal{S}} \alpha(s'|s,a) V^\pi(s')$.

By (2.31) the degree of the reward when varying the policy on an observation $o \in \mathcal{O}$ is upper bounded by the number of states that can cause $o \in \mathcal{O}$ with positive probability.

Definition 2.27 (Compatible states and identifying observations). We call a state $s \in \mathcal{S}$ *compatible* with the observation $o \in \mathcal{O}$ if $\beta(o|s) > 0$. If $o \in \mathcal{O}$ is compatible with at most one state we refer to it as *identifying*.

For the case of an fully observable system (2.31) implies that the rational degree of the reward when varying the policy on a single state $s \in \mathcal{S}$ is one. In Subsection 2.3.2 we study the properties of rational functions with common denominator of degree one. As a direct consequence we obtain the following result that shows that the value functions and state-action frequencies obtained by varying a policy on one state lie on a line, which was first observed and shown in [81]. We use a different proof strategy, which relies on properties of rational functions that provides us with an explicit and interpretable formula for the interpolation speed between the two endpoints of the line.

Theorem 2.28 (Explicit line theorem for MDPs). *Consider an MDP $(\mathcal{S}, \mathcal{A}, \alpha, r)$ and $\gamma \in [0,1)$. Further, let $\pi_0, \pi_1 \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ be two policies that differ on at most k states. For any $\lambda \in [0,1]$ let $V_\lambda \in \mathbb{R}^{\mathcal{S}}$ and $\eta_\lambda \in \Delta_{\mathcal{S}\times\mathcal{A}}$, $\rho_\lambda \in \Delta_{\mathcal{S}}$ denote the value function and state-action frequency belonging to the policy $\pi_0 + \lambda(\pi_1 - \pi_0)$ with respect to the discount factor γ, the initial distribution μ. Then the rational degrees of $\lambda \mapsto V_\lambda$ and $\lambda \mapsto \eta_\lambda$ are at most k. If they differ on at most one state $\tilde{s} \in \mathcal{S}$ then*

$$V_\lambda = V_0 + c(\lambda) \cdot (V_1 - V_0) \quad \text{and} \quad \eta_\lambda = \eta_0 + c(\lambda) \cdot (\eta_1 - \eta_0) \quad \text{for all } \lambda \in [0,1],$$

where

$$c(\lambda) = \frac{\lambda \det(I - \gamma p_1)}{\det(I - \gamma p_\lambda)} = \frac{\lambda \det(I - \gamma p_1)}{(\det(I - \gamma p_1) - \det(I - \gamma p_0))\lambda + \det(I - \gamma p_0)} = \lambda \cdot \frac{\rho_\lambda(\tilde{s})}{\rho_1(\tilde{s})}.$$

The theorem above describes the *interpolation speed* $c(\lambda)$ in terms of the discounted state distribution in $\tilde{s}$. This expressions extends to the case of mean rewards – note that the determinants vanish – and the theorem can be shown to hold in this case as well, if we set $0/0 := 0$. Note that the interpolation speed does not depend on the initial condition μ.

Remark 2.29. Refinements on the upper bound of the rational degree of $\lambda \mapsto V_\lambda$ and $\lambda \mapsto \eta_\lambda$ can be obtained using Proposition 2.34. Indeed, if we write $\eta_\lambda(s,a) = q_{sa}(\lambda)/q(\lambda)$ like in Proposition 2.16 those degrees can be upper bounded by

$$\deg(q_{sa}) \leq \operatorname{rank}(p_1 - p_0)^0_s + \mathbb{1}_S(s) \leq \operatorname{rank}(p_1 - p_0) \quad \text{and} \quad \deg(q) \leq \operatorname{rank}(p_1 - p_0),$$

where $S \subseteq \mathcal{S}$ is the set of states on which the two policies differ; see also the proof of Theorem 2.25 for more details. Hence, the degree of the two curves $\lambda \mapsto V_\lambda$ and $\lambda \mapsto \eta_\lambda$ is upper bounded by $\operatorname{rank}(p_1 - p_0) \leq n_{\mathcal{S}}$.

Theorem 2.30 (Location of reward optimizers). *Let $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$ be a POMDP, $\mu \in \Delta_{\mathcal{S}}$ be an initial distribution and $\gamma \in [0,1)$ a discount factor and let $\pi \in \Delta^{\mathcal{O}}_{\mathcal{A}}$ and denote the set of observations o such that $|\{s \in \mathcal{S} : \beta(o|s) > 0\}| \leq 1$ by O. Then there is a policy $\tilde{\pi}$, which is deterministic on every $o \in O$ and agrees with π on all $o \in \mathcal{O} \setminus O$ such that $R(\tilde{\pi}) \geq R(\pi)$.*

Proof. For $o \in O$, the reward function restricted to the o-component of the policy is a rational function of degree at most one. By Corollary 2.43 degree one rational function with common denominator are maximized at a vertex (see Subsection 2.3.3) and hence there is a policy $\tilde{\pi}$, which is deterministic on o and satisfies $R(\tilde{\pi}) \geq R(\pi)$. Iterating over $o \in O$ yields the result. □

A similar result showing the existence of optimal policies that are deterministic on O was obtained in [203]. In contrast to our algebraic argument their proof relies on a decomposition of the set of state-action frequencies into infinitely many convex pieces.

Remark 2.31 (Semialgebraic structure of level and superlevel sets for POMDPs). Consider a POMDP $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \alpha, \beta, r)$ and fix a discount rate $\gamma \in (0,1)$ as well as an initial condition $\mu \in \Delta_{\mathcal{S}}$. The levelset

$$L_a := \left\{\pi \in \Delta^{\mathcal{O}}_{\mathcal{A}} : R(\pi) = a\right\}$$

of the reward function is the intersection of a variety generated by determinantal polynomials of degree at most $|\mathcal{S}|$ with the policy polytope $\Delta^{\mathcal{O}}_{\mathcal{A}}$. Indeed, by Theorem 2.25 the reward function R is the fraction f/g of two determinantal polynomials f and g of degree at most $|\mathcal{S}|$. The level set L_a consists of all policies, such that $f(\pi) = ag(\pi)$. Thus, the levelset is given by

$$L_a = \Delta^{\mathcal{O}}_{\mathcal{A}} \cap \left\{x \in \mathbb{R}^{\mathcal{O} \times \mathcal{A}} : f(x) - ag(x) = 0\right\}.$$

Analogously, a superlevel set is the intersection

$$\left\{\pi \in \Delta^{\mathcal{O}}_{\mathcal{A}} : R(\pi) \geq a\right\} = \Delta^{\mathcal{O}}_{\mathcal{A}} \cap \left\{x \in \mathbb{R}^{\mathcal{O} \times \mathcal{A}} : f(x) - ag(x) \geq 0\right\}$$

of a basic semialgebraic generated by a difference of two determinantal polynomials of degree at most $|\mathcal{S}|$ with the policy polytope $\Delta^{\mathcal{O}}_{\mathcal{A}}$. In particular, both the levelset and superlevel sets of POMDPs are semialgebraic sets defined by linear inequalities and equations (corresponding to the conditional probability polytope $\Delta^{\mathcal{O}}_{\mathcal{A}}$) and a determinantal (in)equality of degree at most $|\mathcal{S}|$. This description can be used to bound the number of connected components, which captures important properties of the loss landscape of an optimization problem [33, 69]. By a theorem due to Łojasiewicz, level and superlevel sets possess finitely many connected (semialgebraic) components [245, 37] and there exist algorithmic approaches to computing the number of connected components [122] as well as explicit upper bounds, which involve the dimension, the number of defining polynomials as well as their degrees [36, 35]. Those results are generalizations of the classic result due to Milnor and Thom, which bounds the sum of all Betti numbers of a variety. If we apply the Milnor-Thom theorem to the variety

$$\mathcal{V} = \left\{x \in \mathbb{R}^{O\times\mathcal{A}} : f(x) - ag(x) = 0\right\}$$

we obtain that there are at most $|\mathcal{S}|(2|\mathcal{S}| - 1)^{|O||\mathcal{A}|-1}$ many connected components of $\mathcal{V}$. This bound neglects the determinantal nature of the defining polynomial and might therefore be coarse.

Example 2.32 (Crying baby example continued). We return to the crying baby Example 2.2 and compute optimal policies in this case in order to contrast the case of partially observable models to the strong results on Bellman optimal policies for MDPs. Recall, that s_1 corresponds to the baby being hungry, o_1 corresponds to the baby crying and a_1 to feeding the baby. We begin by considering the underlying fully observable model. Here, the Bellman optimal policy $\tau^* \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ is to feed the baby when it is hungry and not feed it when it is not hungry. In this case the reward – irrespective of the discount factor and initial distribution – is[3] $R(\tau^*) = 0$, which shows that this policy is indeed optimal as $r(s, a) \le 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}$ and hence $R(\tau) \le 0$ for all $\tau \in \Delta_{\mathcal{A}}^{\mathcal{S}}$. This reward is however not achievable with observation based policies. In order to see this we assume first that $\pi(a_1|o_1) > 0$ then

$$\rho^\pi(s_2) \ge (1-\gamma)(\mu_{s_2} + \gamma\pi(a_1|o_1)\mu_{s_1}) > 0.$$

Hence, we find that

$$R(\pi) = \langle r_\pi, \rho^\pi\rangle_{\mathcal{S}} \le r_\pi(s_2)\rho^\pi(s_2) = -\frac{1}{2}\left(\pi(a_1|o_1) + \pi(a_1|o_2)\right)\rho^\pi(s_2) < 0.$$

If on the other hand $\pi(a_1|o_1) = 0$ then

$$\rho^\pi(s_1) \ge (1-\gamma)\left(\mu_{s_1} + \frac{\gamma\mu_{s_2}}{2}\right) > 0$$

and consequently

$$R(\pi) = \langle r_\pi, \rho^\pi\rangle_{\mathcal{S}} \le r_\pi(s_2)\rho^\pi(s_2) = -10\pi(a_2|o_1)\rho^\pi(s_1) < 0.$$

In order to study the optimal policies of the POMDP we choose $\gamma = 1/2$. We can use the expression (2.12) of the reward function as a rational function in the entries of the policy. In order to simplify the expression we write $\pi(a_1|o_1) = p$, $\pi(a_1|o_2) = q$ and substitute $\pi(a_2|o_1) = 1-p$ and $\pi(a_2|o_2) = 1-q$. Hence, we consider the parameter dependent policy class

$$\pi_{p,q} = \begin{array}{c|cc} & o_1 & o_2 \\ \hline a_1 & p & q \\ a_2 & 1-p & 1-q \end{array} \in \Delta_{\mathcal{A}}^{O}.$$

[3] Here, in slight abuse of notation we write $R(\pi)$ and $R(\tau)$ for observation-based and state-based policies.

Carrying out these computation we obtain in slightly informal notation

$$R(p,q) = R(\pi_{p,q}) = \frac{-20p^2 - 20pq + 210\mu_{s_1}p + 20p + 10\mu_{s_1}q + 200\mu_{s_1} - 20}{19p - q + 22}.$$

We see that the degree of the reward function $R(p,q)$ in the parameter p corresponding to the observation o_1 is $2 = |\{s \in \mathcal{S} : \beta(o_1|s) > 0\}|$, which agrees with the number of compatible states. On the other hand the degree in the parameter q corresponding to observation o_2 is $1 = |\{s \in \mathcal{S} : \beta(o_2|s) > 0\}|$. Both degrees meet the upper bound (3.21).

By Theorem 2.30 there exists an optimal policy with $q \in \{0,1\}$. In order to compute the optimal observation based policy we make the ansatz of we never feeding the baby when it is not crying since it is never hungry in this case, which corresponds to setting $q = 0$[4]. In this case the reward function simplifies to

$$R(p) = R(\pi_p) = \frac{-20p^2 + (210\mu_{s_1} + 20)p - 200\mu_{s_1} - 20}{19p + 22}.$$

The critical points $p \in (0,1)$ are the parameters p satisfying $R'(p) = 0$, which is equivalent to

$$19p^2 + 44p - 421\mu_{s_1} - 41 = 0.$$

Solving for p yields the two solutions

$$\frac{-44 \pm \sqrt{31996\mu_{s_1} + 5052}}{38} = \frac{\pm\sqrt{421}\sqrt{19\mu_{s_1} + 3} - 22}{19},$$

where the first solution is surely negative and hence not in the feasible domain $p \in [0,1]$. The second solution is surely positive since $421 \cdot 3 = 1263 > 484 = 22^2$ and is at least 2 if $\mu_{s_1} \leq \frac{22}{421}$ and hence we make the choice $\mu = \delta_{s_2}$. In this case the critical point of consideration becomes

$$p^* = \frac{\sqrt{3 \cdot 421} - 22}{19} \approx 0.713,$$

which achieves a reward of

$$R(\pi_{p^*}) \approx -0.448.$$

We computed the reward of the four deterministic policies in Example 2.5, which where $-1, -\frac{10}{11}, -\frac{20}{41}$ and $-\frac{20}{21}$ and hence the optimal reward achievable with a deterministic policy is given by $-\frac{20}{41} \approx -0.488$.

Remark 2.33 (Optimality in POMDPs). We collect the lessons learned from the example above. The optimal reward achievable with memoryless stochastic policies in a POMDP might be strictly smaller than the optimal reward of the underlying MDP [180].

In contrast to MDPs the optimal policy in a POMDPs the optimal policy might be required to be stochastic [180]. However, with a similar approach to the proof of Theorem 2.23 it is possible to strengthen Theorem 3.28 and to show that there always exists an optimal policy $\pi^* \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ such that

$$\left|\{a \in \mathcal{A} : \pi^*(a|o) > 0\}\right| \leq \left|\{s \in \mathcal{S} : \beta(o|s) > 0\}\right|,$$

i.e., there always exists an optimal policy that randomizes between at most as many actions as there are states leading to the observation with positive probability [203, 204].

[4] We do not justify this ansatz here; however, one can also compute the optimal value of p under the assumption $q = 1$ in analogous fashion an see that it leads to a suboptimal reward.

In POMDPs the optimal policy depends on the initial distribution, which is in contrast to MDPs where there always exists a Bellman optimal policy.

2.3.1. Formulas for the degree of determinantal polynomials. Determinantal representation of polynomials play an important role in convex geometry [131, 217] , but often the emphasis is put on symmetric matrices. We complement existing results by studying the non symmetric case here. We call p a *determinantal polynomial* if it admits a representation

$$p(x) = \det\left(A_0 + \sum_{i=1}^{m} x_i A_i\right) \quad \text{for all } x \in \mathbb{R}^m, \tag{2.32}$$

for some $A_0, \dots, A_m \in \mathbb{R}^{n\times n}$. Let us use the notations

$$A(x) := A_0 + \sum_{i=1}^{m} x_i A_i \quad \text{and } B(x) := \sum_{i=1}^{m} x_i A_i. \tag{2.33}$$

Proposition 2.34 (Degree of monic univariate determinantal polynomials). *Let $A, B \in \mathbb{R}^{n\times n}$ and let A be invertible and let $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ denote the eigenvalues of $A^{-1}B$ if repeated according to their algebraic multiplicity. Then,*

$$p\colon \mathbb{R} \to \mathbb{R}, \quad t \mapsto \det(A + tB)$$

is a polynomial of degree

$$\deg(p) = \left|\left\{j \in \{1, \dots, n\} : \lambda_j \neq 0\right\}\right| \leq \operatorname{rank}(B).$$

The roots of p are given by $\{-\lambda_j^{-1} : j \in J\} \subseteq \mathbb{C}$. If further $A^{-1}B$ is symmetric, then we have $\deg(p) = \operatorname{rank}(B)$*.*

Proof. Let $J \subseteq \{1, \dots, n\}$ denote the set of indices j such that $\lambda_j \neq 0$. For $x \neq 0$ we have[5]

$$\begin{aligned} p(t) &= \det(A)\det(I + tA^{-1}B) = x^n \det(A)\det(A^{-1}B + t^{-1}I) = x^n \det(A)\chi_{A^{-1}B}(-t^{-1}) \\ &= t^n \prod_{i=1}^{n}(-t^{-1} - \lambda_i) = (-1)^{n-|J|} \cdot \prod_{j\in J}(-\lambda_j) \cdot \prod_{j\in J}\left(t + \lambda_j^{-1}\right), \end{aligned}$$

which is a polynomial of degree $|J|$. Note that $|J|$ is upper bounded by the complex rank of $A^{-1}B$. Since the rank over $\mathbb{C}$ and $\mathbb{R}$ agree for a real matrix, we have $\deg(p) \leq \operatorname{rank}(A^{-1}B) = \operatorname{rank}(B)$. Assume now that $A^{-1}B$ is symmetric, then the rank of $A^{-1}B$ coincides with the number $|J|$ of non zero eigenvalues. Further, the rank of B and $A^{-1}B$ is the same. □

Remark 2.35. Note that the degree of p can be lower than $\operatorname{rank}(B)$, for example if

$$A = I \quad \text{and } B = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \end{pmatrix}.$$

Then we have $\operatorname{rank}(B) = 1$, but

$$p(\lambda) = \det\begin{pmatrix} 1+\lambda & -\lambda \\ \lambda & 1-\lambda \end{pmatrix} = (1+\lambda)(1-\lambda) + \lambda^2 = 1$$

and therefore $\deg(p) = 0$. Note that in this case $A^{-1}B = B$ has no non-zero eigenvalues.

We use the following Lemma to obtain result for the multivariate case.

[5] Here, $\chi_C(\lambda) = \det(C - \lambda I)$ denotes the characteriztic polynomial of a matrix C.

Lemma 2.36 (Degree of polynomials). *Let $p\colon \mathbb{R}^n \to \mathbb{R}$ be a polynomial. Then there is a direction $x \in \mathbb{R}^n$ such that $t \mapsto p(tx)$ is a polynomial of degree* $\deg(p)$*. Moreover, for any $x \in \mathbb{R}^n$, the univariate polynomial $t \mapsto p(tx)$ has degree at most* $\deg(p)$*.*

Proof. Let without loss of generality p be non trivial. Decompose p into its leading and lower order terms $p = p_1 + p_2$ and choose $x \in \mathbb{R}^n$ such that $p_1(x) \neq 0$. Let $k := \deg(p)$, then we have $p_1(tx) = t^k p_1(x)$ for all $\mu \in \mathbb{R}$. Since the degree of $t \mapsto p_2(tx)$ is at most $k-1$, the degree of $t \mapsto p(tx) = p_1(tx) + p_2(tx)$ is k. □

The following result generalizes Proposition 2.34 to multivariate determinantal polynomials.

Proposition 2.37 (Degree of determinantal polynomials). *Let $0 \neq p\colon \mathbb{R}^m \to \mathbb{R}$ be a non trivial determinantal polynomial with the representation* (2.32) *and fix $x_0 \in \mathbb{R}^m$ with $p(x_0) \neq 0$. Then $A(x_0)$ is invertible and for $x \in \mathbb{R}^m$ we denote the number of non zero eigenvalues counted with (algebraic) multiplicities of the matrix $A(x_0)^{-1}B(x)$ by $N(x) \in \{0, \dots, m\}$. Then*

$$\deg(p) = \max_{x\in\mathbb{R}^m} N(x) \leq \max_{x\in\mathbb{R}^m} \operatorname{rank}(B(x)).$$

Proof. By Lemma 2.36 the degree of p is the maximum of the degrees of the univariate determinantal polynomials

$$t \mapsto p(x_0 + tx) = \det(A(x_0) + tB(x)),$$

which by Propostion 2.34 is equal to $N(x)$. □

Now we come to the proof of Theorem 2.25 that we restate here for convenience.

Theorem 2.25 (Uniqueness of stationary distributions). *Consider a POMDP $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$ and $\gamma \in [0, 1)$. Let $\pi_0 \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ and $O \subseteq \mathcal{O}$ be a subset of observations and let*

$$\Pi_O := \left\{\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}} : \pi(\cdot|o) = \pi_0(\cdot|o) \textit{ for all } o \in \mathcal{O} \setminus O\right\} \subseteq \Delta_{\mathcal{A}}^{\mathcal{O}}$$

denote all policies that agree with π_0 on all observations $o \in \mathcal{O} \setminus O$. The rational degree of $R|_{\Pi_O}$ is upper bounded by

$$\deg(R|_{\Pi_O}) \leq \left|\left\{s \in \mathcal{S} : \beta(o|s) > 0 \textit{ for some } o \in O\right\}\right|. \tag{2.31}$$

Proof. Recall that $R(\pi) = (1-\gamma)p(\pi)/q(\pi) - 1 + \gamma$ where

$$p(\pi) = \det(1 - \gamma p_\pi + r_\pi \mu^T) \quad \text{and } q(\pi) = \det(I - \gamma p_\pi),$$

see (2.12). When restricting q to

$$\Pi_O := \left\{\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}} : \pi(\cdot|o) = \pi_0(\cdot|o) \text{ for all } o \in \mathcal{O} \setminus O\right\} \subseteq \Delta_{\mathcal{A}}^{\mathcal{O}}$$

the matrix $I - \gamma p_\pi$ admits the representation

$$(I - \gamma p_\pi)_{ss'} = \delta_{ss'} - \gamma \sum_{a\in\mathcal{A}, o\in\mathcal{O}\setminus O} \beta(o|s)\pi_0(a|o)\alpha(s'|s,a) - \gamma \sum_{a\in\mathcal{A}, o\in O} \beta(o|s)\pi(a|o)\alpha(s'|s,a),$$

where we denote the last sum by $B(\pi)$ according to the notation (2.33). It remains to estimate $\operatorname{rank}(B(\pi))$. For this we note that $B(\pi)_{ss'} = 0$ if $\beta(o|s) = 0$ for all $o \in O$. By Proposition 2.37 we obtain

$$\deg(q|_{\Pi_O}) \leq \max_{\pi} B(\pi) \leq \left|\left\{s \in \mathcal{S} : \beta(o|s) > 0 \text{ for some } o \in O\right\}\right|.$$

The argument for

$$\deg(p|_{\Pi_O}) \leq \max_{\pi} B(\pi) \leq \Big|\{s \in \mathcal{S} : \beta(o|s) > 0 \text{ for some } o \in O\}\Big|$$

follows analogously. □

2.3.2. A line theorem for degree-one rational functions. First, we show that degree-one rational functions with common denominator map lines to lines, which implies that they map polytopes to polytopes. Further, the extreme points of the range lie in the image of the extreme points, which implies that degree-one rational functions are maximized in extreme points – just like linear functions.

Recall that we have seen that the state-action frequencies, the reward function and the value function of POMDPs are rational functions of degree at most $|\mathcal{S}|$ with common denominator. In the case of MDPs and if a policy is fixed on all but k states, it is a rational function with common denominator of degree at most k.

Proposition 2.38 (A line theorem). *Let $\Omega \subseteq \mathbb{R}^d$ be convex and $f\colon \Omega \to \mathbb{R}^m$ be a rational function of degree at most one with common denominator with $f_i(x) = p_i(x)/q(x)$ for affine linear functions p_i, q. Then, f maps lines to lines. More precisely, if $x_0, x_1 \in \Omega$, then*

$$c\colon [0,1] \to [0,1], \quad \lambda \mapsto \frac{q(x_1)\lambda}{q(x_\lambda)} = \frac{q(x_1)\lambda}{(q(x_1) - q(x_0))\lambda + q(x_0)}$$

is strictly increasing and satisfies

$$f((1-\lambda)x_0 + \lambda x_1) = (1 - c(\lambda))f(x_0) + c(\lambda)f(x_1) = f(x_0) + c(\lambda)(f(x_1) - f(x_0)). \tag{2.34}$$

Further, c is strictly convex if $|q(x_1)| < |q(x_0)|$, strictly concave if $|q(x_1)| > |q(x_0)|$ and linear if $|q(x_0)| = |q(x_1)|$.

Proof. We set $x_\lambda := (1-\lambda)x_0 + \lambda x_1$ and compute

$$f(x_\lambda) = \frac{p(x_\lambda)}{q(x_\lambda)} = \frac{(1-\lambda)p(x_0) + \lambda p(x_1)}{q(x_\lambda)} = \frac{(1-\lambda)q(x_0)}{q(x_\lambda)} \cdot f(x_0) + \frac{\lambda q(x_1)}{q(x_\lambda)} \cdot f(x_1).$$

Noting that

$$\frac{\lambda q(x_1)}{q(x_\lambda)} = \frac{\lambda q(x_1)}{(1-\lambda)q(x_0) + \lambda q(x_1)} = c(\lambda)$$

and

$$\frac{(1-\lambda)q(x_0)}{q(x_\lambda)} + \frac{\lambda q(x_1)}{q(x_\lambda)} = 1$$

yields (2.34). Finally, we differentiate and obtain

$$c'(\lambda) = \frac{q(x_0)q(x_1)}{q(x_\lambda)^2}. \tag{2.35}$$

Since q has no root in Ω it follows that $q(x_0)$ and $q(x_1)$ have the same sign and hence $c'(\lambda) > 0$. Differentiating a second time yields

$$c''(\lambda) = -2q(x_0)q(x_1)(q(x_1) - q(x_0)) \cdot q(x_\lambda)^{-3}.$$

Using that $\operatorname{sgn}(q(x_\lambda)) = \operatorname{sgn}(q(x_0)) = \operatorname{sgn}(q(x_1))$ yields the assertion. □

Remark 2.39. The formula (2.34) holds for all $\lambda \in \mathbb{R}$ for which $x_\lambda = \lambda x_0 + (1-\lambda)x_1 \in \Omega$.

Proposition 2.40 (Level sets of degree one rational functions). *Let $\Omega \subseteq \mathbb{R}^d$ be convex and $f\colon \Omega \to \mathbb{R}$ be a rational function of degree at most one. Then, $L_\alpha \coloneqq \{x \in \Omega : f(x) = \alpha\}$ is the intersection of an affine space with Ω.*

Proof. For any $x, y \in L_\alpha$ the ray $\{x + t(y-x) : t \in \mathbb{R}\} \cap \Omega$ is contained in L_α by the line theorem. □

2.3.3. Extreme points of degree-one rational functions.

It is well known that linear functions obtain their maxima in extreme points. We show that this is also the case for rational functions of degree at most one.

Definition 2.41. Let $\Omega \subseteq \mathbb{R}^d$. Then we call $x \in \Omega$ an *extreme point* of Ω if x is not the strict convex combination of two other points in Ω, i.e., if $x = (1-\lambda)x_0 + \lambda x_1$ for $x_0, x_1 \in \Omega$ and $\lambda \in (0,1)$ implies $x_0 = x_1 = x$. We denote the set of extreme points of Ω by $\operatorname{extr}(\Omega)$.

Proposition 2.42. *Let $\Omega \subseteq \mathbb{R}^d$ be convex and $f\colon \Omega \to \mathbb{R}^m$ be a rational function of degree at most one with common denominator. Then $f(\Omega)$ is convex. If in addition Ω is compact then* $\operatorname{extr}(f(\Omega)) \subseteq f(\operatorname{extr}(\Omega))$.

Proof. Let $y_0 = f(x_0), y_1 = f(x_1) \in f(\Omega)$. Then by the line theorem, the line connecting y_0 and y_1 agrees with the image of the line connecting x_0 and x_1 under f, in particular, it is contained in $f(\Omega)$, which shows the convexity of $f(\Omega)$.

Assume now that Ω is compact and pick an extreme point $y = f(x) \in \operatorname{extr}(f(\Omega))$. If $x \in \operatorname{extr}(\Omega)$, there is nothing to show, so let $x \notin \operatorname{extr}(\Omega)$. By the Krein-Milman theorem a convex and compact set is the closed convex hull of its extreme points [8] and hence by the Carathéodory theorem [318] we can write x as a strict convex combination $\sum_{i=1}^n \lambda_i x_i$ for some $\lambda_i > 0, n \geq 2$ for some extreme points $x_i \in \operatorname{extr}(\Omega)$ (in fact, Carathéodory's theorem ensures that $n \leq d+1$). In particular, it is possible to write x as the strict convex combination $x = (1-\lambda)x_0 + \lambda x_1$, $\lambda \in (0,1)$ by setting $x_0 \coloneqq \sum_{i=2}^n \lambda_i x_i$. By the line theorem we have

$$y = (1 - c(\lambda))f(x_0) + c(\lambda)f(x_1),$$

where $c(\lambda) \in (0,1)$; here we use the strict monotonicity of c. Since y is an extreme point it holds that $f(x_0) = f(x_1) = y$. In particular, this shows that $y = f(x_1) \in f(\operatorname{extr}(\Omega))$. □

Corollary 2.43 (Maximizers of degree-one rational functions). *Let $\Omega \subseteq \mathbb{R}^d$ be a convex and compact set and let $f\colon \Omega \to \mathbb{R}$ be a rational function of degree at most one with common denominator. Then f is maximized in at least one extreme point of Ω. In particular, if Ω is a polytope, f is maximized in at least one vertex.*

Proof. Since Ω is compact and f is continuous, $f(\Omega)$ is a compact interval $f(\Omega) = [\alpha, \beta]$. By the preceding proposition we have $\{\alpha, \beta\} = \operatorname{extr}(f(\Omega)) \subseteq f(\operatorname{extr}(\Omega))$, which shows that f is maximized in at least one extreme point. □

Corollary 2.44. *Let $P \subseteq \mathbb{R}^d$ be a polytope and $f\colon \Omega \to \mathbb{R}^m$ be a rational function of degree at most one with common denominator. Then $f(P)$ is a polytope and we have* $\operatorname{vert}(f(P)) \subseteq f(\operatorname{vert}(P))$.

Proof. By the preceding proposition, $f(P)$ is convex. Further, $f(P)$ has finitely many extreme points since $\operatorname{extr}(f(P)) \subseteq f(\operatorname{extr}(P)) = f(\operatorname{vert}(P))$, which implies the assertion. □

Since the policies form a product $\Delta_{\mathcal{A}}^{\mathcal{S}}$ of simplices we now study products of polytopes.

Proposition 2.45. *Let $f\colon P \longrightarrow \mathbb{R}^m$ be defined on the Cartesian product $P = P_1 \times \cdots \times P_k$ of polytopes, which is a degree-one rational function with common denominator whenever all but one components are fixed. Then $f(P)$ has finitely many extreme points and it holds that*

$$\operatorname{extr}(f(P)) \subseteq f(\operatorname{vert}(P)) = f(\operatorname{vert}(P_1) \times \cdots \times \operatorname{vert}(P_k)).$$

In particular, if $m = 1$ this shows that f is maximized in at least one vertex of P.

Proof. Let now $x = (x^{(1)}, \dots, x^{(k)}) \in P_1 \times \cdots \times P_k$ be such that $f(x) \in \operatorname{extr}(f(P))$. If $x^{(i)} \in \operatorname{vert}(P_i)$, there is nothing to show. Hence, we assume that $x^{(i)} \notin \operatorname{vert}(P_i)$. Let us denote the restriction of f onto P_i by g, where we keep the other components fixed to be $x^{(j)}$. Then we have $g(x^{(i)}) \in \operatorname{extr}(g(P_i))$ and hence by Proposition 2.42 there is $\tilde{x}^{(i)} \in \operatorname{vert}(P_i)$ such that $g(\tilde{x}^{(i)}) = g(x^{(i)}) = f(x)$. Replacing $x^{(i)}$ by $\tilde{x}^{(i)}$ and iterating over i yields the claim. □

We have seen that both the value function as well as the discounted state-action frequencies are degree-one rational functions in the rows of the policy in the case of full observability. Hence, the extreme points of the set of all value functions and of the set of discounted state-action frequencies are described by the proposition above. In fact we will see later that the discounted state-action frequencies form a polytope; further, one can show that the set of value functions is a finite union of polytopes [81].

2.4 Solution methods for Markov decision processes

A variety of exact and approximate solution methods for Markov decision processes have been developed. Here, we provide an overview over classic approaches where we first focus on solution methods for fully observable problems and discuss value iteration, policy iteration, linear programming approaches and policy gradient methods. We conclude this section by reviewing some solution methods for partially observable problems. In order to keep our introduction to Markov decision processes short we restrict our attention within this chapter to discounted value functions and refer [187] for the suitable generalizations for the mean reward case.

2.4.1. Value iteration. Value iteration is a classical solution method and dates back to [256] for stochastics games and was generalized to fully observable Markov decision processes in a series of works [46, 45, 55, 297, 54, 273, 91, 56, 284]. We limit our discussion to the infinite horizon discounted problems and refer to the classic references as well as [235, 135, 236] for variants of value iteration for finite horizon and undiscounted problems.

Definition 2.46 (Bellman optimality operator). Inspired by the principle of Bellman optimality we define the *Bellman optimality operator* by

$$T_\gamma = T\colon \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}, \quad TV(s) := \max_{a\in\mathcal{A}} (1-\gamma)r(s,a) + \gamma \sum_{s'\in\mathcal{S}} \alpha(s'|s,a)V(s'). \tag{2.36}$$

By the Bellman optimality criterion (2.28) a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ is optimal if and only if the value function V^π is a fixed point of T, i.e., if $TV^\pi = V^\pi$.

Theorem 2.47 (Convergence of value iterates). *The operator T_γ is a γ-contraction, i.e., for all $V, W \in \mathbb{R}^{\mathcal{S}}$ it holds that $\|T_\gamma V - T_\gamma W\|_\infty \le \gamma \|V - W\|_\infty$. Hence, T_γ possesses a unique fixed* *point $V^* \in \mathbb{R}^{\mathcal{S}}$, which agrees with the value function of a Bellman optimal policy. For any $V_0 \in \mathbb{R}^{\mathcal{S}}$ the sequence $V_k := T_\gamma^k V_0$ convergence to V^* and it holds that*

$$\|V_k - V^*\|_\infty \leq \frac{\gamma^k}{1-\gamma} \cdot \|V_0 - V_1\|_\infty \quad \textit{and } \|V_k - V^*\|_\infty \leq \gamma^k \cdot \|V_0 - V^*\|_\infty.$$

Proof. Once the γ contraction is established everything else is a direct consequence of Banach's fixed point theorem [118]. Note that in general it holds that

$$\left|\sup_{i\in I} x_i - \sup_{i\in I} y_i\right| \leq \sup_{i\in I} |x_i - y_i|.$$

Hence, we can estimate

$$\begin{aligned} \left|T_\gamma V(s) - T_\gamma W(s)\right| &\leq \max_{a\in\mathcal{A}} \left|\gamma \sum_{s'} \alpha(s'|s,a)(V(s') - W(s'))\right| \\ &\leq \gamma \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \alpha(s'|s,a)|V(s') - W(s')| \\ &\leq \gamma \|V - W\|_\infty. \end{aligned}$$

Taking the maximum over $s \in \mathcal{S}$ yields the assertion. □

The fixed point iteration $V_k := T^k V_0$ described in Algorithm 1 is commonly referred to as *value iteration* and converges exponentially quickly to the optimal value function.

Algorithm 1 Bellman's value iteration (VI)

```
Require: $V_0 \in \mathbb{R}^{\mathcal{S}}$, number of iteration steps $N \in \mathbb{N}$
  for $k = 1, \dots, N$ do
    $V_{k+1} \leftarrow TV_k$                         ▷ Requires $n_{\mathcal{S}}^2 \cdot n_{\mathcal{A}}$ operations
  end for
  return $\pi$ greedy with respect to $V_N$           ▷ Guaranteed to be $O(\frac{\gamma^N}{1-\gamma})$-optimal
```

The next lemma bounds suboptimality of a policy that is greedy with respect to V by the distance of V to the optimal value function V^*.

Lemma 2.48. *Let $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ be a greedy policy with respect to $V \in \mathbb{R}^{\mathcal{S}}$ and let $R^* := \mu^\top V^*$ denote the optimal reward. Then it holds that*

$$R^* - R(\pi) \leq \frac{\|V^* - V\|_\infty}{1-\gamma}. \tag{2.37}$$

Proof. Recall that $Q^* \geq Q^\pi$. By the performance difference Lemma 2.18 we can estimate

$$\begin{aligned}(1-\gamma)(R(\pi^*)-R(\pi)) &= \sum_{s\in\mathcal{S},a\in\mathcal{A}} \eta^{\pi^*}(s,a)\,(Q^\pi(s,a)-V^\pi(s)) \\ &= \sum_{s\in\mathcal{S}} \rho^{\pi^*}(s)\left(\sum_{a\in\mathcal{A}} \pi^*(a|s)Q^\pi(s,a)-V^\pi(s)\right) \\ &\leq \sum_{s\in\mathcal{S}} \rho^{\pi^*}(s)\left(\sum_{a\in\mathcal{A}} \pi^*(a|s)Q^*(s,a)-V^\pi(s)\right) \\ &= \sum_{s\in\mathcal{S}} \rho^{\pi^*}(s)\,(V^*(s)-V^\pi(s)) \\ &\leq \|V^{\pi^*}-V^\pi\|_\infty.\end{aligned}$$

□

Compare the bound (2.37) to the estimate

$$R^*-R(\pi) \leq \frac{2\gamma\|V^*-V\|_\infty}{1-\gamma},$$

which is often used in the literature [301, 268, 3], which is tighter for $\gamma < 1/2$ and looser for $\gamma > 1/2$. Combining Lemma 2.48 and Theorem 2.47 yields the following result.

Corollary 2.49. *Consider $V_0 \in \mathbb{R}^{\mathcal{S}}$ and let $V_k := T^k V_0 \in \mathbb{R}^{\mathcal{S}}$ be the k-th iterate of the value iteration and $\pi_k \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ a corresponding greedy policy and let $\mu \in \Delta_{\mathcal{S}}$ be an arbitrary initial distribution. Then it holds that*

$$R^*-R(\pi_k) \leq \frac{\gamma^k}{1-\gamma}\cdot\|V_0-V^*\|_\infty \quad \textit{and} \tag{2.38}$$

$$R^*-R(\pi_k) \leq \frac{\gamma^k}{(1-\gamma)^2}\cdot\|V_0-V_1\|_\infty. \tag{2.39}$$

In particular, for $\varepsilon > 0$ we have $0 \leq R^*-R(\pi_k) \leq \varepsilon$ if

$$k \geq \frac{\log\left(\varepsilon^{-1}(1-\gamma)^{-1}\|V_0-V^*\|_\infty\right)}{\log(\gamma^{-1})} \quad \textit{or} \tag{2.40}$$

$$k \geq \frac{2\log\left(\varepsilon^{-1}(1-\gamma)^{-1}\|V_0-V_1\|_\infty\right)}{\log(\gamma^{-1})}. \tag{2.41}$$

The computational cost of one application of the Bellman operator T is $n_{\mathcal{S}}^2 \cdot n_{\mathcal{A}}$. Hence, from Corollary 2.49 we can deduce that the computational cost to obtain an accuracy of $\varepsilon > 0$ via value iteration is upper bounded by

$$O\left(n_{\mathcal{S}}^2\cdot n_{\mathcal{A}}\cdot\frac{\log(\varepsilon^{-1}(1-\gamma)^{-1})}{\log(\gamma^{-1})}\right) = O\left(n_{\mathcal{S}}^2\cdot n_{\mathcal{A}}\cdot\frac{\log(\varepsilon^{-1}(1-\gamma)^{-1})}{1-\gamma}\right)$$

for $\varepsilon \longrightarrow 0$ where we used the standard estimate $\log(t) \leq t-1$. Hence, the cost to compute an approximately optimal policy with value iteration is polynomial in size of the problem $n_{\mathcal{S}}, n_{\mathcal{A}}$ and $H_{\gamma,\varepsilon} := \frac{\log(\varepsilon(1-\gamma))}{1-\gamma}$, which has only recently been proven in [114]. The quantity $H_{\gamma,\varepsilon}$ is sometimes referred to as the *(discounted approximate) horizon* corresponding to γ

and ε. Note that the complexity of an algorithm solving an MDP up to accuracy $\varepsilon < \frac{\gamma}{4(1-\gamma)}$ is lower bounded by $\Omega(n_{\mathcal{S}}^2 \cdot n_{\mathcal{A}})$ [75].

Corollary 2.49 guarantees that value iteration produces an approximately optimal policy in a number of operations, which is logarithmic in the desired accuracy ε when γ is fixed. We can use this in order to show that value iteration produces a Bellman optimal policy in finitely many steps.

Theorem 2.50 (Iteration complexity of value iteration). *Consider a vector $V_0 \in \mathbb{R}^{\mathcal{S}}$ and let $V_k := T^k V_0 \in \mathbb{R}^{\mathcal{S}}$ be the k-th iterate of the value iteration and $\pi_k \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ a corresponding greedy policy. Further, if*[6]

$$\delta := \min\left\{ \|V^* - V^\pi\|_\infty : \pi \in \Delta_{\mathcal{A}}^{\mathcal{S}} \textit{ is deterministic and } V^\pi \neq V^* \right\} > 0,$$

then π_k is a Bellman optimal policy if

$$k > \frac{\log(\delta^{-1}(1-\gamma)^{-1}\|V^* - V_0\|_\infty)}{\log(\gamma^{-1})}. \tag{2.42}$$

Proof. By Theorem 2.47 it holds after

$$k > \frac{\log(\delta(1-\gamma)\|V^* - V_0\|_\infty^{-1})}{\log(\gamma)}$$

iterations that $\|V^* - V^{\pi_k}\|_\infty < \delta$ and hence by the definition of δ the deterministic policy π_k is Bellman optimal. □

The upper bound on the iteration complexity for value iteration required in order to return a Bellman optimal policy depends on the *horizon* $H_\gamma := \frac{\log(1-\gamma)}{1-\gamma}$ as well as the quantity δ, which captures geometric information about the set of value functions of the MDP. Bounds similar to (2.42) depending not on the geometry of value functions but rather on the number of bits required to describe the MDP and hence on the size of state and action space have been established in [282, 181].

In addition to the upper bound on the required iterations to return an optimal policy in Theorem 2.50 one can construct an MDP where value iteration takes

$$\frac{\log((1-\gamma)^{-1})}{\log(\gamma^{-1})} \geq \frac{\log((1-\gamma)^{-1})}{2(1-\gamma)}$$

iterations to return an optimal policy [181] as well as an MDP with three states and k actions that requires $e^{k-3}/\log(\gamma^{-1})$ iterations to produce an optimal policy [115]. The examples providing lower bounds rely on the existence of almost optimal deterministic policies in which case the constant $\delta \to 0$ in Theorem 2.50 and our bound grows to $+\infty$.

2.4.2. Policy iteration. Value iteration approximates the optimal value function V^* and then returns a greedy policy with respect to the approximation. In contrast, Ronald A. Howard proposed to work with policies rather than value functions and improve them iteratively through greedy updates [142, 55, 285]. For this a *policy evaluation* step computing the value function of a policy is required. This method is called *policy iteration* and formalized in Algorithm 2. An attractive property of policy iteration is that a policy returned by Algorithm 2 surely is Bellman optimal due to the policy improvement Lemma 2.22. If the greedy policies are chosen to be deterministic then policy iteration is

[6] Note that $\delta > 0$ whenever there is at least one suboptimal deterministic policy.

Algorithm 2 Howard's policy iteration (PI)

```
Require: $\pi_0 \in \Delta_{\mathcal{A}}^{\mathcal{S}}$
  $V_0 \leftarrow V^{\pi_0}$
  $k \leftarrow 0$
  while true do
    Choose $\pi_{k+1}$ greedy with respect to $\pi_k$          ▷ Uses $V = V^{\pi_k}$
    $V_{k+1} \leftarrow V^{\pi_{k+1}}$          ▷ Requires $O(n_{\mathcal{S}}^2 n_{\mathcal{A}} + n_{\mathcal{S}}^3)$ operations
    if $V_{k+1} = V_k$ then
      break
    end if
    $k \leftarrow k + 1$
  end while
  return $\pi_k$          ▷ Guaranteed to be Bellman optimal
```

terminate to converge in at most $n_{\mathcal{A}}^{n_{\mathcal{S}}}$ steps since it can at visit every deterministic policy at most once. Where this naive upper bound ensures a convergence in exponentially many steps we can deduce another convergence result by comparing policy to value iteration.

Lemma 2.51 (Policy vs. value iteration). *Consider a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ and let $\pi' \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ be a greedy improvement of π. Then it holds that $V^{\pi'} \geq TV^{\pi}$ componentwise.*

Proof. By the policy improvement Lemma 2.22 we have $Q^{\pi'} \geq Q^{\pi}$ and for $s \in \mathcal{S}$ we can estimate

$$V^{\pi'}(s) = \sum_{a\in\mathcal{A}} \pi'(a|s)Q^{\pi'}(s,a) \geq \sum_{a\in\mathcal{A}} \pi'(a|s)Q^{\pi}(s,a) = \max_{a\in\mathcal{A}} Q^{\pi}(s,a) = TV^{\pi}(s).$$

□

This lemma lets us borrow from the convergence analysis of value iteration.

Theorem 2.52 (Iteration complexity of policy iteration). *Let $\pi_0, \pi_1, \ldots \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ be a sequence of deterministic policies produced by policy iteration, see Algorithm 2. Then policy iteration terminates in at most $n_{\mathcal{A}}^{n_{\mathcal{S}}}$ steps. Further, if*[7]

$$\Delta := \frac{\min\left\{\|V^* - V^{\pi}\|_\infty : \pi \in \Delta_{\mathcal{A}}^{\mathcal{S}} \text{ is deterministic and } V^{\pi} \neq V^*\right\}}{\max\left\{\|V^* - V^{\pi}\|_\infty : \pi \in \Delta_{\mathcal{A}}^{\mathcal{S}} \text{ is deterministic}\right\}} > 0,$$

then policy iteration terminates in at most

$$\left\lceil \frac{\log(\Delta(1-\gamma))}{\log(\gamma)} \right\rceil + 2 \tag{2.43}$$

steps.

Proof. Policy iteration requires at most $n_{\mathcal{A}}^{\mathcal{S}} - 1$ steps to visit all $n_{\mathcal{A}}^{\mathcal{S}}$ deterministic policies and requires one additional step to certify the Bellman optimality of the policy and to terminate.

[7] Note that $\Delta > 0$ whenever there is at least one suboptimal deterministic policy.

By Theorem 2.50 and Lemma 2.51 the policy π_k is Bellman optimal for

$$k > \frac{\log(\delta^{-1}(1-\gamma)^{-1}\|V^* - V_0\|_\infty)}{\log(\gamma^{-1})}.$$

Now we estimate

$$\|V^* - V_0\|_\infty = \|V^* - V^{\pi_0}\|_\infty \leq \sup\left\{\|V^* - V^\pi\|_\infty : \pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}\right\}.$$

It remains to show that the supremum is attained at a deterministic policy. The supremum however is attained at the value function $\tilde{V}^*$ corresponding to a deterministic Bellman optimal policy corresponding to the reward vector $\tilde{r} := -r$, which satisfies $\tilde{V}^* \leq V^\pi$ for any $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$. Consequently we have

$$\|V^* - V^\pi\|_\infty = \max_{s\in\mathcal{S}} V^*(s) - V^\pi(s) \leq \max_{s\in\mathcal{S}} V^*(s) - \tilde{V}^*(s) = \|V^* - \tilde{V}^*\|_\infty.$$

This shows that the policy π_k is optimal after at most

$$\left\lceil \frac{\log(\Delta(1-\gamma))}{\log(\gamma)} \right\rceil + 1$$

and terminates after one additional step. □

Note that the upper bound from Theorem 2.52 does not directly depend on the size of the state and action space but rather on the constant Δ encoding geometric information of the set of value functions of the MDP. Note that the required number of iterations (2.43) can be upper bounded by

$$\left\lceil \frac{\log(\Delta^{-1}(1-\gamma)^{-1})}{1-\gamma} \right\rceil \tag{2.44}$$

This is on contrast to most bounds in the literature that depend on the problem size including the tightest known upper bound

$$O\left(\frac{n_{\mathcal{A}} \log((1-\gamma)^{-1})}{1-\gamma}\right)$$

that is due to [309, 248]. Note that this bound can neither recover our upper bound nor can it be recovered by our upper bound since Δ can decrease towards 0 for fixed $n_{\mathcal{A}}$ but also remain bounded away from 0 even when $n_{\mathcal{A}}$ is growing to $+\infty$. For non fixed discount factor lower bounds on the iteration complexity of policy iteration that are exponential in the problem size have been established [137, 21].

2.4.3. Linear programming for MDPs. Linear programming methods for the solution of Markov decision processes have been developed since the early 1960s [189, 86, 94, 302, 117, 80, 92, 140, 139] see also [154, 236, 310] for more contemporary overviews. We shortly present the linear program associated to MDPs since it both is connected to the geometry of value functions [304] and its dual formulation recovers the state-action frequencies [154] as its variables, see also Theorem 3.5. The linear programming approach played an important role in the study of the computational complexity of Markov decision processes as a carefully designed interior point method was the first algorithm that was shown to run in polynomial time for fixed discount factor [308]. Further, the simplex method applied to the linear programming formulation is closely related to policy iteration [309].

Recall the definition of the Bellman optimality operator defined in (2.36). The linear programming approach is based on the following observations, see [236, Theorem 6.2.2].

Lemma 2.53 (Order relations of the Bellman operator). *Denote the optimal value function by $V^* \in \mathbb{R}^{\mathcal{S}}$. The following statements hold:*

(*i*) *If $V \le W$ componentwise for $V, W \in \mathbb{R}^{\mathcal{S}}$ then $TV \le TW$ componentwise.*
(*ii*) *If $TV \ge V$ componentwise for $V \in \mathbb{R}^{\mathcal{S}}$ then $V \ge V^*$ componentwise.*
(*iii*) *If $TV \le V$ componentwise for $V \in \mathbb{R}^{\mathcal{S}}$ then $V \le V^*$ componentwise.*

Proof. The statement (*i*) is immediate from the definition of the Bellman optimality operator. To prove (*ii*) we use (*i*) as well as Theorem 2.47 and find that

$$V \ge TV \ge T^2V \ge \cdots \ge T^kV \longrightarrow V^* \quad \text{for } k \to \infty.$$

Finally, (*iii*) follows with an analogue argument. □

Theorem 2.54 (Linear programming formulation of MDPs). *The optimal value function $V^* \in \mathbb{R}^{\mathcal{S}}$ is the unique solution to the following linear program*

$$\text{(LP)} \qquad \text{minimize}\, \mu^\top V \quad \textit{sbj. to } V(s) \ge r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \alpha(s'|s,a)V(s') \textit{ for } s \in \mathcal{S}, a \in \mathcal{A},$$

where $\mu \in (0,\infty)^{\mathcal{S}}$ is a positive vector.

Proof. Note that the linear constraints of the linear program (LP) are equivalent to

$$V(s) \ge \max_{a\in\mathcal{A}} r(s,a) + \gamma \sum_{s'\in\mathcal{S}} \alpha(s'|s,a)V(s') = TV(s) \quad \text{for all } s \in \mathcal{S}$$

and hence equivalent to $TV \ge V$. In particular, this shows that the optimal value function $V^* \in \mathbb{R}^{\mathcal{S}}$ is a feasible point as $TV^* = V^*$. Further, for any feasible point of (LP) Lemma 2.53 guarantees that $V \ge V^*$. This shows that V^* is a solution of (LP). Assume now that $V \in \mathbb{R}^{\mathcal{S}}$ is a solution of the linear program LP. Then by Lemma 2.53 the feasibility $TV \ge V$ yields $V \ge V^*$ and by the optimality we have $\mu^\top V = \mu^\top V^*$. The positivity of μ implies that $V = V^*$, which shows that (LP) has V^* as its unique solution. □

The dual problem to (LP) was studied in the classical works [189, 80, 154] and is given by

(D-LP)

$$\text{maximize}\, r^\top \eta \quad \text{sbj. to} \sum_{a\in\mathcal{A}} \eta(s,a) = (1-\gamma)\mu(s) + \gamma \sum_{s',a'} \alpha(s|s',a')\eta(s',a') \text{ for } s \in \mathcal{S}$$
$$\text{and } \eta(s,a) \ge 0 \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}.$$

It is no coincidence that we use the symbol η for the variables of the dual linear programming formulation. Indeed, we will see in Chapter 3 that the feasible region of (D-LP) is precisely the set of state-action frequencies of the Markov process [93, 154, 236]. In other words (D-LP) describes the reward optimization problem in state-action space (ROP-SA) for the case of fully observable models. In Chapter 3 we give a characterization of the set of feasible state-action frequencies of a partially observable Markov decision process via polynomial inequalities, which yields a generalization of the linear program (D-LP).

2.4.4. Policy gradient methods. In machine learning it is a fundamental paradigm to parametrize search variables and use variants of gradient based optimizers to obtain approximate solutions of the original problem. This approach can also be taken when optimizing the policy of a Markov decision process in order to maximize the reward leading to so called *policy gradient methods* that were pioneered in [277, 40, 39]. Here, we model the policy π_θ as a smoothly element in the polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$ of conditional probability distributions of actions given states, with $\pi_\theta(a|s)$ specifying the probability of selecting action $a \in \mathcal{A}$ when currently in state $s \in \mathcal{S}$, for the parameter value θ. Using the slightly sloppy notation $R(\theta) = R(\pi_\theta)$ the vanilla policy gradient update

$$\theta_{k+1} = \theta_k + \Delta t \cdot \nabla R(\theta_k).$$

Inspired by the seminal works of Amari [13, 16], various natural policy gradient methods have been proposed [153, 206, 208]. In general, they take the form

$$\theta_{k+1} = \theta_k + \Delta t \cdot G(\theta_k)^+ \nabla R(\theta_k),$$

where $\Delta t > 0$ denotes the step size, $G(\theta)^+$ denotes the Moore-Penrose pseudo inverse and $G(\theta)_{ij} = g(dP_\theta e_i, dP_\theta e_j)$ is a Gram matrix defined with respect to some Riemannian metric g and some representation $P(\theta)$ of the parameter. Where the most popular choice is given by the mixture of Fisher information matrices [153]

$$G_K(\theta)_{ij} := \sum_s \rho_\theta(s) \sum_a \pi_\theta(a|s) \partial_{\theta_i} \log(\pi_\theta(a|s)) \partial_{\theta_j} \log(\pi_\theta(a|s))$$

other choices are possible, see also Chapter 4 for a more detailed discussion. In general, policy gradient methods converge globally when applied to fully observable problems where vanilla policy gradient methods converge at a rate of $O(t^{-1})$, which can be increased to an exponential convergence $O(e^{-ct})$ for suitable choices of $G(\theta)$; for a more detailed discussion of existing results we refer to Chapter 4.

2.4.5. Solution methods for POMDPs. Now we turn towards solution methods for partially observable Markov decision processes. Reward optimization with history dependent policies is equivalent to a belief state MDP, i.e., an MDP with continuous state space. Hence, the solution methods presented above like value and policy iteration can be applied to the belief state MDP and we refer to the survey articles [272, 254] for an overview of belief state methods. Here, we focus on methods for solving the reward optimization problem for memoryless stochastic policies although these methods extend to finite memory policies for example by augmenting the state space with an external memory [179, 231, 145]. From the approaches presented for MDPs only policy gradient methods can be applied to POMDPs without significant adjustments, however, without the global convergence guarantees that are available in the fully observable case. We present two other methods for MDPs: Bellman constrained programming, which has been proposed in [17] as well a polynomial programming approach to POMDPs generalizing the dual linear program of MDPs, which we established in [211]. These approaches both reformulate reward optimization as a polynomially constrained linear objective optimization problem and can be combined with any solver designed for such problems. The choice of the solver will crucially influence the convergence properties where many approaches will only yield locally optimal solutions. Note however that (global) reward optimization in POMDPs is NP-hard in general [286].

Bellman constrained programming. It is immediate to see that $R^\mu(\pi) = \langle \mu, V^\pi \rangle_{\mathcal{S}}$ for any policy $\pi \in \Delta_{\mathcal{A}}^{O}$ and any initial distribution $\mu \in \Delta_{\mathcal{A}}$. In the light of the Bellman equation $V^\pi = \gamma p_\pi V^\pi + (1-\gamma) r_\pi$ (see Theorem 2.9), the reward optimization problem (ROP) is equivalent to the following quadratically constrained linear program

$$\text{(BCP)} \qquad \text{maximize } \langle \mu, v \rangle \quad \text{subject to } \pi \in \Delta_{\mathcal{A}}^{O} \text{ and } v = \gamma p_\pi v + (1-\gamma) r_\pi,$$

as pointed out by [17]. We call this optimization problem the *Bellman constrained program* (BCP), which can be approached with any constrained optimization method. Here, the search variable is the tuple (π, v) of a policy and its value function that are coupled by the quadratic constraint given by the value function.

Reward optimization in state-action space and polynomial programming. Motivated by the fact that $R^\mu(\pi) = \langle r, \eta^\pi \rangle_{\mathcal{S}\times\mathcal{A}}$ we have in general introduced the reward optimization in state-action space (ROP-SA)

$$\text{(ROP-SA)} \qquad \text{maximize } \langle r, \eta \rangle_{\mathcal{S}\times\mathcal{A}} \quad \text{subject to } \eta \in \mathcal{N}^\beta.$$

For fully observable problems the set of state-action frequencies agrees with the feasible region of the dual linear program and the reward optimization problem in state-action space (ROP-SA) is precisely given by the dual linear program (D-LP). The main contribution of Chapter 3 is the characterization of the set $\mathcal{N}^\beta$ of feasible state-action frequencies by polynomial inequalities. In particular, this generalizes the dual linear program of MDPs to a polynomial program describing reward optimization in POMDPs.

CHAPTER 3

State-action geometry of partially observable MDPs

The state-action frequency η^π describes the relative (discounted) time the individual states and actions are visited in a Markov decision process when following a given policy π. The reward of a policy can be computed by weighting the reward $r(s, a)$ of a state-action pair with the state-action frequency η^π, see (2.16). Hence, reward optimization can be studied and carried out over the state-action frequencies where the resulting optimization problem

$$\text{(ROP-SA)} \qquad \text{maximize } \langle r, \eta\rangle_{\mathcal{S}\times\mathcal{A}} \quad \text{subject to } \eta \in \mathcal{N}^\beta.$$

is a linear objective problem with the state-action frequencies

$$\mathcal{N}^\beta = \{\eta^\pi : \pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}\} \subseteq \Delta_{\mathcal{S}\times\mathcal{A}}$$

as a feasible set. For the reward optimization problem of maximizing $R(\pi)$ subject to $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ much of the complexity lies in the objective R as the feasible set is given by a polytope and hence described by linear constraints. In contrast, the complexity of the the reward optimization problem (ROP-SA) lies in the geometry of the feasible set $\mathcal{N}^\beta$. The first systematic study of the geometry of the set of state-action frequencies was carried out by Cyrus Derman who showed that for fully observable problems they form a polytope $\mathcal{N}$. that we refer to as the state-action polytope [93]. In particular, this polytope coincides with the feasible region of the dual linear programming formulation (D-LP) that was previously studied [189, 94, 80]. For partially observable systems however, a decomposition of the set of state-action frequencies $\mathcal{N}^\beta$ into infinitely many convex pieces was obtained in [203].

In this chapter we obtain an explicit description of the set $\mathcal{N}^\beta$ of feasible state-action frequencies via polynomial constraints for which we give explicit expression under some conditions, see Section 3.2. For a deterministic observation process we show that the feasible state-action frequencies are described by a product of varieties of rank one matrices. In particular, this yields a description of the reward optimization problem as a polynomially constrained linear objective problem and establishes a connection of POMDPs to the field of (semi-)algebraic statistics and applied algebraic geometry. In Section 3.3 we use the explicit characterizations of the feasible region $\mathcal{N}^\beta$ of the reward optimization problem (ROP-SA) as a polynomially constrained set to gain insight regarding the properties of the reward optimization problem. More precisely, we use the theory of the *algebraic degree* of an optimization problem to bound on the number of critical points of the optimization problem over the individual faces of the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{O}}$. In Section 3.4, we demonstrate that the reward optimization problem in state-action space can be solved with different approaches like interior point methods, numerical algebraic software and a convex relaxation as a semidefinite program. We find that solving the reward maximization problem in state-action space is more stable. A further benefit is that convex relaxations are able to provide globally optimal solutions. Before we study the geometry of the set $\mathcal{N}^\beta$ of feasible state-action frequencies of a POMDP we present the fully observable case in Section 3.1.

Recall that we work under the following ergodicity assumption.

Assumption 2.14 (Uniqueness of stationary distributions). If $\gamma = 1$, we assume that for any policy $\pi \in \Delta_{\mathcal{A}}^{O}$ there exists a unique stationary distribution $\eta \in \Delta_{\mathcal{S}\times\mathcal{A}}$ of P_π.

3.1 The state-action polytope of fully observables systems

The set of all state-action frequencies is known to be a polytope in the fully observable case [93, 154]. We generalize the approach of [203] to incorporate the discounted case for a proof of this result and show that the polytope of state-action frequencies is combinatorially equivalent to the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$ under mild assumptions. We do so by computing the derivative of the state-action distributions, which also yields the well known policy gradient theorem as a consequence.

Let $\nu_\gamma^{\pi,\mu} = \nu^\pi \in \Delta_{\mathcal{S}\times\mathcal{S}}$ denote the expected number of transitions from s to s' given by

$$(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}^{\pi,\mu}(S_t = s, S_{t+1} = s') \quad \text{and} \quad \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{P}^{\pi,\mu}(S_t = s, S_{t+1} = s')$$

respectively. Note that we have

$$\nu^\pi(s,s') = \sum_{a\in\mathcal{A}}\eta^\pi(s,a)\alpha(s'|s,a), \tag{3.1}$$

which can be seen through explicit computation, e.g., in the discounted case as

$$\begin{aligned}
\nu^\pi(s,s') &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}^{\pi,\mu}(S_t = s)p_\pi(s',s) \\
&= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}^{\pi,\mu}(S_t = s)\sum_{a\in\mathcal{A}}(\pi\circ\beta)(a|s)\alpha(s'|s,a) \\
&= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}^{\pi,\mu}(S_t = s, A_t = a)\alpha(s'|s,a) \\
&= \sum_{a\in\mathcal{A}}\eta^\pi(s,a)\alpha(s'|s,a)
\end{aligned}$$

and similarly for the mean reward case.

Hence, ν^π is the image of η^π under the linear transformation

$$f_\alpha\colon \Delta_{\mathcal{S}\times\mathcal{A}} \to \Delta_{\mathcal{S}\times\mathcal{S}}, \quad \eta \mapsto \left(\sum_{a\in\mathcal{A}}\eta(s,a)\alpha(s'|s,a)\right)_{s,s'\in\mathcal{S}}. \tag{3.2}$$

Therefore, we can hope to obtain a characterization of $\mathcal{N}$ using this mapping. In order to do so, we would like to understand the structural properties of $\nu_\gamma^{\pi,\mu}$. For $\gamma = 1$ those distributions have equal marginals since we can compute

$$\sum_{s'\in\mathcal{S}}\nu_1^{\pi,\mu}(s,s') - \sum_{s'\in\mathcal{S}}\nu_1^{\pi}(s',s) = \lim_{T\to\infty}\frac{1}{T}\left(\mathbb{P}^{\pi,\mu}(S_0 = s) - \mathbb{P}^{\pi,\mu}(S_{T+1} = s)\right) = 0. \tag{3.3}$$

Note that $\sum_{s'\in\mathcal{S}} \nu_1^{\pi,\mu}(s,s') = \sum_{s'\in\mathcal{S}} \nu_1^{\pi,\mu}(s',s) = \rho_1^{\pi,\mu}(s)$ is the (unique) stationary state distribution. In the discounted case, we compute similarly

$$\begin{aligned}\sum_{s'\in\mathcal{S}} \nu_\gamma^\pi(s,s') - \gamma \sum_{s'\in\mathcal{S}} \nu_\gamma^\pi(s',s) &= (1-\gamma)\left(\sum_{t=0}^\infty \gamma^t \mathbb{P}^{\pi,\mu}(S_t = s) - \sum_{t=0}^\infty \gamma^{t+1}\mathbb{P}^{\pi,\mu}(S_{t+1} = s)\right) \\ &= (1-\gamma)\mu(s).\end{aligned} \tag{3.4}$$

If we perceive $f_\alpha(\eta_\gamma^\pi) = \nu_\gamma^\pi \in \Delta_{\mathcal{S}\times\mathcal{S}} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$ as a matrix, we have shown that

$$(\nu_\gamma^{\pi,\mu})^T \mathbb{1}_\mathcal{S} = \gamma(\nu_\gamma^{\pi,\mu})\mathbb{1}_\mathcal{S} + (1-\gamma)\mu,$$

which motivates the following definition.

Definition 3.1 (Discounted Kirchhoff polytope). For a distribution $\mu \in \Delta_\mathcal{S}$ and $\gamma \in [0,1]$ we define the *discounted Kirchhoff polytope* (this is a generalization of a definition by [296])

$$\Xi_\gamma^\mu := \left\{\nu \in \Delta_{\mathcal{S}\times\mathcal{S}} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{S}} : \nu^T\mathbb{1}_\mathcal{S} = \gamma\nu\mathbb{1}_\mathcal{S} + (1-\gamma)\mu\right\},$$

where $\mathbb{1}_\mathcal{S} \in \mathbb{R}^\mathcal{S}$ is the all one vector.

So far, we have observed that $f_\alpha(\eta_\gamma^{\pi,\mu}) \in \Xi_\gamma^\mu$ and we will see that $f_\alpha(\eta) \in \Xi_\gamma^\mu$ already implies that there is a policy π such that $\eta_\gamma^{\pi,\mu} = \eta$. This is based on the fact that for $\eta \in f_\alpha^{-1}(\Xi_\gamma^\mu)$ a policy π with state-action frequency $\eta_\gamma^{\pi,\mu} = \eta$ can be constructed by conditioning, which is well known in the context of linear programming [93, 139].

Lemma 3.2. *Let $\gamma \in [0,1]$ and $\eta \in \Delta_{\mathcal{S}\times\mathcal{A}}$ and let $\rho \in \Delta_\mathcal{S}$ denote the state marginal of η and assume that $\nu = f_\alpha(\eta) \in \Xi_\gamma^\mu$. Setting*

$$\pi(\cdot|s) := \begin{cases} \eta(\cdot|s) = \eta(s,\cdot)/\rho(s) & \text{if } \rho(s) > 0 \\ \text{arbitrary element in } \Delta_\mathcal{A} & \text{if } \rho(s) = 0, \end{cases} \tag{3.5}$$

we have $\eta_\gamma^{\pi,\mu} = \eta$. In particular it holds that $\mathcal{N}_\gamma^\mu = f_\alpha^{-1}(\Xi_\gamma^\mu)$.

Proof. By (3.3) and (3.4), it holds that $f_\alpha(\mathcal{N}_\gamma^\mu) \subseteq \Xi_\gamma^\mu$ and thus $\mathcal{N}_\gamma^\mu \subseteq f_\alpha^{-1}(\Xi_\gamma^\mu)$.

In order to show that $\eta^\pi = \eta$ for $\pi \in \Delta_\mathcal{A}^\mathcal{S}$ defined in (3.5) we calculate

$$\begin{aligned}\gamma(P^\pi)^T\eta(s,a) &= \gamma\sum_{s',a'} \alpha(s|s',a')\pi(a|s)\eta(s',a') \\ &= \gamma\pi(a|s)\sum_{s',a'}\alpha(s|s',a')\eta(s',a') \\ &= \gamma\pi(a|s)\sum_{s'}\nu(s',s) \\ &= \pi(a|s)\left(\sum_{s'}\nu(s,s') - (1-\gamma)\mu(s)\right) \\ &= \pi(a|s)\rho(s) - (1-\gamma)\pi(a|s)\mu(s) \\ &= \eta(s,a) - (1-\gamma)(\mu * \pi)(s,a).\end{aligned}$$

The unique characterization of $\eta_\gamma^{\pi,\mu}$ as the discounted stationary distribution from Theorem 2.15 yields the $\eta_\gamma^{\pi,\mu} = \eta$. We have shown that for every $\eta \in f_\alpha^{-1}(\Xi_\gamma^\mu)$ there is a policy $\pi \in \Delta_\mathcal{A}^\mathcal{S}$ such that $\eta_\gamma^{\pi,\mu} = \eta$ and hence it holds that $f_\alpha^{-1}(\Xi_\gamma^\mu) \subseteq \mathcal{N}_\gamma^\mu$. □

It will be convenient later to work under the following assumption in which ensures that policies in $\Delta_{\mathcal{A}}^{\mathcal{S}}$ are one-to-one with state-action frequencies.

Assumption 3.3 (Positivity). For every $s \in \mathcal{S}$ and $\pi \in \Delta_{\mathcal{A}}^{O}$, we assume that $\sum_a \eta_{sa}^{\pi} > 0$.

Note that this positivity assumption holds in particular, if either $\alpha > 0$ and $\gamma > 0$ or $\gamma < 1$ and $\mu > 0$ or componentwise. Indeed, if $\alpha > 0$, then the transition kernel p_π is strictly positive for any policy since

$$p_\pi(s'|s) = \sum_a (\pi \circ \beta)(a|s)\alpha(s'|s,a) > 0,$$

since $(\pi \circ \beta)(a|s) > 0$ for some $a \in \mathcal{A}$. Using that $\rho_\gamma^{\pi,\mu}$ is discounted stationary with respect to p_π (see Theorem 2.15), it holds that

$$\rho_\gamma^{\pi,\mu}(s) = \gamma \sum_{s'} \rho_\gamma^{\pi,\mu}(s')p_\pi(s|s') + (1-\gamma)\mu(s) > 0$$

since $\rho_\gamma^{\pi,\mu}(s') > 0$ for some $s' \in \mathcal{S}$. If $\mu > 0$ and $\gamma < 1$, then $\rho_\gamma^{\pi,\mu}(s) \geq (1-\gamma)\mu(s) > 0$. Assumption 3.3 is standard in linear programming approaches and necessary for the convergence of policy gradient methods in MDPs [154, 200].

Proposition 3.4 (Inverse of state-action map). *Under Assumption 3.3, the mapping*

$$\Psi \colon \Delta_{\mathcal{A}}^{\mathcal{S}} \to \mathcal{N}, \quad \pi \mapsto \eta^\pi$$

is rational and bijective with rational inverse given by conditioning

$$\Psi^{-1} \colon \mathcal{N} \to \Delta_{\mathcal{A}}^{\mathcal{S}}, \quad \eta \mapsto \pi, \quad \textit{where } \pi(a|s) = \frac{\eta(s,a)}{\sum_{a'} \eta(s,a')}.$$

Proof. We have seen in Remark 2.26 that Ψ is a rational map. By Lemma 3.2 it is one to one under Assumption 3.3 with conditioning as an inverse. □

As a consequence of Lemma 3.2, we obtain the following characterization of $\mathcal{N}_\gamma^\mu$ as a polytope, which dates back to the work of Cyrus Derman on state-action frequencies [93].

Theorem 3.5 (State-action polytope). *Let $(\mathcal{S}, \mathcal{A}, \alpha, r)$ be an MDP, $\mu \in \Delta_{\mathcal{S}}$ be an initial distribution and $\gamma \in [0,1]$. The state-action frequencies of the MDP form a polytope given by $\mathcal{N} = \mathcal{N}_\gamma^\mu = \Delta_{\mathcal{S}\times\mathcal{A}} \cap \mathcal{L}$, where*

$$\mathcal{L} := \left\{ \eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \ell_s(\eta) = 0 \, \textit{for } s \in \mathcal{S} \right\} \tag{3.6}$$

is an affine space defined by the functions

$$\ell_s(\eta) := \sum_{a\in\mathcal{A}} \eta_{sa} - \gamma \sum_{s'\in\mathcal{S}, a'\in\mathcal{A}} \eta_{s'a'}\alpha(s|s',a') - (1-\gamma)\mu_s. \tag{3.7}$$

For $\gamma \in [0,1)$, it holds that $\mathcal{N} = \mathcal{N}_\gamma^\mu = \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}} \cap \mathcal{L}$.

Proof. It remains to spell out the characterization $\mathcal{N}_\gamma^\mu = f_\alpha^{-1}(\Xi_\gamma^\mu)$ explicitely. For $\eta \in \Delta_{\mathcal{S}\times\mathcal{A}}$ the statement $\eta \in \mathcal{N}_\gamma^\mu$ is equivalent to $\eta \in \Delta_{\mathcal{S}\times\mathcal{A}}$ and $\nu := f_\alpha(\eta) \in \Xi_\gamma^\mu$. Using the definition of Ξ_γ^μ this equivalent to

$$\sum_{s'} \nu(s,s') = \gamma \sum_{s'} \nu(s',s) + (1-\gamma)\mu(s)$$

for all $s \in \mathcal{S}$. Plugging in the definition of f_α we see that the term on the left hand side is equivalent to

$$\sum_{s'}\sum_{a}\eta(s,a)\alpha(s'|s,a) = \sum_{a}\eta(s,a) = \langle\delta_s\otimes\mathbb{1}_{\mathcal{A}},\eta\rangle_{\mathcal{S}\times\mathcal{A}}$$

and the first term of the right hand side equals

$$\gamma\sum_{s'}\sum_{a}\eta(s',a)\alpha(s|s',a)$$

Hence, we have seen that $f_\alpha(\eta) \in \Xi_\gamma^\mu$ is equivalent to the condition

$$\ell_s(\eta) = \sum_{a\in\mathcal{A}}\eta_{sa} - \gamma\sum_{s'\in\mathcal{S},a'\in\mathcal{A}}\eta_{s'a'}\alpha(s|s',a') - (1-\gamma)\mu(s) = 0 \quad \text{for all } s\in\mathcal{S}, \tag{3.8}$$

which shows that $\mathcal{N} = \Delta_{\mathcal{S}\times\mathcal{A}}\cap\mathcal{L}$.

For $\gamma\in[0,1)$ and $\eta\in\mathcal{L}$ it holds that

$$0 = \sum_{s\in\mathcal{S}}\ell_s(\eta) = (1-\gamma)\sum_{s',a'}\eta(s',a') - (1-\gamma)\mu_s,$$

which implies $\sum_{s',a'}\eta(s',a') = 1$. Hence, it holds that $\mathcal{N} = \Delta_{\mathcal{S}\times\mathcal{A}}\cap\mathcal{L} = \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}}\cap\mathcal{L}$. □

Example 3.6 (A fully observable baby). We continue the crying baby example and compute the state-action frequencies of the underlying MDP, i.e., that we could achieve when we knew whether the baby is hungry or not at the time of our decision. The affine linear functions ℓ_s from (3.7) take the form

$$\begin{aligned}\ell_{s_1}(\eta) &= \eta_{s_1a_1} + \eta_{s_1a_2} - \gamma(\eta_{s_1a_2} + 0.1\eta_{s_2a_2}) - (1-\gamma)\mu_{s_1}\\ \ell_{s_2}(\eta) &= \eta_{s_2a_1} + \eta_{s_2a_1} - \gamma(\eta_{s_1a_1} + \eta_{s_2a_1} + 0.9\eta_{s_2a_2}) - (1-\gamma)\mu_{s_2}.\end{aligned}$$

For the specific choice $\gamma = 1/2$ the state-action frequencies are described by

$$\mathcal{N} = \Delta_{\mathcal{A}}^{\mathcal{S}}\cap\left\{\eta\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \begin{array}{r}20\eta_{s_1a_1} + 10\eta_{s_1a_2} - \eta_{s_2a_2} - 10\mu_{s_1} = 0\\ -10\eta_{s_1a_1} + 10\eta_{s_2a_1} + 11\eta_{s_2a_2} - 10\mu_{s_2} = 0\end{array}\right\}. \tag{3.9}$$

Hence, the reward optimization problem in state-action space (ROP-SA), which corresponds to the dual linear program (D-LP) of the MDP takes the form

$$\text{maximize } -10\eta_{s_1a_2} - \eta_{s_2a_1} \quad \text{subject to } \begin{cases}20\eta_{s_1a_1} + 10\eta_{s_1a_2} - \eta_{s_2a_2} - 10\mu_{s_1} = 0\\ -10\eta_{s_1a_1} + 10\eta_{s_2a_1} + 11\eta_{s_2a_2} - 10\mu_{s_2} = 0\\ \eta_{s_1a_1},\eta_{s_1a_2},\eta_{s_2a_1},\eta_{s_2a_2} \geq 0.\end{cases} \tag{3.10}$$

We can use this formulation to solve the MDP that underlies the crying baby example in state-action space. This is in contrast to the solution based on state policies given in 2.32. Indeed, it is clear from the formulation that 0 is an upper bound of the optimal value and hence, it suffices to construct a feasible point η such that $\eta_{s_1a_2} = \eta_{s_2a_1} = 0$. Setting $\eta_{s_1a_2} = \eta_{s_2a_1} = 0$ and solving for $\eta_{s_1a_1}$ and $\eta_{s_2a_2}$ we obtain

$$\begin{pmatrix}\eta_{s_1a_1}\\ \eta_{s_2a_2}\end{pmatrix} = \begin{pmatrix}20 & -1\\ -10 & 11\end{pmatrix}^{-1} 10\mu = \frac{1}{21}\begin{pmatrix}11\mu_{s_1} + \mu_{s_2}\\ 10\mu_{s_1} + 20\mu_{s_2}\end{pmatrix}$$

and hence a feasible point $\eta \geq 0$. In particular, this yields the deterministic Bellman optimal state policy

$$\tau^* = \begin{array}{c|cc} & s_1 & s_2 \\ a_1 & 1 & 0 \\ a_2 & 0 & 1 \end{array} \in \Delta_{\mathcal{A}}^{\mathcal{S}}$$

that feeds the baby if and only if the baby is hungry.

Improvement paths. We have seen in Subsection 2.4.3 that MDPs can be solved by means of linear programming. However, Proposition 3.2 and Theorem 3.5 together imply the stronger statement that the reward optimization problem in MDPs is up to reparametrization a linear program. In particular, this implies that the non existence of bad strict local minima, which we make precise now.

Theorem 3.7 (Existence of improvement paths in MDPs). *For every policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, there is a continuous path connecting π to an optimal policy along which the reward is monotone. If further $\pi \mapsto \eta^\pi$ is injective, the reward is strictly monotone along this path, if π is suboptimal. In particular, the superlevel sets $L_{\geq\alpha} := \{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}} : R(\pi) \geq \alpha\}$ of MDPs are connected.*

Proof. Let us fix $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ and set $\eta_0 := \eta^\pi$ and η_1 be a global optimum and η_t be the linear interpolation and ρ_t be the corresponding state marginal. Note that for $s \in \mathcal{S}$ it holds that either $\rho_t(s) > 0$ for all $t \in (0,1)$ or $\rho_t(s) = 0$ for all $t \in [0,1]$. In the latter case, we can set $\pi_t(\cdot|s)$ to be an arbitrary element in $\Delta_{\mathcal{A}}$. For the other states and $t \in (0,1)$ we can define the policy through conditioning by $\pi_t(a|s) := \eta_t(s,a)/\rho_t(s)$ and will continuously extend the definition to $t \in \{0,1\}$ in the following. If $\rho_0(s) > 0$ or $\rho_1(s) > 0$, then the definition extends naturally. Suppose that $\rho_0(s) = 0$, then we now that $\rho_1(s) > 0$ since otherwise $\rho_t(s) = 0$ for all $t \in [0,1]$. Now for $t > 0$ it holds that

$$\pi_t(s,a) = \frac{\eta_t(s,a)}{\rho_t(s)} = \frac{(1-t)\eta_0(s,a) + t\eta_1(s,a)}{(1-t)\rho_0(s) + t\rho_1(s)} = \frac{t\eta_1(s,a)}{t\rho_1(s)} = \frac{\eta_1(s,a)}{\rho_1(s)},$$

which extends continuously to $t = 0$. If $\rho_1(s) = 0$, then like before, $\pi_t(\cdot|s)$ does not depend on t and we can extend it to $t = 1$. Now we have constructed a continuous path π_t, such that $\eta^{\pi_t} = \eta_t$ and by Lemma 3.2

$$R(\pi_t) = \langle r, \eta_t \rangle = (1-t)\langle r, \eta_0 \rangle + t\langle r, \eta_1 \rangle = R(\pi_0) + t(R^* - R(\pi_0)),$$

which is strictly increasing if π_0 is suboptimal. It remains to construct a continuous path between π_0 and π. Note that if $\rho_0(s) > 0$, the policies π_0 and π agree on the state s and so does the linear interpolation between the two policies. Now, by Lemma 3.2 we see that every linear interpolation between π_0 and π has the state-action distribution η_0. Gluing the two paths, we obtain a path that first leaves the state-action distribution unchanged and then increases the reward strictly up to optimality. □

Derivative of the discounted state-action frequencies. In this paragraph we discuss the Jacobian of the parametrization $\pi \mapsto \eta^\pi$ of the discounted state-action frequencies. One motivation for this is that this Jacobian plays an important role in the relation of critical points in the policy space and the space of discounted state-action frequencies. Note that $(1-\gamma)(1-\gamma P_\pi^T)^{-1}(\mu * \pi)$ is well defined, whenever $\|P_\pi\|_2 < \gamma^{-1}$. Hence, we can extend $\pi \mapsto \eta^\pi$ onto the neighborhood

$$U := \left\{\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \|P_\pi\|_2 < \gamma^{-1}\right\} \tag{3.11}$$

of $\Delta_{\mathcal{A}}^{\mathcal{S}}$ and consider the mapping to state-action frequencies on this open set

$$\Psi = \Psi_\gamma^\mu \colon U \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}, \quad \pi \mapsto (1-\gamma)(1-\gamma P_\pi^T)^{-1}(\mu * \pi)$$

and compute the Jacobian of Ψ_γ^μ.

Lemma 3.8 (Jacobian of Ψ). *For any policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ and $s \in \mathcal{S}, a \in \mathcal{A}$ it holds that*

$$\partial_{(s,a)}\Psi(\pi) = \rho^\pi(s)(I - \gamma P_\pi^T)^{-1}(\delta_s \otimes \delta_a). \tag{3.12}$$

Hence, $\partial_{(s,a)}\Psi(\pi)$ is identical to the (s,a)-th column of $(I-\gamma P_\pi^T)^{-1}$ up to the scaling factor of $\rho^\pi(s)$. In particular, if $\rho^\pi(s) > 0$ for all $s \in \mathcal{S}$, the Jacobian $D\Psi$ has full rank.

Proof. Recall that for invertible matrices $A(t)$, it holds that

$$\partial_t A(t)^{-1} = -A(t)^{-1}(\partial_t A(t))A(t)^{-1}.$$

We compute

$$\begin{aligned}
(1-\gamma)^{-1}\partial_{(s,a)}\Psi_\gamma^\mu(\pi) &= \partial_{(s,a)}(I-\gamma P_\pi^T)^{-1}(\mu * \pi) \\
&= (\partial_{(s,a)}(I-\gamma P_\pi^T)^{-1})(\mu * \pi) + (I-\gamma P_\pi^T)^{-1}\partial_{(s,a)}(\mu * \pi) \\
&= -(1-\gamma)^{-1}(I-\gamma P_\pi^T)^{-1}\partial_{(s,a)}(I-\gamma P_\pi^T)\eta_\gamma^{\pi,\mu} \\
&\quad + (I-\gamma P_\pi^T)^{-1}(\mu * \partial_{(s,a)}\pi) \\
&= (I-\gamma P_\pi^T)^{-1}\left((1-\gamma)^{-1}\gamma(\partial_{(s,a)}P_\pi^T)\eta_\gamma^{\pi,\mu} + \mu * \partial_{(s,a)}\pi\right).
\end{aligned}$$

Further, direct computation shows

$$\begin{aligned}
((\partial_{(s,a)}P_\pi^T)\eta_\gamma^{\pi,\mu})(s,a) &= \partial_{(s,a)}\pi(a|s)\sum_{s',a'}\alpha(s|s',a')\pi(a'|s')\rho_\gamma^{\pi,\mu}(s') \\
&= (p_\pi^T\rho_\gamma^{\pi,\mu} * \partial_{(s,a)}\pi)(s,a).
\end{aligned}$$

Using the fact that $\rho_\gamma^{\pi,\mu}$ is the discounted stationary distribution, yields

$$\begin{aligned}
(1-\gamma)^{-1}\gamma(\partial_{(s,a)}P_\pi^T)\eta_\gamma^{\pi,\mu} + \mu * \partial_{(s,a)}\pi &= ((1-\gamma)^{-1}\gamma p_\pi^T\rho_\gamma^{\pi,\mu} + \mu) * \partial_{(s,a)}\pi \\
&= (1-\gamma)^{-1}\rho_\gamma^{\pi,\mu} * \partial_{(s,a)}\pi,
\end{aligned}$$

which shows (3.12). We compute

$$\rho_\gamma^{\pi,\mu} * \partial_{(s,a)}\pi)(s',a') = \rho_\gamma^{\pi,\mu}(s')\partial_{(s,a)}\pi(a'|s') = \rho_\gamma^{\pi,\mu}(s)(\delta_s \otimes \delta_a)(s',a').$$

Note that $(I-\gamma P_\pi^T)^{-1}(\delta_s \otimes \delta_a)$ is precisely the (s_0,a_0)-th column of the matrix $(I-\gamma P_\pi^T)^{-1}$. Those columns are linearly independent, and so are the partial derivatives $\partial_{(s,a)}\Psi_\gamma^\mu(\pi)$, given that the discounted stationary distribution $\rho_\gamma^{\pi,\mu}$ vanishes nowhere. □

Corollary 3.9 (Dimension of $\mathcal{N}$). *Assume that $\rho_\gamma^{\pi,\mu} > 0$ entrywise for some policy $\pi \in \operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$. Then we have* $\dim(\mathcal{N}) = \dim(\Delta_{\mathcal{A}}^{\mathcal{S}}) = |\mathcal{S}|(|\mathcal{A}|-1)$.

Proof. By Lemma 3.8 the mapping Ψ^μ_γ is full rank in a neighborhood of π and hence, we have

$$\dim(\mathcal{N}) = \dim(\Psi^\mu_\gamma(\Delta^{\mathcal{S}}_{\mathcal{A}})) = \dim(\Delta^{\mathcal{S}}_{\mathcal{A}}).$$

□

Let us consider a parametrized policy model $\Pi_\Theta = \{\pi_\theta : \theta \in \Theta\} \subseteq \Delta^{\mathcal{S}}_{\mathcal{A}}$ with differentiable parametrization $\theta \mapsto \pi_\theta$ where $\Theta \subseteq \mathbb{R}^p$. When unambiguous, we drop the policy in the notation of reward, state(-action) frequncies and value functions, i.e., simply write $R(\theta), \eta_\theta, Q_\theta$ etc. instead of $R(\pi_\theta), \eta^{\pi_\theta}, Q^{\pi_\theta}$.

Lemma 3.10 (Parameter derivatives of discounted state-action frequencies). *It holds that*

$$\partial_{\theta_i}\eta^\theta_\gamma = (I - \gamma P^T_{\pi_\theta})^{-1}(\rho_\theta * \partial_{\theta_i}\pi_\theta),$$

where

$$(\rho_\theta * \partial_{\theta_i}\pi_\theta)(s, a) = \rho_\theta(s)\partial_{\theta_i}\pi_\theta(a|s).$$

Proof. This follows directly from the application of the chain rule and (3.12). □

Using this expression, we can compute the parameter gradient with respect to the discounted reward and recover the well known policy gradient theorem.

Theorem 3.11 (Policy gradient theorem, [277, 190, 2]). *It holds that*

$$(1-\gamma)\partial_{\theta_i}R(\theta) = \sum_s \rho_\theta(s)\sum_a \partial_{\theta_i}\pi_\theta(a|s)Q_\theta(s,a) = \sum_{s,a}\eta_\theta(s,a)\partial_{\theta_i}\log(\pi_\theta(a|s))Q_\theta(s,a).$$

Proof. Using the preceding lemma, we compute

$$\begin{aligned}
\partial_{\theta_i}R(\theta) &= \langle (I-\gamma P^T_{\pi_\theta})^{-1}\rho_\theta * \partial_{\theta_i}\pi_\theta, r\rangle_{\mathcal{S}\times\mathcal{A}} \\
&= \langle \rho_\theta * \partial_{\theta_i}\pi_\theta, (I-\gamma P_{\pi_\theta})^{-1} r\rangle_{\mathcal{S}\times\mathcal{A}} \\
&= (1-\gamma)^{-1}\langle \rho_\theta * \partial_{\theta_i}\pi_\theta, Q_\theta\rangle_{\mathcal{S}\times\mathcal{A}} \\
&= (1-\gamma)^{-1}\sum_s \rho_\theta(s)\sum_a \partial_{\theta_i}\pi_\theta(a|s)Q_\theta(s,a) \\
&= (1-\gamma)^{-1}\sum_{s,a}\eta_\theta(s,a)\partial_{\theta_i}\log(\pi_\theta(a|s))Q_\theta(s,a).
\end{aligned}$$

□

The policy gradient theorem can be used to estimate the gradient of the reward function by estimating the state frequency and the Q-value function [40, 39, 207, 276] where the derivative of the policy model $\partial_{\theta_i}\pi_\theta$ is often relatively cheap to compute.

Remark 3.12 (Policy gradients for POMDPs). The case of partial observability can be regarded as a special case of parametrized policies. In fact the observation mechanism β induces a linear map $\pi \mapsto \pi \circ \beta$. This interpretation shows that the policy gradient theorem is also valid for partially observable problems.

Corollary 3.13 (Lipschitz continuity of the reward). *It holds that*

$$\|\nabla R(\theta)\|_\infty \le \max_{i=1,\dots,p}\frac{\|r\|_\infty \cdot \|\partial_{\theta_i}\pi_\theta\|_\infty}{1-\gamma} \tag{3.13}$$

In particular, it holds that

$$|R(\pi) - R(\pi')| \le \|\pi - \pi'\|_1 \cdot \frac{\|r\|_\infty}{1-\gamma} \quad \textit{for all } \pi, \pi' \in \Delta^{\mathcal{S}}_{\mathcal{A}} \tag{3.14}$$

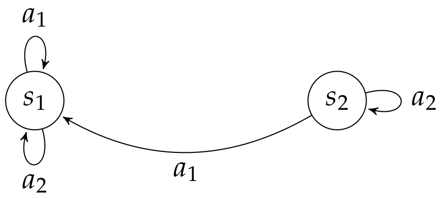

FIGURE 3.1. Transition graph of the example.

and

$$\|\eta^\pi - \eta^{\pi'}\|_\infty \le \frac{\|\pi-\pi'\|_1}{1-\gamma} \quad \textit{for all } \pi, \pi' \in \Delta_{\mathcal{A}}^{\mathcal{S}} \tag{3.15}$$

Proof. The estimate (3.13) follows directly from the policy gradient Theorem 3.11 since $\|Q_\theta\|_\infty \le \|r\|_\infty$ and $\|\rho_\theta\|_\infty \le 1$. Note that (3.13) holds for any parametrization $\Theta \to U$ for the neighborhood U of $\Delta_{\mathcal{A}}^{\mathcal{S}}$ defined in (3.11). For $\pi, \pi' \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ the mean value theorem implies the existence of $\hat{\pi} \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ such that $R(\pi) - R(\pi') = \nabla R(\hat{\pi})^\top(\pi - \pi')$ and hence we can estimate

$$|R(\pi) - R(\pi')| = |\nabla R(\hat{\pi})^\top(\pi - \pi')| \le \|\pi - \pi'\|_1 \cdot \|\nabla R(\hat{\pi})\|_\infty \le \|\pi - \pi'\|_1 \cdot \frac{\|r\|_\infty}{1-\gamma}.$$

The statement about the state-action frequencies follows when we perceive the entry $\eta^\pi(s,a)$ as the reward for $r = \delta_{(s,a)} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. □

Example 3.14 (Blow up of Lipschitz constant). We convince ourselves that the bound on the Lipschitz constant can be attained. Consider a Markov decision process with two states and two actions and deterministic state transitions as depicted in Figure 3.1. As a reward vector we choose $r(s,a) = \delta_{ss_1}$ and hence the reward equals the first entry of the state frequency, i.e., $R(\pi) = \rho^\pi(s_1)$. Hence, the reward is given by

$$R(\pi) = \rho^\pi(s_1) = \delta_{s_1}^\top(1-\gamma)(I - \gamma p_\pi^T)\mu = \mu_{s_1} + \mu_{s_2} \cdot \frac{\gamma\pi(a_1|s_2)}{1-\gamma+\gamma\pi(a_1|s_2)},$$

where we omit the explicit computation. Hence, the reward only depends on $p = \pi(a_1|s_2)$ and we write $R(p)$ instead of $R(\pi)$. Note that

$$R'(p) = \mu_{s_2} \cdot \frac{1-\gamma}{(1-\gamma+\gamma p)^2}$$

and hence $R'(p) \to (1-\gamma)^{-1}$ for $p \to 0$ if $\mu = \delta_{s_2}$. Hence, the Lipschitz constant of the reward R is $(1-\gamma)^{-1}$, which shows that the bound (3.13) can be attained.

The face lattice in the fully observable case. So far, we have seen that the set of state-action frequencies form a polytope in the fully observable case. However, not all polytopes are equally complex and thus we aim to describe the *face lattice* of $\mathcal{N}_\gamma^\mu$, which describes the combinatorial properties of a polytope, see [318].

Theorem 3.15 (Combinatorial equivalence of $\mathcal{N}_\gamma^\mu$ and $\Delta_{\mathcal{A}}^{\mathcal{S}}$). *Let $(\mathcal{A}, \mathcal{S}, \alpha, r)$ be an MDP and $\gamma \in [0,1]$. Then $\Psi \colon \Delta_{\mathcal{A}}^{\mathcal{S}}, \pi \mapsto \eta^\pi$ induces an order preserving surjective morphism between the face lattices of $\Delta_{\mathcal{A}}^{\mathcal{S}}$ and $\mathcal{N}$, such that for every $I \subseteq \mathcal{S} \times \mathcal{A}$ it holds that*

$$\left\{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}} : \pi(a|s) = 0 \textit{ for all } (s,a) \in I\right\} \mapsto \{\eta \in \mathcal{N} : \eta(s,a) = 0 \textit{ for all } (s,a) \in I\}. \tag{3.16}$$

If in addition Assumption 3.3 holds, this is a dimension preserving isomorphism.

Proof. First, we note that the faces of both $\Delta_{\mathcal{A}}^{\mathcal{S}}$ and $\mathcal{N}_\gamma^\mu$ have the structure of the left and right hand side of (3.16) respectively, which follows from Theorem 3.5. Denote now the left and right hand side in (3.16) by F and G respectively, then we need to show that $\Psi(F) = G$. For $\pi \in F$ it holds that

$$\eta^\pi(s,a) = \rho^\pi(s)\pi(a|s) = 0 \quad \text{for all } (s,a) \in I$$

and hence $\eta^\pi \in G$. On the other hand for $\eta \in G$ we can set $\pi(\cdot|s) := \eta(\cdot|s)$ whenever defined and any other element such that $\pi(a|s) = 0$ for all $(s,a) \in I$ otherwise. Then we surely have $\pi \in F$ and by Lemma 3.2 also $\eta^\pi = \eta$. To check that the mapping described in (3.16) is a morphism, consider two faces F_1, F_2 of $\Delta_{\mathcal{A}}^{\mathcal{S}}$. It holds that $\Psi(F_1 \wedge F_2) = \Psi(F_1 \cap F_2) = \Psi(F_1) \cap \Psi(F_2) = G_1 \wedge G_2$, where $G_i := \Psi(F_i)$. Further, we have that

$$\Psi(F_1 \vee F_2) = \Psi\left(\bigcap_{\substack{F \in \mathcal{F}(\Delta_{\mathcal{A}}^{\mathcal{S}}) \\ F_1, F_2 \subseteq F}} F \right) = \bigcap_{\substack{F \in \mathcal{F}(\Delta_{\mathcal{A}}^{\mathcal{S}}) \\ F_1, F_2 \subseteq F}} \Psi(F) = \bigcap_{\substack{G \in \mathcal{F}(\mathcal{N}) \\ G_1, G_2 \subseteq G}} G = G_1 \vee G_2,$$

which shows that the join and meet are respected.

In the case that $\rho^\pi > 0$ entrywise for all policies $\pi \in \operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$, the mapping $\eta \mapsto \eta(\cdot|\cdot)$ defines an inverse to Ψ, which shows that the mapping defined in (3.16) is bijective. The assertion on the dimension follows from basic dimension counting, from the fact that the rank is preserved by a lattice isomorphism or by virtue of Lemma 3.8. □

State-action frequencies of history dependent policies. Consider an MDP with arbitrary history dependent policies. In order to study the state-action frequencies that can be achieved with this larger policy class we follow the same approach like for memoryless policies. This shows that in MDPs history dependent policies achieve the same state-action frequencies as memoryless policies.

Theorem 3.16 (State-action frequencies of history policies). *Consider an MDP $(\mathcal{S}, \mathcal{A}, \alpha, r)$. The set of state-action frequencies that can be achieved by history dependent policies agrees with the state-action frequencies induced by memoryless policies. In particular, they form a polytope with explicit expression given in Theorem 3.5. This shows that memoryless policies achieve the same optimal reward for fully observable problems.*

Proof. For a history dependent policy π we consider the state-state transition frequencies $\nu^\pi = f_\alpha(\eta^\pi)$. Revisiting (3.3) and (3.4) reveals that they do not require the policy to be memoryless and hence it holds that $f_\alpha(\eta^\pi) \in \Xi_\gamma^\mu$. By Lemma 3.2 there is a memoryless policy $\pi' \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ such that $\eta^{\pi'} = \eta^\pi$. This shows that the state-action frequencies that can be achieved by history and memoryless policies agree for fully observable MDPs. □

The approach of using state-action frequencies for the reduction to memoryless (stationary) policies can be traced back to [93] see also [236, 167] for further discussions of history dependent policy classes and the reduction to memoryless policies for fully observable systems.

Beyond finite MDPs. The state-action frequencies of countable MDPs have been studied in [10] and were shown to form a compact and convex set where the extreme points correspond to memoryless deterministic policies. This complements the results that for finite MDPs the state-action frequencies form a polytope with vertices corresponding to the deterministic policies. The reduction of history dependent policies to memoryless policies like in Theorem 3.16 for general state and action spaces can be found in [167].

3.2 State-action geometry of partially observable systems

Now that we have revisited the classic characterization of the state-action frequencies of an MDP as a polytope we study partially observable systems. We have seen in Proposition 2.16 that the mapping from policies to state-action frequencies is a rational function. By the Tarski-Seidenberg theorem the image of a semialgbraic map under a rational function is again semialgebraic, i.e., the finite union of polynomially constrained sets. Since the set of policies $\Delta_{\mathcal{A}}^{\mathcal{O}}$ is a polytope the set of state-action frequencies $\mathcal{N}^{\beta}$ is a semialgebraic set, however, the Tarski-Seidenberg theorem does not provide an explicit description of the range. In this section we provide an explicit characterization of the set $\mathcal{N}_{\beta}$ of feasible state-action frequencies via polynomial inequalities. In particular, this shows that feasible state-action frequencies of a POMDP form a (semi)algebraic statistical model.

3.2.1. General description of state-action frequencies. We start with a general description of the state-action frequencies associated to a constrained policy class of a an MDP. Note that POMDPs can be considered as a special case of constrained MDPs with the policy class being the effective policies $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$, which form a polytope inside $\Delta_{\mathcal{A}}^{\mathcal{S}}$.

Proposition 3.17 (General characterization of state-action frequencies). *Consider an MDP $(\mathcal{S},\mathcal{A},\alpha,r)$, an initial distribution $\mu\in\Delta_{\mathcal{S}}$ and a family of policies $\Pi = X\cap\Delta_{\mathcal{A}}^{\mathcal{S}}$ and denote the set of feasible state-action frequencies by $\mathcal{N}_{\Pi} = \{\eta^{\pi} : \pi\in\Pi\}$. Under the ergodicity Assumption 2.14 and the positivity Assumption 3.3 it holds that*

$$\mathcal{N}_{\Pi} = \left\{\eta\in\mathbb{R}_{>0}^{\mathcal{S}\times\mathcal{A}} : \eta(\cdot|\cdot)\in X\right\}\cap\mathcal{N}, \tag{3.17}$$

where $\mathcal{N}$ is the state-action polytope, see Theorem 3.5.

Proof. This is a direct consequence of Lemma 3.2. □

For polynomially constrained policy classes this general principle implies that also the state-action frequencies are polynomially constrained.

Theorem 3.18 (State-action frequencies of polynomially constrained policy models). *Let $(\mathcal{S},\mathcal{O},\mathcal{A},\alpha,\beta,r)$ be a POMDP and let both the ergodicity Assumption 2.14 and the positivity Assumption 3.3 hold and consider a polynomially constrained policy set*

$$\Pi = \Delta_{\mathcal{A}}^{\mathcal{S}}\cap\left\{\pi\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}} : f_i(\pi)\geq 0 \text{ for } i=1,\dots,k\right\}$$

with defining polynomials

$$f_i(\pi) = \sum_{j=1}^{n} b_j^{(i)}\prod_{s,a}\pi(s,a)^{a_j^{(i)}(s,a)}. \tag{3.18}$$

Then the corresponding state-action frequencies form a polynomially constrained set given by

$$\mathcal{N}_\Pi = \mathcal{N} \cap \left\{ \eta \in \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}} : h_i(\eta) \geq 0 \, \textit{for } i = 1, \dots, k \right\}, \tag{3.19}$$

where

$$h_i(\eta) = \sum_{j=1}^{n} b_j^{(i)} \prod_{s,a} \eta(s,a)^{a_j^{(i)}(s,a)} \prod_{s} \rho(s)^{d_s^{(i)} - a_j^{(i)}(s)} \tag{3.20}$$

and $\rho(s) := \sum_a \eta(s,a), a_j^{(i)}(s) := \sum_a a_j^{(i)}(s,a)$ *and* $d_s^{(i)} := \max_j a_j^{(i)}(s)$*. It holds that*

$$\deg(h_i) \leq \sum_{s\in\mathcal{S}} d_s^{(i)} = \sum_{s\in\mathcal{S}} \max_{j=1,\dots,n} \sum_{a\in\mathcal{A}} a_j^{(i)}(s,a). \tag{3.21}$$

Finally, $\mathcal{N}_\Pi$ *is a basic semialgebraic set and* $\mathcal{N}_\Pi$ *is combinatorially equivalent to* Π*.*

Proof. In order to use Proposition 3.17 we set

$$X := \left\{ \pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : f_i(\pi) \geq 0 \text{ for } i = 1, \dots, k \right\}.$$

By (3.17) the feasible state-action frequencies are described by the inequalities

$$g_i(\eta) := f_i(\eta(\cdot|\cdot)) \geq 0.$$

When f_i is polynomial then g_i is a rational function and takes the form

$$g_i(\eta) = \sum_{j=1}^{n} b_j^{(i)} \prod_{s,a} \left(\frac{\eta(s,a)}{\rho(s)} \right)^{a_j^{(i)}(s,a)} = \sum_{j=1}^{n} b_j \frac{\prod_{s,a} \eta(s,a)^{a_j^{(i)}(s,a)}}{\prod_s \rho(s)^{a_j^{(i)}(s)}},$$

where $a_j^{(i)}(s) := \sum_{a\in\mathcal{A}} a_j^{(i)}(s,a)$ and $\rho(s) = \sum_{a\in\mathcal{A}} \eta(s,a)$ denotes the state marginal. Taking $d_s^{(i)} := \max_j a_j^{(i)}(s)$ we can multiply g by $\prod_s \rho(s)^{d_s^{(i)}}$ to obtain a polynomial h given by

$$h(\eta) = \sum_{j=1}^{n} b_j^{(i)} \prod_{s,a} \eta(s,a)^{a_j^{(i)}(s,a)} \prod_{s} \rho(s)^{d_s^{(i)} - a_j^{(i)}(s)}$$

such that $\{\eta \in \mathbb{R}_{>0}^{\mathcal{S}\times\mathcal{A}} : g_i(\eta) \geq 0\} = \{\eta \in \mathbb{R}_{>0}^{\mathcal{S}\times\mathcal{A}} : h_i(\eta) \geq 0\}$.

Note that the degree of the polynomials h_i is bounded by

$$\begin{aligned} \deg(h_i) &\leq \max_{j=1,\dots,n} \sum_{s,a} a_j^{(i)}(s,a) + \sum_s (d_s^{(i)} - a_j^{(i)}(s)) \\ &= \max_{j=1,\dots,n} \sum_s d_s^{(i)} = \sum_s d_s^{(i)} = \sum_s \max_j \sum_a a_j^{(i)}(s,a). \end{aligned}$$

Regarding the combinatorial equivalence, we note that Proposition 3.2 implies that the mapping $\Psi \colon \Delta_{\mathcal{A}}^{\mathcal{S}} \to \mathcal{N}, \pi \mapsto \eta^\pi$ induces a bijection of the face lattices of Π and $\mathcal{N}_\Pi$ according to

$$F_I := \Pi \cap \left\{ \pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : f_i(\pi) = 0 \text{ for } i \in I \right\} \mapsto G_I := \mathcal{N}_\Pi \cap \left\{ \eta \in \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}} : h_i(\eta) = 0 \text{ for } i \in I \right\}$$

for $I \subseteq \{1, \dots, n\}$. It remains to show that this bijection respects the join and meet, i.e., that $\Psi(F_I \wedge F_J) = G_I \wedge G_J$ as well as $\Psi(F_I \vee F_J) = G_I \vee G_J$ for any $I, J \subseteq \{1, \dots, n\}$ where

$F_I \wedge F_J = F_I \cap F_J$ and $F_I \vee F_J$ is the intersection over all faces containing F_I and F_J. First, we note that

$$\Psi(F_I \wedge F_J) = \Psi(F_I \cap F_J) = \Psi(F_{I\cup J}) = G_{I\cup J} = G_I \cap G_J = G_I \wedge G_J.$$

Further, let us denote the faces of Π and $\mathcal{N}_\Pi$ by $\mathcal{F}(\Pi)$ and $\mathcal{F}(\mathcal{N}_\Pi)$ then we have

$$G_I \vee G_J = \bigcap_{\substack{G\in\mathcal{F}(\mathcal{N}_\Pi)\\ G_I,G_J\subseteq G}} G = \bigcap_{\substack{F\in\mathcal{F}(\Pi)\\ F_I,F_J\subseteq F}} \Psi(F) = \Psi\left(\bigcap_{\substack{F\in\mathcal{F}(\Pi)\\ F_I,F_J\subseteq F}} F\right) = \Psi(F_I \vee F_J).$$

□

In general, the degree of h_i can be higher than the degree of f_i. For example we will see later that for linearly constrained policy models Π the resulting state-action frequencies $\mathcal{N}_\Pi$ do not necessarily form a polytope, see Example 3.23. However, in the case that $\sum_a a_j(s,a) = \max_j \sum_a a_j(s,a)$ for all $j = 1,\dots,n$ it holds that

$$\deg(h_i) \le \sum_s \max_j \sum_a a_j^{(i)}(s,a) = \max_j \sum_{s,a} a_j^{(i)}(s,a) = \deg(f_i). \tag{3.22}$$

Let us now come to the case of partially observable models. By virtue of Theorem 3.18 it suffices to characterize the set of state policies that can be realized for a given observation kernel β.

Definition 3.19 (Effective policies). For an observation policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ we call the state policy $\tau = \pi\circ\beta \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ defined via $\tau(a|s) := \sum_{o\in\mathcal{O}} \pi(a|o)\beta(o|s)$ the corresponding *effective policy*. We denote the set of effective policies by

$$\Delta_{\mathcal{A}}^{\mathcal{S},\beta} = \left\{\pi\circ\beta : \pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}\right\} \subseteq \Delta_{\mathcal{A}}^{\mathcal{S}}$$

and refer to it as the *effective policy polytope*.

Note that $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$ is indeed a polytope since it is the image of the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{O}}$ under the linear mapping $\pi \mapsto \pi\circ\beta$. Hence, the effective policy polytope has a description by linear inequalities

$$\Delta_{\mathcal{A}}^{\mathcal{S},\beta} = \{\tau \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : f_i(\tau) \ge c_i\} \cap \Delta_{\mathcal{A}}^{\mathcal{S}} \tag{3.23}$$

for suitable linear functions $f_i(\tau) = \sum_{s,a} b_{sa}^{(i)}\tau_{sa}$.

Corollary 3.20 (State-action frequencies of POMDPs). *Let $(\mathcal{S},\mathcal{O},\mathcal{A},\alpha,\beta,r)$ be a POMDP, $\mu \in \Delta_{\mathcal{S}}$ and $\gamma \in [0,1]$ and let Assumption 3.3 hold. Then we have the feasible state-action frequencies $\mathcal{N}_\beta$ form a polynomially constrained subset of the state-action polytope $\mathcal{N}$ that is combinatorially equivalent to the effective policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$. Further, if with the description* (3.23) *we have*

$$\mathcal{N}_\beta = \{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : g_i(\eta) \ge 0\} \cap \mathcal{N} \tag{3.24}$$

for the multi-homogeneous polynomials

$$g_i(\eta) := \sum_{s\in S_i}\sum_a b_{sa}^{(i)}\eta_{sa} \prod_{s'\in S^i\setminus\{s\}}\sum_{a'}\eta_{s'a'} - c_i \prod_{s'\in S_i}\sum_{a'}\eta_{s'a'}, \tag{3.25}$$

where $S^i = \{s \in \mathcal{S} : b^{(i)}_{sa} \neq 0$ *for some* $a \in \mathcal{A}\}$. *It holds that*

$$\deg(g_i) \leq |S^i| = \left|\left\{s \in \mathcal{S} : b^{(i)}_{sa} \neq 0 \textit{ for some } a \in \mathcal{A}\right\}\right|. \tag{3.26}$$

Proof. The statement is a direct consequence of the general characterization in Theorem 3.18. To see the statement about the degree we use (3.21) in combination with

$$\max_j \sum_a a^{(i)}_j(s,a) = \begin{cases} 1 & \text{if } b^{(i)}_{sa} \neq 0 \text{ for some } a \in \mathcal{A} \\ 0 & \text{otherwise.} \end{cases}$$

□

According to the preceding corollary, a linear inequality in the state policy polytope $\Delta^{\mathcal{S}}_{\mathcal{A}}$ involving actions of k different states yields a polynomial inequality of degree k in the set of state-action frequencies $\mathcal{N}$. In particular, for a linearly constrained policy model $\Pi \subseteq \Delta^{\mathcal{S}}_{\mathcal{A}}$, where every constraint only addresses a single state, the set of state-action frequencies induced by these policies will still form a polytope. This shows that this type of box constraints are well aligned with the algebraic geometric structure of the problem. The linear constraints arising from partial observability never exhibit this box type structure – unless the system is equivalent to its fully observable version. This is because the projection of the effective policy polytope $\Delta^{\mathcal{S},\beta}_{\mathcal{A}}$ onto a single state always gives the entire probability simplex $\Delta_{\mathcal{A}}$, which is never the case, if there is a non trivial linear constraint concerning only this state.

A polynomial programming formulation of POMDPs. We have seen that the feasible state-action frequencies of a POMDP – and of a MDP with polynomially constrained policy class – are described by polynomial inequalities within the state-action polytope. In particular, this shows that the reward optimization problem in state-action space

$$\text{maximize } \langle r, \eta\rangle_{\mathcal{S}\times\mathcal{A}} \quad \text{subject to } \eta \in \mathcal{N}^{\beta}$$

becomes a linear objective polynomially constrained program. This polynomial optimization problem can be seen as a direct generalization of the dual linear programming formulation (D-LP) of MDPs to partially observable problems.

3.2.2. Explicit formulas for POMDPs with injective β. By Corollary 3.20 it suffices to find the describing linear inequalities of the policy polytope $\Delta^{\mathcal{S},\beta}_{\mathcal{A}}$, which is the image of the policy polytope $\Delta^{\mathcal{O}}_{\mathcal{A}}$ under the linear map $\pi \mapsto \pi \circ \beta$. Obtaining inequality descriptions of the images of polytopes under linear maps is a fundamental problem that is non-trivial in general. It can be approached algorithmically, e.g., by Fourier-Motzkin elimination, block elimination, vertex approaches, and equality set projection [150]. We characterize the image of a polytope under a linear map $x \mapsto Ax$ for the special case where the linear map is injective, corresponding to the case where the matrix A has linearly independent columns. As a polytope is a finite intersection of closed half spaces $H = \{x : v^T x \geq \alpha\}$, it suffices to characterize the image AH. It holds that

$$AH = \left\{y \in \operatorname{range} A : v^T A^+ y \geq \alpha\right\} = \left\{y : ((A^+)^T v)^T y \geq \alpha\right\} \cap \ker(A^T)^{\perp}, \tag{3.27}$$

where A^+ is a pseudoinverse and where we used that for $y \in \operatorname{range} A$ the injectivity of A implies that A^+y is the unique pre-image of y.

Let us now come back to the mapping $\pi \mapsto \pi \circ \beta$. When the Markov kernels π and β are expressed as row stochastic matrices, i.e., when $\pi_{oa} = \pi(a|o)$ and $\beta_{so} = \beta(o|s)$ the mapping takes the form $\pi \mapsto \beta\pi$. In vectorized form, this map corresponds to $\operatorname{vec}(\beta\pi I) = (I^T \otimes \beta)\operatorname{vec}(\pi)$ [244, 4]. Hence the linear map is represented by the matrix $B = I \otimes \beta$. We observe that $(I \otimes \beta)^+ = I \otimes \beta^+$ [166, Section 2.6.3]. Notice that $B = I \otimes \beta$ has linearly independent columns if and only if β does, which leads us to the following assumption.

Assumption 3.21. The matrix $\beta \in \Delta_{\mathcal{O}}^{\mathcal{S}} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{O}}$ has linearly independent columns.

The assumption above does not imply that the system is fully observable. Recall that if β has linearly independent columns, the Moore-Penrose takes the form $\beta^+ = (\beta^T\beta)^{-1}\beta^T$. An interesting special case is when β is deterministic but may map several states to the same observation, which is often referred to as *state aggregation*. In this case,

$$\beta^+ = \operatorname{diag}(n_1^{-1}, \dots, n_{|\mathcal{O}|}^{-1})\beta^T, \tag{3.28}$$

where n_o denotes the number of states with observation o. Here, β^+_{so} agrees with the conditional distribution $\beta(s|o)$ with respect to a uniform prior over the states; however, this is not in general the case since β^+ can have negative entries.

Now we elaborate the implications of our general discussion above for the effective policy polytope.

Theorem 3.22 (H-description of the effective policy polytope). *Let $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$ be a POMDP and let Assumption 3.21 hold. Then it holds that*

$$\Delta_{\mathcal{A}}^{\mathcal{S},\beta} = \mathcal{U} \cap \mathcal{C} \cap \mathcal{D} = \Delta_{\mathcal{A}}^{\mathcal{S}} \cap \mathcal{U} \cap \mathcal{C}, \tag{3.29}$$

where $\mathcal{U} = \ker(I \otimes \beta^T)^\perp$ is a subspace, $\mathcal{C} = \{\tau \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \beta^+\tau \geq 0\}$ is a pointed polyhedral cone and $\mathcal{D} = \{\tau \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \sum_a (\beta^+\tau)_{oa} = 1$ for all $o \in \mathcal{O}\}$ an affine subspace. Further, the face lattices of $\Delta_{\mathcal{A}}^{\mathcal{O}}$ and $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$ are isomorphic.

Proof. Under Assumption 3.3 the mapping $\Delta_{\mathcal{A}}^{\mathcal{O}} \to \Delta_{\mathcal{A}}^{\mathcal{S},\beta}, \pi \mapsto \pi \circ \beta$ is linear and bijective. In particular, this map induces an isomorphism between the face lattices, see [318].

By the above discussion, if β has linearly independent columns, then an inequality $\langle \pi, \nu \rangle \geq 0$ in the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{O}}$ corresponds to an inequality $\langle \tau, (\beta^+)^T\nu \rangle \geq 0$ in the polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$.

We recall the defining linear (in)equalities of the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{O}}$, which are given by

$$\begin{aligned} \pi(a|o) = \langle \delta_o \otimes \delta_a, \pi \rangle_{\mathcal{O}\times\mathcal{A}} &\geq 0 \quad \text{for all } a \in \mathcal{A}, o \in \mathcal{O} \quad \text{and} \\ \sum_a \pi(a|o) = \langle \delta_o \otimes \mathbb{1}_{\mathcal{A}}, \pi \rangle_{\mathcal{O}\times\mathcal{A}} &= 1 \quad \text{for all } o \in \mathcal{O}. \end{aligned}$$

Hence, by the general discussion from above, namely by (3.27), it holds that

$$\Delta_{\mathcal{A}}^{\mathcal{S},\beta} = \ker(I \otimes \beta^T)^\perp \cap \{\tau : \beta^+\tau \geq 0\} \cap \Big\{\tau : \sum_a (\beta^+\tau)_{oa} = 1 \text{ for all } o \in \mathcal{O}\Big\}.$$

Note that the linear inequalities $\sum_a(\beta^+\tau)_{oa} = 1$ are redundant in $\Delta_{\mathcal{A}}^{\mathcal{S}}$. To see this, we note that $\beta^+\mathbb{1}_{\mathcal{S}} = \mathbb{1}_{\mathcal{O}}$ by the injectivity of β and $\beta\mathbb{1}_{\mathcal{O}} = \mathbb{1}_{\mathcal{S}}$. Now we can check that

$$\sum_a(\beta^+\tau)_{oa} = \sum_a\sum_s\beta^+_{os}\tau_{sa} = \sum_s\beta^+_{os}\sum_a\tau_{sa} = \sum_s\beta^+_{os} = 1.$$

This together with $\beta(\Delta_{\mathcal{A}}^{\mathcal{O}}) \subseteq \Delta_{\mathcal{A}}^{\mathcal{S}}$ shows that

$$\Delta_{\mathcal{A}}^{\mathcal{S},\beta} = \Delta_{\mathcal{A}}^{\mathcal{S}} \cap \ker(I\otimes\beta^T)^\perp \cap \{\tau : \beta^+\tau \geq 0\}.$$

□

Explicit formulas for the feasible state-action frequencies of POMDPs. Corollary 3.20 provides a description of the set of feasible state-action frequencies as a polynomially constrained set. The constraints can be computed explicitly by Theorem 3.22. Indeed, the linear inequalities describing the cone C in Theorem 3.22 are are indexed by $a\in\mathcal{A}, o\in\mathcal{O}$ and correspond to polynomial inequalities $p_{ao}(\eta)\geq 0$ given by

$$p_{ao}(\eta) := \sum_{s\in S_o}\left(\beta^+_{os}\eta_{sa}\prod_{s'\in S_o\setminus\{s\}}\sum_{a'}\eta_{s'a'}\right) = \sum_{f\colon S_o\to\mathcal{A}}\left(\sum_{s'\in f^{-1}(\{a\})}\beta^+_{os'}\right)\prod_{s\in S_o}\eta_{sf(s)}, \tag{3.30}$$

where $S_o := \{s\in\mathcal{S} : \beta^+_{os}\neq 0\}$. The polynomials depend only on β and not on γ, μ nor α, and have $|S_o||\mathcal{A}|^{|S_o|-1}$ monomials of degree $|S_o|$ of the form $\prod_{s\in S_o}\eta_{sf(s)}$ for some $f\colon S_o\to\mathcal{A}$. In particular, we can read of the multi-degree, i.e., the vector of the degree in the individual blocks, of p_{ao} with respect to the blocks $(\eta_{sa})_{a\in\mathcal{A}}$, which is given by $\mathbb{1}_{S_o}$. Note that in the fully observable case we have $|S_o| = 1$ for each o. Hence, each of the polynomial inequalities has a single term of degree 1. Indeed, in this case the inequalities are simply $\eta_{sa}\geq 0$, for each a, for each s. In the case of a deterministic β, (3.28) implies that $\beta^+_{os}\neq 0$ if and only if $\beta(o|s) > 0$ and hence we have $S_o = \{s\in\mathcal{S} : \beta(o|s) > 0\}$ in this case. For each o, a, there is an inequality $\sum_{f\colon S_o\to\mathcal{A}}|f^{-1}(a)|\prod_{s\in S_o}\eta_{sf(s)}\geq 0$ of degree $|S_o|$ equal to the number of states that are compatible with o.

Analogously to the defining inequalities, we can compute the defining polynomial equalities in the following way, which arise from the linear equations defining $\mathcal{U}$ in Theorem 3.22. These linear equations occur when $\ker(I\otimes\beta^T)\neq\{0\}$ is non trivial. Note that $I\otimes\beta^T$ is injective if and only if β^T is injective, which is again equivalent to β being surjective. First, we need to compute a basis $\{b^j\}_{j\in J}$ of $\{\beta\pi : \pi\in\mathbb{R}^{\mathcal{O}\times\mathcal{A}}\}^\perp = \ker(I\otimes\beta^T)\subseteq\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, which can easily be done using the Gram-Schmidt process. Note that the defining linear equalities of the effective policy polytope within in the state policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$ are given by $\langle b^j,\tau\rangle_{\mathcal{S}\times\mathcal{A}} = 0$. Hence, by Theorem 3.18 the corresponding polynomial equality inside the state-action polytope $\mathcal{N}$ is given by $q_j(\eta) = 0$, where

$$q_j(\eta) := \sum_{s\in S_j}\sum_{a\in\mathcal{A}} b^j_{sa}\eta_{sa}\prod_{s'\in S^j\setminus\{s\}}\sum_{a'\in\mathcal{A}}\eta_{s'a'}, \tag{3.31}$$

where $S^j := \{s\in\mathcal{S} : b^j_{sa}\neq 0 \text{ for some } a\in\mathcal{A}\}$.

A complete description of the set $\mathcal{N}^\beta$ via (in)equalities follows from the description of the state-action polytope $\mathcal{N}$ via linear (in)equalities given in Theorem (3.5). In Section 3.3 we discuss how the degree of these polynomials controls the complexity of the optimization problem.

Example 3.23 (Crying baby example continued). We revisit our running example and compute the polynomial constraints defining the feasible state-action frequencies of the POMDP within the state-action polytope for which we have given the defining inequalities in (3.9). In the case of the crying baby example the observation kernel takes is given by

$$\beta = \begin{array}{c|cc} & o_1 & o_2 \\ \hline s_1 & 1 & 0 \\ s_2 & 0.5 & 0.5 \end{array} \in \Delta_{\mathcal{O}}^{\mathcal{S}},$$

which is invertible with inverse

$$\beta^{-1} = \begin{pmatrix} 1 & 0 \\ -1 & 2 \end{pmatrix}.$$

Let us first compute the effective policy polytope for which we apply Theorem 3.22. Recall that $\mathcal{U} = \ker(I \otimes \beta^T) = \{0\}$ if β is surjective and hence the effective policies are given by

$$\Delta_{\mathcal{A}}^{\mathcal{S},\beta} = \{\tau \in \Delta_{\mathcal{A}}^{\mathcal{S}} : \beta^{-1}\tau \geq 0\} = \left\{\tau \in \Delta_{\mathcal{A}}^{\mathcal{S}} : \begin{array}{r} -\tau(a_1|s_1) + 2\tau(a_1|s_2) \geq 0 \\ \tau(a_1|s_1) - 2\tau(a_1|s_2) + 1 \geq 0 \end{array}\right\},$$

where we made the substitution $\tau(a_2|s) = 1 - \tau(a_1|s)$, see also Figure 3.2. Hence, with the observation kernel β we have the restriction of selecting an action a in state s_2 at least with the probability $\tau(a|s_1)/2$ since with probability $1/2$ we will hear the baby crying and will act like we the baby was hungry.

Let us now turn towards the feasible state-action frequencies. Again, there are no additional polynomial equalities compared to the fully observable case but only polynomial inequalities $p_{ao}(\eta) \geq 0$ with p_{ao} given in (3.30). Note that $p_{ao_2}(\eta) = \eta_{s_2a}$ and hence the condition $p_{ao_2}(\eta) \geq 0$ is satisfied for $a \in \mathcal{A}$ and any state-action frequency $\eta \in \mathcal{N} \subseteq \Delta_{\mathcal{S}\times\mathcal{A}}$ of the underlying MDP. The two remaining polynomials are given by

$$p_{a_1o_1}(\eta) = \sum_{s\in\mathcal{S}} \beta_{o_1s}^{-1}\eta_{sa} \prod_{s'\in\mathcal{S}\setminus\{s\}} \sum_{a'\in\mathcal{A}} \eta_{s'a'} \quad \text{for } a \in \mathcal{A}.$$

Hence, the feasible state-action frequencies are given by

$$\mathcal{N}^{\beta} = \mathcal{N} \cap \left\{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \begin{array}{r} \eta_{s_1a_1}\eta_{s_2a_1} - \eta_{s_1a_1}\eta_{s_2a_2} + 2\eta_{s_1a_2}\eta_{s_2a_1} \geq 0 \\ 2\eta_{s_1a_1}\eta_{s_2a_2} - \eta_{s_1a_2}\eta_{s_2a_1} + \eta_{s_1a_2}\eta_{s_2a_2} \geq 0 \end{array}\right\}, \tag{3.32}$$

see also Figure 3.2. Together with the linear conditions describing the state-action polytope given in (3.9) the reward optimization problem in state-action space (ROP-SA) of the POMDP takes the form

$$\text{maximize } -10\eta_{s_1a_2} - \eta_{s_2a_1} \quad \text{subject to } \left\{ \begin{array}{r} 20\eta_{s_1a_1} + 10\eta_{s_1a_2} - \eta_{s_2a_2} - 10\mu_{s_1} = 0 \\ -10\eta_{s_1a_1} + 10\eta_{s_2a_1} + 11\eta_{s_2a_2} - 10\mu_{s_2} = 0 \\ \eta_{s_1a_1}\eta_{s_2a_1} - \eta_{s_1a_1}\eta_{s_2a_2} + 2\eta_{s_1a_2}\eta_{s_2a_1} \geq 0 \\ 2\eta_{s_1a_1}\eta_{s_2a_2} - \eta_{s_1a_2}\eta_{s_2a_1} + \eta_{s_1a_2}\eta_{s_2a_2} \geq 0 \\ \eta_{s_1a_1}, \eta_{s_1a_2}, \eta_{s_2a_1}, \eta_{s_2a_2} \geq 0. \end{array}\right.$$

This polynomial program can be seen as an extension of the dual linear program (D-LP) describing reward optimization in the fully observable case, which is given by (3.10) in this specific example.

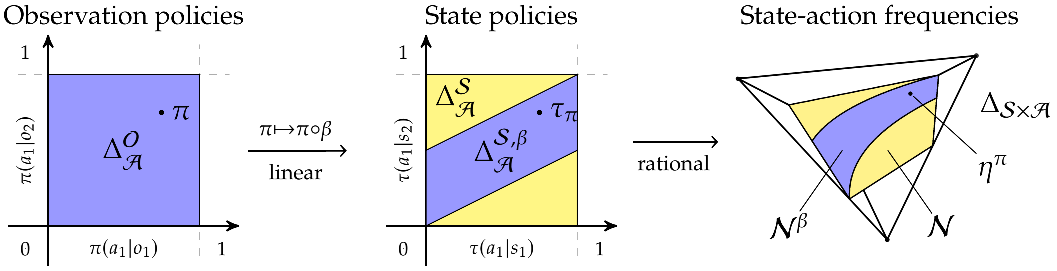

Figure 3.2. Shown are on the left the observation policies $\Delta^{\mathcal{O}}_{\mathcal{A}}$, in the middle the effective policy polytope $\Delta^{\mathcal{S},\beta}_{\mathcal{A}}$ inside the polytope of state policies $\Delta^{\mathcal{S}}_{\mathcal{A}}$ and on the right the corresponding state-action feasible state-action frequencies $\mathcal{N}^\beta$ inside the state-action polytope $\mathcal{N}$ inside the probability simplex $\Delta_{\mathcal{S}\times\mathcal{A}}$, which is a tetrahedron in this case.

3.2.3. State-action frequencies for deterministic observations. Here, we study deterministic observations, i.e., the case where the observation kernel $\beta\in\Delta^{\mathcal{S}}_{\mathcal{O}}\cap\{0,1\}$ has binary entries. This case is often referred to as *state-aggregation* and is a classic way to reduce the size of the state space of MDPs [42, 241, 173]. If β has no zero column – a zero column would correspond to an observation that is observed with zero probability – it satisfies the rank Assumption 3.21. We first elaborate the implications of our general analysis for this important case. Then we provide an alternative characterization describing the feasible state-action frequencies via products of varieties of rank one matrices.

In the case of deterministic observations β corresponds to a mapping $g=g_\beta\colon\mathcal{S}\to\mathcal{O}$ satisfying $\beta(o|s)=\delta_{og(s)}$ and we can compute all polynomial constraints in closed form. Let us again write $S_o:=\{s\in\mathcal{S}:\beta(o|s)>0\}=g_\beta^{-1}(\{o\})\subseteq\mathcal{S}$, then by Theorem 3.22 a policy $\tau\in\Delta^{\mathcal{S}}_{\mathcal{A}}$ belongs to the effective policy polytope $\Delta^{\mathcal{S},\beta}_{\mathcal{A}}$ if and only if

$$\tau(a|s_1)=\tau(a|s_2)\quad\text{for all } s_1,s_2\in S_o, a\in\mathcal{A}, o\in\mathcal{O}. \tag{3.33}$$

Note that this can be encoded in $\sum_o|\mathcal{A}|(|S_o|-1)=|\mathcal{A}|(|\mathcal{S}|-|\mathcal{O}|)$ linear equations; indeed if we fix $s_o\in S_o$, then (3.33) is equivalent to

$$\tau(a|s)-\tau(a|s_o)=0\quad\text{for all } s\in S_o\setminus\{s_o\}, a\in\mathcal{A}, o\in\mathcal{O}. \tag{3.34}$$

By Corollary 3.20 for $\eta\in\mathcal{N}$ it is equivalent to lie in $\mathcal{N}^\beta$ or to satisfy

$$p^o_{sa}(\eta):=\eta_{sa}\sum_{a'}\eta_{s_oa'}-\eta_{s_oa}\sum_{a'}\eta_{sa'}=0\quad\text{for all } s\in S_o\setminus\{s_o\}, a\in\mathcal{A}, o\in\mathcal{O}. \tag{3.35}$$

Note that in this case, there are no polynomial inequalities; this can also be seen from (3.28) and (3.30). We collect this finding.

Corollary 3.24. *For deterministic observation β, fix an arbitrary action $a_o\in\mathcal{A}$ and an arbitrary state $s_o\in S_o=\{s\in\mathcal{S}:\beta(o|s)>0\}$ for every $o\in\mathcal{O}$. The set of feasible state-action frequencies $\mathcal{N}^\beta$ can be described as the intersection $\mathcal{N}^\beta=\mathcal{N}\cap\mathcal{Y}$, where*

$$\mathcal{Y}:=\left\{\eta\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}:\ p^o_{sa}(\eta)=0\,\textit{for all } o\in\mathcal{O}, a\in\mathcal{A}\setminus\{a_o\}, s\in S_o\setminus\{s_o\}\right\},$$

and the polynomials p^o_{sa} are given in (3.35). *Further, $\mathcal{Y}$ is a complete intersection of these polynomials, i.e.,* $\operatorname{codim}(\mathcal{Y}) = |\mathcal{A}|(|\mathcal{S}| - |\mathcal{O}|)$.

Where the corollary above is a consequence of the general theory established before we now provide a new description that is specific to deterministic observations.

Theorem 3.25 (Feasible state-action frequencies). *Let the ergodicity Assumption 2.14 and the positivity Assumption 3.3 hold. For deterministic observation β the set of feasible state-action frequencies $\mathcal{N}^\beta$ is the intersection $\mathcal{N}^\beta = \mathcal{N} \cap \mathcal{X}$, see Theorem 3.5, and $\mathcal{X}$ is the product of real determinantal varieties*

$$\mathcal{X} = \left\{\eta \in \mathbb{R}^{S\times A} : \eta_{sa}\eta_{s'a'} - \eta_{sa'}\eta_{s'a} = 0 \; \forall a, a' \in \mathcal{A} \text{ and } s, s' \in \mathcal{S} \text{ with } g_\beta(s) = g_\beta(s')\right\}.$$

Proof. In the light of Corollary 3.24 it suffices to show $\mathcal{X} \cap \mathcal{N} = \mathcal{Y} \cap \mathcal{N}$. Since p^o_{sa} is a linear combination of 2×2 minors, we have the inclusion $\mathcal{X} \subseteq \mathcal{Y}$. On the other hand, equation (3.35) together with the positivity Assumption 3.3 implies the linear dependence of the two vectors

$$(\eta_{sa})_a, \; (\eta_{s_o a})_a \in \mathbb{R}^{\mathcal{A}}$$

for every observation o and state $s \in S_o$. Consequently, every 2×2 minor in the definition of $\mathcal{X}$ vanishes on $\mathcal{Y} \cap \mathcal{N}$. This shows the desired inclusion $\mathcal{Y} \cap \mathcal{N} \subseteq \mathcal{X} \cap \mathcal{N}$, which finishes the proof. □

The state-action polytope $\mathcal{N}$ is given by the intersection $\mathcal{N} = \Delta_{\mathcal{S}\times\mathcal{A}} \cap \mathcal{L}$, see Theorem 3.5. Hence, the variety $\mathcal{L} \cap \mathcal{X}$ encodes the polynomial equalities defining $\mathcal{N}^\beta$ and we call $\mathcal{L} \cap \mathcal{X}$ the *state-aggregation variety*, where $\mathcal{L}$ is the affine space describing the state-action polytope, see Theorem 3.5. Note that $\mathcal{X}$ is determined by the condition that for every observation o the $d_o \times n_{\mathcal{A}}$ submatrix $(\eta_{sa})_{s\in S_o,\, a\in\mathcal{A}}$ of η, consisting of all entries η_{sa} with $g_\beta(s) = o$. Hence, $\mathcal{X}$ is equal to the product $\prod_{o\in O} \mathcal{D}_1^{d_o\times n_{\mathcal{A}}}$ of determinantal varieties of rank one matrices of size $d_o \times n_{\mathcal{A}}$.

3.2.4. State-action geometry of multi-agent systems. In an MDP only a single agent makes a decision, however, in many settings multiple agents simultaneously act in an environment, which is modelled by multi-agent POMDPs, which are sometimes also referred to as decentralized POMDPs [224, 225]. These models lie at the heart of multi-agent reinforcement where in particular the communication between agents has received huge attention lately. Here, we study the case of decentralized policies, i.e., where groups of agents make their decision collectively but independently from all other groups. However, more general (conditional) independence structures can be studied.

Definition 3.26 (Multi-agent MDPs and decentralized policies). We call an MDP $(\mathcal{S}, \mathcal{A}, \alpha, r)$ a *multi-agent MDP (MA-MDP)* with n agents, if the action space factorizes into n according to $\mathcal{A} = \prod_{i=1}^n \mathcal{A}_i$. Consider a partition $(\nu_i)_{i=1}^k$ of $\{1, \dots, n\}$, i.e., a collection of disjoint subsets $\nu_i \subseteq \{1, \dots, n\}$ such that $\bigcup_{i=1}^k \nu_i = \{1, \dots, n\}$. For $i = 1, \dots, k$ we set $\mathcal{A}_{\nu_i} := \prod_{j\in\nu_i} \mathcal{A}_j$ and $a_{\nu_i} := (a_j)_{j\in\nu_i}$ for $a = (a_1, \dots, a_n) \in \mathcal{A}$. We call a policy $\pi \in \Delta^{\mathcal{S}}_{\mathcal{A}}$ *decentralized* with respect to $(\nu_i)_{i=1,\dots,k}$ if there are Markov kernels $\pi_i \in \Delta^{\mathcal{S}}_{\mathcal{A}_{\nu_i}}$ such that

$$\pi(a|s) = \prod_{i=1}^k \pi_i(a_{\nu_i}|s) \quad \text{for all } a \in \mathcal{A}, s \in \mathcal{S}. \tag{3.36}$$

If $\nu_i = \{i\}$ then we simply call such a policy *decentralized*.

Note that a factorization of the form (3.36) exists if and only if the actions a_{v_i} are made independently from another. We also introduce partially observable multi-agent problems.

Definition 3.27 (Multi-agent POMDPs and decentralized policies). We call a POMDP $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$ a *multi-agent POMDP (MA-POMDP)* with n agents if the action and observation spaces factorize into n factors, i.e., $\mathcal{A} = \prod_{i=1}^n \mathcal{A}_i$ and $\mathcal{O} = \prod_{i=1}^n \mathcal{O}_i$. For a partition $(v_i)_{i=1}^k$ of $\{1, \dots, n\}$ we call a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ *decentralized* with respect to $(v_i)_{i=1,\dots,k}$ if there are policies $\pi_i \in \Delta_{\mathcal{A}_{v_i}}^{\mathcal{O}_{v_i}}$ such that

$$\pi(a|o) = \prod_{i=1}^k \pi_i(a_{v_i}|o_{v_i}) \quad \text{for all } a \in \mathcal{A}, o \in \mathcal{O}, \tag{3.37}$$

where $a = (a_1, \dots, a_n)$, $o = (o_1, \dots, o_n)$ and $a_{v_i} := (a_j)_{j \in v_i}$, $o_{v_i} := (o_j)_{j \in v_i}$. If $v_i = \{i\}$ then we simply call such a policy *decentralized*.

We denote the set of decentralized policies with $\Delta_{\mathcal{A}}^{\mathcal{O},\mathrm{dec}} \subseteq \Delta_{\mathcal{A}}^{\mathcal{O}}$ and similarly we write $\Delta_{\mathcal{A}}^{\mathcal{S},\mathrm{dec}} \subseteq \Delta_{\mathcal{A}}^{\mathcal{S}}$ for the decentralized policies in a multi-agent MDP. We denote the corresponding state-action frequencies by

$$\mathcal{N}^{\beta,\mathrm{dec}} = \left\{\eta^\pi : \pi \in \Delta_{\mathcal{A}}^{\mathcal{O},\mathrm{dec}}\right\} \subseteq \mathcal{N}^\beta \quad \text{and } \mathcal{N}^{\mathrm{dec}} = \left\{\eta^\pi : \pi \in \Delta_{\mathcal{A}}^{\mathcal{S},\mathrm{dec}}\right\} \subseteq \mathcal{N}.$$

In the following we provide characterizations of the sets $\mathcal{N}^{\beta,\mathrm{dec}}$ and $\mathcal{N}^{\mathrm{dec}}$.

Notation. For a subset $I \subseteq \{1, \dots, n\}$ and $a \in \mathcal{A}$ we use the notation $a_I := (a_i)_{i \in I}$ as well as $\mathcal{A}_I := \prod_{i \in I} \mathcal{A}_i$. We denote the marginal policies by

$$\pi_{I,+}(a_I|o) := \sum_{\substack{\tilde{a} \in \mathcal{A} \\ \tilde{a}_I = a_I}} \pi(a|o) \quad \text{for all } a_I \in \mathcal{A}_I, o \in \mathcal{O}. \tag{3.38}$$

If $I = \{i\}$ we write a_i and $\pi_{i,+}$ and for $I = \{1, \dots, n\} \setminus \{i\}$ we write a_{-i} and $\pi_{-i,+}$ for a_I and $\pi_{I,+}$ respectively. We adopt a similar notation for multiple index sets, e.g., for disjoint $I, J \subseteq \{1, \dots, n\}$ we write $\pi_{I,J,+}$ for the corresponding marginal policy.

In a multi-agent problem can be captured by the following sequence of mappings

$$\begin{array}{ccccccccc} \prod_{i=1}^k \Delta_{\mathcal{A}_{v_i}}^{\mathcal{O}_{v_i}} & \to & \Delta_{\mathcal{A}}^{\mathcal{O}} & \to & \Delta_{\mathcal{A}}^{\mathcal{S}} & \to & \mathcal{N} & \to & \mathbb{R} \\ (\pi_i)_{i=1,\dots,k} & \mapsto & \pi & \mapsto & \tau & \mapsto & \eta & \mapsto & R(\pi). \end{array} \tag{3.39}$$

For a decentralized policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ it holds that $\pi_{v_i,+}(a_{v_i}|o) = \pi_i(a_{v_i}|o_{v_i})$, which shows in particular that the parametrization $(\pi_i) \mapsto \pi$ of decentralized policies is a Lipschitz homeomorphism. In particular this implies that the dimension of decentralized policies is given by $\dim(\prod_{i=1}^k \Delta_{v_i}^{\mathcal{A}_{v_i}})$ Further, this yields that

$$\pi(a|o) = \prod_{i=1}^k \pi_{v_i,+}(a_{v_i}|o) \quad \text{for all } a \in \mathcal{A}, o \in \mathcal{O}. \tag{3.40}$$

Hence, since for a decentralized policy $\pi_{v_i,+}(a_{v_i}|o)$ does only depend on o_{v_i} we can choose $\pi_i(a_{v_i}|o_{v_i}) := \pi_{v_i,+}(a_{v_i}|o')$ for some $o' \in \mathcal{O}$ with $o'_{v_i} = o_{v_i}$. Note however that (3.40) alone does not imply that π is decentralized. However, if in addition

$$\pi_{v_i,+}(a_{v_i}|o) = \pi_{v_i,+}(a_{v_i}|o') \quad \text{for all } o, o' \in \mathcal{O} \text{ with } o_{v_i} = o'_{v_i} \tag{3.41}$$

is satisfied then surely π is decentralized with $\pi_i = \pi_{\nu_i,+}$. Overall, this implies that a poliy $\pi \in \Delta_{\mathcal{A}}^{O}$ is decentralized if and only if it satisfies (3.40) and (3.41).

The fully observable case. For a decentralized policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, all agents indexed by ν_i select their actions independently from all other groups ν_j of agents. Such independence relationships can be described by the vanishing of 2×2 minors of marginals of the probability distribution [275]. In our case this yields the following characterization.

Proposition 3.28 (Characterization of decentralized policies). *Consider a multi-agent MDP $(\mathcal{S}, \mathcal{A}, \alpha, r)$ with n agents and a partition $(\nu_i)_{i=1,\dots,k}$ of $\{1, \dots, n\}$. Then $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ is decentralized with respect to $(\nu_i)_{i=1,\dots,k}$ if and only if*

$$\pi_{I,J,+}(a_I, a_J|s)\pi_{I,J,+}(a'_I, a'_J|s) - \pi_{I,J,+}(a_I, a'_J|s)\pi_{I,J,+}(a_I, a'_J|s) = 0 \tag{3.42}$$

for all $a_I, a'_I \in \mathcal{A}_I$, $a_J, a'_J \in \mathcal{A}_J$, $s \in \mathcal{S}$ and $I, J \in \{\nu_i\}_{i=1,\dots,n}$.

For the state-action frequencies we use an analogue notation to policies, e.g.,

$$\eta_I(s, a_I) = \sum_{\substack{\tilde{a} \in \mathcal{A} \\ \tilde{a}_I = a_I}} \eta(s, a)$$

and analogously for multiple index sets.

Theorem 3.29 (State-action frequencies of decentralized policies in a MDP). *Consider a multi-agent MDP $(\mathcal{S}, \mathcal{A}, \alpha, r)$ with n agents and a partition $(\nu_i)_{i=1,\dots,k}$ of $\{1, \dots, n\}$ and let the positivity Assumption 3.3 hold. Then $\eta \in \mathcal{N}$ is the state-action frequency of a decentralized policy if and only if*

$$\eta_{I,J,+}(s, a_I, a_J) \cdot \eta_{I,J,+}(s, a'_I, a'_J) - \eta_{I,J,+}(s, a_I, a'_J) \cdot \eta_{I,J,+}(s, a'_I, a_J) = 0 \tag{3.43}$$

for all $a, a' \in \mathcal{A}$, $s \in \mathcal{S}$, $I, J \in (\nu_i)_{i=1,\dots,n}$. Hence, we have $\mathcal{N}^{\mathrm{dec}} = \mathcal{X} \cap \mathcal{N}$, where

$$\mathcal{X} = \left\{ \eta \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}} : \text{(3.43) } \textit{holds for all } a, a' \in \mathcal{A}, s \in \mathcal{S}, I, J \in (\nu_i)_{i=1,\dots,n} \right\}. \tag{3.44}$$

Proof. The condition (3.42) characterizes decentralization in fully observable multi agent problems. Note that this is homogeneous polynomial condition that only addresses $\pi(\cdot|s)$. Computing the corresponding equations in state-action space using the substitution $\pi(a|s) = \eta(s, a)/\rho(s)$ and multiplying the resulting equation by $\rho(s)^2 > 0$ yields the claim. □

We have seen that the state-action frequencies of decentralized policies are described by the same polynomial equations as the decentralized policies, in particular, they are also of degree 2. This is a due to the homogeneity of the constraints, see also (3.22). Note that the dimension of the set of decentralized policies and under Assumption 3.3 also of the corresponding state-action frequencies is given by

$$\dim \left(\prod_{i=1}^{k} \Delta_{\mathcal{A}_{\nu_i}}^{\mathcal{S}} \right) = \sum_{i=1}^{k} |\mathcal{S}|(|\mathcal{A}_{\nu_i}| - 1) = \sum_{i=1}^{k} |\mathcal{S}| \left(\prod_{j \in \nu_i} |\mathcal{A}_j| - 1 \right). \tag{3.45}$$

If all agents share the same action space $\mathcal{A}_i = A$ then the dimension is given by

$$\sum_{i=1}^{k} |\mathcal{S}| \left(|A|^{|\nu_i|} - 1 \right). \tag{3.46}$$

In particular, in the two extreme cases of centralized policies, i.e., $k = 1$ and decentralized policies, i.e., $k = n$ we obtain the dimensions $|\mathcal{S}|\left(|A|^n - 1\right)$ and $n|\mathcal{S}|\left(|A| - 1\right)$. Note that the dimension grows exponentially in the number of agents if they act centralized and only linearly when they act in a decentralized way.

Characterization for multi-agent POMDPs. For the partially observable case we can obtain an analogous description for the feasible state-action frequencies $\mathcal{N}^{\beta,\mathrm{dec}}$.

Proposition 3.30 (Characterization of decentralized policies). *Consider a multi-agent POMDP $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \alpha, \beta, r)$ with n agents and a partition $(\nu_i)_{i=1,\dots,k}$ of $\{1, \dots, n\}$. Then $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ is decentralized with respect to $(\nu_i)_{i=1,\dots,k}$ if and only if*

$$\pi_{I,J,+}(a_I, a_J|o)\pi_{I,J,+}(a'_I, a'_J|o) - \pi_{I,J,+}(a_I, a'_J|o)\pi_{I,J,+}(a_I, a'_J|o) = 0 \tag{3.47}$$

for all $a_I, a'_I \in \mathcal{A}_I$, $a_J, a'_J \in \mathcal{A}_J$, $o \in \mathcal{S}$ and $I, J \in \{\nu_i\}_{i=1,\dots,n}$ and

$$\pi_{I,+}(a_I|o) = \pi_{I,+}(a_I|o') \tag{3.48}$$

for all $a_I \in \mathcal{A}_I$ and $o, o' \in \mathcal{O}$ with $o_I = o'_I$ for all $I \in \{\nu_i\}_{i=1,\dots,n}$.

Proof. It is immediate to check that (3.47) and (3.48) hold for decentralized policies.

Let us now assume that (3.47) and (3.48) hold. It is well known that (3.47) is equivalent to the independence

$$\pi(a|o) = \prod_{i=1}^{k} \pi_{\nu_i,+}(a_{\nu_i}|o)$$

for all $a \in \mathcal{A}$, $o \in \mathcal{O}$, see for example [275]. Together with (3.48) it is clear that a factorization of the form (3.37) can be obtained via $\pi_{\nu_i}(a_{\nu_i}|o_{\nu_i}) := \pi_{\nu_i,+}(a_{\nu_i}|o')$ for some $o' \in \mathcal{O}$ with $o'_{\nu_i} = o_{\nu_i}$. □

Theorem 3.31 (State-action frequencies of decentralized policies). *Consider a multi-agent POMDP $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$ with n agents and a partition $(\nu_i)_{i=1,\dots,k}$ of $\{1, \dots, n\}$ and let the positivity Assumption 3.3 and the rank Assumption 3.21 hold. Then $\eta \in \mathcal{N}^{\beta}$ is the state-action frequency of a decentralized policy if and only if*

$$p^{o}_{a_I,a_J}(\eta)p^{o}_{a'_I,a'_J}(\eta) - p^{o}_{a_I,a'_J}(\eta)p^{o}_{a'_I,a_J}(\eta) = 0 \tag{3.49}$$

for all $a, a' \in \mathcal{A}$, $o \in \mathcal{O}$, $I, J \in (\nu_i)_{i=1,\dots,n}$ and

$$q^{o,o'}_{a_I}(\eta) = q^{o',o}_{a_I}(\eta) \tag{3.50}$$

for all $a_I \in \mathcal{A}_I$, $o, o' \in \mathcal{O}$ with $o_I = o'_I$ and $I \in (\nu_i)_{i=1,\dots,n}$, where

$$p^{o}_{a_I,a_J}(\eta) = \sum_{\substack{a' \in \mathcal{A} \\ a'_I = a_I, a'_J = a_J}} \sum_{s \in \mathcal{S}} \beta^{+}_{os}\eta_{sa'} \left(\prod_{s' \in S_o \setminus \{s\}} \sum_{\tilde{a} \in \mathcal{A}} \eta_{s'\tilde{a}} \right) \tag{3.51}$$

and

$$q^{o,o'}_{a_I}(\eta) = \sum_{\substack{a' \in \mathcal{A} \\ a'_I = a_I}} \sum_{s \in \mathcal{S}} \beta^{+}_{os}\eta_{sa'} \cdot \prod_{s' \in (S_o \cup S_{o'}) \setminus \{s\}} \sum_{\tilde{a} \in \mathcal{A}} \eta_{s'\tilde{a}}, \tag{3.52}$$

where $S_o = \{s \in \mathcal{S} : \beta^+_{os} \neq 0\}$. Hence, we have $\mathcal{N}^{\beta,\mathrm{dec}} = \mathcal{X} \cap \mathcal{N}^{\beta}$, where

$$\mathcal{X} = \left\{ \eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \begin{array}{l} \text{(3.49) holds for all } a, a' \in \mathcal{A}, o \in \mathcal{O}, I, J \in (\nu_i)_{i=1,\dots,n} \\ \text{(3.50) holds for all } a_I \in \mathcal{A}_I, o, o' \in \mathcal{O} \text{ with } o_I = o'_I, I \in (\nu_i)_{i=1,\dots,n} \end{array} \right\}.$$

Proof. Just like in the case of single agent partially observable Markov decision processes we can compute the policy $\pi \in \Delta^{\mathcal{O}}_{\mathcal{A}}$ from the state-action frequency $\eta \in \mathcal{N}$ according to

$$\pi(a|o) = \sum_{s\in\mathcal{S}} \beta^+_{os} \cdot \frac{\eta_{sa}}{\sum_{a'\in\mathcal{A}} \eta_{sa'}}, \tag{3.53}$$

where β^+ denotes a pseudoinverse of the observation kernel $\beta \in \Delta^{\mathcal{S}}_{\mathcal{O}} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{O}}$. Substituting π in (3.47) and (3.48) according to (3.53) and multiplying the resulting rational equations by the suitable marginals $\sum_{a'\in\mathcal{A}} \eta_{sa'}$ to obtain a polynomial condition we obtain (3.49) and (3.50). □

In contrast to the fully observable case the degree of the defining equations (3.43) in state-action space do not necessarily agree with the degree of the defining equations (3.47) and (3.48) in policy space. Indeed, we have $\deg(p^o_{a_I,a_J}) \leq |S_o|$ and $\deg(q^{o,o'}_{a_I}) \leq |S_o \cup S_{o'}|$ and hence the degree of (3.49) is upper bounded by $2|S_o|$ where the degree of (3.50) is upper bounded by $|S_o \cup S_{o'}|$.

Like in the fully observable case we can compute the dimension of the state-action frequencies of the decentralized policies

$$\dim\left(\prod_{i=1}^{k} \Delta^{\mathcal{O}_{\nu_i}}_{\mathcal{A}_{\nu_i}}\right) = \sum_{i=1}^{k} \prod_{j\in\nu_i} |\mathcal{O}_j| \left(\prod_{j\in\nu_i} |\mathcal{A}_j| - 1\right). \tag{3.54}$$

In the case that all agents share the same action and observation space, i.e., $\mathcal{A}_i = A$ and $\mathcal{O}_i = O$ for all $i = 1, \dots, k$ the dimension is given by

$$\sum_{i=1}^{k} |O|^{|\nu_i|} \left(|A|^{|\nu_i|} - 1\right). \tag{3.55}$$

In particular, in the two extreme cases of centralized and decentralized policies the respective dimension given by $|O|^n\,(|A|^n - 1)$ and $n|O|\,(|A| - 1)$.

Proposition 3.32 (Effective decentralized policies are decentralized). *Consider a multi-agent POMDP $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$ and let $\pi \in \Delta^{\mathcal{O}}_{\mathcal{A}}$ be decentralized with respect to $(\nu_i)_{i=1,\dots,k}$ for a partition $(\nu_i)_{i=1,\dots,k}$ of $\{1, \dots, n\}$. Then $\tau = \pi \circ \beta \in \Delta^{\mathcal{S}}_{\mathcal{A}}$ is decentralized with respect to $(\nu_i)_{i=1,\dots,k}$. In particular, we have $\mathcal{N}^{\beta,\mathrm{dec}} \subseteq \mathcal{N}^{\mathrm{dec}}$.*

Proof. We perceive the policy τ as a graphical model and borrow from the theory of conditional independence for graphical models to show that $\{\nu_i\}$ are independent under $\tau(\cdot|s)$ for every $s \in \mathcal{S}$. Consider the case of complete decentralization, i.e., $\nu_i = \{i\}$. Then for $\mu \in \mathrm{int}(\Delta_{\mathcal{S}})$ the decentralized policy π induces a joint distribution $\xi \in \Delta_{\mathcal{S}\times\mathcal{O}\times\mathcal{A}}$ on $\mathcal{S} \times \mathcal{O} \times \mathcal{A}$ according to $\xi(s, o, a) = \mu(s)\beta(o|s) \prod_{i=1}^{n} \pi_i(a_i|o_i)$. This distribution is Markov with respect to the graph shown in Figure 3.3. Then a_i and a_j are d-separated given s and hence a_i and a_j are independent given s, see [158, Theorem 3.3]. Note that a factorization of the form (3.36) exists if and only if the i-th and j-th action are taken independently, which shows that τ is decentralized. For general partitions the argument is analogue. □

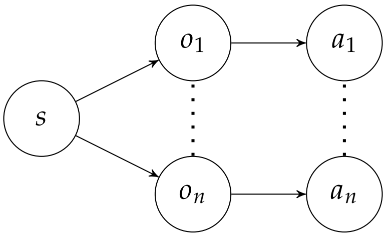

Figure 3.3. Graph describing the decentralized policy.

Factored jointly fully observable models. Finally, we consider the special case of multi-agent POMDPs where we assume that also the state space factorizes according to $\mathcal{S} = \prod_{i=1}^n \mathcal{S}_i$. We call such problems *factored*. For example, $s_i \in \mathcal{S}_i$ could encode the position of the i-th agent. We call a factored multi-agent POMDP *jointly fully observable* if the i-th agent observes its own state, i.e, if $\mathcal{O} = \mathcal{S}$ and $\beta = I$.

Note that in the context of multi-agent POMDPs this is not equivalent to the corresponding multi-agent MDP since for a decentralized policy in the sense of the POMDP $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ every agent selects its action solely based on its individual state. Hence, policies $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ that are decentralized in the sense of the POMDP admit a representation of the form

$$\pi(a|s) = \prod_{i=1}^{k} \pi_i(a_{v_i}|s_{v_i}) \tag{3.56}$$

for some policies $\pi_i \in \Delta_{\mathcal{A}_{v_i}}^{\mathcal{S}_{v_i}}$. In contrast if a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ is decentralized in the underlying multi-agent MDP every agent has access to the entire state tuple $s = (s_1, \dots, s_n)$ and hence we require

$$\pi(a|s) = \prod_{i=1}^{k} \pi_i(a_{v_i}|s) \tag{3.57}$$

for some $\pi_i \in \Delta_{\mathcal{A}_{v_i}}^{\mathcal{S}}$. If we assume that all agents share the same state and action space, i.e., $\mathcal{S}_i = S$ and $\mathcal{A}_i = A$ then in the jointly observable case the dimension of the state-action frequencies and the decentralized policies is given by

$$\sum_{i=1}^{k} |S|^{|v_i|} \left(|A|^{|v_i|} - 1 \right). \tag{3.58}$$

In the case of centralized and completely decentralized policies the dimension is given by $|S|^n(|A|^n - 1)$ and $n|S|(|A| - 1)$ respectively. Note that in the corresponding fully observable MDP the respective dimensions are given by $|S|^n(|A|^n - 1)$ and $n|S|^n(|A| - 1)$.

For a jointly observable multi-agent problem the defining equations in state-action space simplify significantly as $\beta^+ = I$ and $S_o = \{o\}$ for $o \in \mathcal{O} = \mathcal{S}$. Hence, in the notation of Theorem 3.31 we obtain

$$p^s_{a_I, a_J}(\eta) = \sum_{\substack{a' \in \mathcal{A} \\ a'_I = a_I, a'_J = a_J}} \eta_{sa'} = \eta_{I,J,+}(s, a_I, a_J)$$

and

$$q_{a_I}^{s,s'}(\eta) = \sum_{\substack{a'\in\mathcal{A}\\ a'_I=a_I}} \eta_{sa'} \cdot \left(\sum_{\tilde{a}\in\mathcal{A}} \eta_{s'\tilde{a}}\right)^{1-\delta_{ss'}} = \eta_{I,+}(s,a_I) \cdot \left(\sum_{\tilde{a}\in\mathcal{A}} \eta_{s'\tilde{a}}\right)^{1-\delta_{ss'}} .$$

Hence, the defining equations take the form

$$\eta_{I,J,+}(s,a_I,a_J)\eta_{I,J,+}(s,a'_I,a'_J) - \eta_{I,J,+}(s,a_I,a'_J)\eta_{I,J,+}(s,a'_I,a_J) = 0 \tag{3.59}$$

and

$$\eta_{I,+}(s,a_I)\left(\sum_{\tilde{a}\in\mathcal{A}} \eta_{s'\tilde{a}}\right) = \eta_{I,+}(s',a_I)\left(\sum_{\tilde{a}\in\mathcal{A}} \eta_{s\tilde{a}}\right) \tag{3.60}$$

for all $s,s'\in\mathcal{S}$ with $s\neq s'$ and $s_I = s'_I$. In particular, other than in the general partially obseravble case these are quadratic equations.

3.2.5. Conclusion and outlook. In this section we have showed that the feasible state-action frequencies of partially observable Markov decision processes as well as of (fully and partially observable) multi-agent Markov decision processes can be characterized by polynomial inequalities. This characterizes these frequencies as a polynomially constrained subset of the probability simplex and hence as a semialgebraic statistical model. Further, this implies that reward optimization in any of these models is equivalent to a polynomially constrained linear objective program that generalizes the dual linear program associated to an MDP. We provide an overview over the different characterization and correspondences between inequalities regarding policies and state-action frequencies obtained in this section in Table 3.1.

We give a complete characterization of the state-action frequencies achievable with memoryless stochastic policies and believe that the following two directions provide natural continuations of this work:

- *Geometry of memory:* Where we have focused on memoryless stochastic policies a complementary study for finite memory policies could provide important insights into the design of memory. An obvious approach to this is to augment the state space with a finite memory and apply the results obtained here.
- *Geometry of value functions of POMDPs:* Further, we believe that studying the geometry of the set

$$\mathcal{V}^\beta = \{V^\pi : \pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}\} \subseteq \mathbb{R}^{\mathcal{S}}$$

  of value functions of partially observable Markov decision processes would complement our analysis of the state-action frequencies nicely. Note that for fully observable problems the set of value functions has been characterized as a finite union of polytopes [81, 304, 289], which has been used to design optimal representations [44]. In contrast to state-action frequencies the policy can not be reconstructed from its (state-action) value function. Therefore, the approach taken here for the characterization of the feasible state-action frequencies can not be transferred naively.

	(In)equalities of policies	(In)equalities of state-action frequencies
MDPs	$\Delta_{\mathcal{A}}^{\mathcal{S}}$ is described by	$\mathcal{N}_{\gamma}^{\mu}$ is described by
	$\tau(a\mid s) \geq 0$ Row normalization: $\sum_a \tau(a\mid s) - 1 = 0$ – –	$\eta(s,a) \geq 0$ – Discounted stationarity: $\ell_s(\eta) = 0$ For $\gamma = 1$: $\sum_{s,a} \eta_{sa} - 1 = 0$
POMDPs	$\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$ is described in $\Delta_{\mathcal{A}}^{\mathcal{S}}$ by	$\mathcal{N}^{\beta}$ is described in $\mathcal{N}$ by
	Linear (in)equalities See Subsection 3.2.1 Closed form under Assumption 3.21: See Theorem 3.22 Closed form for deterministic observ.: See (3.33)	Polynomial (in)equalities See Subsection 3.2.1 Closed form under Assumption 3.21: See (3.30) and (3.31) Vanishing of some 2 × 2 minors See Theorem 3.25
MA-MDPs	$\Delta_{\mathcal{A}}^{\mathcal{S},\mathrm{dec}}$ is described in $\Delta_{\mathcal{A}}^{\mathcal{S}}$ by	$\mathcal{N}^{\mathrm{dec}}$ is described in $\mathcal{N}$ by
	Vanishing of 2 × 2 minors of marginals See Proposition 3.28	Vanishing of 2 × 2 minors of marginals See Theorem 3.29
MA-POMDPs	$\Delta_{\mathcal{A}}^{\mathcal{O},\mathrm{dec}}$ is described in $\Delta_{\mathcal{A}}^{\mathcal{O}}$ by	$\mathcal{N}^{\beta,\mathrm{dec}}$ is described in $\mathcal{N}^{\beta}$ by
	Vanishing of 2 × 2 minors of marginals and linear equations See Proposition 3.30	Polynomial equations See Theorem 3.31

Table 3.1. Correspondence of the defining linear and polynomial inequalities of the policies and the (feasible) state-action frequencies for MDPs, POMDPs, MA-MDPs and MA-POMDPs respectively.

3.3 Number and location of critical points

In this section we use the reformulation of the reward optimization problem as a polynomially constraint linear objective problem to gain regarding the optimization problem encountered in POMDPs. We apply tools from algebraic statistics and applied algebraic geometry to describe the (algebraic) complexity of the reward optimization problem.

Again, we perceive the reward maximization problem as the maximization of a linear function p_0 over the set of feasible state-action frequencies $\mathcal{N}^{\beta}$, which is a polynomially constrained subset of the state-action polytope $\mathcal{N}$, see Corollary 3.20 and Theorem 3.5.

Since under Assumption 3.3 the parametrization $\pi \mapsto \eta^\pi$ is injective and has a full-rank Jacobian, see Lemma 3.8, the critical points of the reward function R in the policy polytope $\Delta_{\mathcal{A}}^{O}$ correspond to the critical points of p_0 on $\mathcal{N}^\beta$ [281]. In general, critical points of this linear function can occur on every boundary component of the semialgebraic set $\mathcal{N}^\beta$. The optimization problem thus has a combinatorial and a geometric component, corresponding to the number of boundary components of each dimension and the number of critical points in the interior of any given boundary component. We have discussed the combinatorial part in Theorem 3.18 and focus now on the geometric part. Writing

$$\mathcal{N}^\beta = \left\{ \eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \colon p_i(\eta) \geq 0 \text{ for } i \in I \right\},$$

we are interested in the number of critical points of the reward optimization problem, which in state-action space takes the form

$$\text{maximize } p_0(\eta) \quad \text{subject to} \quad p_i(\eta) \geq 0 \text{ for } i \in I. \tag{3.61}$$

We call a point η *(primal) feasible* if it satisfies $p_i(\eta) \geq 0$ for all $i \in I$ and further we call a feasible point η *critical* if there exists $\kappa \in \mathbb{R}^I_{\geq 0}$ such that (η, κ) solves the Karush-Kuhn-Tucker conditions (KKT conditions)

$$\nabla p_0(\eta) + \sum_{i \in I_a(\eta)} \kappa_i \nabla p_i(\eta) = 0, \tag{KKT}$$

where $I_a(\eta) \coloneqq \{i \in I : p_i(\eta) = 0\}$ denotes the constraints that are active at η. We refer to the non negativity condition $\kappa_i \geq 0$ as the *dual feasibility* condition.

The number of critical points on the interior of a boundary component

$$\operatorname{int}(F_J) = \left\{ \eta \in \mathcal{N}^\beta : p_j(\eta) = 0 \text{ for } j \in J, p_i(\eta) > 0 \text{ for } i \in I \setminus J \right\},$$

provides an upper bounded by the critical points over the variety

$$\mathcal{V}_J \coloneqq \left\{ \eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : p_j(\eta) = 0 \text{ for } j \in J \right\}.$$

For the case of an equality constrained optimization problem the KKT system (KKT) reduces to the Lagrange system

$$\nabla p_0(\eta) + \sum_{j \in J} \lambda_j \nabla p_j(\eta) = 0 \tag{L}$$

for some $\lambda \in \mathbb{R}^J$, where the Lagrange multipliers λ_j are allowed to have arbitrary signs. Hence, we can upper bound the number of critical points in the interior of the face F_J by the number of critical points of the polynomial optimization problem

$$\text{maximize } p_0(\eta) \quad \text{subject to} \quad p_j(\eta) = 0 \text{ for } j \in J. \tag{3.62}$$

An upper bound on the number of critical points of the original inequality constrained problem (3.61) can be obtained by iterating over the individual boundary components.

Solutions of the KKT system vs the Lagrange systems. First, we note that every primal feasible point η that is critical does indeed solve the Lagrange system (L) for $J = I_a(\eta)$. Hence, when obtaining upper bounds on the number of solutions of the Lagrange systems over the individual boundary components does in fact yield an upper bound on the number of critical points.

However, not every solution $(\eta, \lambda) \in \mathcal{V}_J \times \mathbb{R}^J$ of one of the Lagrange systems is a critical point. First, it is not clear whether the Lagrange multipliers λ_j can be chosen non negative and further η might not satisfy the primal feasibility conditions $p_i(\eta) \geq 0$.

If $(\eta, \lambda) \in \mathcal{V}_J \times \mathbb{R}^J$ solves the Lagrange system (L) then it surely solves the KKT system (KKT) without the dual feasibility condition $\kappa \geq 0$ and vice versa.

Note that there are choices $J \subseteq I$ such that the corresponding variety $\mathcal{V}_J$ is non-trivial but contains no feasible points. Such choices of J can be excluded when combining the bounds on the number of solutions of the Lagrange systems. If this is done then then the upper bounds aggregated over the choices of J such that $\mathcal{V}_J$ contain feasible points will provide a tighter upper bound on the number of critical points than the number of solutions of the KKT system (KKT). In a general setting it is hard to decide whether $\mathcal{V}_J$ contains a feasible point. When the boundary components of the feasible set $\{\eta : p_i(\eta) \geq 0 \text{ for } i \in I\}$ are known then this can be done efficiently. In the case of state-action frequencies the boundary components are one to one to the faces of the policy polytope $\Delta^{\mathcal{O}}_{\mathcal{A}}$, see Corollary 3.20.

Overall we have the following chain of inclusions, which are also visualized in Figure 3.4:

$$\begin{aligned}\{\text{critical points}\} &= \{\text{primal and dual feasible solutions of (KKT)}\} \\ &\subseteq \{\text{primal feasible solutions of (KKT)}\} \\ &= \{\text{primal feasible solutions of (L)}\} \\ &\subseteq \{\text{solutions of (L) for } J \subseteq I \text{ such that } \mathcal{V}_J \text{ contains feasible points}\} \\ &\subseteq \{\text{solutions of (L) for some } J \subseteq I\} \\ &= \{\text{solutions of (KKT)}\}.\end{aligned} \tag{3.63}$$

3.3.1. The algebraic degree of polynomial optimization.

For the sake of notation, let us assume that $J = \{1, \dots, m\}$ from now on. We try to present the results from the theory of algebraic degrees that we use here and refer the interested reader to the excellent introduction in [64] and the references therein. Although in practice, we might be interested in the number of real critical points we consider the problem over the complex numbers, which gives an upper bound on the number of real critical points. Working over the complex number has the advantage that the number of solutions of a system of polynomial equations is constant under suitable genericity assumptions. Here, we say that a property holds for a *generic point* if there is an open and dense subset (usually of full measure) such that the property holds for all points within this set; we say a property holds for a *generic polynomial* if it holds for generic coefficients.

Let us consider the polynomial optimization problem (3.62), where we do not require p_0 to be linear. Further, denote the number of variables by n (in the case of state-action

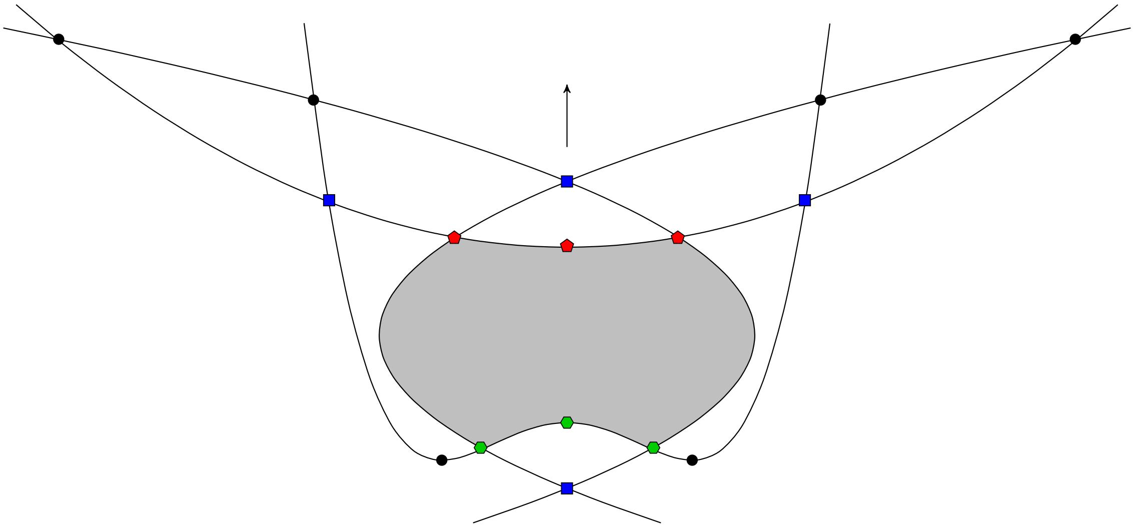

FIGURE 3.4. Schematic illustration of the feasible region (gray) and objective gradient (arrow) of a polynomially constrained linear program showing (i) the primal and dual feasible solutions of (KKT) (red pentagons), (ii) the ptimal feasible solutions of (KKT) or equivalently (L) (red pentagons and green hexagons), (iii) solutions of (L) for some $J \subseteq I$ such that $\mathcal{V}_J$ contains feasible points (red pentagons, green hexagons and black points) (iv) solutions of (L) for some $J \subseteq I$ or equivalently (4.11) (any marked point).

frequencies $n = n_{\mathcal{S}} n_{\mathcal{A}}$) and denote the degrees of $p_0, \dots, p_m$ by $d_0, \dots, d_m$. Again, a point is critical, if it satisfies the KKT conditions

$$\nabla p_0(x) + \sum_{i=1}^{m} \lambda_i \nabla p_i(x) = 0, \quad p_1(x) = \cdots = p_m(x) = 0, \tag{KKT}$$

for some $\lambda \in \mathbb{C}^n$, which is a system of polynomial equations in (x, λ). The number of complex solutions to those criticality equations, when finite, is called the *algebraic degree* of the problem. The algebraic degree is determined by the nature of the polynomials $p_0, \dots, p_m$ and captures the complexity of the optimization problem as the coordinates of critical points can be shown to be roots of some univariate polynomials whose degree equals the algebraic degree and whose coefficients are rational functions of the coefficients of $p_0, \dots, p_m$, see [163, 31]. A special case of (3.62) is when $m = n$ and the polynomials $p_1, \dots, p_m$ are generic. Then by Bézout's theorem there are exactly $d_1 \cdots d_n$ isolated points satisfying the polynomial constraints and all of them are critical and hence the algebraic degree is precisely $d_1 \cdots d_n$ [280]. If the polynomials $p_0, \dots, p_m$ define a complete intersection, i.e., the co-dimension of their induced variety is $m + 1$, the algebraic degree of (3.62) is upper bounded by

$$d_1 \cdots d_m \sum_{i_0 + \cdots + i_m = n - m} (d_0 - 1)^{i_0} \cdots (d_m - 1)^{i_m}, \tag{3.64}$$

and this bound is attained for generic polynomials [221, 64]. For non-complete intersections, the expression (3.64) does not need to yield an upper bound if some constraints are redundant. However, we can modify the expression to obtain a valid upper bound.

Indeed, if l and $c = n - l$ denote the dimension and co-dimension of

$$\mathcal{V} := \{x : p_1(x) = \cdots = p_m(x) = 0\}$$

and if p_0 is generic and if the degrees are ordered, i.e., $d_1 \geq \cdots \geq d_m$, then the algebraic degree is upper bounded by

$$d_1 \cdots d_c \sum_{i_0 + \cdots + i_c = l} (d_0 - 1)^{i_0} \cdots (d_c - 1)^{i_c}. \tag{3.65}$$

When the polynomials are not generic, then this provides an upper bound on the number of isolated critical points. To see this, fix a subset $J \subseteq \{1, \dots, m\}$ of cardinality c, such that

$$\mathcal{V} = \{x : p_j(x) = 0 \text{ for } j \in J\}.$$

Then we can apply the bound from (3.64) and evaluate it to be

$$\prod_{j \in J} d_j \sum_{i_0 + \sum_{j \in J} i_j = n - c} (d_0 - 1)^{i_0} \cdot \prod_{j \in J} (d_j - 1)^{i_j},$$

which is clearly upper bounded by (3.65). If p_0 is linear, then $d_0 = 1$ and the expression simplifies to

$$d_1 \cdots d_c \sum_{i_1 + \cdots + i_c = l} (d_1 - 1)^{i_0} \cdots (d_c - 1)^{i_c}.$$

If further $d_i = 1$ for $i \geq k$ for some $k \leq c$, then we obtain

$$d_1 \cdots d_k \sum_{i_1 + \cdots + i_k = l} (d_1 - 1)^{i_1} \cdots (d_k - 1)^{i_k}. \tag{3.66}$$

3.3.2. Upper bounds for invertible observation matrix.

Here, we apply results from the general theory of algebraic degrees to the case of invertible observation matrices, which yield an explicit expression of the polynomials constraints defining the set of state-action frequencies. If the observation matrix is invertible we have seen in Subsection 3.2.2 that there are no polynomial equalities but only inequalities with polynomials given in (3.30).

Theorem 3.33. *Consider a POMDP $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$, $\gamma \in [0, 1)$ that $\beta \in \mathbb{R}^{\mathcal{S} \times \mathcal{O}}$ is invertible, and that the positivity Assumption 3.3 holds. For $I \subseteq \mathcal{A} \times \mathcal{O}$ consider the following set of policies*

$$\operatorname{int}(F_I) = \left\{\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}} : \pi(a|o) = 0 \textit{ if and only if } (a, o) \in I\right\},$$

which is the relative interior of a face F_I of the policy polytope. Let

$$O := \{o \in \mathcal{O} : (a, o) \in I \textit{ for some } a\}$$

and set $k_o := |\{a \in \mathcal{A} : (a, o) \in I\}|$ as well as $d_o := |\{s \in \mathcal{S} : \beta_{os}^{-1} \neq 0\}|$. Then, the number of isolated critical points of the reward function on $\operatorname{int}(F)$ *is at most*

$$\left(\prod_{o \in O} d_o^{k_o}\right) \cdot \sum_{\sum_{o \in O} i_o = l} \prod_{o \in O} (d_o - 1)^{i_o}, \tag{3.67}$$

where $l = n_{\mathcal{S}}(n_{\mathcal{A}} - 1) - |I|$.

Proof. The face G_I of the effective policy polytope corresponding to F_I is given by

$$\mathrm{int}(G_I) = \left\{\tau \in \Delta_{\mathcal{A}}^{\mathcal{S},\beta} : (\beta^{-1}\tau)_{oa} = 0 \Leftrightarrow (a,o) \in I\right\}.$$

Using the explicit formulas from Subsection 3.2.2 and in particular (3.30) it holds that

$$\mathcal{N}^\beta = \{\eta \in \mathcal{N} : p_{ao}(\eta) \geq 0 \text{ for all } a \in \mathcal{A}, o \in \mathcal{O}\},$$

where

$$p_{ao}(\eta) = \sum_{s \in S_o} \left(\beta_{os}^{-1} \eta_{sa} \prod_{s' \in S_o \setminus \{s\}} \sum_{a'} \eta_{s'a'} \right)$$

and $S_o := \{s \in \mathcal{S} : \beta_{os}^{-1} \neq 0\}$. Then, F_I and G_I correspond to the boundary component

$$\begin{aligned}\mathrm{int}(H_I) &= \left\{\eta \in \mathcal{N}^\beta : p_{ao}(\eta) = 0 \Leftrightarrow (a,o) \in I\right\} \\ &= \left\{\eta \in \mathcal{N} : p_{ao}(\eta) \geq 0 \text{ and equality if and only if } (a,o) \in I\right\}\end{aligned}$$

of the set $\mathcal{N}^\beta$ of feasible state-action frequencies. In order to use the explicit description of the state-action polytope $\mathcal{N}$ given in Theorem (3.5), we remind the reader that

$$\ell_s(\eta) := \sum_{a \in \mathcal{A}} \eta_{sa} - \gamma \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \eta_{s'a'} \alpha(s|s',a') - (1-\gamma)\mu_s.$$

Then, it holds that

$$\mathrm{int}(H_I) = \left\{ \eta \in \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}} : \begin{array}{c} p_{ao}(\eta) \geq 0 \text{ and equality if and only if } (a,o) \in I \\ \ell_s(\eta) = 0 \text{ for } s \in \mathcal{S} \end{array} \right\}.$$

Since the state frequencies are all positive by Assumption 3.3, for $\eta \in \mathrm{int}(H)$ it holds $\eta_{sa} = 0$ if and only if $\tau(a|s) := \eta(a|s) = 0$. Note that $\tau = \pi \circ \beta$ for some $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ by assumption and thus for $\eta \in \mathrm{int}(H)$ it holds that $\eta_{sa} = 0$ if and only if

$$0 = \tau(a|s) = \sum_o \beta(o|s)\pi(a|o),$$

which holds if and only if $(a,o) \in I$ for every $o \in \mathcal{O}$ with $\beta(o|s) > 0$. Hence, if we write

$$J := \{(s,a) \in \mathcal{S} \times \mathcal{A} : (a,o) \in I \text{ for all } o \in \mathcal{O} \text{ with } \beta(o|s) > 0\},$$

we obtain

$$\mathrm{int}(H_I) = \left\{ \eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \begin{array}{c} \eta_{sa} \geq 0 \text{ and equality if and only if } (s,a) \in J, \\ \ell_s(\eta) = 0 \text{ for } s \in \mathcal{S}, \\ p_{ao}(\eta) \geq 0 \text{ and equality if and only if } (a,o) \in I \end{array} \right\}.$$

The number of critical points over this surface is upper bounded by the number of critical points over

$$\mathcal{V}_I = \left\{ \eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \begin{array}{c} \eta_{sa} = 0 \text{ for } (s,a) \in J, \\ \ell_s(\eta) = 0 \text{ for } s \in \mathcal{S}, \\ p_{ao}(\eta) = 0 \text{ for } (a,o) \in I \end{array} \right\}.$$

Now we want to apply (3.66) and note that the objective $p_0 = \langle r, \cdot \rangle_{\mathcal{S}\times\mathcal{A}}$ is generic. Further, we see that there are $|I|$ non-linear constraints and hence in the notation of (3.66) have

$k = |I|$. Further, we can calculate to dimension and co-dimension of $\mathcal{V}_I$ as follows. Note that $F_I \longrightarrow \mathcal{V}_I, \pi \mapsto \eta^\pi$ is a parametrization of $\mathcal{V}$ (it parametrizes a full dimensional subset of $\mathcal{V}_I$), which is injective and has full rank Jacobian everywhere. Hence, we have

$$l = \dim(\mathcal{V}_I) = \dim(F_I) = n_{\mathcal{S}}(n_{\mathcal{A}} - 1) - |I| = n_{\mathcal{S}} n_{\mathcal{A}} - n_{\mathcal{S}} - |I|.$$

The co-dimension of $\mathcal{V}_I$ is given by $n_{\mathcal{S}} n_{\mathcal{A}} - \dim(\mathcal{V}_I) = n_{\mathcal{S}} + |I|$ and with the notation from (3.66), we have $c = n_{\mathcal{S}} + |I| \geq k$. Further, it holds that $\deg(p_{ao}) \leq d_o$ and using (3.66) yields an upper bound of

$$\prod_{(s,o)\in I} d_o \cdot \sum_{\sum_{(a,o)\in I} j_{ao} = l} \prod_{(a,o)\in I} (d_o - 1)^{j_{ao}} = \prod_{o\in O} d_o^{k_o} \cdot \sum_{\sum_{o\in O} i_o = l} \prod_{o\in O} (d_o - 1)^{i_o},$$

which finishes the proof. □

Remark 3.34 (The mean reward case). Theorem 3.33 can be generalized to the mean reward case, i.e., to the case of $\gamma = 1$ with some adjustments. Indeed, the proof can be carried out analogously, however, the characterization of $\mathcal{N}$ has the extra linear condition that $\sum_{sa} \eta_{sa} = 1$, see also Theorem 3.5. In the mean reward case we have with the notation from the proof above

$$\operatorname{int}(H_I) = \left\{ \eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \begin{array}{c} \eta_{sa} \geq 0 \text{ and equality if and only if } (s,a) \in J, \\ \ell_s(\eta) = 0 \text{ for } s \in \mathcal{S}, \sum_{sa} \eta_{sa} = 1, \\ p_{ao}(\eta) \geq 0 \text{ and equality if and only if } (a,o) \in I \end{array} \right\}.$$

Hence, the upper bound in (3.67) remains valid if we set

$$l := \dim \left\{ \eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \begin{array}{c} \eta_{sa} = 0 \text{ for } (s,a) \in J, \ell_s(\eta) = 0 \text{ for } s \in \mathcal{S}, \sum_{sa} \eta_{sa} = 1, \\ p_{ao}(\eta) = 0 \text{ for } (a,o) \in I \end{array} \right\}. \tag{3.68}$$

In the discounted case we obtained an explicit formulation for l. In the mean case the value obeys a case distinction depending, in particular, on whether the constraint $\sum_{sa} \eta_{sa} = 1$ is redundant with respect to the constraints $\ell_s(\eta) = 0$. However, the value can be computed from the above expression (3.68) in any given specific case.

Corollary 3.35 (Critical points of MDPs). *Consider an MDP $(\mathcal{S}, \mathcal{A}, \alpha, r)$, $\gamma \in [0,1)$ and let Assumption 3.3 holds. Then, every isolated critical point $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ of the discounted expected reward function is deterministic.*

Proof. We evaluate the bound of Equation (3.67) and have $O = \mathcal{S}$ in this fully observable case. If the face is not a vertex, then the corresponding index set $I \subsetneq \mathcal{A} \times O$ satisfies $|I| < n_O(n_{\mathcal{A}} - 1)$ and thus in the notation from Theorem 3.33 it holds that $l > 0$. Note that $d_o = 1$ for every $o \in O$ and hence there is at least one factor in the product in (3.67) that vanishes and so does the whole expression in (3.67). □

The result above strengthens Theorem 2.23, which ensures the existence of a deterministic optimal policy.

Example 3.36 (Crying baby example continued). Let us revisit the crying baby example and discuss the implications of the bound on the number of critical points given in

Theorem 3.33. First, recall that the observation matrix $\beta \in \Delta_{\mathcal{O}}^{\mathcal{S}}$ is invertible with inverse

$$\beta^{-1} = \begin{array}{c c} & \begin{array}{cc} s_1 & s_2 \end{array} \\ \begin{array}{c} o_1 \\ o_2 \end{array} & \begin{pmatrix} 1 & 0 \\ -1 & 2 \end{pmatrix} \end{array}$$

and hence in the notation of Theorem 3.33 we have $d_{o_1} = 1$ and $d_{o_2} = 2$. Consider first the interior of a face F for which we assume that $\pi(a|o_2) > 0$ for $a \in \mathcal{A}$ or in other words $k_{o_2} = 0$ and consequently $O = \{o_1\}$. Then $l = n_{\mathcal{S}}(n_{\mathcal{A}} - 1) - |I| = 2 - k_{o_1} \geq 1 > 0$ since $k_{o_1} = 2$ would imply $\pi(a_1|o_1) = \pi(a_2|o_1) = 0$, which would correspond to the empty face. Now the second factor of (3.67) becomes $(d_{o_1} - 1)^l = 0$ and hence there is no critical point if $k_{o_2} = 0$. Note that for any non empty face $k_{o_2} \leq 1$ and hence for generic reward vector r there can only be a critical policy if $k_{o_2} = 1$, i.e., if the face only contains policies, which are deterministic on o_2. In this particular case this strengthens Theorem 2.30 that assures the existence of an optimal policy, which is deterministic on o_2.

Let us now consider the case $k_{o_2} = 1$, then either $k_{o_1} = 0$ or $k_{o_1} = 1$. In the case that $k_{o_1} = 0$ we have $O = \{o_2\}$, $l = 1$ and hence the bound evaluates to

$$d_{o_2}^{k_{o_2}}(d_{o_2} - 1)^l = 2.$$

If $k_{o_1} = 1$, the face consists of a single deterministic policy and the bound evaluates to 2.

Remark 3.37 (Geometry around the critical points). The key argument in the proof of Theorem 3.33 is that a critical point $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ of the reward function corresponds to a critical point η of a linear function over a multi-homogeneous variety $\mathcal{V}$. A closer study of this variety would shed light into the geometry of the loss landscape around the critical points, which has important implications for gradient based methods.

Remark 3.38 (Efficient design of observation mechanisms). The bound (3.67) could be used to design observation mechanisms in such a way that the reward function has the least critical points, which would potentially make the system more approachable for gradient based methods. Consider two observation kernels $\beta, \beta' \in \Delta_{\mathcal{O}}^{\mathcal{S}}$ satisfying $\|\beta(\cdot|s) - \beta'(\cdot|s)\|_{TV} = \sum_o |\beta(o|s) - \beta'(o|s)|/2 \leq \varepsilon$ for every $s \in \mathcal{S}$. Then if $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ is an optimal policy of $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \alpha, \beta', r)$, then it is a $2\varepsilon\gamma\|r\|_\infty/(1-\gamma)$-optimal policy of $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \alpha, \beta, r)$, see [239]. Hence, if β does not fulfill the invertibility assumption made in Theorem 3.33 an arbitrary small perturbation of it does (given that β is a square matrix) and hence Theorem 3.33 provides an upper bound on the number of critical points of an approximate problem. Further, note that the faces, which are guaranteed to contain an optimal policy by [204] might be considerably fewer for the POMDP $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \alpha, \beta', r)$. The bound (3.67) could be used to identify the best perturbations of a given magnitude to obtain a problem with a minimal number of critical points.

3.3.3. Upper bounds for deterministic observations. In this section, we study the critical points of the reward optimization problem with deterministic observations.

The description of the set of feasible state-action frequencies $\mathcal{N}^\beta$ obtained in Corollary 3.24 implies that the reward maximization problem in state action space (ROP-SA) is

the following constrained polynomial optimization problem:
(3.69)

$$\text{maximize } \langle r, \eta \rangle \quad \text{subject to } \begin{cases} \ell_s(\eta) = 0 & \text{for } s \in \mathcal{S}, \\ p^o_{sa}(\eta) = 0 & \text{for } o \in \mathcal{O}, a \in \mathcal{A} \setminus \{a_o\}, s \in S_o \setminus \{s_o\}, \\ \eta_{sa} \geq 0 & \text{for } s \in \mathcal{S}, a \in \mathcal{A}, \end{cases}$$

where the linear constraints ℓ_s are given in Proposition 3.5, the polynomial constraints $p^o_{sa}(\eta)$ are provided in (3.35) taking a fixed action $a_o \in \mathcal{A}$ and a fixed state $s_o \in S_o$ for each observation $o \in \mathcal{O}$, and the inequality constraints simply ensure the entries of η being nonnegative. Observe that problem (3.69) is in fact a quadratically constrained linear program.

We bound the number of critical points individually for each boundary component of the feasible set. A boundary component consists of all feasible points for which a given subset of the inequality constraints are active. The boundary components of the feasible set $\mathcal{N}^\beta$ are in one-to-one correspondence with the faces of $\Delta^{\mathcal{O}}_{\mathcal{A}}$ according to
(3.70)

$$B = \left\{ \pi \in \Delta^{\mathcal{O}}_{\mathcal{A}} : \pi(a|o) = 0 \text{ for } a \in A_o, o \in \mathcal{O} \right\} \leftrightarrow F = \left\{ \eta \in \mathcal{N}^\beta : \eta_{sa} = 0 \text{ for } a \in A_{g_\beta(s)} \right\},$$

where A_o is a proper subset of $\mathcal{A}$ for every $o \in \mathcal{O}$, and $g_\beta(s)$ is the observation associated with state s. In particular, there is a boundary component associated to each tuple $(A_o)_{o \in \mathcal{O}}$ with $A_o \subsetneq \mathcal{A}$, $o \in \mathcal{O}$.

We point out the following result, which allows us to ignore high-dimensional boundary components when searching for a maximizer of the reward. Recall that for an observation $o \in \mathcal{O}$, the cardinalities of the fibers of g_β are denoted by $d_o = |S_o|$.

Theorem 3.39 (Existence of maximizers in low dimensional faces, [204]). *There exist $A_o \subsetneq \mathcal{A}$ with $|A^c_o| \leq d_o$, $o \in \mathcal{O}$, such that the set B described in* (3.70) *contains a (globally optimal) solution of the problem* (3.69).

Remark 3.40. Instead of considering the critical points in all $(2^{n_{\mathcal{A}}} - 1)^{n_{\mathcal{O}}}$ boundary components, it is enough to consider those in the boundary components with $A_o \subsetneq \mathcal{A}$ satisfying $|A_o| \geq n_{\mathcal{A}} - d_o$. This reduces the number of boundary components that need to be checked to

$$\prod_{o \in \mathcal{O}} \left(\sum_{k_o = \max(n_{\mathcal{A}} - d_o, 0)}^{n_{\mathcal{A}} - 1} \binom{n_{\mathcal{A}}}{k_o} \right),$$

which we call *relevant* boundary components. Note that this number only depends on the number of actions $n_{\mathcal{A}}$ and d_o (the cardinality of the fibers of g_β).

With the description of the boundary components of the feasible set at hand, we can deduce upper bounds on the number of critical points over each of them based on the degrees of the defining equations and the degree of the objective function.

Theorem 3.41 (Bound on the algebraic degree). *Consider a POMDP $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$, $\gamma \in [0, 1)$ with deterministic observations and with $d_o := \{s \in \mathcal{S} : \beta(o|s) > 0\}$ we denote the number of states that are mapped to o and let the positivity Assumption 3.3 hold. Fix $A_o \subsetneq \mathcal{A}$ for every $o \in \mathcal{O}$ and set $n := n_{\mathcal{S}} n_{\mathcal{A}} - n_{\mathcal{S}} - \sum_o d_o |A_o|$ and $m := \sum_o (d_o - 1)(|A^c_o| - 1)$, where we assume n is not zero. Then the number of isolated critical points the reward function over the* *interior of the face*

$$\mathrm{int}(F) = \{\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}} : \ \pi(a|o) = 0 \textit{ if and only if } a \in A_o\}$$

is upper bounded by $2^m \binom{n-1}{m-1}$.

Proof. First, note that the face $F = \{\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}} : \ \pi(a|o) = 0 \text{ for } a \in A_o\}$ corresponds to the boundary component

$$B = \{\eta \in \mathcal{L} \cap \mathcal{X} : \ \eta_{sa} = 0 \text{ for } a \in A_{g_\beta(s)}\}. \tag{3.71}$$

Since the parametrization $\pi \mapsto \eta^\pi$ has full rank Jacobian everywhere, see Lemma 3.1, the number of critical points of the reward function on $\mathrm{int}(F)$ is upper bounded by the number of critical points of the linear function $\eta \mapsto \langle r, \eta\rangle_{\mathcal{S}\times\mathcal{A}}$ over B.

Recall from Corollary 3.24 that $\mathcal{N}^\beta$ is defined in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as an intersection of $n_{\mathcal{S}}$ linear equations, $\sum_o(d_o - 1)(n_{\mathcal{A}} - 1)$ quadratic equations of the form (3.35), and the linear inequalities $\eta \geq 0$. We start by showing that the family of linear equations $\ell_s(\eta) = 0, s \in \mathcal{S}$ and $\eta_{sa} = 0$, $a \in A_o$, $s \in S_o$, $o \in \mathcal{O}$ is linearly independent for any choice of $A_o \subsetneq \mathcal{A}$, $o \in \mathcal{O}$. For this we first note that the linear equations $\ell_s(\eta) = 0$, $s \in \mathcal{S}$ define the space $\mathcal{L} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ of dimension $\dim(\mathcal{L}) = \dim(\mathrm{affine}(\Delta_{\mathcal{A}}^{\mathcal{S}})) = n_{\mathcal{S}}(n_{\mathcal{A}} - 1)$, see Proposition 3.4, which implies their linear independence. It now suffices to see that the restrictions of η_{sa} to $\mathcal{L}$ are linearly independent. Note that the pullback of the equations $\eta_{sa} = 0$ restricted to $\mathcal{L}$ along the birational map Ψ are the equations $\tau_{as} = 0, a \in A_o, s \in S_o$, which are linearly independent on $\mathrm{affine}(\Delta_{\mathcal{A}}^{\mathcal{S}})$.

On the set B given in (3.71) there are $\sum_o d_o|A_o|$ active linear inequalities with $A_o \subsetneq \mathcal{A}$ for each $o \in \mathcal{O}$, and hence B is contained in an affine space of dimension

$$n = n_{\mathcal{S}}n_{\mathcal{A}} - n_{\mathcal{S}} - \sum_o d_o|A_o|.$$

Further, given these linear equations, the quadratic equations

$$p_{sa}^o(\eta) = \eta_{sa}\sum_{a'\in\mathcal{A}}\eta_{s_oa'} - \eta_{s_oa}\sum_{a'\in\mathcal{A}}\eta_{sa'} = 0$$

are redundant for all $a \in A_o, s \in S_o$. By choosing $a_o \in A_o^c$ in Corollary 3.24 for every $o \in \mathcal{O}$ there remain $n_{\mathcal{A}} - |A_o| - 1$ non-redundant quadratic equalities for every $s \in S_o \setminus \{s_o\}$. Therefore, we get $m = \sum_o(d_o - 1)(|A_o^c| - 1)$ non-redundant quadratic equalities. By Theorem 2.2 and Corollary 2.5 in [147] the algebraic degree for the optimization of the linear function $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ over an n-dimensional affine space subject to m non-redundant quadratic constraints is upper bounded by $2^m\binom{n-1}{m-1}$. □

With Theorem 3.41 we can provide upper bounds for the number of critical points of the optimization problem (3.69). Indeed, the number of critical points over the interior

$$\{\eta \in \mathcal{L} \cap \mathcal{X} : \ \eta_{sa} = 0 \text{ for all } a \in A_{g_\beta(s)}, \eta_{sa} > 0 \text{ otherwise}\} \tag{3.72}$$

of a boundary component is clearly upper bounded by the number of critical points over B defined in (3.71). This bound over the individual boundary components can be summed to obtain an upper bound on the number of critical points of the polynomial optimization problem (3.69), see also [147]. Note that the Zariski closure of the interior of a boundary component defined in (3.72) is contained in B but might be a strict subset. Similarly, a bound on the number of critical points over the relevant boundary components can be established.

In Table 3.2 we present the upper bounds on the number of critical points for problems of different sizes. We compare the bound on the total number of critical points obtained by iterating Theorem 3.41 over all boundary components and the one iterating only over the relevant components described in Theorem 3.39. In addition, we report the total and relevant number of boundary components discussed in Remark 3.40. Both the number of boundary components and the upper bound on the number of critical points, depend on $n_{\mathcal{S}}, n_{\mathcal{A}}$, and the tuple $(d_o)_{o\in\mathcal{O}}$. The two extreme cases for the tuple $(d_o)_{o\in\mathcal{O}}$, namely $(n_{\mathcal{S}})$ and $(1,\dots,1)$, correspond to a *numb controller*, i.e., all states map to the same observation, and the *fully observable case*, i.e., states and observations are in one-to-one correspondence, respectively. The bounds are independent of the specific transition kernel $\alpha \in \Delta^{\mathcal{S}}_{\mathcal{S}\times\mathcal{A}}$.

$n_{\mathcal{S}}$	$n_{\mathcal{A}}$	partitions of $n_{\mathcal{S}}$: $(d_o)_{o\in\mathcal{O}}$	Number of boundary components		Bound on number of critical points	
			total	relevant	total	relevant
3	2	(3)	3	3	10	10
		$(2,1)$	9	6	10	8
		$(1,1,1)$	27	8	8	8
4	3	(4)	7	7	1419	1419
		$(3,1)$	49	21	2237	561
		$(2,2)$	49	36	1265	153
		$(2,1,1)$	343	54	1189	81
		$(1,1,1,1)$	2401	81	81	81
5	3	(5)	7	7	9411	9411
		$(4,1)$	49	21	23745	4257
		$(3,2)$	49	42	13431	4371
		$(3,1,1)$	343	63	24363	1683
		$(2,2,1)$	343	108	12159	459
		$(2,1,1,1)$	2401	162	9195	243
		$(1,1,1,1,1)$	16807	243	243	243

Table 3.2. Listed are the number of boundary components and the upper bound on the number of critical points from Theorem 3.41 both over all boundary components and over the subset of relevant boundary components from Theorem 3.39 for problems of different size.

In these examples, we observe that restricting to the relevant boundary components significantly reduces the upper bound. This is reflected in the last two columns in Table 3.2. The difference is most notable when the fibers of g_β have a small cardinality, i.e., only few states lead to the same observation. In the fully observable case, the relevant boundary components correspond to the vertices of $\Delta^{\mathcal{O}}_{\mathcal{A}}$. This is consistent with the fact that in the fully observable case the feasible set of state-action frequencies $\mathcal{N}$ is a polytope [93] and hence the optimization problem (3.69) is a linear program, for which the solutions are attained at the vertices. On the other hand, in the case of a numb controller (with a single observation o), all boundary components are relevant since $d_o = n_{\mathcal{S}}$.

3.3.4. A tighter bound for one observation and two actions. Already in the Example 3.36 of the crying baby we have seen that the bound from Theorem 3.33 on the number of critical points are not tight. One reason for this is that the bound does not consider the specific structure of the problem and only takes the degree of the polynomials describing the feasible state-action frequencies into account. Here, we provide a tighter bound on the number of critical points for numb controllers with two actions. Our proof relies on the expression of the reward function as a rational function and does not easily generalize to larger problems. We offer an alternative argument based on polar degrees that we believe could be extended to the general case of deterministic observations.

Let us begin by evaluating the bound from Theorem 3.41 for the case of one observation and two actions. In this case the policy polytope is equivalent to the line segment $\Delta_{\mathcal{A}} \cong [0,1]$ and hence there are two zero dimensional and one full dimensional face. On the full dimensional face the bound evaluates to $n_{\mathcal{S}}2^{n_S-1}$, which is (essentially) exponential in the size of the state space. This can be improved to the following linear bound.

Proposition 3.42. *Let $(\mathcal{S},\mathcal{O},\mathcal{A},\alpha,\beta,r)$ be a POMDP describing a numb controller with two actions, i.e., $\mathcal{O}=\{o\}$ and $\mathcal{A}=\{a_1,a_2\}$ and let $\gamma\in[0,1)$. Then the reward function R has at most $2n_{\mathcal{S}}-2$ isolated critical points in the interior $\operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{O}})\cong(0,1)$ of the policy polytope and hence at most $2n_{\mathcal{S}}$ isolated critical points.*

Proof. We associate the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{O}}$ with $[0,1]$ and for $p\in[0,1]$ we write π_p and η^p for the associated policy and the state-action frequency. From Theorem 2.25 we know that the reward function $R=f/g\colon[0,1]\to\mathbb{R}$ is a rational function of degree at most $k:=n_{\mathcal{S}}$. The critical points of this function satisfy $f'(p)g(p)-g'(p)f(p)=0$. It is immediate that the degree of $h:=f'g-g'f$ is at most $2k-1$. However, writing $f(p)=\sum_{i=0}^k c_ip^i$, $g(p)=\sum_{i=0}^k d_ip^i$ we obtain

$$h(p)=\sum_{l=0}^{2k-1}\left(\sum_{i=0}^{k-1}c_{i+1}d_{l-i}-\sum_{j=0}^{k-1}c_{l-j}d_{j+1}\right)p^l$$

and see that the coefficient of p^{2k-1} cancels. Since h has at most $\deg(h)\le 2k-2$ isolated roots this shows that there are at most $2k-2$ isolated critical points of the reward function in the interior $(0,1)$. □

The proof given above does not easily generalize to arbitrarily many actions and observations. Thus, we provide a different ansatz that could potentially be generalized to deterministic observations. If $p_{k+1},\dots,p_m$ are affine linear and $p_1,\dots,p_k$ are homogeneous polynomials then the algebraic degree of (3.62) is given by the $(m-k-1)$-th *polar degree* $\delta_{m-k-1}(\mathcal{V})$ of the projective variety

$$\mathcal{V}:=\{\eta : p_1(\eta)=\cdots=p_k(\eta)=0\}\subseteq\mathbb{P}^{n-1},$$

see [104, 70, 195]. This relation is particularly useful, since for state-action frequencies there are always active linear equations as described in (3.6). The polar degrees of certain interesting cases (Segre-Veronese varieties) have been recently computed by [270] and we use those formulas and their presentation by [70].

In this case, the combinatorial part is simple, since there are only two zero-dimensional faces of the state-action frequencies (corresponding to the endpoints of the unit interval) and one one-dimensional face (corresponding to the interior of the unit interval). Let us set

$$\mathcal{L} = \left\{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \ell_s(\eta) = 0 \text{ for all } s \in \mathcal{S}\right\},$$

where $\ell_s(\eta) = \sum_{a\in\mathcal{A}} \eta_{sa} - \gamma \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} \eta_{s'a'}\alpha(s|s',a') - (1-\gamma)\mu_s$. By Theorem 3.25 the set of discounted state-action frequencies is given by

$$\mathcal{N}^\beta = \mathcal{N} \cap \mathcal{D}_1^{n_\mathcal{S}\times 2} = \mathbb{R}_{\ge 0}^{\mathcal{S}\times\mathcal{A}} \cap \mathcal{L} \cap \mathcal{D}_1^{n_\mathcal{S}\times 2},$$

where $\mathcal{D}_1^{n_\mathcal{S}\times 2}$ denotes the determinantal variety of rank one matrices of size $n_\mathcal{S} \times 2$. Like above, we associate the policy polytope $\Delta_\mathcal{A}^O$ with $[0,1]$ and for $p \in [0,1]$ we write π_p and η^p for the associated policy and the state-action frequency. We aim to bound the number of critical points of the reward function over $(0,1)$ or equivalently the number of critical over $\{\eta^p : p \in (0,1)\}$ if Assumption 3.3 holds. Denoting the state marginal of η^p with ρ_s^p, recall that $\eta^p(a|s) = \eta_{sa}^p/\rho_s^p$, we have that

$$\begin{aligned}\{\eta^p : p \in (0,1)\} &= \{\eta \in \mathcal{N}^\beta : \eta(a|s) > 0 \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}\}\\ &= \{\eta \in \mathcal{N}^\beta : \eta_{sa} > 0 \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}\}\\ &= \mathbb{R}_{>0}^{\mathcal{S}\times\mathcal{A}} \cap \mathcal{U} \cap \mathcal{D}_1^{n_\mathcal{S}\times 2}.\end{aligned}$$

Thus the number of critical points over $\{\eta^p : p \in (0,1)\}$ are upper bounded by the number of critical points on $\mathcal{U} \cap \mathcal{D}_1$. Since the co-dimension of $\mathcal{L}$ is $n_\mathcal{S}$, the number of complex solutions to the KKT conditions over $\mathcal{U} \cap \mathcal{D}_1^{n_\mathcal{S}\times 2}$ are given by the $(n_\mathcal{S}-1)$-th polar degree $\delta_{n_\mathcal{S}-1}(\mathcal{D}_1)$, which we can compute using the formula presented in [70, Corollary 5.4]. This yields

$$\begin{aligned}\delta_{n_\mathcal{S}-1}(\mathcal{D}_1) &= \sum_{l=0}^{n_\mathcal{S}-2n_\mathcal{S}+n_\mathcal{S}+1} (-1)^l \binom{n_\mathcal{S}-l+1}{2n_\mathcal{S}-n_\mathcal{S}} (n_\mathcal{S}-l)! \left(\sum_{i+j=l} \frac{\binom{n_\mathcal{S}}{i}}{(n_\mathcal{S}-1-i)!} \cdot \frac{\binom{2}{j}}{(2-1-j)!}\right)\\ &= \sum_{l=0}^{1} (-1)^l \binom{n_\mathcal{S}-l+1}{n_\mathcal{S}} (n_\mathcal{S}-l)! \left(\sum_{i+j=l} \frac{\binom{n_\mathcal{S}}{i}}{(n_\mathcal{S}-1-i)!} \cdot \frac{\binom{2}{j}}{(1-j)!}\right)\\ &= \binom{n_\mathcal{S}+1}{n_\mathcal{S}} n_\mathcal{S}! \left(\frac{\binom{n_\mathcal{S}}{0}}{(n_\mathcal{S}-1)!} \cdot \frac{\binom{2}{0}}{1!}\right) - \binom{n_\mathcal{S}}{n_\mathcal{S}} (n_\mathcal{S}-1)! \left(\frac{\binom{n_\mathcal{S}}{1}}{(n_\mathcal{S}-2)!} + \frac{\binom{2}{1}}{(n_\mathcal{S}-1)!}\right)\\ &= n_\mathcal{S}(n_\mathcal{S}+1) - 2 - n_\mathcal{S}(n_\mathcal{S}-1)\\ &= 2n_\mathcal{S} - 2\end{aligned}$$

and hence obtain the same bound as in Proposition 3.42. We believe that this ansatz can be extended to cover general deterministic observations, which would yield tighter bounds compared to Theorem 3.33. For this one would need to study the polar degrees of the product of determinantal varieties, see Theorem 3.25.

Example 3.43 (An example with multiple smooth and non-smooth critical points). It is the goal of this example to demonstrate that for a numb controller multiple critical points can occur in the interior $(0,1) \cong \operatorname{int}(\Delta_\mathcal{A}^O)$ as well as at the two endpoints of $[0,1] \cong \Delta_\mathcal{A}^O$ of the policy polytope. We refer to such points as smooth and non-smooth critical points. We consider a numb controller with one observation, two actions a_1, a_2 and three states s_1, s_2, s_3 and a deterministic transition kernel α and reward described by the graph shown

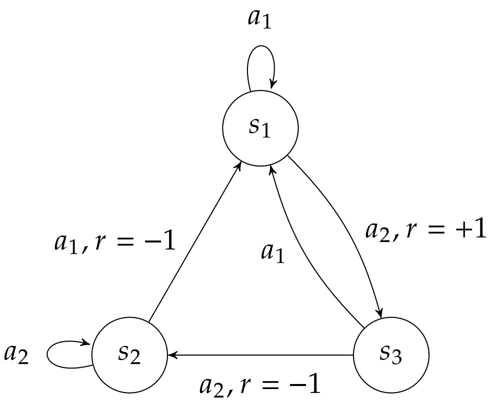

Figure 3.5. Graph describing the deterministic transition kernel α and the associated instantaneous rewards.

in Figure 3.5. We choose $\gamma = 0.8$ and $\mu \in \Delta_{\mathcal{S}}$ to be the uniform distribution and plot the reward function in Figure 3.6. We see that there are two critical points in the interior of the policy polytope $\Delta_A^O \cong [0, 1]$. The bound from Proposition 3.42 evaluate to $2n_{\mathcal{S}} - 2 = 4$ and therefore it is not tight in this example.

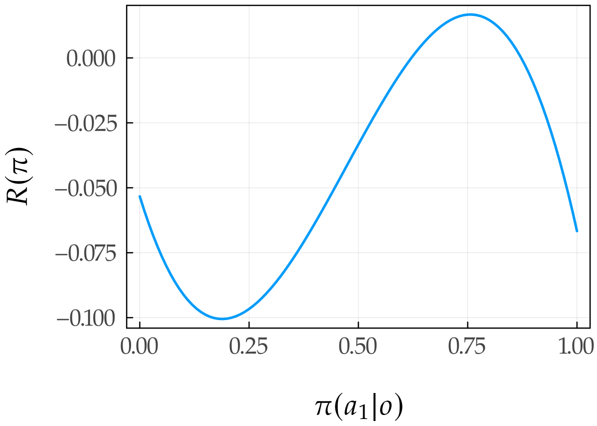

Figure 3.6. Plot of the reward function; note that there are two critical points in the interior of the policy polytope $\Delta_{\mathcal{A}}^O \cong [0, 1]$.

3.3.5. Outlook. In this section we have studied the number and location of critical points of the reward function. For this we used the explicit description of the geometry of the feasible state-action frequencies for invertible and deterministic observation kernel β obtained in Section 3.2. This allowed us to employ tools from the theory of polynomial optimization to obtain upper bounds on the number of critical points on every face of the policy polytope $\Delta_{\mathcal{A}}^O$. We have seen in examples that these bounds are not tight. One reason for this is that the results used to obtain the upper bounds only consider the degree but not the specific nature of the polynomial constraints. For the specific case of a numb controller with two actions obtained an improved bound with only linear dependence on the size of the state space. Whereas the argument builds on the description of the reward as a rational function we also provide a second argument based on polar degrees.

We consider the following questions to be relevant for future research:

- *Tighter bounds via polar degrees:* We have seen in one particular case the bounds on the number of critical points can be tightened significantly by working with polar degrees. A generalization of this approach has the potential to greatly improve the bounds established in this work.
- *Multi-agent problems:* Where we have described the feasible state-action frequencies of multi-agent MDPs, the optimization problem arising in multi-agent problems has not been studied conclusively. In particular, studying the number of critical points could yield inside into the role of the degree of decentralization for the algebraic complexity of the problem. In particular, insight regarding the number of critical points for different degrees of decentralization would be a valuable contribution.
- *Information theoretic objectives:* We think that it is useful to study the critical points for objective functions that are not linear in state-action space. Important examples for such settings include apprenticeship learning, where the state-action objective is the Euclidean distance, as well as unsupervised Reinforcement Learning, where the state-action objective is given by certain information theoretic quantities, for example the entropy [312]. The number of critical points for these objectives have been studied in the algebraic statistics community under the names *Euclidean distance degree (ED degree)* [104] and *maximum likelihood degree (ML degree)* [68].

3.4 Reward optimization in state-action space (ROSA)

We have seen that the feasible state-action frequencies of a POMDP form a polynomially constrained subset of the simplex, see Corollary 3.20. This implies that reward optimization in POMDPs is equivalent to a polynomially constrained linear objective program, which can be seen as a extension of the dual linear program associated to fully observable MDPs to partially observable problems. We refer to this approach as **R**eward **O**ptimization in **S**tate-**A**ction space (ROSA) and present its pseudo code in Algorithm 3. The two non-trivial steps in the algorithm are the computation of the defining linear inequalities of the effective policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$ in line 4 and the solution of the constrained optimization problem in line 6. The defining linear inequalities of $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$ can either be computed relatively simple for injective β, see Subsection 3.2.2, or algorithmically, e.g., by Fourier-Motzkin elimination, block elimination, vertex approaches, and equality set projection [150].

For the solution of the constrained optimization problem we first use an interior point method. Further, we solve the critical equations using a polynomial systems solver as well as a convex SDP relaxation. We find that using an interior point method in state-action space offers stability benefits to existing methods working with policies. In addition the numerical algebraic approaches provide globally optimal solutions.

3.4.1. Interior point methods and stability improvements. Here, we investigate the practical viability of this approach to optimize the reward in POMDPs. We consider navigation tasks in random mazes of different sizes, for which we solve the constrained

Algorithm 3 **R**eward **O**ptimization in **S**tate-**A**ction space (ROSA)

```
Require: $\alpha \in \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}, \beta \in \Delta_{O}^{\mathcal{S}}, \gamma \in [0,1), \mu \in \Delta_{\mathcal{S}}$
for all $s \in \mathcal{S}$ do
    $\ell_s(\eta) \longleftarrow \sum_{a\in\mathcal{A}} \eta_{sa} - \gamma \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} \eta_{s'a'}\alpha(s|s',a') - (1-\gamma)\mu_s$
end for
Compute the defining linear inequalities of $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$
Compute the defining polynomial inequalities $p_i(\eta) \geq 0$ of $\mathcal{N}_\beta$    ▷ According to (3.25)
$\eta^* \longleftarrow \arg\max\langle r, \eta\rangle$ sbj to $\eta \geq 0$, $\ell_s(\eta) = 0$ , $p_i(\eta) \geq 0$
$R^* \longleftarrow \langle r, \eta^*\rangle$    ▷ Evaluate the optimal value
$\tau^* \longleftarrow \eta^*(\cdot|\cdot) \in \Delta_{\mathcal{A}}^{\mathcal{S}}$    ▷ Compute an optimal state policy
$\pi^* \longleftarrow$ solution of $\beta\pi = \tau^*$    ▷ Compute an optimal observation policy
    return $\eta^*, R^*, \pi^*$    ▷ maximizer, optimal value, optimal policy
```

optimization problem using the interior point method `Ipopt`. Our experiments show that this can yield significant computational savings compared to several baselines, while remaining numerically stable across values of γ where other methods fail.

Baselines: Policy gradients and Bellman constrained programming. A very popular approach in Reinforcement Learning dating back to [277] are policy gradients methods that are variants of the gradient descent algorithm. Typically, policies are parametrized $\theta \mapsto \pi_\theta$ and in its simplest form the iterates are given by

$$\theta_{k+1} = \theta_k + \Delta t \cdot \nabla R(\theta_k),$$

where $\Delta t > 0$ is the step size. In our experiments we use tabular softmax policies given by

$$\pi_\theta(a|o) := \frac{e^{\theta_{oa}}}{\sum_{a'} e^{\theta_{oa'}}} \quad \text{for all } a \in \mathcal{A}, o \in O$$

and use limited memory version of the Broyden–Fletcher–Goldfarb–Shanno (L-BFGS), which is a quasi Newton method. In comparison to a naive policy gradient, we observed L-BFGS to converge faster. We refer to this approach as direct policy optimization (DPO).

As a second baseline we consider a reformulation of the reward maximization problem as a quadratically constrained linear program [17] that we refer to as Bellman constrained programming (BCP), see also Subsection 2.4.5. Recall that $R^\mu(\pi) = \langle \mu, V^\pi\rangle_{\mathcal{S}}$ for any policy $\pi \in \Delta_{\mathcal{A}}^{O}$ and any initial distribution $\mu \in \Delta_{\mathcal{A}}$, see (2.9). In the light of the Bellman equation $V^\pi = \gamma p_\pi V^\pi + (1-\gamma)r_\pi$, Theorem 2.9, the reward optimization problem (ROP) is equivalent to the following quadratically constrained linear program

$$\text{(BCP)} \qquad \text{maximize } \langle \mu, v\rangle_{\mathcal{S}} \quad \text{subject to } \pi \in \Delta_{\mathcal{A}}^{O} \text{ and } v = \gamma p_\pi v + (1-\gamma)r_\pi.$$

In our experiment we use the interior point method `Ipopt` to solve (BCP).

Value and policy iteration can not directly be used to when optimizing stochastic memoryless policies in a POMDP and hence we do not consider these in our experiments.

The regime for $\gamma \to 1$ for MDPs. We recall how the complexity of solving fully observable MDPs depends on the discount factor γ. In fully observable problems, the iteration complexity of policy gradient methods behaves like $O((1-\gamma)^{-\kappa})$, where $\gamma \in [0,1)$ is the discount factor and $\kappa \in \mathbb{N}$ depends on the specific method [71]. This is reminiscent of the Lipschitz constant of the reward function (as a function of the policy), which behaves like $O((1-\gamma)^{-1})$, see Corollary 3.13 and Example 3.14. This leads to increasingly ill-conditioned problems as $\gamma \to 1$ and can cause undesired oscillations during optimization [288]. However, choosing a discount factor close to 1 is desirable as one often wishes to optimize the mean reward rather than a discounted reward. This is also required to prevent vanishing policy gradients in sparse reward MDPs, where, denoting $n_{\mathcal{S}}$ the number of states, gradients can of order $O(2^{-n_{\mathcal{S}}/2})$ if $\gamma \leq n_{\mathcal{S}}/(n_{\mathcal{S}}+1)$ [2]. In principle the ill-conditioning problem can be addressed by introducing an appropriate metric, as in natural policy gradients or trust region policy optimization, which can be costly, however.

We have seen that the iteration complexity bounds for value and policy iteration degrade for $\gamma \to 1$ as they scale like $O(\log(1-\gamma)/\gamma)$. This is also complemented by a lower bound of $\log(1-\gamma))/\gamma$ for value iteration. In MDPs, the state-action frequencies form a polytope and hence the reward optimization problem in state-action space becomes a linear program [93, 154], see also Subsection 2.4.3. This yields a strongly polynomial algorithmic approach, i.e., does not degrade for $\gamma \to 1$ [234].

Overall, the complexity of solving fully observable MDPs in state-action space does not suffer diverge for $\gamma \to 1$. We extend this by empirically showing that reward optimization in state-action space for POMDPs has similar benefits: We find that using an interior point method to solve the resulting polynomially constrained optimization problem offers an efficient approach that does not deteriorate for $\gamma \to 1$.

Description and discussion of the experiments. To demonstrate the performance of ROSA combined with the interior point method `Ipopt` we test it on navigation problems in mazes. For this, we generate connected mazes using a random depth first search [196], see Figure 3.7 for an example of such a maze with 199 states. Then we randomly select a state as the goal state at which a reward of $n_{\mathcal{S}}$ is picked up and from which the agent transitions to a uniform state. For all other states the four possible actions move the agent right, left, up or down. The agent can only observe the 8 neighboring cells and starts at a uniform position.

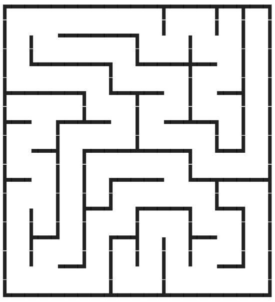

Figure 3.7. Shown is one of the mazes with 199 states that is used in the experiments.

In order to compare the running times of the three approaches, we generate square mazes of side length $2n-1$ and $2n^2-1$ states, for $n = 2, \dots, 10$. We solve the POMDPs for a discount factor of $\gamma = 0.9999$ using ROSA, BCP and DPO for 10^2 different mazes of

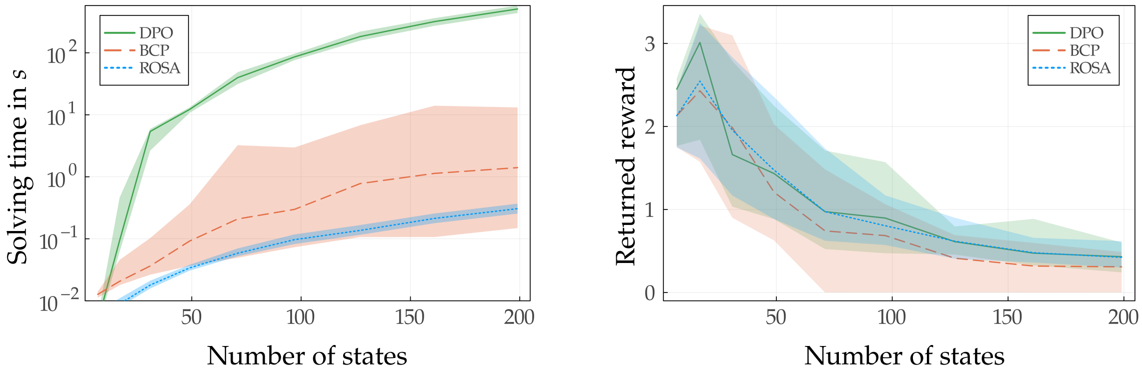

FIGURE 3.8. Shown is the solution time and cumulative reward obtained by different methods solving navigation tasks depending on the number of states; ROSA reaches competitive reward in less time compared to the other methods.

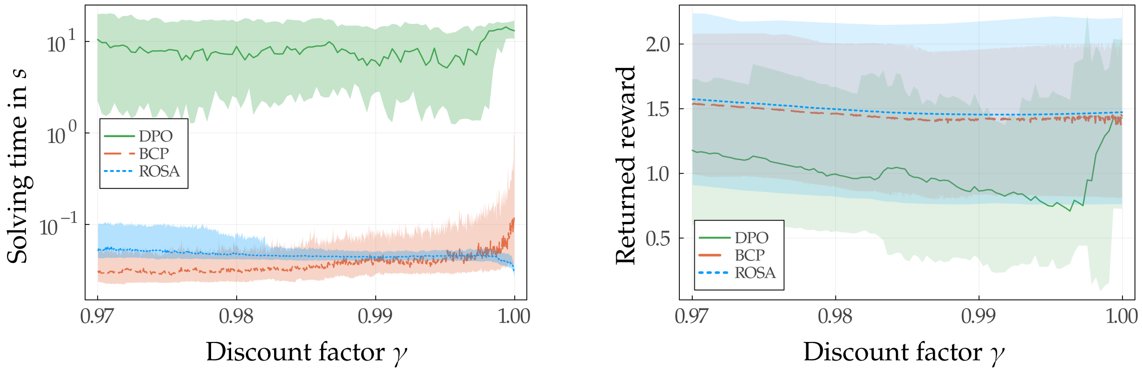

FIGURE 3.9. Shown is the solution time and cumulative reward obtained by different methods solving navigation tasks depending on discount factor; ROSA reaches higher reward in competitive time with stability benefits compared to DPO for $\gamma \to 1$.

each size[1] and report the mean solution times and achieved rewards as well as their 16% and 84% quantiles in Figure 3.8. We observe that all three methods achieve comparable rewards. However, DPO becomes inefficient even for problems of moderate size and the running time of BCP grows significantly faster compared to ROSA.

To evaluate the performance of ROSA for $\gamma \to 1$ we solve 10^2 mazes[2] with side length 9 and 49 states for increasing discount factors. We report the average solution times and achieved reward in Figure 3.8. In the comparison of the rewards, examples where BCP did not converge are excluded. In these experiments we see that BCP becomes unstable, whereas the solution time of ROSA appears to be very robust and even decrease for $\gamma \to 1$. In fact, in the solution of (BCP) `Ipopt` fails to converge to local optimality for about 15% of all problems with discount factor at least 0.9999.

[1] For DPO we solved only 20 mazes of each size due to the long solution time.

[2] For DPO we consider only 30 mazes and 10^2 values of γ.

Reproducability statement. We provide a `Julia` [49] implementation of ROSA for deterministic observations. In general, problem (3.69) can be solved with any constrainted optimization solver. Our implementation is built on `Ipopt`, an interior point line search method [287]. We call `Ipopt` via the modeling language `JuMP` in which the constraints are easy to implement [108]. The implementation is available under `https://github.com/muellerjohannes/POMDPs-ROSA`.

3.4.2. Solution via numerical algebraic approaches. We provide an implementation that solves the reward optimization problem in state-action space by computing the critical points via the Karush-Kuhn-Tucker conditions. For this we reduce the combinatorial complexity of the problem by focusing on relevant boundary components, see Theorem 3.39. We implement this approach using numerical algebra methods [63] that automatically certify the correctness of the results [62]. We use a convex relaxation of the polynomial problem to certify the global optimality of the results. Moreover, we observe that in specific instances this numerical algebraic approach leads to superior results when compared to an interior point method for the solution of the reward optimization problem in state-action space. Finally, we compare the number of critical points obtained in numerical experiments with our theoretical bounds obtained in Subsection 3.3.3.

The KKT critical point equations. A standard approach for constrained optimization problems are the KKT conditions [162], which provide necessary conditions of stationary points under certain regularity conditions; see, e.g., [1, 48, 41]. If both the constraints and objective function are polynomial, the KKT conditions form a polynomial system, which can be solved using various numerical algebraic methods.

Applied to our problem, the KKT conditions reduce to the following polynomial system in $\eta \in \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}}$ with multipliers $\lambda \in \mathbb{R}^{\mathcal{S}}$, $\nu_{sa}^{o} \in \mathbb{R}$, $\kappa \in \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}}$:

$$\begin{aligned} &\text{Primal feasibility:} && \ell_s(\eta) = 0 \text{ for } s \in \mathcal{S}, \\ & && p_{sa}^{o}(\eta) = 0 \text{ for } o \in \mathcal{O}, a \in \mathcal{A} \setminus \{a_o\}, s \in S_o \setminus \{s_o\}, \\ &\text{Complementary slackness:} && \kappa_{s_o a}\eta_{s_o a} = 0 \text{ for all } s_o, a, \\ &\text{Stationarity:} && r + \sum_s \lambda_s \nabla \ell_s(\eta) + \sum_{o,s,a} \nu_{sa}^{o} \nabla p_{sa}^{o}(\eta) + \kappa = 0, \end{aligned} \tag{3.73}$$

where $a_o \in \mathcal{A}$ and $s_o \in S_o$ for every $o \in \mathcal{O}$ are fixed arbitrarily. Here we have included the primal feasibility $\eta_{sa} \geq 0$ for $s \in \mathcal{S}, a \in \mathcal{A}$ and the dual feasibility $\kappa_{sa} \geq 0$ for $s \in \mathcal{S}, a \in \mathcal{A}$ in the definition of the search space for η and κ.

The number of linear constraints ℓ_s is $n_{\mathcal{S}}$, while the number of polynomial constraints p_{sa}^{o} is $(n_{\mathcal{A}} - 1)\sum_{o\in\mathcal{O}}(d_o - 1) = (n_{\mathcal{A}} - 1)(n_{\mathcal{S}} - n_{\mathcal{O}})$. Due to the symmetry of the effective policies, there are only $n_{\mathcal{O}}n_{\mathcal{A}}$ inequalities, $\eta_{s_o a} \geq 0$ for each $a \in \mathcal{A}, o \in \mathcal{O}$. Hence the dimension of the square KKT system (3.73) is

$$n_{\mathcal{S}}n_{\mathcal{A}} + n_{\mathcal{S}} + (n_{\mathcal{A}} - 1)(n_{\mathcal{S}} - n_{\mathcal{O}}) + n_{\mathcal{O}}n_{\mathcal{A}} = 2n_{\mathcal{S}}n_{\mathcal{A}} + n_{\mathcal{O}}.$$

In this setting, we can verify that the linear independence constraint qualification is satisfied. Given an element η^* in the feasible set $\mathcal{N}^{\beta}$, it suffices to verify the linear independence of the gradients of the active inequality constraint functions and the equality constraint functions at η^*. Notice that under the pullback along the birational morphism

Ψ^{-1} the equality constraints in (3.35) are identified with the affine-linear functions l^o_{sa} defined in the proof of Theorem 3.25. Checking the linear independence of their gradients can be done by counting the dimension of the faces.

The Lagrange critical point equations over boundary components. Alternatively to solving the KKT system, one can compute the critical equations given by the Lagrange criterion over every boundary component individually. If there are no inequality constraints, the KKT equations specialize to the Lagrange multiplier equations. Consider a boundary component B in (3.71) for a choice of $A_o \subsetneq \mathcal{A}$ for every $o \in \mathcal{O}$, and consider the optimization problem over B. This amounts to setting $\eta(s,a) = 0$ for $a \in A_o$ whenever $g_\beta(s) = o, o \in \mathcal{O}$, which reduces optimization to a subspace of $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. We denote the new primal variables by $\hat{\eta}$. Similarly, we denote the restriction of ℓ_s and p^o_{sa} to this space by $\hat{\ell}_s$ and $\hat{p}^o_{sa}$ and the projection of r onto this space (i.e., the vector obtained by dropping the indices, which are set to zero in η) by $\hat{r}$. In the lower dimensional variables $\hat{\eta}$ for a given B the Lagrange system becomes

$$\begin{aligned} &\text{Feasibility:} && \hat{\ell}_s(\hat{\eta}) = 0 \text{ for } s \in \mathcal{S}, \\ &&& \hat{p}^o_{sa}(\hat{\eta}) = 0 \text{ for } o \in \mathcal{O}, a \in \mathcal{A} \setminus \{a_o\}, s \in S_o \setminus \{s_o\}, \\ &\text{Stationarity:} && \hat{r} + \sum_s \lambda_s \nabla \hat{\ell}_s(\hat{\eta}) + \sum_{o,s,a} \nu^o_{sa} \nabla \hat{p}^o_{sa}(\hat{\eta}) = 0, \end{aligned} \tag{3.74}$$

where $a_o \in A^c_o$ and $s_o \in S_o$ are fixed arbitrarily for every $o \in \mathcal{O}$. The dimension of the primal variable $\hat{\eta}$ is $n_\mathcal{S} n_\mathcal{A} - \sum_o d_o|A_o|$, the dimension of the Lagrange multipliers λ is $n_\mathcal{S}$ and of ν^o_{sa} is $\sum_o (d_o - 1)(|A^c_o| - 1)$, see also the proof of Theorem 3.41. Overall, the Lagrange system (3.74) is a square polynomial system of dimension

$$2n_\mathcal{S} n_\mathcal{A} - (n_\mathcal{A} - 1)n_\mathcal{O} - \sum_o (2d_o - 1)|A_o|.$$

Note that we have discussed the relation between the solutions of the Lagrange equations and the KKT system in Section 3.3.

Description of the experiments. We test our computational approach on random POMDPs of different sizes with deterministic observations. To this end, we first specify the number of states $n_\mathcal{S}$, the number of actions $n_\mathcal{A}$, and the number of states aggregated in each observation $(d_o)_{o\in\mathcal{O}}$ with $\sum_o d_o = n_\mathcal{S}$. For each specification of these values, we generate 20 random problems as follows. We sample the initial state distribution μ and the transition probabilities $\alpha(\cdot|s,a)$, $(s,a) \in \mathcal{S} \times \mathcal{A}$ from a symmetric Dirichlet distribution, and sample the instantaneous reward vector $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ from a standard Gaussian distribution. We use the same random data for both approaches, KKT and Lagrange over boundary components.

Computation. The optimization problem (3.69) can be solved using several methods:

- First, we use the numerical algebra package `HomotopyContinuation.jl` [63] to solve the KKT system (3.73) and the Lagrange system (3.74) of each boundary component. This automatically certifies the results [62], meaning that for every returned solution, a unique true solution is guaranteed in a small neighborhood. From the returned solutions to the critical equations, we then just need to select the real ones that satisfy the primal inequality constraints $\eta_{s,a} \geq 0$, and among them the one that has the maximum objective value.

- Alternatively, we solve a convex relaxation of the polynomial optimization problem (3.69). Namely, we relax the problem to a semidefinite program (SDP) via the moment-SOS-approach that is implemented in the freeware `GloptiPoly3` [133], and solve the SDPs using the numerical solver `Mosek`; see [82] for details. We note that `GloptiPoly3` builds upon a hierarchy of moment/SOS programs (also called Lasserre hierarchy), which allows to approximate the optimal value arbitrarily close, and can be used to test optimality and extract global optimizers [134, 220]. We use this key feature to check if our methods reach global optimality.
- We may also solve the optimization problem (3.69) using the interior point solver `Ipopt` [287], which is a local optimization method for large-scale nonlinear optimization, an approach recently pursued in [210], see also Subsection 3.4.1.

$n_{\mathcal{S}}$	$n_{\mathcal{A}}$	$(d_o)_{o\in\mathcal{O}}$	KKT			Lagrange (all)			Lagrange (relevant)		
			complex	real	positive	complex	real	positive	complex	real	positive
3	2	(3)	6±0	4.4±1.2	2.1±0.3	6±0	4.4±1.2	2.1±0.3	6±0	4.4±1.2	2.1±0.3
		(2,1)	12±0	10.1±1.9	4.25±0.44	10±0	8.2±1.9	4.25±0.44	8±0	6.7±1.6	4.25±0.44
		(1,1,1)	20±0	20±0	8±0	8±0	8±0	8±0	8±0	8±0	8±0
4	3	(4)	45±0	17.1±4.3	4.3±1.3	45±0	17.1±4.3	4.3±1.3	45±0	17.1±4.3	4.3±1.3
		(3,1)	150±0	79±11	11±1.9	129±0	68.7±9.7	11±1.9	81±0	41.6±8.5	10.9±1.8
		(2,2)	281.6±0.75	154±16	13.9±4.7	263±0	136±16	13.9±4.7	153±0	89±10	13.65±4.3
		(2,1,1)	381.2±0.7	292±23	31.5±4.3	216±0	168±16	31.5±4.3	81±0	68±11	30.9±4.0
		(1,1,1,1)	495±0	495±0	81±0	81±0	81±0	81±0	81±0	81±0	81±0
5	3	(5)	71±0	21.4±6	3.7±0.98	71±0	21.4±6	3.7±0.98	71±0	21.4±6	3.7±0.98
		(3,2)	637.95±0.76	219±28	12.60±2.9	626±0	213±29	12.6±2.9	477±0	171±24	12.6±2.9
		(4,1)	269.85±0.49	99±20	11.9±3.3	234±0	87±18	11.9±3.3	144±0	52±13	11.55±2.6
		(3,1,1)	881.95±0.22	436±68	36±10	558±0	285±47	36±10	243±0	117±20	35.3±9.2
		(2,2,1)	1717.3±2.5	890±49	35.6±5.3	1260±0	624±56	36.5±7.1	459±0	244±25	35.7±6.6
		(2,1,1,1)	2269.9±3.9	1712±142	89±12	810±0	624±74	89.3±12.3	243±0	195±37	88.1±9.5
		(1,1,1,1,1)	3002.9±0.31	3002.9±0.3	243±0	243±0	243±0	243±0	243±0	243±0	243±0

Table 3.3. Mean and standard deviation of the number of solutions of the KKT system (3.73), the Lagrange system (3.74) over all boundary components, and the Lagrange system over the relevant boundary components, for 20 random POMDPs with the indicated number of states $n_{\mathcal{S}}$, actions $n_{\mathcal{A}}$, and state-aggregation partition $(d_o)_{o\in\mathcal{O}}$. In our setting, positive solutions are feasible solutions.

Discussion of the results. In this section, we discuss the experimental results on the number of solutions obtained by solving the KKT and Lagrange systems introduced above. In Table 3.3, we report the average and standard deviation of the number of complex, real, and positive solutions returned in each case. Note that in our setting, positive solutions (i.e., solutions satisfying $\eta \geq 0$) are (primal) feasible solutions. We also compare these methods' performance and computational times with convex relaxations and interior point methods.

We start by comparing the number of solutions of the KKT and the Lagrange systems. In Table 3.3 we see that the KKT system has at least as many complex solutions as the Lagrange systems over all boundary components. This is consistent with our previous discussion, since, as we have pointed out, any solution of the Lagrange system over a boundary component is a solution of KKT. Moreover, KKT and Lagrange over all boundary components generally have the same number of positive, i.e., primal feasible, solutions, see (3.63) and Figure 3.4.

The difference between the number of complex, real, and positive solutions is also worth noting. Table 3.3 reveals a drop between the number of complex solutions and the number of real and positive solutions of the three types of systems. We find an exception to this in the Lagrange system for fully observable systems ($d_o = (1, \ldots, 1)$), where the number of complex, real, and positive solutions coincide. In this case all boundary components are affine spaces, so only the zero-dimensional boundary components have a solution, and these correspond to the $n_{\mathcal{A}}^{n_S}$ vertices of the feasible set.

We also observe that the number of complex solutions has a much smaller variance than the number of real or positive ones. This is expected, since choosing the coefficients of polynomial systems randomly gives the same number of complex solutions with probability one. In fact, the number of complex solutions for the Lagrange system has no variance across the different random parameters. Still, we see a small variance in the number of complex KKT solutions, which we attribute to numerical instability: this can prevent the software package `HomotopyContinuation.jl` from finding all solutions to the KKT system. In contrast to the complex case, the variance on the number of real and positive solutions is not due to numerical errors. This is a typical phenomenon in polynomial system solving and is one of the possible limitations of classic algebraic methods when one wants to estimate the number of real solutions of a system.

In the following we compare the experimental results presented in Table 3.3 with the theoretical upper bounds shown in Table 3.2 and highlight two particular facts. First notice that in most cases the theoretical bound is significantly larger than the number of solutions of the Lagrange system. Moreover, this gap becomes particularly pronounced for problems where the fibers of g_β are large. This indicates that there is a discrepancy between the theoretical bounds and the algebraic degree of the optimization problem. Indeed, our bounds are based on the theory for generic polynomials. Hence, we do not expect that they provide a tight estimate of the algebraic degree for the particular polynomials we are dealing with. Here we also observe a particular behavior in the case of fully observable systems where the number of critical points of the Lagrange systems agrees with our bounds. On the other hand, we see that in some cases the number of solutions of KKT is larger than the bound, which agrees with our discussion on solutions of KKT and Lagrange systems in (3.63).

In addition to analyzing the number of solutions of the KKT and Lagrange systems, we are interested in comparing the different solution methods for the optimization problem. Therefore, we compare the optimal solution found by solving these systems with `HomotopyContinuation.jl` with the one found by `Ipopt` and `GloptiPoly3`. Although `HomotopyContinuation.jl` is not guaranteed to find all solutions to the KKT and Lagrange systems, we observe that this approach yields a reward that is at least as high as the one obtained by the interior point method `Ipopt` and, in a few instances, strictly higher.

In fact, solving the optimization problem with `GloptiPoly3` returns a certificate for the optimality of the result, which in all computed instances coincides with the optimal value obtained by solving the KKT and Lagrange systems with `HomotopyContinuation.jl`. That is, `GloptiPoly3` offers numerical evidence that they always provide globally optimal solutions. In all computed instances using `GloptiPoly3`, the optimal value of the optimization problem was already attained at the first-order relaxation of the Lasserre hierarchy [148]. We conjecture that objective value exactness for the first order relaxation of (3.69) holds with high probability for generic input data. Since the size of the SDP depends very sensitively on the order of the relaxation, this conjecture would remedy one of the major drawbacks of the SDP relaxation method. In more detail, the t-th order relaxation for both, the moment and the SOS relaxation of a polynomial optimization problem, can be computed via an SDP of size $\binom{n+t}{t}$, where n is the number of variables of the involved polynomials.

As described in [220], finite convergence of the Lasserre hierarchy, i.e., convergence after finitely many relaxation steps, is closely related to certifying the flat truncation property. In fact, finite convergence holds generically [219]. However, studying exactness properties of the SOS and the moment relaxation is still an ongoing topic of current research, see e.g. [32].

Finally, in Table 3.4 we report the computation times of the different approaches. The KKT and Lagrange systems as well as Ipopt were computed on a server with a 2x 32-Core AMD Epyc 7601 at 2.2 GHz and 1024 GB RAM, whereas the SDP relaxation was computed on a Intel(R) Core(TM) i7-8550U CPU with 4 cores at 1.8 GHz and 16GB RAM. Solving the Lagrange equations only over the relevant boundary components was up to two orders of magnitude faster than solving them over all boundary components. The improvements are more pronounced when g_β has small fibers in which we can exclude more faces by means of Theorem 3.39; see also Table 3.2. The computation times for the solution of the KKT system are in the same order to magnitude as the computation time of the solution of the Lagrange systems over all boundary components. KKT is slightly faster when g_β has small fibers and slightly slower when g_β has large fibers. The SDP approach is several orders of magnitude faster compared to the solution of the KKT and Lagrange systems with the gap becoming more pronounced for problems of increasing size. The interior point method Ipopt is again several orders of magnitude faster. Ipopt and SDP return one candidate solution, whereas homotopy continuation attempts to return all critical points. Note however that in contrast to the SDP relaxation the interior point method only guarantees locally optimal solutions. In our experiments we consistently observed that Ipopt yields less accurate solutions and sometimes converges to suboptimal points. The maximum difference of the reward obtained by Ipopt and SDP is 9.7×10^{-2}, where the maximum difference between either of KKT and Lagrange methods and SDP is 3.0×10^{-7}.

Reproducibility statement. The computer code for our experiments is publicly available at `https://github.com/marinagarrote/Algebraic-Optimization-of-Sequential-Decision-Rules`. We conducted our experiments using `Julia` [49] version 1.7.0, an open source programming language under the MIT license. We used the Julia package `HomotopyContinuation.jl` version 2.6.3, which is freely available for personal use under

	Partitions of $n_{\mathcal{S}}$	Ipopt	SDP	KKT	Lagrange (all)	Lagrange (relevant)
$n_{\mathcal{S}} = 3, n_{\mathcal{A}} = 2$	(3)	0.01	0.213	1.575	0.046	1.175
	(2,1)	0.009	0.168	1.551	3.563	2.757
	(1,1,1)	0.006	0.171	0.114	0.119	0.03
$n_{\mathcal{S}} = 4, n_{\mathcal{A}} = 3$	(4)	0.011	1.167	19.885	7.642	10.407
	(3,1)	0.01	1.114	76.071	43.759	22.17
	(2,2)	0.011	1.278	173.644	114.208	48.52
	(2,1,1)	0.009	1.292	79.775	191.394	27.004
	(1,1,1,1)	0.007	1.184	13.82	32.637	0.693
$n_{\mathcal{S}} = 5, n_{\mathcal{A}} = 3$	(5)	0.011	7.394	62.321	31.257	31.501
	(3,2)	0.01	6.338	1768.722	509.877	259.054
	(4,1)	0.011	7.256	307.524	163.88	69.5
	(3,1,1)	0.01	6.608	895.701	704.813	91.901
	(2,2,1)	0.011	6.078	2831.482	2175.098	313.557
	(2,1,1,1)	0.009	6.22	899.981	2058.912	188.536
	(1,1,1,1,1)	0.006	5.159	172.621	319.165	3.667

TABLE 3.4. Average run times for the different approaches reported in seconds. KKT and Lagrange are computed with homotopy continuation.

the MIT license, and `Ipopt.jl` version 0.7.0, which is a Julia interface to the `COIN-OR` nonlinear solver `Ipopt` available under the EPL (Eclipse Public License) open-source license. The convex relaxation is computed via the freeware `GloptiPoly3` implemented in Matlab, for which there also exists an Octave implementation.

3.4.3. CONCLUSION AND OUTLOOK. In this section we have presented potential benefits of reward optimization in state-action space (ROSA) over methods working directly with policies or parametrizations of policies. First, we have seen that solving the resulting polynomially constrained linear objective problem with the interior point method Ipopt remains effective for $\gamma \to 1$, which is in contrast to existing approaches.

Further, we employed numerical algebraic approaches to solve the polynomial optimization problem. For this we considered KKT equations and the Lagrange system over individual boundary components, where we leveraged knowledge about the location of maximizers on lower dimensional boundary components. The relatively small number of solutions observed in the experiments indicate that there is room for refining the theory either to obtain tighter estimates of the algebraic degree or also tighter descriptions of the possible number of feasible solutions. Using a convex relaxation to an SDP we obtained empirical evidence that our approach of solving the critical equations provides global maximizers of the reward. This is in strong contrast to naive gradient optimization, which yields only locally optimal solutions for this problem. In our experiments, the first order relaxation produced exact objective values.

We highlight the following questions that arose during our analysis:

- *Exactness of SDP relaxations:* In our experiments, we observed that the sequence of convex relaxations given by the moment-SOS hierarchy provided exact global solutions in the first order relaxation. We believe that this observation deserves a closer theoretical analysis.

- *Riemannian optimization for POMDPs:* Our description of the state-action frequencies of POMDPs with deterministic observations via products of varieties of rank one matrices in sTheorem 3.25 could provide a starting point for a Riemannian optimization technique for POMDPs.
- *Reinforcement learning:* In this thesis we study the planning problem in MDPs, i.e., assume knowledge of the Markov decision process and it is a central question to understand how reinforcement learning can benefit our insights.

CHAPTER 4

Geometry and convergence of natural policy gradient methods

So far we have studied the geometry of partial observability and have seen that partial observability induces polynomial constraints on the feasible state-action frequencies of a Markov decision process. In this chapter, we study the geometry of several natural policy gradient (NPG) methods in infinite-horizon discounted fully observable Markov decision processes with regular policy parametrizations. We model the policy π_θ as a differentiably parametrized element in the polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$ of conditional probability distributions of actions given states, with $\pi_\theta(a|s)$ specifying the probability of selecting action $a \in \mathcal{A}$ when currently in state $s \in \mathcal{S}$, for the parameter value θ. It is the goal to maximizer the reward function $R(\theta) = R(\pi_\theta)$ and we study gradient-based policy optimization methods and more specifically natural policy gradient (NPG) methods. Inspired by the seminal works of Amari [13, 16], various NPG methods have been proposed [153, 206, 208]. In general, they take the form

$$\theta_{k+1} = \theta_k + \Delta t \cdot G(\theta_k)^+ \nabla R(\theta_k),$$

where $\Delta t > 0$ denotes the step size, $G(\theta)^+$ denotes the Moore-Penrose inverse and $G(\theta)_{ij} = g(dP_\theta e_i, dP_\theta e_j)$ is a Gram matrix defined with respect to some Riemannian metric g and some representation $P(\theta)$ of the parameter. Most of our analysis does not actually depend on the specific choice of the pseudoinverse, but in Section 4.4 we will use the Moore-Penrose inverse. The most traditional natural gradient method is the special case where $P(\theta)$ is a probability distribution and g is the Fisher information in the corresponding space of probability distributions. However, the terminology may be used more generally to refer to a Riemannian gradient method where the metric is in some sense natural. Kakade [153] proposed using $P(\theta) = \pi_\theta$ and taking for g a product of Fisher metrics weighted by the state frequencies resulting from running the Markov process with policy π_θ. Although this is a natural choice for P, the choice of a Riemannian metric on $\Delta_{\mathcal{A}}^{\mathcal{S}}$ is a non trivial problem. Peters et al. as well as Bagnell and Schneider [232, 28] offered an interpretation of Kakade's metric as the limit of Fisher metrics defined on the finite horizon path measures, but other choices of the weights can be motivated by axiomatic approaches to defining a Fisher metric of conditional probabilities [169, 205]. From our perspective, a main difficulty is that it is not clear how to choose a Riemannian metric on $\Delta_{\mathcal{A}}^{\mathcal{S}}$ that interacts nicely with the objective function $R(\pi)$, which is a non-convex rational function of $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$. An alternative choice for $P(\theta)$ is the vector of state-action frequencies η_θ, whose components $\eta_\theta(s, a)$ are the probabilities of state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$ resulting from running the Markov process with policy π_θ. Morimura et al. [206] proposed using $P(\theta) = \eta_\theta$ and the Fisher information on the state-action probability simplex $\Delta_{\mathcal{S}\times\mathcal{A}}$ as a Riemannian metric. We will study both approaches and variants from the perspective of Hessian geometry.

Contributions. We study the natural policy gradient dynamics inside the polytope $\mathcal{N}$ of state-action frequencies, which provides a unified treatment of several existing NPG methods. We focus on finite state and action spaces and the expected infinite-horizon discounted reward optimized over the set of memoryless stochastic policies. For an overview of the convergence rates established in this work see Table 4.1 in Section 4.5 where our main contributions can be summarized as follows:

- We show that the dynamics of Kakade's NPG and Morimura's NPG solve a gradient flow in $\mathcal{N}$ with respect to the Hessian geometries of conditional entropic and entropic regularization of the reward (Sections 4.2.2 and 4.2.3 and Proposition 4.13).
- Leveraging results on gradient flows in Hessian geometries, we derive linear convergence rates for Kakade's and Morimura's NPG flow for the unregularized reward, which is a linear and hence not strictly concave function in state-action space, and also for the regularized reward (Theorems 4.26 and 4.27 and Corollaries 4.33 and 4.34).
- Further, for a class of NPG methods, which correspond to β-divergences and which generalize Morimura's NPG, we show sub-linear convergence in the unregularized case and linear convergence in the regularized case (Theorem 4.27 and Corollary 4.34, respectively).
- We complement our theoretical analysis with experimental evaluation, which indicates that the established linear and sub-linear rates for unregularized problems are essentially tight.
- For discrete-time gradient optimization, our ansatz in state-action space yields an interpretation of the regularized NPG method as an inexact Newton iteration if the step size is equal to the inverse regularization strength. This yields a relatively short proof for the local quadratic convergence of regularized NPG methods with Newton step sizes (Theorem 4.36). This recovers as a special case the local quadratic convergence of Kakade's NPG under state-wise entropy regularization previously obtained in [71].

Related work. The application of natural gradients to optimization in MDPs was first proposed by Kakade [153], taking as a metric on $\Delta_{\mathcal{A}}^{\mathcal{S}} = \prod_{s\in\mathcal{S}} \Delta_{\mathcal{A}}$ the product of Fisher metrics on the individual components $\Delta_{\mathcal{A}}^{s} \cong \Delta_{\mathcal{A}}$, $s \in \mathcal{S}$, weighted by the stationary state distribution. The relation of this metric to finite-horizon Fisher information matrices was studied by Bagnell and Schneider [28] as well as by Peters et al. [232]. Later, Morimura et al. [206] proposed a natural gradient using the Fisher metric on the state-action frequencies, which are probability distributions over states and actions.

There has been a growing number of works studying the convergence properties of policy gradient methods. It is well known that reward optimization in MDPs is a challenging problem, where both the non-convexity of the objective function with respect to the policy and the particular parametrization of the policies can lead to the existence of suboptimal critical points [50]. Global convergence guarantees of gradient methods require assumptions on the parametrization. Most of the existing results are formulated for tabular softmax policies, but more general sufficient criteria have been given in [50, 316, 317].

Vanilla PGs have been shown to converge sublinearly at rate $O(t^{-1})$ for the unregularized reward and linearly for entropically regularized reward. For unregularized problems, the convergence rate can be improved to a linear rate by normalization [200, 199]. For continuous state and action spaces, vanilla PG converges linearly for entropic regularization and shallow policy networks in the mean-field regime [168].

For Kakade's NPG, [2] established sublinear convergence rate $O(t^{-1})$ for unregularized problems, and the result has been improved to a linear rate of convergence for step sizes found by exact line search [51], constant step sizes [156, 6, 311], and for geometrically increasing step sizes [305, 7]. For regularized problems, the method converges linearly for small step sizes and locally quadratically for Newton-like step size [71, 172]. These results have been extended to more general frameworks using state-mixtures of Bregman divergences on the policy polytope [165, 315, 172, 7], which however do not include NPG methods defined in state-action space such as Morimura's NPG. For projected PGs, [2] shows sublinear convergence at a rate $O(t^{-1/2})$, and the result has been improved to a sublinear rate $O(t^{-1})$ [305], and to a linear rate for step sizes chosen by exact line search [51]. Apart from the works on convergence rates for policy gradient methods for standard MDPs, a primal-dual NPG method with sublinear global convergence guarantees has been proposed for constrained MDPs [98, 97]. For partially observable systems, a gradient domination property has been established in [26]. NPG methods with dimension-free global convergence guarantees have been studied for multi-agent MDPs and potential games [5]. The sample complexity of a Bregman policy gradient arising from a strongly convex function in parameter space has been studied in [143]. For the linear quadratic regulator, global linear convergence guarantees for vanilla, Gauss-Newton and Kakade's natural policy gradient methods are provided in [113]; this setting is different to reward optimization in MDPs, where the objective at a fixed time is linear and not quadratic. A lower bound of $O(\Delta t^{-1}|\mathcal{S}|^{2^{\Omega((1-\gamma)^{-1})}})$ on the iteration complexity for softmax PG method with step size Δt has been established in [171].

4.1 Natural gradients

In this section we provide some background on the concept of natural gradients.

4.1.1. Definition and general properties of natural gradients. In many applications, one aims to optimize a model parameter θ with respect to an objective function ℓ that is defined on a model space $\mathcal{M}$, as illustrated in Figure 4.1. This general setup, with an objective function that factorizes as $L(\theta) = \ell(P(\theta))$, covers parameter estimation and supervised learning cases, and also problems such as the numerical solution of PDEs with neural networks or policy optimization in MDPs and reinforcement learning. Naively, the optimization problem can be approached with first order methods, computing the gradients in parameter space with respect to the Euclidean geometry. However, this neglects the geometry of the parametrized model $\mathcal{M}_\Theta = P(\Theta)$, which is often seen as a disadvantage since it may lead to parametrization-dependent plateaus in the optimization landscape. At the same time, the biases that particular parametrizations can introduce into the optimization can be favorable in some cases. This is an active topic of investigation particularly in deep learning, where P is often a highly non-linear function

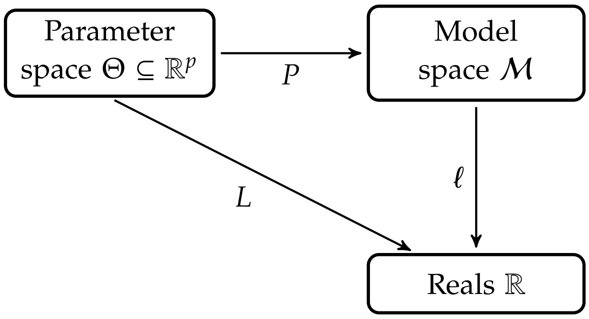
Parameter space $\Theta \subseteq \mathbb{R}^p$
P
Model space $\mathcal{M}$
L
ℓ
Reals $\mathbb{R}$

Figure 4.1. Schematic drawing of parametric models with an objective function ℓ and resulting parameter objective function L; note that neither the choice of geometry in the model space nor the factorization or the model space itself is uniquely determined by the objective function L.

of θ. At any rate, there is a good motivation to study of the effects of the parametrization and the possible advantages from incorporating the geometry of model space into the optimization procedure in parameter space.

The natural gradient as introduced in [13] is a way to incorporate the geometry of the model space into the optimization procedure and to formulate iterative update directions that are invariant under reparametrizations. Although it has been most commonly applied in the context of parameter estimation under the maximum likelihood criterion, the concept of natural gradient has been formulated for general parametric optimization problems and in combination with arbitrary geometries. In particular, natural gradients have been applied to neural network training [229, 193, 95, 146], policy optimization [153, 232, 206] and inverse problems [223]. Especially in the latter case, different notions of natural gradients have been introduced. A version that incorporates the geometry of the sample space are natural gradients based on an optimal transport geometry in model space [175, 188, 18]. We shall discuss natural gradients in a way that emphasizes that even for a specific problem there may not be a unique natural gradient. This is because both the factorization $L(\theta) = \ell(P(\theta))$ of the objective as well as what should be considered a natural geometry in model space may not be unique.

But what is it that makes a gradient or update direction *natural*? The general consensus is that it should be invariant under reparametrization to prevent artificial plateaus and provide consistent stopping criteria, and it should (approximately) correspond to a gradient update with respect to the geometry in the model space. We now give the formal definition of the natural gradient with respect to a given factorization and a geometry in model space that we adopt in this work, which can be shown to satisfy the desired properties.

Definition 4.1 (General natural gradient). Consider the problem of optimizing an objective $L\colon \Theta \to \mathbb{R}$, where the parameter space $\Theta \subseteq \mathbb{R}^p$ is an open subset. Further, assume that the objective factorizes as $L = \ell \circ P$, where $P\colon \Theta \to \mathcal{M}$ is a *model parametrization* with $\mathcal{M}$ a Riemannian manifold with Riemannian metric g, and $\ell\colon \mathcal{M} \to \mathbb{R}$ is a *loss in model space*, as shown in Figure 4.1. For $\theta \in \Theta$ we define the Gram matrix $G(\theta)_{ij} := g_{P(\theta)}(dP_\theta e_i, dP_\theta e_j)$ and call $\nabla^N L(\theta) := G(\theta)^+ \nabla L(\theta)$ the *natural gradient (NG) of L at θ with respect to the factorization $L = \ell \circ P$ and the metric g*.

Natural gradient as best improvement direction. Let us consider a parametrization $P\colon \Theta \to \mathcal{M}$ with image $\mathcal{M}_\Theta = P(\Theta)$, where $\mathcal{M}$ is a Riemannian manifold with metric g. Let us fix a parameter $\theta \in \Theta$ and set $p := P(\theta)$. Moving into the direction $v \in \mathbb{R}^p$ in the parameter space results in moving in the direction $w = dP_\theta v \in T_p\mathcal{M}$ in model space. The space of all directions that can result in this way is the *generalized tangent space* $T_\theta\mathcal{M}_\Theta := \operatorname{range}(d_\theta P) \subseteq T_p\mathcal{M}$. Hence, the best direction one can take on $\mathcal{M}_\Theta$ by infinitesimally varying the parameter θ is given by

$$\underset{w\in T_\theta\mathcal{M}_\Theta, g_p(w,w)=1}{\arg\max} \quad \partial_w \ell(p),$$

which is equal (up to normalization) to the projection $\Pi_{T_\theta\mathcal{M}_\Theta}(\nabla^g \ell(p))$ of the Riemannian gradient $\nabla^g \ell(p)$ onto $T_\theta\mathcal{M}_\Theta$. Moving in the direction of the natural gradient in parameter space results in the optimal update direction over the generalized tangent space $T_\theta\mathcal{M}_\Theta$ in model space.

Theorem 4.2 (Natural gradient leads to steepest descent in model space). *Consider the settings of Definition 4.1, where $\mathcal{M}$ is a Riemannian manifold with metric g and enote the natural gradient with respect to this factorization by $\nabla^N L(\theta) = G(\theta)^+\nabla_\theta L(\theta)$. Then it holds that*

$$dP_\theta(\nabla^N L(\theta)) = \Pi_{T_\theta\mathcal{M}_\Theta}(\nabla^g \ell(P(\theta))).$$

For invertible Gram matrices $G(\theta)$ this result is well known [14, Subsection 12.1.2]; for singular Gram matrices we refer to [226, Theorem 1] and provide a proof for Hilbert spaces in Chapter 6.

4.1.2. Choice of a geometry in model space.

Neither the factorization of the objective function nor the choice of the geometry in the model space is unique. Here, we discuss different approaches for the choice of the geometry in the model space.

Invariance axiomatic geometries. A celebrated theorem by Chentsov [72] characterizes the Fisher metric of statistical manifolds with finite sample spaces as the unique metric (up to multiplicative constants) that is invariant with respect to congruent embeddings by Markov mappings. A generalization of Chentsov's result for arbitrary sample spaces was given by Ay et al. [24].

Given two Riemannian manifolds $(\mathcal{E}, g)$, $(\mathcal{E}', g')$ and an embedding $f\colon \mathcal{E} \to \mathcal{E}'$, the metric is said to be invariant if f is an isometry, meaning that

$$g_p(u,v) = (f^*g')_p(u,v) := g'_{f(p)}(f_*u, f_*v), \quad \text{for all } p \in \mathcal{E} \text{ and } u,v \in T_p\mathcal{E},$$

where $f_*\colon T_p\mathcal{E} \to T_{f(p)}\mathcal{E}'$ is the pushforward of f. A congruent Markov mapping is in simple terms a linear map $p \mapsto M^T p$, where M is a row stochastic partition matrix, i.e., a matrix of non-negative entries with a single non-zero entry per column and entries of each row adding to one. Such a mapping has the natural interpretation of splitting each elementary event into several possible outcomes with fixed conditional probabilities. By Chentsov's theorem, requiring invariance with respect to these mappings results in a single possible choice for the metric (up to multiplicative constants). We recall that on the interior of the probability simplex $\Delta_\mathcal{S}$ the Fisher metric is given by

$$g_p(u,v) = \sum_{s\in\mathcal{S}} \frac{u_s v_s}{p_s}, \quad \text{for all } u,v \in T_p\Delta_\mathcal{S}.$$

Based on this approach, Campbell [67] characterized the set of invariant metrics on the set of non-normalized positive measures with respect to congruent embeddings by Markov mappings. This results in a family of metrics, which restrict to the Fisher metric (up to a multiplicative constant) over the probability simplex. Following this line of ideas, Lebanon [169] characterized a class of invariant metrics of positive matrices that restrict to products of Fisher metrics over stochastic matrices.[1] The maps considered by Lebanon do not map stochastic matrices to stochastic matrices, which motivated [205] to investigate a natural class of mappings between conditional probabilities. They showed that requiring invariance with respect to their proposed mappings singles out a family of metrics that correspond to products of Fisher metrics on the interior of the conditional probability polytope,

$$g_\pi(u, v) = \sum_{s\in\mathcal{S}} \frac{1}{|\mathcal{S}|} \sum_{a\in\mathcal{A}} \frac{u_{sa}v_{sa}}{\pi_{sa}}, \quad \text{for all } u, v \in T_\pi\Delta_{\mathcal{A}}^{\mathcal{S}},$$

up to multiplicative constants. This work also offered a discussion of metrics on general polytopes and weighted products of Fisher metrics, which correspond to the Fisher metric when the conditional probability polytope is embedded in the joint probability simplex by way of providing a marginal distribution.

Hessian geometries. Instead of characterizing the geometry of model space $\mathcal{M}$ via an invariance axiomatic, one can select a metric based on the optimization problem at hand. For example, it is well known that the Fisher metric is the local Riemannian metric induced by the Hessian of the KL-divergence in the probability simplex. Hence, if the objective function is a KL-divergence, choosing the Fisher metric yields preconditioners that recover the inverse of the Hessian at the optimum, which can yield locally quadratic convergence rates. More generally, if the objective $\ell\colon \mathcal{M} \to \mathbb{R}$ has a positive definite Hessian at every point, it induces a Riemannian metric via

$$g_p(v, w) = v^\top \nabla^2 \ell(p) w$$

in local coordinates, which we call the *Hessian geometry* induced by ℓ on $\mathcal{M}$; see [15, 259].

Example 4.3 (Hessian geometries). The following Riemannian geometries are induced by strictly convex functions.

(*i*) *Euclidean geometry:* The Euclidean geometry on $\mathbb{R}^d$ is induced by the squared Euclidean norm $x \mapsto \frac{1}{2}\sum_i x_i^2$.

(*ii*) *Fisher geometry:* The Fisher metric on $\mathbb{R}^d_{>0}$ is induced by the negative entropy $x \mapsto \sum_i x_i \log(x_i)$.

(*iii*) *Itakura-Saito:* The logarithmic barrier function $x \mapsto -\sum_i \log(x_i)$ of the positive cone $\mathbb{R}^d_{>0}$ yields the Itakura-Saito metric (see the next item).

(*iv*) *σ-geometries:* All of the above examples can be interpreted as special cases of a parametric family of Hessian metrics. More precisely, if we let

$$\phi_\sigma(x) := \begin{cases} \sum_i x_i \log(x_i) & \text{if } \sigma = 1 \\ -\sum_i \log(x_i) & \text{if } \sigma = 2 \\ \frac{1}{(2-\sigma)(1-\sigma)} \sum x_i^{2-\sigma} & \text{otherwise,} \end{cases} \tag{4.1}$$

[1] For Riemannian manifolds $(\mathcal{M}_1, g_1)$ and $(\mathcal{M}_2, g_2)$, the product metric on $\mathcal{M}_1 \times \mathcal{M}_2$ is defined by $g(u_1 + u_2, v_1 + v_2) = g_1(u_1, v_1) + g_2(u_2, v_2)$.

then the resulting Riemannian metric on $\mathbb{R}^d$ for $\sigma \in (-\infty, 0]$ and on $\mathbb{R}^d_{>0}$ for $\sigma \in (0, \infty)$ is given by

$$g^\sigma_x(v, w) = \sum_i \frac{v_i w_i}{x_i^\sigma}. \tag{4.2}$$

This recovers the Euclidean geometry for $\sigma = 0$, the Fisher metric for $\sigma = 1$, and the Itakura-Saito metric for $\sigma = 2$. Note that these geometries are closely related to the so-called β-divergences [15], which are the Bregman divergences of the functions ϕ_σ for $\beta = 1 - \sigma$. We use σ instead of β in order to avoid confusion with our notation for the observation kernel β in a POMDP.

(v) *Conditional entropy:* Given two finite sets $\mathcal{X}, \mathcal{Y}$ and a probability distribution μ in $\Delta_{\mathcal{X}\times\mathcal{Y}}$ we can consider the conditional entropy

$$\phi_C(\mu) = H(\mu|\mu_X) := -\sum_{x,y} \mu(x, y) \log \frac{\mu(x, y)}{\mu_X(x)} = H(\mu) - H(\mu_X). \tag{4.3}$$

This is a convex function on the simplex $\Delta_{\mathcal{X}\times\mathcal{Y}}$ [218]. The Hessian of the conditional entropy is given by

$$\partial_{(x,y)}\partial_{(x',y')}\phi_C(\mu) = \delta_{xx'}\left(\delta_{yy'}\mu(x, y)^{-1} - \mu_X(x)^{-1}\right), \tag{4.4}$$

as can be verified by explicit computation or the chain rule for Hessian matrices (see also proof of Theorem 4.8). This Hessian does not induce a Riemannian geometry on the entire simplex since it is not positive definite on the tangent space $T\Delta_{\mathcal{X}\times\mathcal{Y}}$, as can be seen from the specific choice $\mathcal{X} = \mathcal{Y} = \{1, 2\}$, $\mu_{ij} = 1/4$ for all $i, j = 1, 2$ and the tangent vector $v \in T_\mu\Delta_{\mathcal{X}\times\mathcal{Y}}$ given by $v_{ij} = (-1)^i$. However, when fixing a marginal distribution $\nu \in \Delta_{\mathcal{X}}, \nu > 0$, then the conditional entropy ϕ_C induces a Riemannian metric on the interior of $P = \{\mu \in \Delta_{\mathcal{X}\times\mathcal{Y}} : \mu_X = \nu\}$. To see this we consider the diffeomorphism given by conditioning $\operatorname{int}(P) \to \operatorname{int}(\Delta^{\mathcal{X}}_{\mathcal{Y}}), \mu \mapsto \mu_{Y|X}$. It can be shown by explicit computation (analogous to the proof of Theorem 4.8) that the Hessian $\nabla^2\phi_C(\mu)$ is the metric tensor of the pull back of the Riemmanian metric

$$g\colon T\Delta^{\mathcal{X}}_{\mathcal{Y}} \times T\Delta^{\mathcal{X}}_{\mathcal{Y}} \to \mathbb{R}, \quad g_{\mu(\cdot|\cdot)}(v, w) := \sum_x \nu(x) \sum_y \frac{v(x, y)w(x, y)}{\mu(y|x)}.$$

This argument can be adapted to sets $\{\mu \in \Delta_{\mathcal{X}\times\mathcal{X}} : \mu_X = \nu(\mu_{Y|X})\}$, where $\nu\colon \operatorname{int}(\Delta^{\mathcal{X}}_{\mathcal{Y}}) \to \operatorname{int}(\Delta_{\mathcal{X}})$ depends smoothly on the conditional $\mu_{Y|X} \in \Delta^{\mathcal{X}}_{\mathcal{Y}}$.

We note that the Bregman divergence induced by the conditional entropy is the conditional relative entropy [218],

$$\begin{aligned} D_{\phi_C}(\mu^{(1)}, \mu^{(2)}) &= D_{KL}(\mu^{(1)}, \mu^{(2)}) - D_{KL}(\mu^{(1)}_X, \mu^{(2)}_X) \\ &= \sum_x \mu^{(1)}_X(x) D_{KL}(\mu^{(1)}(\cdot|x), \mu^{(2)}(\cdot|x)). \end{aligned}$$

Local Hessian of Bregman divergences. Let ϕ be a twice differentiable convex function and denote its Bregman divergence with $D_\phi(x, y) = \phi(x) - \phi(y) - \langle\nabla\phi(y), x - y\rangle$. Then it holds that

$$\nabla^2_y D_\phi(x, y)|_{y=x} = \nabla^2_y D_\phi(y, x)|_{y=x} = \nabla^2\phi(x). \tag{4.5}$$

To see this, we set $f(y) \coloneqq D_\phi(x, y)$. It is immediate that $\nabla^2 f(y) = \nabla^2\phi(y)$. Further, one can compute

$$\begin{aligned} \partial_{y_j} f(y) &= \partial_{y_j}\left(\phi(x) - \phi(y) - \sum_k \partial_{y_k}\phi(y)(x_k - y_k)\right) \\ &= -\partial_{y_j}\phi(y) + \sum_k \partial_{y_j}\partial_{y_k}\phi(y)(y_k - x_k) + \partial_{y_j}\phi(y). \end{aligned}$$

Hence, we obtain

$$\partial_{y_i}\partial_{y_j} f(y) = -\partial_{y_i}\partial_{y_j}\phi(y) + \sum_k \partial_{y_i}\partial_{y_j}\partial_{y_k}\phi(y)(y_k - x_k) + \partial_{y_i}\partial_{y_j}\phi(y) + \partial_{y_i}\partial_{y_j}\phi(y),$$

and hence $\nabla^2 f(x) = \nabla^2\phi(x)$.

Connection to generalized Gauss-Newton methods. Let ϕ be a twice differentiable strictly convex function. Then the Gram matrix of the Hessian geometry is given by

$$G(\theta) = DP(\theta)^\top \nabla^2\phi(P(\theta))DP(\theta).$$

Hence $G^{-1}(\theta)$ can be interpreted as a generalized Gauss-Newton matrix of the objective function $\phi \circ P$ [192]. In particular, for the square loss we have $\phi(x) = \|x\|_2^2$, in which case $G(\theta)^{-1} = (DP(\theta)^\top DP(\theta))^{-1}$ is the usual nonlinear least squares Gauss-Newton matrix. Note that this is only the case when choosing the Hessian geometry of the objective function. Later, in the case of Markov decision processes this is the case when applying (the right) natural policy gradient to a regularized reward optimization problem. In contrast, the unregularized case, the objective in state-action space is linear and does not induce a Hessian geometry and hence the natural policy gradients do not agree with a generalized Gauss-Newton method.

4.2 Natural policy gradient methods

In this section we give a brief overview of different notions of policy gradient methods that have been proposed in the literature and study their associated geometries in state-action space. Policy gradient methods [300, 159, 277, 190, 40] offer a flexible approach to reward optimization. They have been used in robotics [232] and have been combined with deep neural networks [265, 266, 255]. In the context of MDPs there are multiple notions of natural policy gradients. For instance, one may choose to use an optimal transport geometry in model space resulting in Wasserstein natural policy gradients [208]. Most important to our discussion, there are different possible choices for the model space. One obvious candidate is the policy space $\Delta_{\mathcal{A}}^{\mathcal{S}}$, which was used by Kakade [153]. However the objective function $R(\pi)$ is a rational non-convex function over this space an thus requires a delicate analysis. A second candidate, which was proposed by Morimura et al. [206], is the state-action space $\mathcal{N} \subseteq \Delta_{\mathcal{S}\times\mathcal{A}}$, for which the objective becomes a rather simple, linear function. We recall the following result, which allow us to study any NPG method defined with respect to the policy space in state-action space.

Assumption 3.3 (Positivity). For every $s \in \mathcal{S}$ and $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$, we assume that $\sum_a \eta_{sa}^\pi > 0$.

Proposition 3.4 (Inverse of state-action map). *Under Assumption 3.3, the mapping*

$$\Psi\colon \Delta_{\mathcal{A}}^{\mathcal{S}} \to \mathcal{N}, \quad \pi \mapsto \eta^\pi$$

is rational and bijective with rational inverse given by conditioning

$$\Psi^{-1}\colon \mathcal{N} \to \Delta_{\mathcal{A}}^{\mathcal{S}}, \quad \eta \mapsto \pi, \quad \text{where } \pi(a|s) = \frac{\eta(s,a)}{\sum_{a'} \eta(s,a')}.$$

Because of the simplicity of the objective function in state-action space, we propose to study the evolution of NPG methods in this space. As we will see, this has the added benefit that it allows us to interpret several of the existing NPG methods as being induced by Hessian geometries. Based on this observation we can conduct a relatively simple convergence analysis for these methods. Finally, we propose a class of policy gradients closely related to β-divergences that interpolate between NPG arising from logarithmic barriers, entropic regularization and the Euclidean geometry.

4.2.1. Policy gradients. Throughout the section, we consider parametric policy models $P\colon \Theta \to \Delta_{\mathcal{A}}^{\mathcal{S}}$ and write $\pi_\theta = P(\theta) \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ for the policy arising from the parameter θ. We denote the corresponding state-action and state frequencies by η_θ and ρ_θ. Finally, in slight abuse of notation we write $R(\theta)$ for the expected infinite-horizon discounted reward obtained by the policy π_θ. The *vanilla policy gradient (vanilla PG)* method is given by the iteration

$$\theta_{k+1} := \theta_k + \Delta t \cdot \nabla R(\theta_k), \tag{4.6}$$

where $\Delta t > 0$ is the step size.

In principle, the reward function can be differentiated using automatic or numerical differentiation methods. A different approach is to use the celebrated policy gradient theorem and use matrix inversion to compute the state-action value function Q_θ. We restate the policy gradient theorem here for convience.

Theorem 3.11 (Policy gradient theorem, [277, 190, 2]). *It holds that*

$$(1-\gamma)\partial_{\theta_i} R(\theta) = \sum_s \rho_\theta(s) \sum_a \partial_{\theta_i} \pi_\theta(a|s) Q_\theta(s,a) = \sum_{s,a} \eta_\theta(s,a) \partial_{\theta_i} \log(\pi_\theta(a|s)) Q_\theta(s,a).$$

In a reinforcement learning setup, one does not have direct access to α and hence to state and state-action transition kernels p_π and P_π nor Q^π, and sometimes even the state space $\mathcal{S}$ is not known a priori. In this case, one has to estimate the gradient from interactions with the environment [40, 39, 207, 276]. In this work, however, we study the planning problem in MDPs, i.e., we assume access to exact gradient evaluations.

Policy parametrizations. Many results on the convergence of policy gradient methods have been provided for *tabular softmax policies*. The tabular softmax parametrization is given by

$$\pi_\theta(a|s) := \frac{e^{\theta_{sa}}}{\sum_{a'} e^{\theta_{sa'}}} \quad \text{for all } a \in \mathcal{A}, s \in \mathcal{S}, \tag{4.7}$$

for $\theta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. One benefit of tabular softmax policies is that they parametrize the interior of the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$ in a regular way, i.e., such that the Jacobian has full rank everywhere, and the parameter is unconstrained in an affine space.

Definition 4.4 (Regular parametrization). We call $\mathbb{R}^p \to \operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{S}}), \theta \mapsto \pi_\theta$ a *regular policy parametrization* if it is differentiable, surjective and satisfies

$$\operatorname{span}\{\partial_{\theta_i}\pi_\theta : i = 1, \dots, p\} = T_{\pi_\theta}\Delta_{\mathcal{A}}^{\mathcal{S}} \quad \text{for every } \theta \in \mathbb{R}^p. \tag{4.8}$$

We will focus on regular policy parametrizations, which cover softmax policies as well as escort transformed policies [198]. Nonetheless, we observe that policy optimization with constrained search variables can also be an attractive option and refer to [210] for a discussion in context of POMDPs.

Remark 4.5 (Partially observable systems). Although we will only consider parametric policies in fully observable MDPs, our discussion covers the case of POMDPs in the following way. Any parametric family of observation-based policies $\{\pi_\theta : \theta \in \Theta\} \subseteq \Delta_{\mathcal{A}}^{\mathcal{O}}$ induces a parametric family of state-based policies $\{\pi_\theta \circ \beta : \theta \in \Theta\} \subseteq \Delta_{\mathcal{A}}^{\mathcal{S}}$. Hence, the policy gradient theorem as well as the definitions of natural policy gradients directly extend to the case of partially observable systems. However, the global convergence guarantees in Section 4.3 and Section 4.4 do not carry over to POMDPs since they assume regular parametrization of the policies.

Regularization in MDPs. In practice, the reward function is often regularized as

$$R_\lambda = R - \lambda\psi.$$

This is often motivated to encourage exploration [300] and has also been shown to lead to fast convergence for strictly convex regularizers ψ [200, 71]. One popular regularizer is the conditional entropy in state-action space, see [218, 200, 71],

$$\psi_C(\theta) = \sum_s \rho_\theta(s) \sum_a \pi_\theta(a|s) \log(\pi_\theta(a|s)) = H(\eta_\theta) - H(\rho_\theta), \tag{4.9}$$

which has also been used to successfully design trust region and proximal methods for reward optimization [251, 250]. It is also possible to take the functions ϕ_σ defined in (4.1) as regularizers. This includes the entropy function, which is studied in state-action space in [218] and logarithmic barriers, which are studied in policy space in [2].

Projected policy gradients. An alternative to using parametrizations with the property that any unconstrained choice of the parameter leads to a policy, is to use constrained parametrizations and projected gradient methods. For instance, one can parametrize policies in $\Delta_{\mathcal{A}}^{\mathcal{S}}$ by their constrained entries and use the iteration

$$\pi_{k+1} := \Pi_{\Delta_{\mathcal{A}}^{\mathcal{S}}}(\pi_k + \Delta t G(\pi_k)^+ \nabla R(\pi)),$$

where $\Pi_{\Delta_{\mathcal{A}}^{\mathcal{S}}}$ is the (Euclidean) projection to $\Delta_{\mathcal{A}}^{\mathcal{S}}$. We will not study projected policy gradient methods and refer to [2, 305] for convergence rates of these methods.

4.2.2. Kakade's natural policy gradient. Kakade [153] proposed a natural policy gradient based on a Riemannian geometry in the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$. We will see that Kakade's NPG can be interpreted as the NPG induced by the Hessian geometry in state-action space arising from conditional entropy regularization of the linear program associated to MDPs. Kakade's idea was to mix the Fisher information matrices of the policy over the individual states according to the state frequencies, i.e., to use the following Gram matrix:

$$
\begin{aligned}
G_K(\theta)_{ij} &= \sum_s \rho_\theta(s) \sum_a \pi_\theta(a|s)\partial_{\theta_i} \log(\pi_\theta(a|s))\partial_{\theta_j} \log(\pi_\theta(a|s)) \\
&= \sum_{s,a} \eta_\theta(s,a)\partial_{\theta_i} \log(\pi_\theta(a|s))\partial_{\theta_j} \log(\pi_\theta(a|s)) \\
&= \sum_s \rho_\theta(s) \sum_a \frac{\partial_{\theta_i}\pi_\theta(a|s)\partial_{\theta_j}\pi_\theta(a|s)}{\pi_\theta(a|s)}.
\end{aligned} \tag{4.10}
$$

Definition 4.6 (Kakade's NPG and geometry in policy space). We refer to the natural gradient $\nabla^K R(\theta) := G_K(\theta)^+ \nabla_\theta R(\pi_\theta)$ as *Kakade's natural policy gradient (K-NPG)*, where G_K is defined in (4.10). Hence, Kakade's NPG is the NPG induced by the factorization $\theta \mapsto \pi_\theta \mapsto R(\theta)$ and the Riemannian metric on $\operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$ given by

$$
g_\pi^K(v,w) := \sum_s \rho^\pi(s) \sum_a \frac{v(s,a)w(s,a)}{\pi(a|s)} \quad \text{for all } v, w \in T_\pi \Delta_{\mathcal{A}}^{\mathcal{S}}. \tag{4.11}
$$

Due to its popularity, this method is often referred to simply as *the* natural policy gradient. We will call it Kakade's NPG in order to distinguish it from other NPGs.

Remark 4.7. In [153] the definition of G_K was heuristically motivated by the fact that the reward is also a mix of the one step rewards according to the state frequencies, $R(\pi) = \sum_s \rho^\pi(s) \sum_a \pi(a|s) r(s,a) = \sum_s \rho^\pi(s) r_\pi(s)$. The invariance axiomatic approaches discussed in [169, 205] also yield mixtures of Fisher metrics over individual states, which however do not fully recover Kakade's metric, since this would require a way to account for the particular process that gives rise to the stationary state distribution ρ^π. The works [232, 28, 216] argued that the Gram matrix G_K corresponds to the limit of the Fisher information matrices of finite-path measures as the path length tends to infinity.

Interpration as Hessian geometry of conditional entropy. The metric g^K on the conditional probability polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$ has been studied in terms of its invariances and its connection to the Fisher metric on finite-horizon path space [28, 232, 205]. We offer a different interpretation of Kakade's geometry by studying its counterpart in state-action space, which we show to be the Hessian geometry induced by the conditional entropy.

Theorem 4.8 (Kakade's geometry as conditional entropy Hessian geometry). *Consider an MDP $(\mathcal{S}, \mathcal{A}, \alpha, r)$ and fix $\mu \in \Delta_{\mathcal{S}}$ and $\gamma \in [0,1)$ such that Assumption 3.3 holds. Then, Kakade's geometry on $\Delta_{\mathcal{A}}^{\mathcal{S}}$ is the pull back of the Hessian geometry induced by the conditional entropy on the state-action polytope $\mathcal{N} \subseteq \Delta_{\mathcal{S}\times\mathcal{A}}$ along $\pi \mapsto \eta^\pi$.*

Proof. We can pull back the Riemannian metric on the policy polytope proposed by Kakade along the conditioning map to define a corresponding geometry in state-action space. The metric tensor in state-action space is given by

$$\begin{aligned}
G(\eta)_{(s,a),(s',a')} &= g^K_\pi(\partial_{(s,a)}\eta(\cdot|\cdot), \partial_{(s',a')}\eta(\cdot|\cdot)) \\
&= \sum_{\tilde{s},\tilde{a}} \rho(\tilde{s}) \frac{\partial_{(s,a)}\eta(\tilde{a}|\tilde{s})\partial_{(s',a')}\eta(\tilde{a}|\tilde{s})}{\eta(\tilde{a}|\tilde{s})} \\
&= \sum_{\tilde{s},\tilde{a}} \rho(\tilde{s})^2 \frac{\partial_{(s,a)}\eta(\tilde{a}|\tilde{s})\partial_{(s',a')}\eta(\tilde{a}|\tilde{s})}{\eta(\tilde{s},\tilde{a})}.
\end{aligned} \tag{4.12}$$

Using $\partial_{(s,a)}\eta(\tilde{a}|\tilde{s}) = \partial_{(s,a)}(\eta(\tilde{s},\tilde{a})\rho(\tilde{s})^{-1}) = \delta_{s\tilde{s}}(\delta_{a\tilde{a}}\rho(\tilde{s})^{-1} - \eta(\tilde{s},\tilde{a})\rho(\tilde{s})^{-2})$ we obtain

$$G(\eta)_{(s,a),(s',a')} = \delta_{ss'}\left(\delta_{aa'}\eta(s,a)^{-1} - \rho(s)^{-1}\right). \tag{4.13}$$

We aim to show that $G(\eta) = \nabla^2\phi_C(\eta)$, where $\phi_C(\eta) = H(\eta) - H(\rho)$ denotes the relative entropy and $\rho(s) = \sum_a \eta(s,a)$ denotes the state-marginal. Note that $\nabla^2 H(\eta) = \operatorname{diag}(\eta)$, which is the first term appearing in (4.13). For linear maps $g_A(x) = Ax$ the chain rule yields the expression

$$\partial_i\partial_j(f \circ g_A)(x) = \sum_{k,l} A_{ki}\partial_k\partial_l f(g_A(x))A_{lj}.$$

Noting that ρ is a linear function of η we obtain

$$\partial_{(s,a)}\partial_{(s',a')}H(\rho) = \sum_{\tilde{s},\hat{s}} \delta_{\tilde{s},s}\partial_{\tilde{s}}\partial_{\hat{s}}H(\rho)\delta_{\hat{s},s'} = \delta_{ss'}\rho(s)^{-1},$$

which is the second term in (4.13). Overall this implies $G(\eta) = \nabla^2\phi_C(\eta)$. □

The above theorem shows that Kakade's natural policy gradient is the natural policy gradient induced by the factorization $\theta \mapsto \eta_\theta \mapsto R(\theta)$ with respect to the conditional entropy Hessian geometry, i.e.,

$$\begin{aligned}
G_K(\theta)_{ij} &= \sum_{s,a} \frac{\partial_{\theta_i}\eta_\theta(s,a)\partial_{\theta_j}\eta_\theta(s,a)}{\eta_\theta(s,a)} - \sum_s \frac{\partial_{\theta_i}\rho_\theta(s)\partial_{\theta_j}\rho_\theta(s)}{\rho_\theta(s)} \\
&= \sum_{s,a} \partial_{\theta_i}\log(\eta_\theta(s,a))\partial_{\theta_j}\log(\eta_\theta(s,a))\eta_\theta(s,a) \\
&\quad - \sum_s \partial_{\theta_i}\log(\rho_\theta(s))\partial_{\theta_j}\log(\rho_\theta(s))\rho_\theta(s).
\end{aligned} \tag{4.14}$$

It is also worth noting that the Bregman divergence of the conditional entropy is the conditional relative entropy and has been studied as a regularizer for the linear program associated to MDPs in [218].

Remark 4.9. Kakade's NPG is known to converge at a locally quadratic rate under conditional entropy regularization [71], a regularizer, which in policy space takes the form

$$\psi(\pi) = \sum_s \rho^\pi(s) \sum_a \pi(a|s)\log(\pi(a|s)) = \sum_s \rho^\pi(s)H(\pi(\cdot|s)).$$

Note however, by direct calculation, that Kakade's geometry in policy space g^K defined in (4.11) is not the Hessian geometry induced by ψ in policy space, which would take the form

$$\nabla^2\psi(\pi) = \sum_s \rho^\pi(s)\nabla^2 H(\pi(\cdot|s)) + \sum_s(\nabla H(\cdot|s)^\top \nabla\rho^\pi(s) + \nabla H(\cdot|s)\nabla\rho^\pi(s)^\top) \\ + \sum_s H(\pi(\cdot|s))\nabla^2\rho^\pi(s).$$

Instead, the metric proposed by Kakade only considers the contribution of the first term; see (4.11). As we will see in Sections 4.3 and 4.4, the interpretation of Kakade's NPG as a Hessian natural gradient induced by the conditional entropic regularization in state-action space allows for a great simplification of its convergence analysis. One can show that Kakade's metric is not a Hessian metric in policy space. By Schwarz's theorem the metric tensor of a Hessian Riemannian metric satisfies $\partial_i g_{jk} = \partial_j g_{ik}$. However, we have

$$\partial_{(\tilde{s},\tilde{a})} G(\pi)_{(s,a),(s',a')} = \delta_{ss'}\delta_{aa'}\left(-\delta_{s\tilde{s}}\delta_{a\tilde{a}}\rho^\pi(s)\pi(a|s)^{-2} + \pi(a|s)\partial_{(\tilde{s},\tilde{a})}\rho^\pi(s)\right),$$

which does not satisfy this symmetry property in general. This shows that the Riemannian metric on the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$ proposed by Kakade does not arise from a Hessian.

4.2.3. Morimura's natural policy gradient. In contrast to Kakade's approach, who proposed a mixture of Fisher metrics to obtain a metric on the conditional probability polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$, Morimura and co-authors [206] proposed to work with the Fisher metric in state-action space $\Delta_{\mathcal{S}\times\mathcal{A}}$ to define a natural gradient for reward optimization. The resulting Gram matrix is given by the Fisher information matrix induced by the state-action distributions, that is $P(\theta) = \eta_\theta$ and

$$G_M(\theta)_{ij} = \sum_{s,a} \partial_{\theta_i}\log(\eta_\theta(s,a))\partial_{\theta_j}\log(\eta_\theta(s,a))\eta_\theta(s,a). \tag{4.15}$$

Definition 4.10 (Morimura's NPG). We refer to the $\nabla^M R(\theta) := G_M(\theta)^+\nabla_\theta R(\pi_\theta)$ as *Morimura's natural policy gradient (M-NPG)*, where G_M is defined in (4.15). Hence, Morimura's NPG is the NPG induced by the factorization $\theta \mapsto \eta_\theta \mapsto R(\theta)$ and the Fisher metric on $\mathrm{int}(\Delta_{\mathcal{S}\times\mathcal{A}})$.

By (4.14) the Gram matrix proposed by Morimura and co-authors and the Gram matrix proposed by Kakade are related to each other by

$$G_K(\theta) = G_M(\theta) - F_\rho(\theta),$$

where $F_\rho(\theta)_{ij} = \sum_s \rho_\theta(s)\partial_{\theta_i}\log(\rho_\theta(s))\partial_{\theta_j}\log(\rho_\theta(s))$ denotes the Fisher information matrix of the state distributions. This relation is reminiscent of the chain rule for the conditional entropy and can be verified by direct computation; see [206]. Where we have seen that Kakade's geometry in state-action space is the Hessian geometry of conditional entropy, the Fisher metric is known to be the Hessian metric of the entropy function [15]. Hence, we can interpret the Fisher metric as the Hessian geometry of the entropy regularized reward $\eta \mapsto \langle r, \eta\rangle - H(\eta)$.

4.2.4. General Hessian natural policy gradient. Generalizing the above definitions, we define general state-action space Hessian NPGs as follows. Consider a twice differentiable function $\phi\colon \mathbb{R}_{>0}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}$ such that $\nabla^2\phi(\eta)$ is positive definite on $T_\eta\mathcal{N} = T\mathcal{L} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$

for every $\eta \in \operatorname{int}(\mathcal{N})$. Then we set

$$G_\phi(\theta)_{ij} := \sum_{s,s',a,a'} \partial_{\theta_i}\eta_\theta(s,a)\partial_{(s,a)}\partial_{(s',a')}\phi(\eta_\theta)\partial_{\theta_j}\eta_\theta(s',a'),$$

which is the Gram matrix with respect to the Hessian geometry in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$.

Definition 4.11 (Hessian NPG). We refer to $\nabla^\phi R(\theta) := G_\phi(\theta)^+\nabla_\theta R(\pi_\theta)$ as *Hessian natural policy gradient with respect to ϕ* or shortly *ϕ-natural policy gradient (ϕ-NPG)*.

Leveraging results on gradient flows in Hessian geometries we will later provide global convergence guarantees including convergence rates for a large class of Hessian NPG flows covering Kakade's and Morimura's natural gradients as special cases. Further, we consider the family ϕ_σ of strictly convex functions defined in (4.1). With $G_\sigma(\theta)$ we denote the Gram matrix associated with the Riemannian metric g^σ, i.e.,

$$G_\sigma(\theta)_{ij} = \sum_{s,a} \frac{\partial_{\theta_i}\eta_\theta(s,a)\partial_{\theta_j}\eta_\theta(s,a)}{\eta_\theta(s,a)^\sigma}.$$

Definition 4.12 (σ-NPG). We refer to the natural gradient $\nabla^\sigma R(\theta) := G_\sigma(\theta)^+\nabla_\theta R(\pi_\theta)$ as the *σ-natural policy gradient (σ-NPG)*. Hence, the σ-NPG is the NPG induced by the factorization $\theta \mapsto \eta_\theta \mapsto R(\theta)$ and the metric g^σ on $\operatorname{int}(\Delta_{\mathcal{S}\times\mathcal{A}})$ defined in (4.2).

For $\sigma = 1$ we recover the Fisher geometry and hence Morimura's NPG; for $\sigma = 2$ we obtain the Itakura-Saito metric; and for $\sigma = 0$ we recover the Euclidean geometry. Later, we show that the Hessian gradient flows exist globally for $\sigma \in [1,\infty)$ and provide convergence rates depending on σ.

4.3 Convergence of natural policy gradient flows

In this section we study the convergence properties of natural policy gradient flows arising from Hessian geometries in state-action space for fully observable systems and regular parametrizations of the interior of the policy polytope $\Delta^{\mathcal{S}}_{\mathcal{A}}$. Leveraging methods from the theory of gradient flows in Hessian geometries established in [11] we show $O(t^{-1})$ convergence of the objective value for a large class of Hessian geometries and unregularized reward. We strengthen this general result and establish linear convergence for Kakade's and Morimura's NPG flows and $O(t^{-1/(\sigma-1)})$ convergence for σ-NPG flows for $\sigma \in (1,2)$. We provide empirical evidence that these rates are tight and that the rate $O(t^{-1/(\sigma-1)})$ also holds for $\sigma \geq 2$. Under strongly convex penalization, we obtain linear convergence for a large class of Hessian geometries.

Reduction to state-action space dynamics. For a solution $\theta(t)$ of the natural policy gradient flow, the corresponding state-action frequencies $\eta(t)$ solve the gradient flow with respect to the Riemannian metric. This is made precise in the following result, which shows that it suffices to study Riemannian gradient flows in state-action space.

Proposition 4.13 (Evolution in state-action space). *Consider an MDP $(\mathcal{S},\mathcal{A},\alpha,r)$, a Riemannian metric g on $\operatorname{int}(\mathcal{N}) = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$ and a differentiable objective function $\mathfrak{R}\colon \operatorname{int}(\Delta_{\mathcal{S}\times\mathcal{A}}) \to \mathbb{R}$. Consider a regular policy parametrization and the parameter objective $R(\theta) := \mathfrak{R}(\eta_\theta)$ and a solution $\theta\colon [0,T] \to \Theta = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ of the natural policy gradient flow*

$$\partial_t\theta(t) = \nabla^N R(\theta(t)) = G(\theta(t))^+\nabla R(\theta(t)), \tag{4.16}$$

where $G(\theta)_{ij} = g_\eta(\partial_{\theta_i}\eta_\theta, \partial_{\theta_j}\eta_\theta)$ and $G(\theta)^+$ denotes some pseudoinverse of $G(\theta)$. Then, setting $\eta(t) := \eta_{\theta(t)}$ we have that $\eta\colon [0,T] \to \Delta_{\mathcal{S}\times\mathcal{A}}$ is the gradient flow with respect to the metric g and the objective $\mathfrak{R}$, i.e., solves

$$\partial_t\eta(t) = \nabla^g\mathfrak{R}(\eta(t)). \tag{4.17}$$

Proof. This is a direct consequence of Theorem 4.2. □

The preceding result covers the commonly studied tabular softmax parametrization. For general parametrizations, the result does not hold. However, if for any two parameters θ, θ' with $\eta_\theta = \eta_{\theta'}$ it holds that

$$\operatorname{span}\{\partial_{\theta_i}\pi_\theta : i = 1, \dots, p\} = \operatorname{span}\{\partial_{\theta_i}\pi_{\theta'} : i = 1, \dots, p\},$$

then a similar result can be established. An important special case of such parametrizations occurs in partially observable problems with memoryless policies parametrized in a regular way, e.g., through the softmax or escort transform; see also Remark 4.5.

By Proposition 4.13 it suffices to study solutions $\eta\colon [0,T] \to \mathcal{N}$ of the gradient flow in state-action space. We have seen before that a large class of natural policy gradients arise from Hessian geometries in state-action space. In particular, this covers the natural policy gradients proposed by Kakade [153] and Morimura et al. [206]. We study the evolution of these flows in state-action space and leverage results on Hessian gradient flows of convex problems in [11, 290] to obtain global convergence rates for different NPG methods.

4.3.1. Convergence and existence for general rewards.

First, we study the convergence and well posedness for general reward functions under the following assumptions.

Setting 4.14. *Let $(\mathcal{S}, \mathcal{A}, \alpha, r)$ be an MDP, $\gamma \in [0,1)$, $\mu \in \Delta_{\mathcal{S}}$ and let the positivity Assumption 3.3 hold. We denote the state-action polytope by $\mathcal{N} = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\ge 0} \cap \mathcal{L}$ (see Theorem 3.5) and the (relative) interior and boundary of $\mathcal{N}$ by $\operatorname{int}(\mathcal{N}) = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0} \cap \mathcal{L}$ and $\partial\mathcal{N} = \partial\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\ge 0} \cap \mathcal{L}$ respectively. We consider an objective function $\mathfrak{R}\colon \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R} \cup \{-\infty, +\infty\}$ that is finite and differentiable on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$ and we assume that $\mathfrak{R}$ continuous on its domain $\operatorname{dom}(\mathfrak{R}) = \{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \mathfrak{R}(\eta) \in \mathbb{R}\}$. We consider a real-valued function $\phi\colon \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R} \cup \{+\infty\}$, which we assume to be finite and twice continuously differentiable on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$ and such that $\nabla^2\phi(\eta)$ is positive definite on $T_\eta\mathcal{N} = T\mathcal{L} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ for every point $\eta \in \operatorname{int}(\mathcal{N})$ and denote the induced Hessian metric on $\operatorname{int}(\mathcal{N})$ by g. Further, with $\eta\colon [0,T) \to \mathcal{N}$ we denote a solution of the Hessian gradient flow*

$$\partial_t\eta(t) = \nabla^g\mathfrak{R}(\eta(t)) \tag{4.18}$$

with initial condition $\eta(0) = \eta_0$. We set $R^* := \sup_{\eta\in\mathcal{N}}\mathfrak{R}(\eta) \in \mathbb{R} \cup \{+\infty\}$ and by $\eta^* \in \mathcal{N}$, we denote a maximizer, if one exists, of $\mathfrak{R}$ over $\mathcal{N}$. We denote the policies corresponding to $\eta(t)$, η_0 and η^* by $\pi(t)$, π_0 and π^*, see Proposition 3.4.

We observe that the Hessian of the conditional entropy only defines a Riemannian metric on $\operatorname{int}(\mathcal{N})$ and not over all of $\Delta_{\mathcal{S}\times\mathcal{A}}$. Note that η^* might lie on the boundary and for linear $\mathfrak{R}$ corresponding to unregularized reward it necessarily lies on the boundary.

We repeatedly make use of the following identity

$$\langle\nabla^2\phi(\eta)\nabla^g\mathfrak{R}(\eta), v\rangle = g_\eta(\nabla^g\mathfrak{R}(\eta), v) = d\mathfrak{R}(\eta)v = \langle\nabla\mathfrak{R}(\eta), v\rangle, \tag{4.19}$$

which holds for any $v \in T\mathcal{L}$.

Sublinear rates for general rewards. We begin by providing a sublinear rate of convergence for general NPG flows, which we then specialize to Kakade and σ-NPGs.

Lemma 4.15 (Convergence of Hessian natural policy gradient flows). *Consider Setting 4.14 and assume that $\mathfrak{R}$ is concave on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$ and that there exists a solution $\eta\colon [0,T) \to \operatorname{int}(\mathcal{N})$ of the NPG flow* (4.18) *with initial condition $\eta(0) = \eta_0$. Then for any $\eta' \in \mathcal{N}$ and $t \in [0,T)$ it holds that*

$$\mathfrak{R}(\eta') - \mathfrak{R}(\eta(t)) \le (D_\phi(\eta', \eta_0) - D_\phi(\eta', \eta(t)))t^{-1} \le D_\phi(\eta', \eta_0)t^{-1}, \tag{4.20}$$

where D_ϕ denotes the Bregman divergence of ϕ, in particular, $\mathfrak{R}(\eta(t)) \to R^*$ as $T \to \infty$. If there is a maximizer $\eta^* \in \mathcal{N}$ of $\mathfrak{R}$ with $\phi(\eta^*) < \infty$ the convergence happens at a rate $O(t^{-1})$.

Proof. This is precisely the statement of Proposition 4.4 in [11]; note however, that they assume a globally defined objective $\mathfrak{R}\colon \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}$ and hence for completeness we provide a quick argument. If $\phi(\eta') = +\infty$ the statement is trivial and hence we assume $\phi(\eta') < +\infty$. It holds that

$$\begin{aligned} \partial_t D_\phi(\eta, \eta(t)) &= -\partial_t \phi(\eta(t)) - \partial_t \langle \nabla\phi(\eta(t)), \eta - \eta(t)\rangle \\ &= -\langle \nabla\phi(\eta(t)), \partial_t \eta(t)\rangle - \langle \nabla^2\phi(\eta(t))\partial_t\eta(t), \eta - \eta(t)\rangle + \langle \nabla\phi(\eta(t)), \partial_t\eta(t)\rangle \\ &= -\langle \nabla\mathfrak{R}(\eta(t)), \eta - \eta(t)\rangle, \end{aligned}$$

where we used $\partial_t\eta(t) = \nabla^g\mathfrak{R}(\eta(t))$ as well as (4.19). Using the concavity of $\mathfrak{R}$ we can estimate

$$\partial_t D_\phi(\eta, \eta(t)) = -\langle \nabla\mathfrak{R}(\eta(t)), \eta - \eta(t)\rangle \le \mathfrak{R}(\eta(t)) - \mathfrak{R}(\eta). \tag{4.21}$$

Integration and the monotonicity of $t \mapsto \mathfrak{R}(\eta(t))$ yields the claim. □

Well posedness of NPG flows. The previous result holds for general state-space objective and hence covers both regularized and unregularized rewards and reduces the problem of showing convergence of the natural gradient flow to the problem of well-posedness. However, well-posedness is not always given, such as for example in the case of an unregularized reward and the Euclidean geometry in state-action space. In this case, the gradient flow in state-action space will reach the boundary of the state-action polytope $\mathcal{N}$ in finite time at which point the gradient is not classically defined anymore and the softmax parameters blow up; see Figure 4.3. An important class of Hessian geometries that prevent a finite hitting time of the boundary are induced by the class of Legendre-type functions, which curve up towards the boundary.

Definition 4.16 (Legendre type functions). We call $\phi\colon \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R} \cup \{+\infty\}$ a *Legendre type function* if it satisfies the following properties:

(*i*) *Domain:* It holds that $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0} \subseteq \operatorname{dom}(\phi) \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\ge 0}$, where the domain is given by $\operatorname{dom}(\phi) = \{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \phi(\eta) < \infty\}$.

(*ii*) *Smoothness and convexity:* We assume ϕ to be continuous on $\operatorname{dom}(\phi)$ and twice continuously differentiable on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$ and such that $\nabla^2\phi(\eta)$ is positive definite on $T_\eta\mathcal{N} = T\mathcal{L} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ for every $\eta \in \operatorname{int}(\mathcal{N})$.

(*iii*) *Gradient blowup at boundary:* For any $(\eta_k)_{k\in\mathbb{N}} \subseteq \operatorname{int}(\mathcal{N})$ with $\eta_k \to \eta \in \partial\mathcal{N}$ we have $\|\nabla\phi(\eta_k)\| \to \infty$.

We note that the above definition differs from [11], who consider Legendre functions on arbitrary open sets but work with more restrictive assumptions. More precisely, they require the gradient blowup on the boundary of the entire cone $\mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}}$ and not only on the boundary of the feasible set $\mathcal{N}$ of the optimization problem. However, this relaxation is required to cover the case of the conditional entropy, which corresponds to Kakade's NPG, as we see now.

Example 4.17. The class of Legendre type functions covers the functions inducing Kakade's and Morimura's NPG via their Hessian geometries. More precisely, the following Legendre type functions will be of great interest in the remainder:

(i) The functions ϕ_σ defined in (4.1) used to define the σ-NPG are Legendre type functions for $\sigma \in [1, \infty)$. Note that this includes the Fisher geometry, corresponding to Morimura's NPG for $\sigma = 1$ but excludes the Euclidean geometry, which corresponds to $\sigma = 0$.

(ii) The conditional entropy ϕ_C defined in (4.9) is a Legendre type function. The Hessian geometry of this function induces Kakade's NPG. Note that in this case the gradient blowup holds on the boundary $\mathcal{N}$ but not on the boundary of $\Delta_{\mathcal{S}\times\mathcal{A}}$ or even $\mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}}$.

The definition of a Legendre function with the gradient blowing up at the boundary of the feasible set prevents the gradient flow from reaching the boundary in finite time and thus ensures the global existence of the gradient flow. Here, we provide a global existence result similar to [11, Theorem 4.1], where they require the objective function $\mathfrak{R}$ to admit a smooth extension onto the ambient space and ϕ to be globally strictly convex. Note that the regularized reward $\mathfrak{R}(\eta) = \langle r, \eta\rangle - \lambda\phi(\eta)$ never admits a smooth extension if ϕ is a Legendre type function like the entropy or conditional entropy. Therefore, we require a different assumption (4.22) that bounds the curvature of the objective in terms of the curvature of the function ϕ as we discuss in Example 4.19.

Theorem 4.18 (Well posedness of Hessian NPG flows). *Consider Setting 4.14, let ϕ be a Legendre type function and assume that there exists $c \in \mathbb{R}$ such that*

$$\langle \nabla^2\mathfrak{R}(\eta)\nabla^g\mathfrak{R}(\eta), \Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta)\rangle \leq c \cdot \|\Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta)\|^2 \quad \textit{for all } \eta \in \operatorname{int}(\mathcal{N}), \tag{4.22}$$

where $\Pi_{T\mathcal{L}}$ denotes the Euclidean projection onto the tangent space of $\mathcal{L}$. Then for any $\eta_0 \in \operatorname{int}(\mathcal{N})$ there exists a unique global solution $\eta\colon [0, \infty) \to \operatorname{int}(\mathcal{N})$ of the Hessian natural policy gradient flow (4.18) *with $\eta(0) = \eta_0$. In particular, this covers the unregularized reward $\mathfrak{R}(\eta) = \langle r, \eta\rangle$, the regularized reward $\mathfrak{R}(\eta) = \langle r, \eta\rangle - \lambda\phi(\lambda)$ and the case $\mathfrak{R}(\eta) = -\lambda\phi(\eta)$ for $\lambda \in \mathbb{R}$.*

Before we present the proof we give insight into its arguments and provide some intuition for the condition (4.22), which compares the Hessians of $\mathfrak{R}$ and ϕ.

Example 4.19 (A one dimensional example). Let us consider a Hessian gradient flow in one dimension given by

$$\partial_t x(t) = \phi''(x(t))^{-1} f'(x(t))$$

for a suitably convex function $\phi\colon \mathbb{R}_{>0} \to \mathbb{R}$ such that $\phi'(x) \to +\infty$ for $x \searrow 0$ and a function $f\colon \mathbb{R}_{>0} \to \mathbb{R}$. It is our goal to show the well posedness in this case, where the condition (4.22) takes the form $f''(x)\phi''(x)^{-1} \leq c$. The only scenario how the Hessian gradient flow might not exist globally is if $x(t) \searrow 0$ for $t \nearrow T < +\infty$, in which case

$\phi'(x(t)) \to +\infty$. Note that $\partial_t\phi'(x(t)) = \phi''(x(t))\partial_t x(t) = f'(x(t))$ and hence

$$\phi'(x(t)) = \phi'(x(0)) + \int_0^t f'(x(s))\mathrm{d}s.$$

In order to bound this we estimate

$$\partial_t f'(x(t))^2 = 2f'(x(t))f''(x(t))\partial_t x(t) = 2f''(x(t))\phi''(x(t))^{-1}f'(x(t))^2 \leq 2cf'(x(t))^2$$

and Grönwall's inequality implies $|f'(x(t))|\leq |f'(x(0))|\cdot \exp(ct)$. This yields

$$\phi'(x(t)) \leq \phi'(x(0)) + T|f'(x(0))| \cdot \exp(cT) < +\infty$$

in contradiction to $\phi'(x(t)) \to +\infty$ for $t \nearrow T$.

The condition $f''(x) \leq c\phi''(x)$ can be interpreted as a bound on the curvature of f in terms of the curvature of ϕ. This bound is indeed necessary for long time existence as for example the choice $\phi(x) := x^{-1}$ and $f(x) := x^{-2}$ leads to the flow $x(t) = x(0) - t$, which only exists in $\mathbb{R}_{>0}$ until time $T = x(0)$.

The proof of Theorem 4.18 follows the same ideas as Example 4.19 but we require the following auxiliary result.

Lemma 4.20 (Lemma 4.3 in [11]). *Let $C \subseteq \mathbb{R}^d$ be a nonempty open convex subset, let $U \subseteq \mathbb{R}^d$ be a subspace of $\mathbb{R}^d$, fix $\hat{x} \in \partial C$ such that $(\hat{x} + U) \cap C \neq \emptyset$ and denote the normal cone of $\overline{C}$ at $\hat{x}$ by $NC_{\overline{C}}(\hat{x}) = \{v \in \mathbb{R}^d : \langle v, y - \hat{x}\rangle \leq 0$ for all $y \in \overline{C}\}$. Then $NC_{\overline{C}}(\hat{x}) \cap U^{\perp} = \{0\}$.*

Proof. Fix $v \in NC_{\overline{C}}(\hat{x}) \cap U^{\perp}$ and $y \in (\hat{x} + U) \cap C$. Then $\langle v, y - \hat{x}\rangle = 0$ since $y - \hat{x} \in U$ and $v \in U^{\perp}$. As C is open, there is $\varepsilon > 0$ such $y + w \in C$ if $\|w\| \leq \varepsilon$. As $v \in NC_{\overline{C}}(\hat{x})$ it holds that $\langle v, w\rangle = \langle v, y + w - \hat{x}\rangle \leq 0$ whenever $\|w\| \leq \varepsilon$, which shows that $v = 0$. □

Proof of Theorem 4.18. We follow the arguments given outlined in Example 4.19 and generalize them to the multi-dimenstional setting.

Let us denote the Euclidean distance of a point η to the boundary $\partial\mathcal{N}$ by $\operatorname{dist}(\eta, \partial\mathcal{N})$. The existence and uniqueness of the Hessian gradient flow on $[0, T)$, where

$$T := \lim_{\varepsilon\to 0} \sup\left\{t > 0 : \inf_{s\in[0,t]} \operatorname{dist}(\eta(s), \partial\mathcal{N}) \geq \varepsilon\right\},$$

follows from standard arguments, i.e., the Picard-Lindelöf theorem and by gluing solutions if they stay within a compact subset of $\operatorname{int}(\mathcal{N})$.

To show $T = +\infty$ we assume that $T < +\infty$. This implies $\inf_{t\in[0,T)} \operatorname{dist}(\eta(t), \partial\mathcal{N}) = 0$. Now we choose a sequence $(t_n)_{n\in\mathbb{N}} \subseteq [0, T)$ such that $\operatorname{dist}(\eta(t_n), \partial\mathcal{N}) \to 0$ for $n \to \infty$. Then surely $t_n \to T$ for $n \to \infty$ since $\inf_{s\in[0,t]} \operatorname{dist}(\eta(s), \partial\mathcal{N}) > 0$ for $t < T$. By compactness of $\mathcal{N}$ and the sphere we can assume without loss of generality that $\eta(t_n) \to \hat{\eta} \in \mathcal{N}$ and $\frac{\nabla\phi(\eta(t_n))}{\|\nabla\phi(\eta(t_n))\|} \to v \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. Note that $\hat{\eta} \in \partial\mathcal{N}$ since $\operatorname{dist}(\hat{\eta}, \partial\mathcal{N}) = \lim_{n\to\infty} \operatorname{dist}(\eta(t_n), \partial\mathcal{N}) = 0$. Since ϕ is a Legendre type function this yields $\|\nabla\phi(\eta(t_n))\| \to \infty$.

Next, we show that $v \in NC_{\mathcal{N}}(\hat{\eta})$, where $NC_{\mathcal{N}}(\hat{\eta})$ denotes the normal cone of $\mathcal{N}$ at $\hat{\eta}$. For any $\eta' \in \operatorname{int}(\mathcal{N})$ the convexity of ϕ implies that $\langle\nabla\phi(\eta(t_n)) - \nabla\phi(\eta'), \eta' - \eta(t_n)\rangle \leq 0$. Devision by $\|\nabla\phi(\eta(t_n))\|$ and taking the limit $n \to \infty$ yields $\langle v, \eta' - \hat{\eta}\rangle \leq 0$ showing $v \in NC_{\mathcal{N}}(\hat{\eta})$. By Lemma 4.20 we have $v \notin (T\mathcal{L})^{\perp}$, where $\mathcal{L}$ is the linear space such that

$\mathcal{N} = \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}} \cap \mathcal{L}$, see Theorem 3.5 and hence $v_0 := \Pi_{T\mathcal{L}} v \neq 0$, where $\Pi_{T\mathcal{L}}$ denotes the Euclidean projection onto $T\mathcal{L}$. For $n \to \infty$ we have

$$\left\langle \frac{\nabla\phi(\eta(t_n))}{\|\nabla\phi(\eta(t_n))\|}, v_0 \right\rangle \to \langle v, v_0\rangle = \|v_0\| > 0$$

and therefore

$$\langle \nabla\phi(\eta(t_n)), v_0\rangle \to \infty \quad \text{for } n \to +\infty. \tag{4.23}$$

Note that $\partial_t \nabla\phi(\eta(t)) = \nabla^2\phi(\eta(t))\partial_t\eta(t) = \nabla^2\phi(\eta(t))\nabla^g\mathfrak{R}(\eta(t))$. Integration and (4.19) yields

$$\begin{aligned}\langle \nabla\phi(\eta(t_n)), v_0\rangle &= \left\langle \int_0^{t_n} \nabla^2\phi(\eta(s))\nabla^g\mathfrak{R}(\eta(s))\mathrm{d}s + \nabla\phi(\eta_0), v_0\right\rangle \\ &= \langle\nabla\phi(\eta_0), v_0\rangle + \int_0^{t_n} \langle \nabla\mathfrak{R}(\eta(s)), v_0\rangle \mathrm{d}s \\ &= \langle\nabla\phi(\eta_0), v_0\rangle + \int_0^{t_n} \langle \Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta(s)), v_0\rangle \mathrm{d}s \\ &\leq \|\nabla\phi(\eta_0)\| + \int_0^{t_n} \|\Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta(s))\|\mathrm{d}s.\end{aligned} \tag{4.24}$$

To bound $\|\Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta(s))\|$ we use (4.22) and estimate

$$\begin{aligned}\partial_t\|\Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta(t))\|^2 &= 2\langle \Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta(t)), \partial_t\Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta(t))\rangle \\ &= 2\langle \Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta(t)), \Pi_{T\mathcal{L}}\partial_t\nabla\mathfrak{R}(\eta(t))\rangle \\ &= 2\langle \Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta(t)), \partial_t\nabla\mathfrak{R}(\eta(t))\rangle \\ &= 2\langle \Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta(t)), \nabla^2\mathfrak{R}(\eta(t))\partial_t\eta(t)\rangle \\ &= 2\langle \Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta(t)), \nabla^2\mathfrak{R}(\eta(t))\nabla^g\mathfrak{R}(\eta(t))\rangle \\ &\leq 2c\|\Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta(t))\|^2.\end{aligned}$$

Now Grönwall's inequality yields $\|\Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta(t))\| \leq \|\Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta_0)\| \cdot \exp(ct)$. Together with (4.24) this implies

$$\langle \nabla\phi(\eta(t_n)), v_0\rangle \leq \|\nabla\phi(\eta_0)\| + T\cdot\|\nabla\mathfrak{R}(\eta_0)\|\cdot\exp(cT) < +\infty$$

contradicting (4.23). Therefore, we have shown that $T = +\infty$.

We now verify the condition (4.22) for the individual cases. If $\mathfrak{R}$ is the unregularized reward, then $\nabla^2\mathfrak{R} = 0$ and hence (4.22) holds with $c = 0$. If $\mathfrak{R}(\eta) = \langle r, \eta\rangle - \lambda\phi(\eta)$ or $\mathfrak{R}(\eta) = -\lambda\phi(\eta)$ for some $\lambda \in \mathbb{R}$, then $\nabla^2\mathfrak{R} = -\lambda\nabla^2\phi$ and (4.19) yields

$$\begin{aligned}\langle \nabla^2\mathfrak{R}(\eta)\nabla^g\mathfrak{R}(\eta), \Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta)\rangle &= -\lambda\langle\nabla^2\phi(\eta)\nabla^g\mathfrak{R}(\eta), \Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta)\rangle \\ &= -\lambda\langle\nabla\mathfrak{R}(\eta), \Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta)\rangle \\ &= -\lambda\|\Pi_{T\mathcal{L}}\nabla\mathfrak{R}(\eta)\|^2\end{aligned}$$

and (4.22) holds with $c = -\lambda$. □

4.3.2. Convergence of unregularized NPG flows. Now, we study the case of unregularized reward, i.e., where the objective in state-action space is linear and given by $\mathfrak{R}(\eta) = \langle r, \eta \rangle$ for Kakade's as well as σ-NPG flows, which include Morimura's NPG as a special case. In this case we obtain global convergence guarantees including rates.

Sublinear rates. Recall that Kakade's natural policy gradient is induced by the conditional entropy ϕ_C defined in (4.3). The Bregman divergence of the conditional entropy, see [233], is given by the conditional relative entropy

$$\begin{aligned} D_{\phi_C}(\eta_1, \eta_2) &= \sum_{s,a} \eta_1(s,a) \log\left(\frac{\eta_1(s,a)}{\eta_2(s,a)}\right) - \sum_{s,a} \eta_1(s,a) \log\left(\frac{\sum_{a'} \eta_1(s,a')}{\sum_{a'} \eta_2(s,a')}\right) \\ &= D_{KL}(\eta_1, \eta_2) - D_{KL}(\rho_1, \rho_2) = \sum_s \rho_1(s) D_{KL}(\eta_1(\cdot|s), \eta_2(\cdot|s)), \end{aligned}$$

which has been studied in the context of mirror descent algorithms of the linear programming formulation of MDPs in [218].

Theorem 4.21 (Convergence of Kakade's NPG flow for unregularized reward). *Consider Setting 4.14 with $\phi = \phi_C$ being the conditional entropy defined in* (4.9) *and denote the unregularized reward by $\mathfrak{R}(\eta) = \langle r, \eta \rangle$ and fix an element $\eta_0 \in \operatorname{int}(\mathcal{N})$. Then there exists a unique global solution $\eta\colon [0,\infty) \to \operatorname{int}(\mathcal{N})$ of Kakade's NPG flow with initial condition $\eta(0) = \eta_0$, i.e., of* (4.18) *with $\phi = \phi_C$, and it holds that*

$$R^* - \mathfrak{R}(\eta(t)) \le t^{-1} D_{\phi_C}(\eta^*, \eta_0) = t^{-1} \sum_s \rho^*(s) D_{KL}(\pi^*(\cdot|s), \pi_0(\cdot|s)), \tag{4.25}$$

where D_{ϕ_C} denotes the conditional relative entropy and η^* an arbitrary maximizer. In particular, $\operatorname{dist}(\eta(t), S) \in O(t^{-1})$, *where $S = \{\eta \in \mathcal{N} : \langle r, \eta \rangle = R^*\}$ and* dist *denotes the Euclidean distance.*

Proof. The flow is well posed by Theorem 4.18 and Example 4.17. Now the result follows directly from Lemma 4.15. □

Now we elaborate the consequences of the general convergence result Lemma 4.15 for the case of σ-NPG flows. Here, the study is more delicate since for $\sigma > 2$ we typically have $\phi_\sigma(\eta^*) = \infty$ since the maximizer η^* lies at the boundary unless the reward is constant.

Theorem 4.22 (Convergence of σ-NPG flow for unregularized reward). *Consider Setting 4.14 with $\phi = \phi_\sigma$ for some $\sigma \in [1, \infty)$ being defined in* (4.1). *Denote the unregularized reward by $\mathfrak{R}(\eta) = \langle r, \eta \rangle$ and fix an element $\eta_0 \in \operatorname{int}(\mathcal{N})$. Then there exists a unique global solution $\eta\colon [0,\infty) \to \operatorname{int}(\mathcal{N})$ of the Hessian NPG flow* (4.18) *with inital condition $\eta(0) = \eta_0$ and and there is a constant $c = c(\sigma) > 0$ such that*

$$R^* - \mathfrak{R}(\eta(t)) \le \begin{cases} t^{-1} D_\sigma(\eta^*, \eta_0) & \text{for } \sigma \in [1,2) \\ c \log(t) t^{-1} & \text{for } \sigma = 2 \\ c t^{\sigma - 3} & \text{for } \sigma \in (2, \infty) \end{cases} \tag{4.26}$$

for an abitrary maximizer η^*. In particular, we have

$$\operatorname{dist}(\eta(t), S) \in \begin{cases} O(t^{-1}) & \text{for } \sigma \in [1,2) \\ O(\log(t) t^{-1}) & \text{for } \sigma = 2 \\ O(t^{\sigma-3}) & \text{for } \sigma \in (2, \infty), \end{cases} \tag{4.27}$$

where $S = \{\eta \in \mathcal{N} : \langle r, \eta\rangle = R^*\}$ denotes the solution set and dist *denotes the Euclidean distance. This result covers Morimura's NPG flow as the special case with $\sigma = 1$.*

Proof. The global existence follows from Theorem 4.18 and Example 4.17 and hence we can apply Lemma 4.15. For $\sigma \in [1, 2)$ we have that $\phi_\sigma(\eta^*) < \infty$ and hence $R^* - \Re(\eta(t)) \leq D_{\phi_\sigma}(\eta^*, \eta_0)t^{-1}$. Consider now the case $\sigma = 2$ and pick $v \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ such that $\eta_\delta \coloneqq \eta^* + \delta v \in \operatorname{int}(\mathcal{N})$ for small $\delta > 0$. Then it holds that

$$\begin{aligned} R^* - \Re(\eta(t)) &= R^* - \langle r, \eta_\delta\rangle + \langle r, \eta_\delta\rangle - \Re(\eta(t)) = O(\delta) + D_{\phi_\sigma}(\eta_\delta, \eta_0)t^{-1} \\ &= O(\delta) + \left(\phi_\sigma(\eta_\delta) - \phi_\sigma(\eta_0) - \langle\nabla\phi_\sigma(\eta_0), \eta_\delta - \eta_0\rangle\right) t^{-1} \\ &= O(\delta) + O(\log(\delta) + 1)t^{-1}. \end{aligned}$$

Setting $\delta = t^{-1}$ we obtain $R^* - \Re(\eta(t)) = O(t^{-1}) + O((\log(t^{-1}) + 1)t^{-1}) = O(\log(t)t^{-1})$ for $t \to \infty$. For $\sigma \in (2, \infty)$ the calculation follows in analogue fashion. Noting that $\operatorname{dist}(\eta(t), S) \sim R^* - \Re(\eta(t))$ finishes the proof. □

Theorem 4.21 and Theorem 4.22 show global convergence of σ-NPG and Kakade's NPG flows for unregularized rewards. Note that the reason why this is possible is that one does not work with a regularized objective but rather with a geometry arising from a regularization but with the original linear objective. For $\sigma < 1$ the flow may reach a face of the feasible set in finite time; see Figure 4.3. For $\sigma \geq 3$ Theorem 4.22 is uninformative since $t^{\sigma-3}$ is non decreasing but $\Re(\eta(t))$ is non increasing as

$$\partial_t\Re(\eta(t)) = g_{\eta(t)}(\partial_t\eta(t), \nabla^g\Re(\eta(t))) = g_{\eta(t)}(\nabla^g\Re(\eta(t)), \nabla^g\Re(\eta(t))) \geq 0.$$

However, by Lemma 4.15 the flows converge in objective value since they are well posed as the functions ϕ_σ are Legendre-type functions for $\sigma \geq 1$; see Example 4.17. It would be interesting to expand the theoretical analysis to clarify the convergence rate in this particular case. For larger σ the plateau problem becomes more pronounced, as can be seen in Figure 4.3. Further, if multiple maximizers exist and the objective $\Re$ is linear one can show that the trajectory converges towards the maximizer that is closest to the initial point η_0 with respect to the Bregman divergence [11, Corollary 4.8].

Faster rates for $\sigma \in [1, 2)$ and Kakade's NPG. Now we obtain improved and even linear convergence rates for Kakade's and Morimura's NPG flow for unregularized problems. To this end, we first formulate the following general result.

Lemma 4.23 (Convergence rates for gradient flow trajectories). *Consider Setting 4.14, assume that there is a global solution $\eta\colon [0, \infty) \to \operatorname{int}(\mathcal{N})$ of the Hessian gradient flow* (4.18). *Assume that there is an optimizer $\eta^* \in \mathcal{N}$ and assume that $\eta(t) \to \eta^*$ for $t \to \infty$ and that there exist $\omega \in (0, \infty)$, $\tau \in [1, \infty)$ and $T \geq 0$ such that*

$$\Re(\eta^*) - \Re(\eta(t)) \geq \omega D_\phi(\eta^*, \eta(t))^\tau \quad \text{for all } t \geq T. \tag{4.28}$$

Then there is a constant $c > 0$, possibly depending on T and $\eta(0)$, such that

(*i*) *if $\tau = 1$, then $D_\phi(\eta^*, \eta(t)) \leq ce^{-\omega t}$ for all $t \geq 0$,*

(*ii*) *if $\tau > 1$, then $D_\phi(\eta^*, \eta(t)) \leq ct^{-1/(\tau-1)}$ for all $t \geq 0$.*

The lower bound (4.28) can be interpreted as a form of strong convexity under which the objective value controls the Bregman divergence and hence convergence in objective value implies convergence of the trajectories in the sense of the Bregman divergence.

Proof. The statement of this result can be found in [11, Proposition 4.9], where however stronger assumptions are made and hence we provide a short proof. Using (4.21) and (4.28) we find that for $t \geq T$ it holds that

$$\partial_t D_\phi(\eta^*, \eta(t)) \leq \Re(\eta(t)) - \Re(\eta^*) \leq -\omega D_\phi(\eta^*, \eta(t))^\tau.$$

A standard integration of the differential inequality yields the claim. □

In order to apply this result to natural policy gradient flows we need to bound the respective Bregman divergences by the suboptimality in the linear objective. For this we establish the following two lemmata where we bound the 1-norm by the suboptimality gap as well as the KL-divergence between probability distributions by the 1-norm.

Lemma 4.24. *Consider a polytope $P = \mathbb{R}^d$ and a vertex $x^* \in P$ that is the unique maximizer of the linear function $x \mapsto \langle v, x\rangle$ over P. Let us denote the set of neighboring vertices*[2] *of x^* by $N(x^*)$. Then*

$$\Delta := \min\left\{\frac{\langle v, x^* - x\rangle}{\|x^* - x\|_1} : x \in N(x^*)\right\} > 0 \tag{4.29}$$

and satisfies

$$\langle v, x^*\rangle - \langle v, x\rangle \geq \Delta \cdot \|x^* - x\|_1 \quad \textit{for all } x \in P. \tag{4.30}$$

Proof. Note that since x^* is the unique maximizer of $x \mapsto \langle v, x\rangle$ over P it holds that $\langle v, x^* - x\rangle > 0$ for every neighboring vertex $x \in N(x^*)$, which implies $\Delta > 0$. Further, the polytope P is contained in the cone

$$C = \left\{x^* + \sum_{y \in N(x^*)} \alpha_y (y - x^*) : \alpha_y \geq 0 \text{ for all } y \in N(x^*)\right\}$$

generated by the edges containing x^*, see [318, Lemma 3.6]. Hence, for $x \in P$ there are non negative weights $\alpha_y \geq 0$ for $y \in N(x^*)$ such that

$$x = x^* + \sum_{y \in N(x^*)} \alpha_y (y - x^*).$$

Now we compute

$$\langle v, x^*\rangle - \langle v, x\rangle = \sum_{y \in N(x^*)} \alpha_y \langle v, x^* - y\rangle \geq \Delta \cdot \sum_{y \in N(x^*)} \alpha_y \|x^* - y\|_1. \tag{4.31}$$

Further, by the triangle inequality we have

$$\|x^* - x\|_1 = \left\| \sum_{y \in N(x^*)} \alpha_y (y - x^*)\right\|_1 \leq \sum_{y \in N(x^*)} \alpha_y \|y - x^*\|_1,$$

which completes the proof. □

Lemma 4.25. *Consider a finite set $\mathcal{X}$ and a probability distribution $\mu \in \Delta_{\mathcal{X}}$ as well $\varepsilon \in (0, 1)$ and set*

$$\delta := \frac{\varepsilon}{1 + \varepsilon} \cdot \min\left\{\mu_x : x \in \mathcal{X} \textit{ and } \mu_x > 0\right\} > 0. \tag{4.32}$$

[2] Two vertices of a polytope P are called *neighbors* if their convex hull is a face of P.

Then for all $\nu \in \Delta_{\mathcal{X}}$ satisfying $\|\mu - \nu\|_\infty \leq \delta$ it holds that

$$D_{KL}(\mu, \nu) \leq \left(\frac{1}{2} + \varepsilon\right) \cdot \|\mu - \nu\|_1. \tag{4.33}$$

Proof. We bound the individual summands in the KL-divergence

$$D_{KL}(\mu, \nu) = \sum_{x \in \mathcal{X}} \mu_x \log\left(\frac{\mu_x}{\nu_x}\right) = \sum_{x \in X} \mu_x \log\left(\frac{\mu_x}{\nu_x}\right),$$

where $X := \{x \in \mathcal{X} : \mu_x > 0\}$. If $\mu_x, \nu_x > 0$ then

$$\begin{aligned} \mu_x \log\left(\frac{\mu_x}{\nu_x}\right) &= \mu_x \left(\log(\nu_x + (\mu_x - \nu_x)) - \log(\nu_x)\right) \\ &\leq \mu_x \left(\log(\nu_x) + \frac{\mu_x - \nu_x}{\nu_x} - \log(\nu_x)\right) \\ &= (\mu_x - \nu_x) \cdot \frac{\mu_x}{\nu_x}, \end{aligned} \tag{4.34}$$

where we used the convexity $\log(t + h) \leq \log(t) + h/t$ for $t > 0, t + h > 0$. If $\|\mu - \nu\|_\infty \leq \delta$ then

$$\nu_x \geq \mu_x - \delta \geq \mu_x \left(1 - \frac{\varepsilon}{1 + \varepsilon}\right) = \frac{\mu_x}{1 + \varepsilon}$$

as well as

$$\nu_x \leq \mu_x + \delta \leq \mu_x \left(1 + \frac{\varepsilon}{1 + \varepsilon}\right) \leq \mu_x \left(1 + \frac{\varepsilon}{1 - \varepsilon}\right) = \frac{\mu_x}{1 - \varepsilon}$$

and therefore $1 - \varepsilon \leq \frac{\mu_x}{\nu_x} \leq 1 + \varepsilon$. If $\mu_x \geq \nu_x$ then

$$(\mu_x - \nu_x) \cdot \frac{\mu_x}{\nu_x} \leq (1 + \varepsilon)(\mu_x - \nu_x) = \mu_x - \nu_x + \varepsilon|\mu_x - \nu_x|$$

and if $\mu_x < \nu_x$ then

$$(\mu_x - \nu_x) \cdot \frac{\mu_x}{\nu_x} \leq (1 - \varepsilon)(\mu_x - \nu_x) = \mu_x - \nu_x + \varepsilon|\mu_x - \nu_x|. \tag{4.35}$$

Together with (4.34) summing over x yields

$$D_{KL}(\mu, \nu) \leq \sum_{x \in X} (\mu_x - \nu_x) + \varepsilon \sum_{x \in X} |\mu_x - \nu_x| \leq \sum_{x \in X} (\mu_x - \nu_x) + \varepsilon \cdot \|\mu - \nu\|_1 \tag{4.36}$$

and hence it remains to estimate the first part. Set $X^c := \mathcal{X} \setminus X$ then

$$\sum_{x \in X} (\mu_x - \nu_x) = \sum_{x \in \mathcal{X}} (\mu_x - \nu_x) - \sum_{x \in X^c} (\mu_x - \nu_x) = -\sum_{x \in X^c} (\mu_x - \nu_x) = \sum_{x \in X^c} |\mu_x - \nu_x|$$

since $\mu_x = 0$ for $x \in X^c$. Now we can estimate

$$2 \sum_{x \in X} (\mu_x - \nu_x) = \sum_{x \in X} (\mu_x - \nu_x) + \sum_{x \in X^c} |\mu_x - \nu_x| \leq \|\mu - \nu\|_1.$$

Together with (4.36) this yields (4.33). □

Combining Lemma 4.24 and Lemma 4.25 yields $cD_{KL}(\mu^*, \mu) \leq \langle v, \mu^* \rangle - \langle v, \mu \rangle$ for all $c \in (0, 2\Delta)$ if $P \subseteq \Delta_{\mathcal{X}}$. This estimate of the KL-divergence of a point to the solution of a linear program in terms of the optimality gap establishes the condition (4.28) in Lemma 4.23 and hence we obtain $O(e^{-ct})$ convergence of the gradient flow with respect to the Fisher metric.

Theorem 4.26 (Linear convergence of unregularized Kakade's NPG flow). *Consider Setting 4.14, where $\phi = \phi_C$ is the conditional entropy defined in (4.9) and assume that there is a unique maximizer η^* of the unregularized reward $\mathfrak{R}$ and denote its state-marginal by ρ^*. Then for any $c_1 \in (0, 2\Delta)$ there is a constant $c_2 = c_2(\eta_0, c_1)$ such that*

$$D_{\phi_C}(\eta^*, \eta(t)) = \sum_s \rho^*(s) D_{KL}(\pi^*(\cdot|s), \pi_t(\cdot|s)) \leq c_2 e^{-c_1 t} \tag{4.37}$$

and

$$R^* - \mathfrak{R}(\eta(t)) \leq \frac{2c_2|\mathcal{S}||\mathcal{A}| \cdot \|r\|_\infty}{(1-\gamma)\min_s \rho^*(s)} \cdot e^{-c_1 t} \tag{4.38}$$

for all $t \geq 0$, where

$$\Delta = \min \left\{ \frac{\langle r, \eta^* - \eta \rangle}{\|\eta^* - \eta\|_1} : \eta \in N(\eta^*) \right\} > 0 \tag{4.39}$$

and $N(\eta^*)$ denotes the neighboring vertices of η^* in the state-action polytope $\mathcal{N}$.

Proof. Let ϕ_C denote the conditional entropy, so that

$$D_{\phi_C}(\eta^*, \eta) = D_{KL}(\eta^*, \eta) - D_{KL}(\rho^*, \rho) \leq D_{KL}(\eta^*, \eta)$$

By Lemma 4.24 and Lemma 4.25 for any $c_1 \in (0, 2\Delta)$ it holds that

$$c_1 D_{\phi_C}(\eta^*, \eta) \leq \mathfrak{R}(\eta^*) - \mathfrak{R}(\eta).$$

Hence, Lemma 4.23 guarantees that

$$D_{\phi_C}(\eta^*, \eta(t)) = c_2 e^{-c_1 t}$$

for some $c_2 = c_2(\eta_0, c_1) > 0$ and it remains to estimate $\mathfrak{R}(\eta^*) - \mathfrak{R}(\eta) = \|r\|_\infty \cdot \|\eta^* - \eta\|_1$ by the conditional relative entropy $D_{\phi_C}(\eta^*, \eta)$. Note that π^* is a deterministic policy and hence we can write $\pi^*(a_s^*|s) = 1$ and estimate

$$\begin{aligned} D_{\phi_C}(\eta^*, \eta) &= \sum_s \rho^*(s) D_{KL}(\pi^*(\cdot|s), \pi^*(\cdot|s)) = -\sum_s \rho^*(s) \log(\pi(a_s^*|s)) \\ &\geq \sum_s \rho^*(s)(1 - \pi(a_s^*|s)) = 2^{-1} \sum_s \rho^*(s) \|\pi^*(\cdot|s) - \pi(\cdot|s)\|_1 \\ &\geq 2^{-1} \left(\min_s \rho^*(s) \right) \cdot \|\pi^* - \pi\|_1, \end{aligned} \tag{4.40}$$

where we have used $\log(t) \leq t - 1$ as well as

$$\begin{aligned} \|\pi^*(\cdot|s) - \pi(\cdot|s)\|_1 &= \sum_{a \neq a_s^*} |\pi^*(a|s) - \pi(a|s)| + |\pi^*(a|s) - \pi(a|s)| \\ &= \sum_{a \neq a_s^*} \pi(a|s) + (1 - \pi(a_s^*|s)) = 2(1 - \pi(a_s^*|s)). \end{aligned}$$

By Lemma 3.13 it holds that

$$\|\eta^\pi - \eta^{\pi'}\|_1 \leq |\mathcal{S}||\mathcal{A}| \cdot \|\eta^\pi - \eta^{\pi'}\|_\infty \leq \frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma} \cdot \|\pi - \pi'\|_1$$

and hence

$$\|\eta^* - \eta(t)\|_1 \leq \frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma} \cdot \|\pi^* - \pi(t)\|_1 \leq \frac{2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\min_s \rho^*(s)} \cdot D_{\phi_C}(\eta^*, \eta(t)). \tag{4.41}$$

Together with $R^* - \mathfrak{R}(\eta(t)) = \langle r, \eta^* - \eta(t)\rangle \leq \|r\|_\infty \cdot \|\eta^* - \eta(t)\|_1$ this concludes the proof. □

In the proof we rely on the estimate $c_1 D_{\phi_C}(\eta^*, \eta) \leq c_1 D_{KL}(\eta^*, \eta) \leq R^* - \langle r, \eta\rangle$, which might not be tight and an estimate of the form $cD_{\phi_C}(\eta^*, \eta) \leq R^* - \langle r, \eta\rangle$ for a constant $c > c_1$ would improve Theorem 4.26.

Theorem 4.27 (Improved convergence rates for σ-NPG flow). *Consider Setting 4.14, where $\phi = \phi_\sigma$ for some $\sigma \in [1, 2)$ as defined in* (4.1)*, and assume that there is a unique maximizer η^* of the unregularized reward $\mathfrak{R}$ with state-marginal ρ^* and consider $\Delta > 0$ defined in* (4.39)*. Then for any $c_1 \in (0, 2\Delta)$ there are constants $c_2 = c_2(\eta_0, c_1), c_3 = c_3(\eta_0) > 0, c_4 = c_4(\eta_0) > 0$ such that*

$$D_{KL}(\eta^*, \eta(t)) \leq c_2 e^{-c_1 t} \quad \textit{and } R^* - \mathfrak{R}(\eta(t)) \leq \frac{2c_2|\mathcal{S}||\mathcal{A}|\cdot\|r\|_\infty}{(1-\gamma)\min_s \rho^*(s)} \cdot e^{-c_1 t} \quad \textit{if } \sigma = 1 \tag{4.42}$$

for all $t \geq 0$ and

$$D_\sigma(\eta^*, \eta(t)) \leq c_3 t^{-(2-\sigma)/(\sigma-1)} \quad \textit{and } R^* - \mathfrak{R}(\eta(t)) \leq c_4 t^{-1/(\sigma-1)} \quad \textit{if } \sigma \in (1, 2) \tag{4.43}$$

for all $t \geq 0$, where D_σ denotes the Bregman divergence induced by ϕ_σ.

Proof. The case $\sigma = 1$ can be treated similarly to the case of Kakade's NPG, where by Lemma 4.24 and Lemma 4.25 one obtains

$$c_1 \cdot D_{KL}(\eta^*, \eta) \leq \mathfrak{R}(\eta^*) - \mathfrak{R}(\eta)$$

for $c_1 \in (0, 2\Delta)$ and η in a neighborhood of η^* and hence Lemma 4.23 yields the existence of $c_2 = c_2(\eta_0, c_1) > 0$ such that

$$D_{KL}(\eta^*, \eta(t)) \leq c_2 e^{-c_1 t}.$$

Since $D_{\phi_C} \leq D_{KL}$, (4.41) yields

$$\mathfrak{R}(\eta^*) - \mathfrak{R}(\eta(t)) \leq \frac{2c_2|\mathcal{S}||\mathcal{A}| \cdot \|r\|_\infty}{(1-\gamma)\min_s \rho^*(s)} \cdot D_{KL}(\eta^*, \eta(t)). \tag{4.44}$$

For $\sigma \in (1, 2)$ we show that (4.28) holds for $\tau = (2-\sigma)^{-1} \geq 1$. Recall that

$$D_\sigma(\eta^*, \eta) = \sum_{s,a} \frac{\eta^*(s,a)^{2-\sigma}}{(1-\sigma)(2-\sigma)} - \sum_{s,a} \frac{\eta(s,a)^{2-\sigma}}{(1-\sigma)(2-\sigma)} - \sum_{s,a} \frac{\eta(s,a)^{1-\sigma}(\eta^*(s,a) - \eta(s,a))}{1-\sigma}.$$

We can bound every individual summand by $O(|\eta^*(s,a) - \eta(s,a)|)$ if $\eta^*(s,a) > 0$ and $O(|\eta^*(s,a) - \eta(s,a)|^{2-\sigma})$ if $\eta^*(s,a) = 0$ for $\eta \to \eta^*$ respectively. Overall, this shows that

$$D_\sigma(\eta^*, \eta) = O(\|\eta^* - \eta\|^{2-\sigma}) = O((\mathfrak{R}(\eta^*) - \mathfrak{R}(\eta))^{2-\sigma}) \quad \text{for } \eta \to \eta^*,$$

where the last estimate holds since η^* is the unique minimizer of the linear function $\mathfrak{R}$ over the polytope $\mathcal{N}$. By Lemma 4.23 we obtain

$$D_\sigma(\eta^*, \eta(t)) = O(t^{-1/(\tau-1)}) = O(t^{-(2-\sigma)/(\sigma-1)}).$$

It remains to estimate the value of $\mathfrak{R}$ by means of the Bregman divergence D_σ. For this, we note that $\mathfrak{R}(\eta^*) - \mathfrak{R}(\eta) = \|r\|_\infty \cdot \|\eta^* - \eta\|_1)$ and estimate the individual terms. First, note that for $x \to y$ (with $x, y \geq 0$) it holds that

$$|x - y| = O\left(\left(\frac{y^{2-\sigma}}{(1-\sigma)(2-\sigma)} - \frac{x^{2-\sigma}}{(1-\sigma)(2-\sigma)} - \frac{x^{1-\sigma}(y-x)}{1-\sigma}\right)^{1/(2-\sigma)}\right).$$

For $y = 0$ this is immediate and for $y > 0$ the local strong convexity of $x \mapsto x^{2-\sigma}$ around y implies

$$\begin{aligned}|x-y| &= O\left(\left(y^{2-\sigma} - x^{2-\sigma} - (2-\sigma)x^{1-\sigma}(y-x)\right)^{1/2}\right) \\ &= O\left(\left(y^{2-\sigma} - x^{2-\sigma} - (2-\sigma)x^{1-\sigma}(y-x)\right)^{1/(2-\sigma)}\right)\end{aligned}$$

for $x \longrightarrow y$. Now, Jensen's inequality yields

$$\|\eta^* - \eta\|_1 = O(D_\sigma(\eta^*, \eta)^{1/(2-\sigma)}).$$

Overall, we obtain

$$\Re(\eta^*) - \Re(\eta(t)) = O(\|\eta^* - \eta(t)\|_1^{1/(2-\sigma)}) = O(t^{-1/(1-\sigma)}).$$

□

Compared to Theorem 4.22 the above Theorem 4.27 improves the convergence rate of $O(t^{-1})$ for parameters $\sigma \in [1, 2)$. Later, we conduct numerical experiments that indicate that the rates $O(t^{-1/(\sigma-1)})$ also hold for $\sigma \geq 2$ and are tight.

Remark 4.28 (Alternative bound). When bounding the suboptimality in terms of the KL divergence in order to obtain the bound (4.42) we use (4.44), i.e., we estimate $\|\eta^* - \eta\|_1$ in terms of the conditional relative entropy D_{ϕ_C}, which is trivially bounded by the KL divergence. Hence, if we directly bound $\|\eta^* - \eta\|_1$ by the KL divergence, we obtain a different bound, which can be tighter in certain instances. The optimal policy π^* corresponding to η^* is deterministic and hence for $s \in \mathcal{S}$ there is $a_s^* \in \mathcal{A}$ such that $\pi^*(a_s^*|s) = 1$. Using $\log(t) \leq \log(s) + \frac{t-s}{s}$ we estimate

$$D_{KL}(\eta^*, \eta) = \sum_s \eta^*(s, a_s^*) \log\left(\frac{\eta^*(s, a_s^*)}{\eta(s, a_s^*)}\right) \geq \sum_s \eta^*(s, a_s^*) \cdot \frac{\eta^*(s, a_s^*) - \eta(s, a_s^*)}{\eta^*(s, a_s^*)} = 1 - \sum_s \eta(s, a_s^*).$$

Note that η^* is the unique maximizer of the linear function

$$\ell\colon \mathcal{N} \longrightarrow \mathbb{R}, \quad \eta \mapsto \sum_s \eta(s, a_s^*)$$

since for suboptimal $\eta^\pi \in \mathcal{N}$ there is $s \in \mathcal{S}$ such that $\pi(a_s^*|s) < 1$ and hence $\eta(s, a_s^*) = \rho^\pi(s)\pi(a_s^*|s) < \rho^\pi(s)$ and thus $\ell(\eta) < 1$. Therefore, we can apply Lemma 4.24 and obtain

$$\|\eta^* - \eta\|_1 \leq \delta^{-1}(1 - \ell(\eta)) \leq \delta^{-1} D_{KL}(\eta^*, \eta),$$

where

$$\delta = \min\left\{\frac{1 - \sum_s \eta(s, a_s^*)}{\|\eta^* - \eta\|_1} : \eta \in N(\eta^*)\right\}.$$

In order to estimate δ we fix $\eta^\pi \in N(\eta^*)$ and note that $\pi \in N(\pi^*)$ is a deterministic policy with $\pi(a_s|s) = 1$ and $|\{s \in \mathcal{S} : a_s \neq a_s^*\}| = 1$, lets say $a_{s_0} \neq a_{s_0}^*$. Then

$$1 - \ell(\eta) = 1 - \sum_s \rho^\pi(s)\pi(a_s^*|s) = 1 - \sum_{s \neq s_0} \rho^\pi(s) = \rho^\pi(s_0) \geq (1-\gamma)\mu(s_0) \geq (1-\gamma)\min_s \mu(s).$$

Further, we have $\|\eta^* - \eta\|_1 \leq 2$ and thus $\delta \geq \frac{(1-\gamma)\min_s \mu(s)}{2}$. If $\min_s \mu(s) > 0$ this implies together with $D_{KL}(\eta^*, \eta(t)) \leq c_2 e^{-c_1 t}$ that

$$R^* - \Re(\eta(t)) \leq \|r\|_\infty \cdot \|\eta - \eta(t)\|_1 \leq \frac{2c_2\|r\|_\infty}{(1-\gamma)\min_s \mu(s)} \cdot e^{-c_1 t}. \tag{4.45}$$

This bound becomes tightest if we choose μ to be the uniform distribution where it evaluates to

$$\frac{2c_2|\mathcal{S}| \cdot \|r\|_\infty}{1-\gamma} \cdot e^{-c_1 t}. \tag{4.46}$$

In comparison, the bound (4.42) becomes tightest if ρ^* is the uniform norm in which it would evaluate to

$$\frac{2c_2|\mathcal{S}|^2|\mathcal{A}|\cdot\|r\|_\infty}{1-\gamma} \cdot e^{-c_1 t},$$

which is bigger compared to (4.46). However, the tighter bound (4.46) and also (4.45) requires the initial distribution μ to be the uniform distribution or have full support, repsectively. However, there are instances where this is not satisfied and the positivity Assumption 3.3 still holds, in which case the bound (4.42) remains valid.

Remark 4.29 (Non-unique optimizers). Both Theorem 4.26 and Theorem 4.27 are formulated under the assumption of unique optimizers since we use Lemma 4.24. The assumption that the linear function $\eta \mapsto \langle r, \eta \rangle$ possesses a unique optimizer over the state-action polytope $\mathcal{N}$ is satisfied for almost all $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. If there is a non trivial face F^* of optimizer the bound on the exponent 2Δ of the linear convergence deteriorates. However, in the case of non unique optimizers the gradient flow converges towards the Bregman projection η^* of the initial condition η_0 onto the set F^* of optimizers, which describes the implicit bias of the different methods, see [11, Corollary 4.8]. Further, the bounds (4.37), (4.38), (4.42) and (4.43) remain valid with the constant

$$\Delta = \min\left\{\frac{\langle r, \eta^* - \eta\rangle}{\|\eta^* - \eta\|_1} : \eta^* \in \operatorname{vert}(F^*), \eta \in N(\eta^*) \setminus F^*\right\} > 0, \tag{4.47}$$

where $\operatorname{vert}(F^*)$ denotes the set of vertices of F^*.

To see this, we follow the same strategy as in the proofs above and generalize Lemma 4.24 to non unique optimizers. Indeed, for

$$\Delta := \min\left\{\frac{\langle v, x^* - x\rangle}{\|x^* - x\|_1} : x \in N(x^*) \setminus F^*, x^* \in \operatorname{vert}(F^*)\right\},$$

where $\operatorname{vert}(F^*)$ denotes the vertex set of the face of optimizers we obtain

$$\Delta \cdot \min_{x* \in F^*} \|x^* - x\|_1 \leq f^* - \langle v, x\rangle,$$

where f^* denotes the optimal value attained on the face F^*. To simplify notation let us define the set $E := \{x - x^* : x \in N(x^*) \setminus F^*, x^* \in \operatorname{vert}(F^*)\}$ of edges such that exactly one of the two endpoints is contained in F^*. Then, the polytope P is contained in the cone

$$C = \left\{x^* + \sum_{e \in E} \alpha_e e : x^* \in F^*, \alpha_e \geq 0 \text{ for all } e \in E\right\}$$

and hence we can write $x \in P$ as $x = x^* + \sum_e \alpha_e e$ for some $x^* \in F^*$. Just like in the proof of Lemma 4.24 we obtain

$$\Delta \|x^* - x\|_1 \leq \Delta \sum_e \alpha_e \|e\|_1 \leq \sum_e \langle v, -e \rangle = \langle v, x^* \rangle - \langle v, x \rangle = f^* - \langle v, x \rangle.$$

Taking the minimum over x^* yields the claim.

Now we come back to the convergence of the flow $\eta(t)$ towards its Bregman projection η^* and consider the case $\sigma = 1$. Let $\hat{\eta}(t) \in F^*$ denote the $\|\cdot\|_1$ projection of $\eta(t)$ onto F^*, i.e., be such that

$$\|\eta(t) - \hat{\eta}(t)\|_1 = \min_{\eta' \in F^*} \|\eta' - \eta(t)\|_1.$$

Note that $\hat{\eta}(t) \to \eta^*$ since $\eta(t) \to \eta^*$. Using by Lemma 4.25 one can show that for any $c < 2$ and t large enough it holds that $D_{KL}(\hat{\eta}(t), \eta(t)) \leq c^{-1} \|\hat{\eta}(t) - \eta(t)\|_1$ and now we obtain

$$D_{KL}(\eta^*, \eta(t)) \leq D_{KL}(\hat{\eta}(t), \eta(t)) \leq c^{-1} \min_{\eta' \in F^*} \|\eta' - \eta(t)\|_1 \leq c^{-1} \Delta^{-1} (R^* - \langle r, \eta(t) \rangle)$$

for t large enough, which yields the strong convexity condition (4.28) along the trajectory. For $\sigma \in (1, 2)$ and Kakade's natural policy gradient the proof can be adapted analogously.

Numerical examples. We use the following example proposed by Kakade [153] and which was also used in [28, 206]. Computer code for all experiments is made available in `https://github.com/muellerjohannes/geometry-natural-policy-gradients`. We consider an MDP with two states s_1, s_2 and two actions a_1, a_2, with the transitions and instantaneous rewards shown in Figure 4.2.

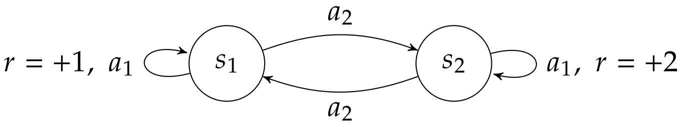

Figure 4.2. Transition graph and reward of the MDP example.

We adopt the initial distribution $\mu(s_1) = 0.2, \mu(s_2) = 0.8$ and work with a discount factor of $\gamma = 0.9$, whereas Kakade studied the mean reward case. Note however that the experiments can be performed for arbirtrarily large discount factor, where we chose a smaller factor since the correspondence between the policy polytope and the state-action polytope is clearer to see in the illustrations. We consider tabular softmax policies and plot the trajectories of vanilla PG, Kakade's NPG, and σ-NPG for the values

$$\sigma \in \{0, 0.5, 1, 1.5, 2, 3, 4\}$$

for 30 random (but the same for every method) initializations. We plot the trajectories in the state-action space (Figure 4.3) and in the policy polytope (Figure 4.4). In order to put the convergence results from this section into perspective, we plot the evolution of the optimality gap $R^* - R(\theta(t))$ (Figure 4.5). We use an adaptive step size Δt_k, which prevents the blowup of the parameters for $\sigma < 1$, and hence we do not consider the number of iterations but rather the sum of the step sizes as a measure for the time, $t_n = \sum_{k=1}^n \Delta t_k$. For vanilla PG and $\sigma \in (1, 2)$ we expect a decay at rate $O(t^{-1})$ [200] and $O(t^{-1/(\sigma-1)})$ by Theorem 4.27. Therefore we use a logarithmic (on both scales) plot for vanilla PG and $\sigma > 1$ and also indicate the predicted rate using a dashed gray line. For Kakade's and

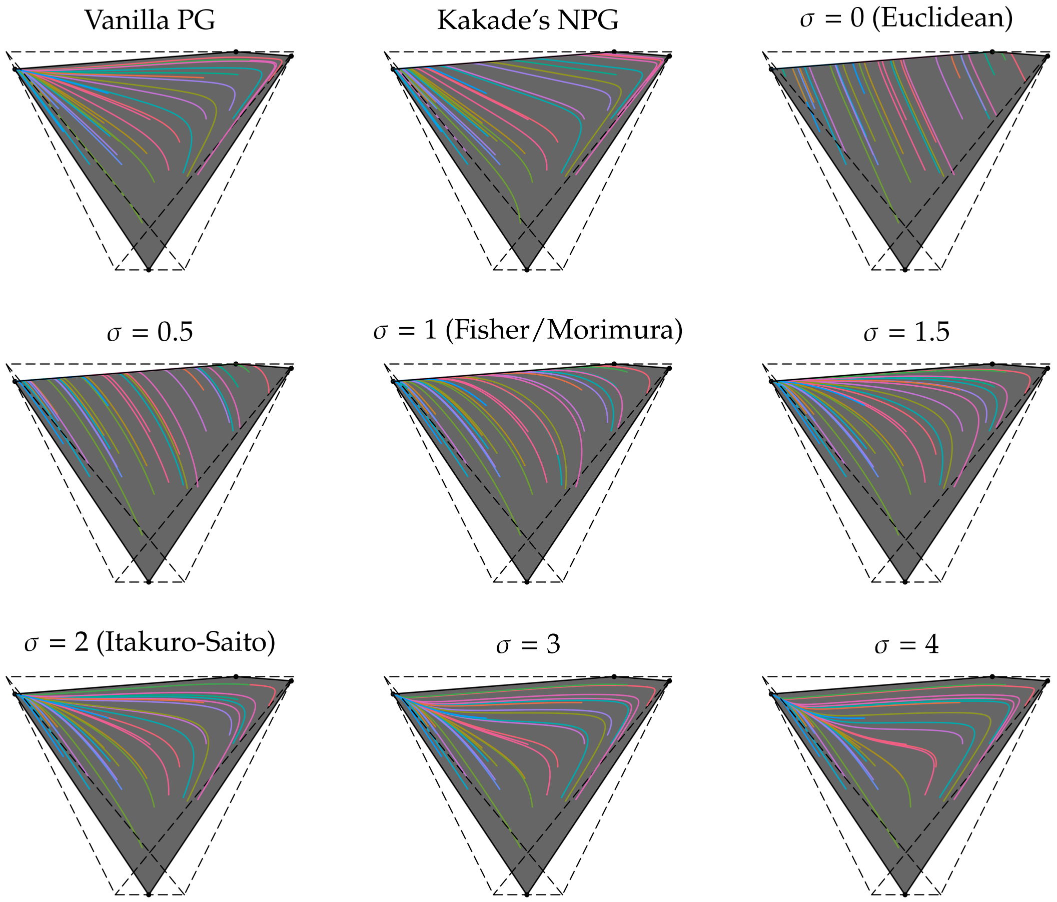

FIGURE 4.3. State-action trajectories for different PG methods, which are vanilla PG, Kakade's NPG and σ-NPG, where Morimura's NPG corresponds to $\sigma = 1$; the state-action polytope is shown in gray inside a three dimensional projection of the the simplex $\Delta_{\mathcal{S}\times\mathcal{A}}$; shown are trajectories with the same random 30 initial values for every method; the maximizer η^* is located at the upper left corner of the state-action polytope.

Morimuras NPG we expect linear convergence by Theorem 4.26 and 4.27 respectively and hence use a semi-logarithmic plot.

First, we note that for $\sigma \in \{-0.5, 0, 0.5\}$ the trajectories of σ-NPG flow hit the boundary of the state-action polytope $\mathcal{N}$, which is depicted in gray inside the simplex $\Delta_{\mathcal{S}\times\mathcal{A}}$. This is consistent with our analysis, since the functions ϕ_σ are Legendre type functions only for $\sigma \in [1, \infty)$ and hence only in this case the NPG flow is guaranteed to admit long time solutions. However, we observe finite-time convergence of the trajectories towards the global optimum (see Figure 4.5), which we suspect to be due to the error of temporal discretization.

For the other methods, namely vanilla PG, Kakade's NPG and σ-NPG with $\sigma \in [1, \infty)$, Theorem 4.22 and Theorem 4.21 show the global convergence of the gradient flow trajectories, which we also observe both in state-action space and in policy space (see Figures 4.3 and 4.4 respectively). When considering the convergence in objective value we

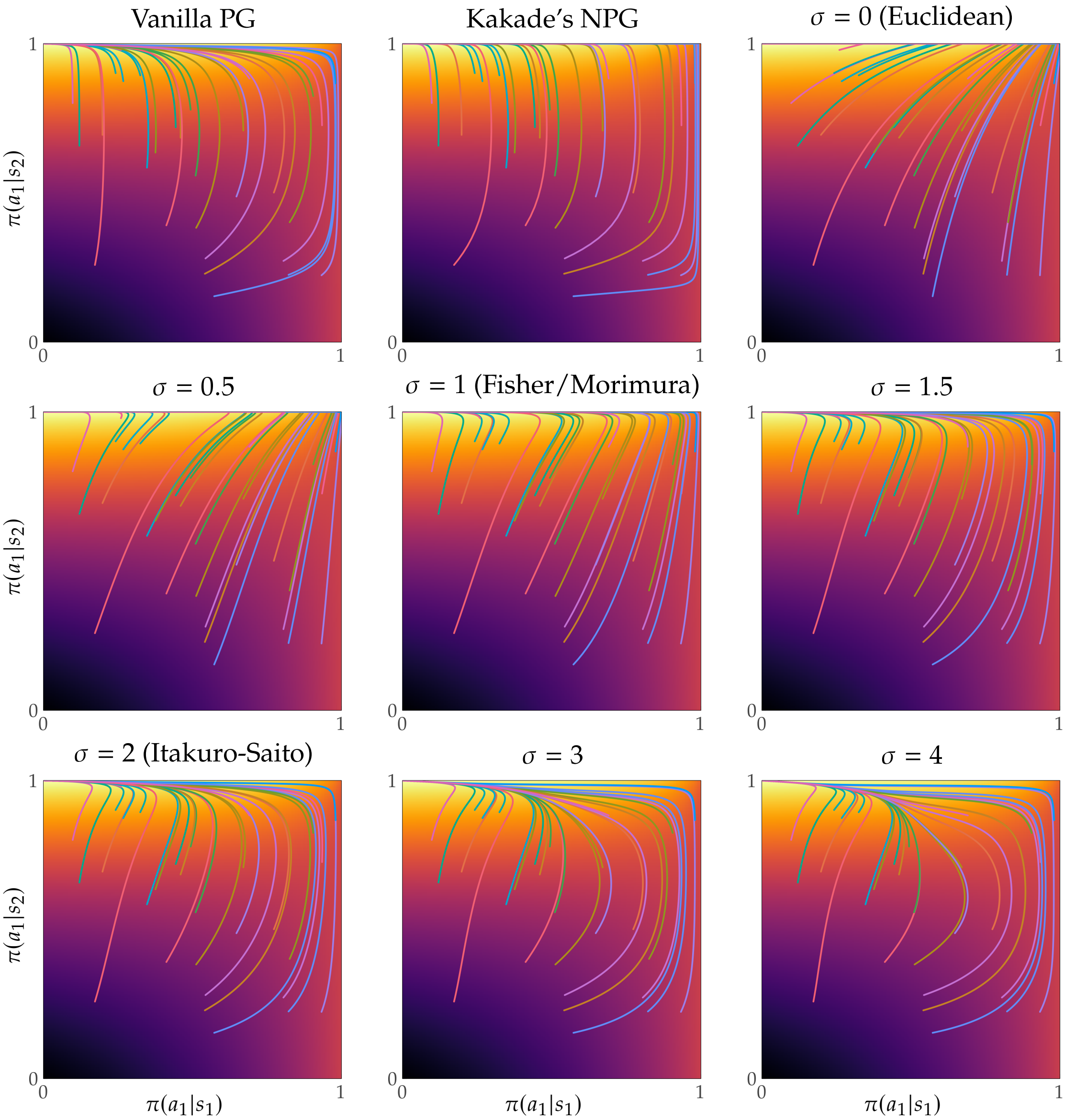

FIGURE 4.4. Plots of the trajectories of the individual methods inside the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S}} \cong [0,1]^2$; additionally, a heatmap of the reward function $\pi \mapsto R(\pi)$ is shown; the maximizer π^* is located at the upper left corner of the policy polytope.

observe that both Kakade's and Morimura's NPG exhibit a linear rate of convergence as asserted by Theorem 4.26 and Theorem 4.27, whereby Kakade's NPG appears to have more severe plateaus in some examples. Further, we compute the constant Δ given in (4.39). To do this, we note that the optimal policy $\pi^*\Delta_{\mathcal{A}}^{\mathcal{S}}$ is the deterministic policy given by $\pi^*(a_2|s_1) = \pi^*(a_1|s_2) = 1$. The two neighboring policies of π^* are the two policies $\pi_1, \pi_2 \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ that agree with π^* on one of the two states and the two neighboring vertices

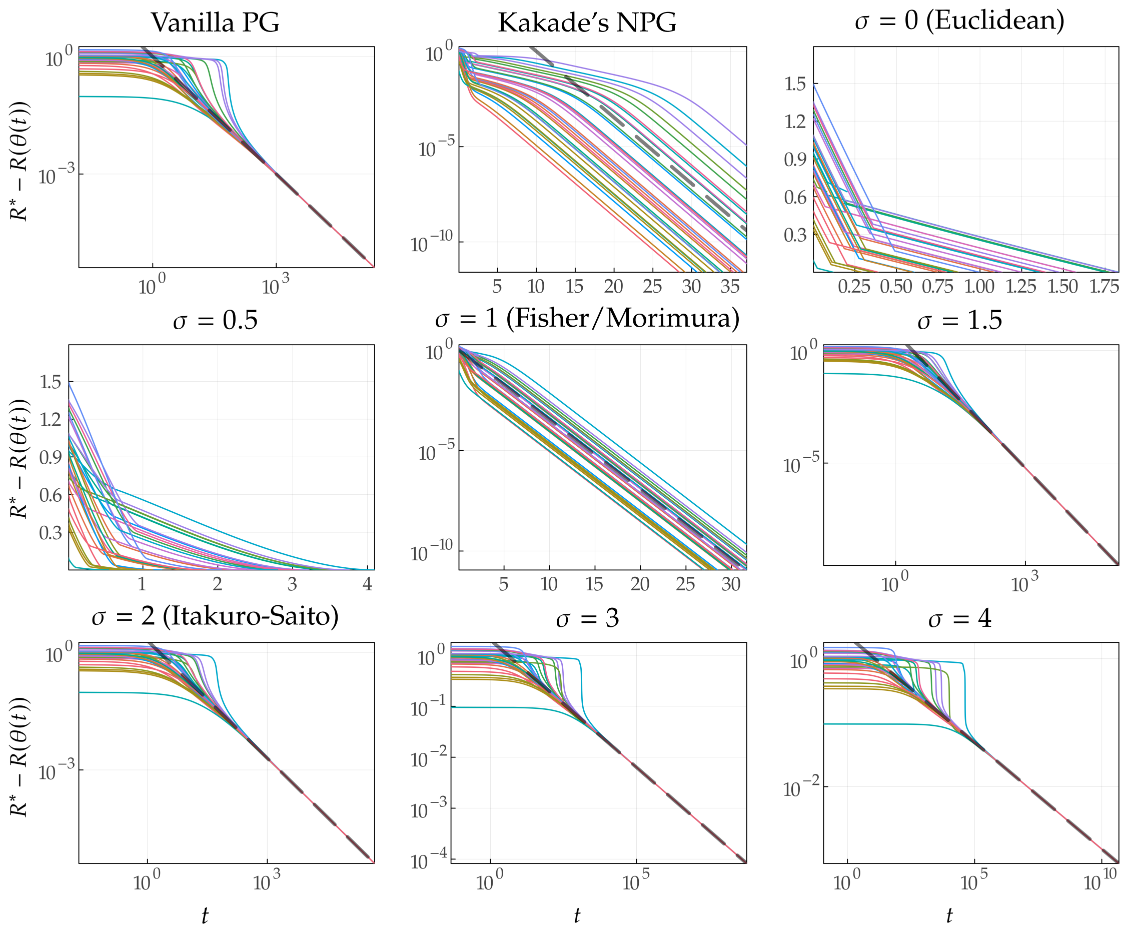

Figure 4.5. Plot of the optimality gaps $R^* - R(\theta(t))$ during optimization; note that for vanilla PG and $\sigma > 1$ these are log-log plots since we expect a decay like t^{-1} and $t^{-1/(\sigma-1)}$ respectively, which are shown as a dashed gray line; Kakade's and Morimura's NPG are at a log plot since we expect a linear convergence of (almost) $O(e^{-2\Delta t})$; finally, for $\sigma < 1$ we observe finite time convergence.

of η^* are the corresponding state-action frequencies η^{π_1} and η^{π_2} and hence

$$\Delta = \min\left(\frac{R^* - R(\pi_1)}{\|\eta^* - \eta^{\pi_1}\|_1}, \frac{R^* - R(\pi_2)}{\|\eta^* - \eta^{\pi_2}\|_1}\right) = 0.4$$

in our example. Since Theorem 4.26 and Theorem 4.27 guarantee the exponential convergence $O(e^{-ct})$ of Kakade's and Morimura's NPG flow for $c \in (0, 2\Delta)$ we show the guaranteed decay $O(e^{-2\Delta t}) = O(e^{-0.8t})$ in Figure 4.5 and observe that it matches the observed convergence. Kakade's natural policy gradient with constant step size $\delta > 0$ was shown to converge linearly at speed $O(e^{-\kappa\delta k})$, where κ depends on the minimal suboptimality of individual actions [156]. More precisely, in the case of a unique optimizer corresponding to a deterministic policy selecting a_s^* in state s we have

$$\kappa = (1-\gamma)^{-1} \min_s \min_{a \neq a_s^*} V^*(s) - Q^*(s,a),$$

which evaluates to 0.8 in our specific example. It is unclear, whether the two rates agree in general as our method always guarantees the same rates for Kakade's and Morimura's natural policy gradient.

For vanilla PG and $\sigma > 1$ we observe a sublinear convergence rate of $O(t^{-1})$ and $O(t^{-1/(\sigma-1)})$ respectively, which are shown via dashed gray lines in each case. This confirms the convergence rate $O(t^{-1})$ for vanilla PG [200] and indicates that the rate $O(t^{-1/(\sigma-1)})$ shown for $\sigma \in (1,2)$ is also valid in the regime $\sigma \geq 2$. Finally, we observe that larger σ appears to lead to more severe plateaus, which is apparent in the convergence in objective and also from the evolution in policy space and in state-action space.

4.3.3. Linear convergence of regularized Hessian NPG flows. It is known both empirically and theoretically that strictly convex regularization in state-action space yields linear convergence in reward optimization for vanilla and Kakade's natural policy gradients [200, 71]. Using Lemma 4.23 we generalize the result for Kakade's NPG and provide a result giving the linear convergence for general Hessian NPG.

Theorem 4.30 (Linear convergence for regularized problems). *Consider Setting 4.14 and let ϕ be a Legendre type function and by $\mathfrak{R}_\lambda(\eta) = \langle r, \eta\rangle - \lambda\phi(\eta)$ denote the regularized reward for some $\lambda > 0$ and fix an $\eta_0 \in \operatorname{int}(\mathcal{N})$ and assume that the global maximizer η^*_λ of $\mathfrak{R}_\lambda$ over $\mathcal{N}$ lies in the interior $\operatorname{int}(\mathcal{N})$. Denote the global solution of the Hessian gradient flow* (4.18) *with respect to the regularized reward $\mathfrak{R}_\lambda$ and the Hessian geometry induced by ϕ by $\eta\colon [0,\infty) \to \mathcal{N}$. For any $c_1 \in (0,\lambda)$ there exists a constant $c_2 = c_2(\eta_0, c_1) > 0$ such that*

$$D_\phi(\eta^*_\lambda, \eta(t)) \leq c_2 e^{-c_1 t} \quad \textit{for all } t \geq 0. \tag{4.48}$$

In particular, for any $\kappa \in (\kappa_c, \infty)$ this implies

$$R^*_\lambda - \mathfrak{R}_\lambda(\eta(t)) \leq \kappa\lambda c_2 e^{-c_1 t}, \tag{4.49}$$

for t large enough, where κ_c denotes the condition number of $\nabla^2\phi(\eta^*_\lambda)$.

Proof. We first recall that by Lemma 4.15 it holds that $\mathfrak{R}(\eta(t)) \to \mathfrak{R}(\eta^*_\lambda)$ and the uniqueness of the maximizer implies $\eta(t) \to \eta^*_\lambda \in \operatorname{int}(\mathcal{N})$. Note that

$$D_\phi(\eta^*_\lambda, \eta) = \lambda^{-1} D_{\lambda\phi}(\eta^*_\lambda, \eta) = \lambda^{-1} D_{-\mathfrak{R}_\lambda}(\eta^*_\lambda, \eta).$$

By Lemma 4.31 for $\omega \in (0,1)$ there is a neighborhood N_ω of η^* such that

$$\mathfrak{R}_\lambda(\eta^*_\lambda) - \mathfrak{R}_\lambda(\eta) \geq \omega D_{-\mathfrak{R}_\lambda}(\eta^*_\lambda, \eta) = \lambda\omega D_\phi(\eta^*_\lambda, \eta), \tag{4.50}$$

for $\eta(t) \in N_\omega$ and hence for t large enough. Now Lemma 4.23 shows the linear convergence

$$D_\phi(\eta^*_\lambda, \eta(t)) \leq c_2(\eta_0, c_1) e^{-c_1 t}$$

of the trajectory in the Bregman divergence. For $m, M > 0$ such that $mI \prec \nabla^2\phi(\eta^*_\lambda) \prec MI$ we can estimate

$$R^*_\lambda - \mathfrak{R}_\lambda(\eta(t)) = \mathfrak{R}_\lambda(\eta^*_\lambda) - \mathfrak{R}_\lambda(\eta(t)) \leq \frac{\lambda M}{2} \cdot \|\eta^*_\lambda - \eta(t)\|^2 \leq \frac{\lambda M}{m} \cdot D_\phi(\eta^*_\lambda, \eta)$$

for $\eta(t)$ close to η^*_λ, where we used that ϕ is m strongly convex in a neighborhood of η^*_λ. □

In the proof of the previous theorem we used the following lemma.

Lemma 4.31. *Let ϕ be a strictly convex function defined on an open convex set $\Omega \subseteq \mathbb{R}^d$ with unique minimizer x^*. Then for any $\omega \in (0,1)$ there is a neighborhood N_ω of x^* such that*

$$\phi(x) - \phi(x^*) \geq \omega D_\phi(x^*, x) \quad \textit{for all } x \in N_\omega. \tag{4.51}$$

Proof. Set $f(x) := D_\phi(x^*, x)$ and $g(x) := D_\phi(x, x^*)$. It holds that $f(x^*) = g(x^*) = 0$ and since both functions are non-negative $\nabla f(x^*) = \nabla g(x^*) = 0$. Further, since $\nabla\phi(x^*) = 0$ we have $g(x) = \phi(x) - \phi(x^*)$. By (4.5) we have $\nabla^2 f(x^*) = \nabla^2 g(x^*) = \nabla^2\phi(x^*)$ and Taylor extension yields

$$\begin{aligned} f(x) &= (x - x^*)^\top \nabla^2\phi(x^*)(x - x^*) + o(\|x - x^*\|^2) \\ &= g(x) + o(\|x - x^*\|^2) \\ &= \phi(x) - \phi(x^*) + o(\|x - x^*\|^2). \end{aligned}$$

Hence, for any $\varepsilon > 0$ there is $\delta > 0$ such that for $x \in B_\delta(x^*)$ it holds that

$$f(x) \le \phi(x) - \phi(x^*) + \varepsilon\|x - x^*\|^2 \le \left(1 + \frac{2\varepsilon}{m}\right)(\phi(x) - \phi(x^*))$$

for any $m \in (0, \lambda_{\min}(\nabla^2\phi(x^*))$ in a possible smaller neighborhood as ϕ is m-strongly convex in a neighborhood around x^*. Setting $\varepsilon := m(\omega^{-1} - 1)/2$ yields the claim. □

Remark 4.32 (Location of maximizers). The condition that $\eta^*_\lambda \in \operatorname{int}(\mathcal{N})$ assumed in Theorem 4.30 is satisfied if the gradient blow-up condition from Definition 4.16 is slightly strengthened. Indeed, suppose that for any $\eta \in \partial\mathcal{N}$ there is a direction v such that $\eta + tv \in \operatorname{int}(\mathcal{N})$ for small t and such that $\partial_v\phi(\eta + tv) = v^\top\nabla\phi(\eta + tv) \to -\infty$ for $t \to 0$. If $\phi(\eta) = \infty$, surely $\eta \neq \eta^*$. To argue in the case that $\phi(\eta) < +\infty$, we note that $\partial_v\mathfrak{R}_\lambda(\eta + tv) \to +\infty$ and choose $t_0 > 0$ such that $\partial_v\mathfrak{R}_\lambda(\eta + t_0v) > 0$. Then by the concavity of $\mathfrak{R}_\lambda$ and continuity of $\mathfrak{R}_\lambda$ we have

$$\mathfrak{R}_\lambda(\eta) \le \mathfrak{R}_\lambda(\eta + t_0v) - t_0\partial_v\mathfrak{R}_\lambda(\eta + t_0v) < \mathfrak{R}_\lambda(\eta + t_0v),$$

and hence $\eta \neq \eta^*_\lambda$.

Now we elaborate the consequences of this general convergence result given in Theorem 4.30 for Kakade and σ-NPG flows.

Corollary 4.33 (Linear convergence of regularized Kakade's NPG flow). *Consider Setting 4.14, where $\phi = \phi_C$ is the conditional entropy defined in* (4.9) *and consider the regularized reward $\mathfrak{R}_\lambda = \langle r, \eta\rangle - \lambda\phi_C(\eta)$ for some $\lambda > 0$ and denote the maximizer of $\mathfrak{R}$ by η^*_λ and denote the global solution of the Hessian gradient flow* (4.18) *by $\eta\colon [0, \infty) \to \mathcal{N}$. For any $c_1 \in (0, \lambda)$ there exists a constant $c_2 = c_2(\eta_0, c_1) > 0$ such that*

$$D_\phi(\eta^*_\lambda, \eta(t)) \le c_2e^{-c_1t} \quad \textit{for all } t \ge 0. \tag{4.52}$$

In particular, for any $\kappa \in (\kappa_c, \infty)$ this implies

$$R^*_\lambda - \mathfrak{R}_\lambda(\eta(t)) \le \kappa\lambda c_2e^{-c_1t} \tag{4.53}$$

for t large enough, where κ_c denotes the condition number of $\nabla^2\phi_C(\eta^*_\lambda)$.

Proof. We want to use Remark 4.32. Recall that

$$\phi_C(\eta) = H(\eta) - H(\rho) = \sum_{s,a}\eta(s, a)\log(\eta(s, a)) - \sum_s \rho(s)\log(\rho(s)),$$

where $\rho(s) = \sum_a \eta(s, a)$ is the state marginal. Note that by Assumption 3.3 it holds that $\rho(s) > 0$. Let us consider a point on the boundary $\eta \in \partial\mathcal{N}$ then surely $\eta(s, a) = 0$ for some $s \in \mathcal{S}, a \in \mathcal{A}$ since $\eta \ge 0$ are the only inequalities of the state-action polytope, see

Theorem 3.5. Fix now any $v \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ such that $\eta_\varepsilon := \eta + \varepsilon v \in \operatorname{int}(\mathcal{N})$ for small $\varepsilon > 0$. Writing ρ_ε for the associated state marginal, we obtain

$$\partial_v \phi_C(\eta_\varepsilon) = \sum_{s,a} \log(\eta_\varepsilon(s,a)) + |\mathcal{S}|(|\mathcal{A}| - 1) - \sum_s \log(\rho_\varepsilon(s)) \to -\infty$$

for $\varepsilon \longrightarrow 0$. □

Corollary 4.34 (Linear convergence for regularized σ-NPG flow). *Consider Setting 4.14 with $\phi = \phi_\sigma$ for some $\sigma \in [1,\infty)$ and let $\mathfrak{R}_\lambda(\eta) = \langle r, \eta\rangle - \lambda\phi(\eta)$ denote the regularized reward and denote the maximizer of $\mathfrak{R}_\lambda$ by η^*_λ and fix an element $\eta_0 \in \operatorname{int}(\mathcal{N})$. Denote the global solution of the Hessian gradient flow* (4.18) *with respect to the regularized reward $\mathfrak{R}_\lambda$ and the Hessian geometry induced by ϕ by $\eta\colon [0,\infty) \to \mathcal{N}$. For any $c_1 \in (0,\lambda)$ there exists a constant $c_2 = c_2(\eta_0, c_1) > 0$ such that*

$$D_\phi(\eta^*_\lambda, \eta(t)) \le c_1 2 e^{-c_1 t} \quad \textit{for all } t \ge 0. \tag{4.54}$$

In particular, for any $\kappa \in (\kappa(\eta^*_\lambda), \infty)$ this implies

$$R^*_\lambda - \mathfrak{R}_\lambda(\eta(t)) \le \kappa^\sigma \lambda c_2 e^{-c_1 t} \tag{4.55}$$

for t large enough, where $\kappa(\eta^*_\lambda) = \frac{\max \eta^*_\lambda}{\min \eta^*_\lambda}$.

Proof. Again, we use Remark 4.32 to see that for the Legendre type functions ϕ_σ the unique maximizer η^*_λ of $\mathfrak{R}_\lambda$ lies in the interior of $\mathcal{N}$. Hence, it remains to compute the condition number, for which we note that $\nabla^2\phi_\sigma(\eta^*_\lambda) = \operatorname{diag}(\eta^*_\lambda)^{-\sigma}$, which yields the result. □

Remark 4.35 (Regularization error). Regularizing an optimization problem changes the optimization problem and usually also the optimizer. The introduced error can be estimated in terms of the regularization strength λ. For logarithmic barriers in state-action space this can be done using standard techniques for interior point methods [58, 2]. For entropic regularization in state-action space, the regularization error is studied in [294], and for the conditional entropy this is done in [200, 71].

The results above do not cover arbitrary combinations of Hessian geometries and regularizers. However, the proof of Theorem 4.30 can be adapted to this case, where the only part that requires adjustments is (4.50) that couples the regularized reward to the Bregman divergence. In principle, this can be extended to the case of regularizers that are different from the function inducing the Hessian geometry.

Numerical examples: The $\lambda \longrightarrow 0$ regime. Theorem 4.30 and its corollaries yield a linear convergence rate of order $O(e^{-\lambda t})$, where the bound deteriorates when the regularization strength λ is sent to zero, $\lambda \longrightarrow 0$. The bound $R^*_\lambda - R_k = O((1-\lambda\Delta t)^k)$ for entropy regularized NPG descent [71] exhibits a similar degradation for $\lambda \longrightarrow 0$. It is natural to expect that the convergence behavior for $\lambda \longrightarrow 0$ is similar to the convergence behavior for $\lambda = 0$, i.e., the unregularized case. Recall that Theorem 4.26 and Theorem 4.27 establish linear rates without regularization for Kakade's and Morimura's NPG and a sublinear rate $O(t^{-1/(\sigma-1)})$ for $\sigma \in (1,2)$.

To evaluate the convergence behavior for $\lambda \longrightarrow 0$ for a specific NPG method we apply it to a collection of small regularization strengths with 10 different random initializations. Here, we revisit Kakade's example that was already used in Subsection 4.3.2 for unregularized problems. For every individual run we estimate the exponent c in the linear convergence rate $R^* - R(\theta(t)) = O(e^{-ct})$ via linear regression after a logarithmic transformation. Here, we take the iterates where the optimality gap $R^* - R(\theta)$ lies between 10^{-10} and 10^{-5}. In Figure 4.6 we present the mean of the estimated convergence rates for Kakade's and Morimura's NPG as well as for σ NPG for $\sigma = 1.5$.

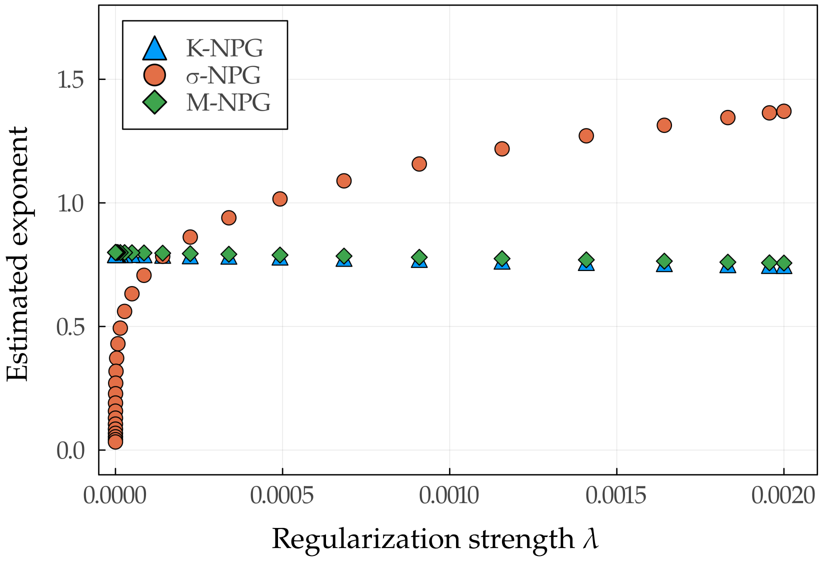

FIGURE 4.6. Shown are the estimated exponents $c > 0$ when fitting an exponential decay $O(e^{-ct})$ to the suboptimality gap $R^* - R(\theta(t))$ for different NPG methods – Kakade, Morimura and $\sigma = 1.5$ – and for different regularization strengths λ.

For both Kakade's and Morimura's NPG method we find that the estimated exponents do not decrease towards zero but rather improve, seamingly linearly in λ, towards the estimated exponents of the corresponding unregularized cases. This indicates that the guarantees in Corollary 4.33 and Corollary 4.34 for these NPG methods are not tight. In contrast for the σ-NPG with $\sigma = 1.5$ we observe that the convergence rates deteriorate for $\lambda \to 0$, which conforms with the sublinear convergence $O(t^{-2})$ of the unregularized problem. However, the exponent seems to decrease slower than linearly in the regularization strength λ like it is the case in our guarantee. Theorem 4.30 shows linear convergence based on the strong convexity of the regularizer. The convergence rate of the unregularized NPG methods however is determined by the behavior of the convex function inducing the Hessian geometry at the boundary rather than the convexity of the loss. We believe that a theoretical analysis combining these two effects could improve the linear rate in Theorem 4.30 for small regularization strength.

4.4 LOCALLY QUADRATIC CONVERGENCE FOR REGULARIZED PROBLEMS

It is known that Kakade's NPG method and more generally quasi-Newton policy gradient methods with suitable regularization and step sizes converge at a locally quadratic rate [71, 172]. Whereas these results regard the NPG method as an inexact Newton method in the parameter space, we regard it as an inexact Newton method in state-action space, which allows us to directly leverage results from the optimization literature and therefore leading to relatively short proofs. Our result extends the locally quadratic convergence rate to general Hessian-NPG methods, which include in particular Kakade's and Morimura's NPG. Note that the result holds when the step size is equal to the inverse penalization strength, which is reminiscent of Newton's method converging for step size 1.

Theorem 4.36 (Locally quadratic convergence of regularized NPG methods). *Consider a real-valued function $\phi\colon \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}\cup\{+\infty\}$, which we assume to be finite and twice continuously differentiable on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$ and such that $\nabla^2\phi(\eta)$ is positive definite when restricted to $T_\eta\mathcal{N} = T\mathcal{L} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ for every $\eta \in \operatorname{int}(\mathcal{N})$. Further, consider a regular policy parametrization and the regularized reward $R_\lambda(\theta) := R(\theta) - \lambda\phi(\eta_\theta)$ and assume that $\eta^* \in \operatorname{int}(\mathcal{N})$, i.e., the maximizer lies in the interior of the state-action polytope. Consider the NPG induced by the Hessian geometry of ϕ with step size $\Delta t = \lambda^{-1}$, i.e.,*

$$\theta_{k+1} = \theta_k + \Delta t \cdot G(\theta_k)^+\nabla R_\lambda(\theta_k),$$

where $G(\theta_k)^+$ denotes the Moore-Penrose inverse. Assume that $\theta_k \to \theta^*$ for some maximizer θ^*, then $\theta_k \to \theta^*$ at a (locally) quadratic rate, i.e., it holds that

$$\|\theta_k - \theta^*\| = O(e^{-k^2}) \quad \text{for } k \to \infty \tag{4.56}$$

and hence $R_\lambda(\theta_k) \to R^*_\lambda$ at a (locally) quadratic rate.

The proof of relies on the following convergence result for inexact Newton methods.

Theorem 4.37 (Theorem 3.3 in [90]). *Consider an objective function $f \in C^2(\mathbb{R}^d)$ with $\nabla^2 f(x) \in \mathbb{S}^{sym}_{>0}$ for any $x \in \mathbb{R}^d$ and assume that f admits a minimizer x^*. Let (x_k) be inexact Newton iterates given by*

$$x_{k+1} = x_k - \nabla^2 f(x_k)^{-1}\nabla f(x_k) + \varepsilon_k,$$

and assume that they converge towards the minimum x^*. If $\|\varepsilon_k\| = O(\|\nabla f(x_k)\|^\omega)$, then $x_k \to x^*$ at rate ω, i.e., $\|x_k - x^*\| = O(e^{-k^\omega})$.

We take this approach and show that the iterates of the regularized NPG method can be interpreted as an inexact Newton method in state-action space. For this, we first make the form of the Newton updates in state-action space explicit.

Lemma 4.38 (Newton iteration in state-action space). *The iterates of Newton's method in state-action space are given by*

$$\eta_{k+1} = \eta_k + \lambda^{-1}\Pi^E_{T\mathcal{L}}(\nabla^2\phi(\eta_k))^{-1}\Pi^E_{T\mathcal{L}}(\nabla\mathfrak{R}_\lambda(\eta_k)), \tag{4.57}$$

where $\mathfrak{R}_\lambda(\eta) = \langle r, \eta\rangle + \lambda\phi(\eta)$ is the regularized reward and $\Pi^E_{T\mathcal{L}}$ the Euclidean projection onto the tangent space of the affine space $\mathcal{L}$ defined in (3.5).

Proof. The domain of the optimization problem is $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\ge 0} \cap \mathcal{L}$ an hence, we perform Newton's method on the affine subspace L. Writing $L = \eta_0 + X$ for a linear subspace X we can equivalently perform Newton's method on X since the method is affine invariant. We denote the canonical embedding $\iota\colon X \hookrightarrow \mathcal{L}, x \mapsto x + \eta_0$ and set $f(x) := \mathfrak{R}_\lambda(\iota x)$. Then, we obtain the Newton iterates x_k and $\eta_k = \iota x_k$ by

$$x_{k+1} = x_k + \nabla^2 f(x_k)^{-1}\nabla f(x_k).$$

Straight up computation yields $\nabla f(x) = \iota^\top \nabla \mathfrak{R}_\lambda(\iota x)$ and $\nabla^2 f(x) = \iota^\top \nabla^2 \mathfrak{R}_\lambda(\iota x)\iota$. Hence, we obtain

$$\begin{aligned}\eta_{k+1} - \eta_k &= \iota(\eta_{k+1} - \eta_k) = \iota \nabla^2 f(x_k)^{-1} \nabla f(x_k) = \iota \iota^+ \nabla^2 \mathfrak{R}_\lambda(\eta_k)^{-1} (\iota^\top)^+ \iota^\top \nabla \mathfrak{R}_\lambda(\eta_k) \\ &= \Pi^E_{T\mathcal{L}} (\nabla^2 \mathfrak{R}_\lambda(\eta_k))^{-1} \Pi^E_{T\mathcal{L}} (\nabla \mathfrak{R}_\lambda(\eta_k)),\end{aligned}$$

where we used $AA^+ = \Pi_{\mathrm{range}(A)}$ and $(A^\top)^+ A^\top = \Pi_{\ker(A^\top)} = \Pi_{\mathrm{range}(A)}$. □

Lemma 4.39. *Let $(\theta_k)_{k\in\mathbb{N}}$ denote the iterates of a Hessian NPG induced by a stricly convex function ϕ and with step size Δt, i.e,*

$$\theta_{k+1} = \theta_k + \Delta t \cdot G(\theta_k)^+ \nabla R_\lambda(\theta_k),$$

where the Gram matrix is given by $G(\theta) = DP(\theta)^\top \nabla^2 \phi(\eta_\theta) DP(\theta)$. Then the state-action iterates satisfy

$$\eta_{\theta_{k+1}} = \eta_{\theta_k} + \Delta t \cdot \Pi^E_{T\mathcal{L}} (\nabla^2 \phi(\eta_k)^{-1} \Pi^E_{T\mathcal{L}} (\nabla \mathfrak{R}_\lambda(\eta_k))) + O(\Delta t^2 \|G(\theta_k)^+ \nabla R_\lambda(\theta_k)\|^2) \tag{4.58}$$

for $\|G(\theta_k)^+ \nabla R_\lambda(\theta_k)\| \to 0$.

Proof. Writing P for the mapping $\theta \mapsto \eta_\theta$ and an application of Taylor's theorem implies that

$$\eta_{\theta_{k+1}} - \eta_{\theta_k} = \Delta t \cdot DP(\theta_k) G(\theta_k)^+ \nabla R_\lambda(\theta_k) + O(\Delta t^2 \|G(\theta_k)^+ \nabla R_\lambda(\theta_k)\|^2).$$

The first term is equal to

$$\Delta t \cdot DP(\theta_k) DP(\theta)^+ \nabla^2 \phi(\eta_k)^{-1} (DP(\theta_k)^\top)^+ \nabla DP(\theta_k)^\top \nabla \mathfrak{R}_\lambda(\eta_k),$$

which again is equal to

$$\Delta t \cdot \Pi^E_{T\mathcal{L}} (\nabla^2 \phi(\eta_k)^{-1} \Pi^E_{T\mathcal{L}} (\nabla \mathfrak{R}_\lambda(\eta_k)))$$

since $DP(\theta_k) DP(\theta_k)^+ = (DP(\theta_k)^\top)^+ DP(\theta_k)^\top = \Pi_{\mathrm{range}(DP(\theta_k))}$ like before and

$$\mathrm{range}(DP(\theta_k)) = T\mathcal{L}.$$

□

Proof of Theorem 4.36. We want to apply Theorem 4.37 to the sequence $(\eta_{\theta_k})_{k\in\mathbb{N}}$ where $f = \mathfrak{R}_\lambda$ and

$$\varepsilon_k = O(\|G(\theta_k)^+ \nabla R_\lambda(\theta_k)\|^2) \quad \text{for } \|G(\theta_k)^+ \nabla R_\lambda(\theta_k)\| \to 0$$

by Lemma 4.38 and Lemma 4.39. Surely, since $\theta_k \to \theta^*$ it holds that $\eta_{\theta_k} \to \eta^*$ for $k \to \infty$. Further, $\theta_k \to \theta^*$ implies

$$G(\theta_k)^+ \nabla R_\lambda(\theta_k) = (\Delta t)^{-1} (\theta_{k+1} - \theta_k) \to 0 \quad \text{for } k \to \infty.$$

Since $\theta \mapsto \pi_\theta$ is a regular policy parametrization the rank of $G(\theta)$ does not depend on θ and is equal to $\dim(\mathcal{N}) = \dim(\Delta^{\mathcal{S}}_{\mathcal{A}}) = n_{\mathcal{S}}(n_{\mathcal{A}} - 1)$. This implies that $G(\theta_k)^+ \to G(\theta^*)^+$ and hence $G(\theta_k)^+$ remains bounded for $k \to \infty$, see [238]. Now we can estimate

$$\begin{aligned}\|\varepsilon_k\| &= O(\Delta t^2 \|G(\theta_k)^+ \nabla R_\lambda(\theta_k)\|^2) \\ &= O(\Delta t^2 \|\nabla R_\lambda(\theta_k)\|^2) \\ &= O(\Delta t^2 \|DP(\theta_k)^\top \nabla \mathfrak{R}_\lambda(\eta_k)\|^2) \\ &= O(\|\nabla \mathfrak{R}_\lambda(\eta_k)\|^2),\end{aligned}$$

where we used that $DP(\theta_k)$ stays bounded as $DP(\theta_k) \to DP(\theta^*)$. Now, Theorem 4.37 proves the claim. □

Remark 4.40. A benefit of regarding the iteration as an inexact Newton method in state-action space is that the problem is strongly convex in state-action space. In contrast, in policy space the problem is non-convex, which makes the analysis in that space more delicate. Further, the corresponding Riemannian metric might not be the Hessian metric of the regularizer in policy space (see also Remark 4.9). In the parameter θ, the NPG algorithm can be perceived as a generalized Gauss-Newton method; however, the reward function is non-convex in parameter space. Further, for overparametrized policy models, i.e., when $\dim(\Theta) > \dim(\Delta_{\mathcal{A}}^{\mathcal{S}}) = |\mathcal{S}|(|\mathcal{A}| - 1)$ the Hessian $\nabla^2 R(\theta^*)$ can not be positive definite, which makes the analysis in parameter space less immediate. Note that the tabular softmax policies in (4.7) are overparametrized since in this case $\dim(\Theta) = |\mathcal{S}||\mathcal{A}|$.

Remark 4.41 (Behavior for $\Delta t < \lambda^{-1}$). For Newton's method the locally quadratic convergence only holds at exactly the step size $\Delta t = \lambda^{-1}$. Consider for example $f(x) = x^2/2$, where Newton's method with step size $\Delta t \in [0, 1]$ will produce the iterates $x_k = (1-\Delta t)^k x_0$. If $\Delta t \neq 1$, this will only converge linearly at a rate of $1 - \Delta t$ that decreases to 0 for $\Delta t \to 1$. Hence, we expect that also for regularized NPG methods the locally quadratic convergence is only achieved for the exact Newton step size $\Delta t = \lambda^{-1}$ and linear convergence for $\Delta t < \lambda^{-1}$ with a rate decreasing towards 0 for $\Delta t \to \lambda^{-1}$.

4.5 Discussion and outlook

We study a general class of natural policy gradient methods arising from Hessian geometries in state-action space. This covers, in particular, the notions of NPG due to Kakade and Morimura and co-authors, which are induced by the conditional entropy and entropy respectively. Leveraging results on gradient flows in Hessian geometries we obtain global convergence guarantees of NPG flows for tabular softmax policies and show that both Kakade's and Morimura's NPG converge linearly, and obtain sublinear convergence rates for NPG associated with β-divergences. We provide experimental evidence of the tightness of these rates. Finally, we perceive the NPG with respect to the Hessian geometry induced by the regularizer, with step size equal to the inverse regularization strength, as an inexact Newton method in state-action space, which allows for a very compact argument of the locally quadratic convergence of this method. An overview of the established results in relation to existing works is presented in Table 4.1.

The following questions arose during our analysis and can provide directions for future research:

- *Improved bounds for Kakade's NPG:* Our analysis guarantees the same speed of convergence for Kakade's and Morimura's natural policy gradient flows. This is because in the proof of Theorem 4.26 we rely on the estimate

$$c_1 D_{\phi_C}(\eta^*, \eta) \leq c_1 D_{KL}(\eta^*, \eta) \leq R^* - \langle r, \eta \rangle,$$

  which might not be tight. An estimate of the form $cD_{\phi_C}(\eta^*, \eta) \leq R^* - \langle r, \eta \rangle$ for a constant $c > c_1$ would sharpen our convergence guarantee.

	Unregularized		Regularized	
	Discr. time	Cts. time	Discr. time	Cts. time
Vanilla	$O(t^{-1})$ [200] linear for normalized gradients [199]	–	linear	–
Kakade	linear [156, 305]	**linear**	linear [71, 165, 315] **loc. quadratic** [71]	**linear**
Morimura	–	**linear**	**loc. quadratic**	**linear**
$\sigma > 1$	–	$\mathbf{O(t^{-\frac{1}{\sigma-1}})}$	**loc. quadratic**	**linear**

TABLE 4.1. Bold results are established in this work; for known results the initial works are referenced; results showing locally quadratic convergence use $\Delta t = \lambda^{-1}$.

- *Improved bounds for regularized problems:* Our linear convergence guarantees for regularized problems degrade when the regularization strength decreases where our experiments indicate that the actual convergence does not. This gap could be filled with an improved theoretical analysis.
- *Plateau-free NPG methods:* Our experiments indicate that various NPG methods suffer from plateaus, which are induced by the Riemannian geometry on the state-action polytope. The design of methods that reduce the influence of these plateaus could have great a great impact in the field of reinforcement learning where policy gradient methods are currently under most popular approaches.
- *Estimation:* Where we have studied convergence behavior under the assumption of exact gradient evaluations it would be interesting to characterize the number of samples required to estimate the respective notions of natural policy gradients.
- *Partially observable problems:* Policy gradient methods are known to not converge globally in partially observable problems, however, a better understanding of their convergence properties remains elusive.

Part II

Neural network based PDE solvers

CHAPTER 5

Theoretical analysis of the boundary penalty method for neural network based PDE solvers

Following the works of [109, 129, 269, 110, 237, 176] neural network based PDE solvers have recently experienced an enormous growth in popularity and attention within the scientific community, overviews over existing methods and advances can be found in the articles [43, 57, 295, 78]. We focus on methods, which parametrize the solution of the PDE by a neural network and use a formulation of the PDE in terms of a minimization problem to construct a loss function used to train the network. Two prominent approaches here are the so called deep Ritz method and physics informed neural networks (PINNs), which are often easy to implement compared to finite element methods and promise great success for high dimensional and parametric problems. Despite this both the deep Ritz method as well as physics informed networks often fail to produce highly accurate solutions [110, 264, 292, 293, 161, 84, 314]. And hence an improvement of the optimization pipeline is required to make these methods applicable at an industrial scale. This leads the recent survey [78] to the conclusion that there are

> *numerous questions for future [theoretical] PINN research, the most important of which is whether or not PINN converges to the correct solution of a differential equation,*

which we also believe to be important for other neural network based approaches that fail to exhibit the desired convergence behavior.

We focus on the aspect of boundary values in those approaches, which pose a greater challenge compared to finite element methods as exact Dirichlet boundary values are often intractable to enforce in a neural network directly. Here we give a short description required for the description of the contributions of this section and refer to [43, 295, 78] for in-depth reviews of these methods. For expository reasons we consider the specific case of the Poisson equation with Dirichlet boundary values

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \subseteq \mathbb{R}^d \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{PE}$$

where $f \in L^2(\Omega)$ is a square integrable function.

The deep Ritz method. It is well known that a weakly differentiable function $u \in H_0^1(\Omega)$ is a (weak) solution of (PE) if and only if it minimizes the so-called variational energy

$$\text{minimize} \, \frac{1}{2}\int_\Omega |\nabla u|^2 \mathrm{d}x - \int_\Omega f u \mathrm{d}x \quad \text{subject to } u \in H_0^1(\Omega). \tag{5.1}$$

In the year of his early death, Walter Ritz proposed to a general method for the approximate solution of variational problems [242]. Ritz proposed to work with a parametric class of functions $\mathcal{F} = \{u_\theta : \theta \in \mathbb{R}^p\} \subseteq H_0^1(\Omega)$ and to minimize

$$\theta \mapsto \frac{1}{2}\int_\Omega |\nabla u_\theta|^2 \mathrm{d}x - \int_\Omega f u_\theta \mathrm{d}x \tag{5.2}$$

in order to get an approximate solution of the original problem. Ritz used this approach to determine the coefficients of polynomials by hand[1] and later this method found great success in the context of finite element methods [61]. More recently, it was suggested to use function classes parametrized by deep neural networks [110] and this approach is commonly referred to as the *deep Ritz method*. It can directly be used if the functions u_θ computed by neural networks have zero boundary values.

Physics informed neural networks (PINNs). A different ansatz from the variational formulation of the Poisson equation (PE) is to use the solution of (PE) is the unique solution of the problem

$$\text{minimize} \int_\Omega |\Delta u + f|^2 \mathrm{d}x \quad \text{subject to } u \in H_0^1(\Omega) \cap H^2(\Omega) \tag{5.3}$$

since it is the only element attaining value 0. Note that the function space objective is the squared L^2 norm $\|\Delta u + f\|_{L^2(\Omega)}^2$ of the residual $-\Delta u - f$. In similar fashion to the deep Ritz method, one can use the following objective function for the optimization of the parameters of a neural network

$$\theta \mapsto \int_\Omega |\Delta u_\theta + f|^2 \mathrm{d}x \tag{5.4}$$

if the functions u_θ computed by neural networks have zero boundary values. This approach is known as *residual minimization* in the finite element literature. In the context of neural networks, it can be traced back to [100, 164] and was recently popularized in [269, 237] under the names *deep Galerkin method* and *physics informed neural networks (PINNs)* although not being a Galerkin method in the traditional sense.

Discretization and incorporation of data. In practice, the integrals in the objective functions (5.2) and (5.4) have to be discretized, which can be done in various ways. For high dimensional problems stochastic integration techniques are typically used. For PINNs the choice of the collocation points in the discretization of the loss has been investigated in a variety of works [183, 215, 85, 313, 291, 303]. In general, both loss functions can be augmented with data-fitting terms by adding

$$\sum_i |u_\theta(x_i) - y_i|^2, \tag{5.5}$$

where $((x_i, y_i))_{i=1,\dots,N} \subseteq \mathbb{R}^d \times \mathbb{R}$ are data points corresponding to measurements or approximate function evaluations. This procedure is standard for PINNs [237].

[1] Ritz considered a Poisson equation with Neumann boundary values rather then Dirichlet zero boundary values, which allowed Ritz to work with polynomials without boundary penalty.

Boundary values. We have introduced both the DRM and PINNs for neural networks that have zero boundary values, which is not the case for standard feedforward architectures. Two approaches to deal with this problem are common in the literature: the construction of neural network based classes with exact boundary values and a relaxation of the problem together with a penalization of boundary values.

One can transform any unconstrained neural network architecture into an ansatz space with the desired boundary conditions the following way [164]. Assume we want to solve the Poisson problem (PE) on Ω with zero boundary values and consider a smooth function $h\colon \Omega \to [0, \infty)$ that satisfies $h|_{\partial\Omega} = 0$ and $h|_{\Omega} \neq 0$. The function h is often referred to as a smooth approximation of the distance function to $\partial\Omega$. For any family $\{u_\theta : \theta \in \Theta\} \subseteq H^1(\Omega)$ of functions we can consider the associated family

$$\{h \cdot u_\theta : \theta \in \Theta\} \subseteq H^1_0(\Omega) \tag{5.6}$$

and use these functions to approximate the solution of (PE) using either (5.2) or (5.4) as an objective function for the optimization of the network parameters. For complex domains it is difficult to obtain h analytically and thus the approximation via neural networks was proposed by [47]. For time-dependent problems, a similar construction to (5.6) using a smoothed distance function to the parabolic boundary of the space-time domain can be used, see [185] for an explicit example.

Another approach is to relax the problem and allow ansatz functions that do not satisfy the boundary values exactly and to augment the objective function (5.2) and (5.4) with the *boundary penalty* term

$$\lambda \cdot \int_{\partial\Omega} u_\theta^2 \mathrm{d}s \tag{5.7}$$

for some $\lambda \geq 0$. This approach is applicable to all domains and is easy to implement if one can (uniformly) sample points on the boundary $\partial\Omega$. The resulting objective functions for neural network training are

$$L = L^\lambda_{\text{DRM}}\colon \Theta \to \mathbb{R}, \quad \theta \mapsto \frac{1}{2}\int_\Omega |\nabla u_\theta|^2 \mathrm{d}x - \int_\Omega f u_\theta \mathrm{d}x + \lambda \cdot \int_{\partial\Omega} u_\theta^2 \mathrm{d}s \tag{5.8}$$

for the deep Ritz method and

$$L = L^\lambda_{\text{PINN}}\colon \Theta \to \mathbb{R}, \quad \theta \mapsto \int_\Omega |\Delta u_\theta + f|^2 \mathrm{d}x + \lambda \cdot \int_{\partial\Omega} u_\theta^2 \mathrm{d}s \tag{5.9}$$

for physics informed neural networks. The corresponding function space problems are

$$\text{minimize } \frac{1}{2}\int_\Omega |\nabla u|^2 \mathrm{d}x - \int_\Omega f u \mathrm{d}x + \lambda \cdot \int_{\partial\Omega} u_\theta^2 \mathrm{d}s \quad \text{subject to } u \in H^1(\Omega) \tag{5.10}$$

for the deep Ritz method with boundary penalty and

$$\text{minimize } \int_\Omega |\Delta u + f|^2 \mathrm{d}x + \lambda \cdot \int_{\partial\Omega} u_\theta^2 \mathrm{d}s \quad \text{subject to } u \in H^2(\Omega) \tag{5.11}$$

for physics informed neural networks with boundary penalty. Note that for any $\lambda > 0$ the solution of the Poisson problem (PE) is still the unique solution of (5.11), which makes the theoretical analysis of the boundary penalty method in PINNs significantly easier. However, the error estimates of boundary penalized PINNs and PINNs with exact boundary values have different qualitative properties as we discuss in Section 5.5. For the deep Ritz method, however, the minimizer of the penalized problem (5.10) does not have zero boundary values and depends on the penalty λ. Hence, the penalization introduces a new source of error, which makes the theoretical analysis more delicate.

Contributions. In our work we focus on theoretical guarantees for both the deep Ritz method as well as physics informed networks that can be made and characterize the error in different norms in different settings. We do not study the optimization process itself but rather develop an understanding of the implications of successful training. The main contributions contained in this chapter can be summarized as follows:

- We provide a convergence guarantee in H^1 norm for the deep Ritz method for nonlinear problems, when the boundary penalty strength is coupled to the boundary values required to approximate the solution of the problem; this is uniform over right hand sides of the PDE (see Theorem 5.1 and 5.21).
- We provide an error estimate in H^1 norm for the deep Ritz method with boundary penalty for elliptic PDEs that depends on the optimization error, the approximation error and the penalization strength (see Theorem 5.3). For the specific case of ReLU networks and for right hand side $f \in H^r(\Omega)$ this implies that under perfect training and with penalization strength $\lambda_n \sim n^{\frac{2r+3}{2d}}$ the error made by the deep Ritz method decays like $O(n^{-\rho})$ for any $\rho < \frac{2r+3}{4d}$ (see Theorem 5.38). Note the difference compared to the convergence rate $O(n^{-\frac{2r+3}{2d}})$, which is due to the error introduced by the penalization.
- For residual minimization with boundary penalty, the $H^{1/2}$ convergence rate is known to agree with the H^2 approximation rate; we show that this result is sharp (see Theorem 5.46) and show that for ansatz classes with exact boundary values a similar estimate in H^2 hold (see Theorem 5.44).

In Section 5.1 we present these contributions in more detail and discuss their relation to related works. We defer all proofs to the appendix of the chapter. In Section 5.2 we present background material on neural networks and Sobolev spaces, where we also point the reader to our overall notation. In Section 5.3 we present the proofs regarding the uniform convergence guarantees for possibly nonlinear problems. In Section 5.4 we provide the proofs for the error estimate for the deep Ritz method with boundary penalty for the case of linear problems. In Section 5.5 we establish the error estimates regarding physics informed neural networks.

5.1 Presentation and discussion of the main results

The following section presents the main results of our theoretical analysis, provides brief insights into the underlying arguments where appropriate, and discusses the relationships with existing works.

5.1.1. Uniform convergence for the deep Ritz method for nonlinear problems. We consider an open and bounded set $\Omega \subseteq \mathbb{R}^d$ with Lipschitz boundary $\partial\Omega$. For $n \in \mathbb{N}$ let Θ_n denote the parameter space of a neural network and let $\mathcal{F}_n \coloneqq \{u_\theta : \theta \in \Theta_n\}$ denote the family of functions parametrized by this network, which we assume to be contained in the Sobolev space $\mathcal{F}_n \subseteq W^{1,p}(\Omega)$ for some $p \in (1, \infty)$. Let $E\colon W^{1,p}(\Omega) \to (-\infty, \infty]$ be a functional and $(\lambda_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$ be a sequence of real numbers. Furthermore, let $f \in W^{1,p}(\Omega)^*$ be fixed and define the functional $E^f : W^{1,p}(\Omega) \to (-\infty, \infty]$ by

$$E^f(x) := \begin{cases} E(u) - f(u) & \text{for } u \in W_0^{1,p}(\Omega), \\ \\ +\infty & \text{otherwise .} \end{cases}$$

Further, we define the following boundary penalized loss functions

$$L_n : \Theta_n \to \mathbb{R}, \quad \theta \mapsto E^f(u_\theta) + \lambda_n \|u_\theta\|^p_{L^p(\partial\Omega)}.$$

We make the following assumptions:

(A1) *Universal approximation:* For every $u \in W_0^{1,p}(\Omega)$ there are parameters $\theta_n \in \Theta_n$ such that $\|u_{\theta_n} - u\|_{W^{1,p}(\Omega)} \to 0$ and $\lambda_n \|u_{\theta_n}\|^p_{L^p(\partial\Omega)} \to 0$ for $n \to \infty$.

(A2) The functional E is bounded from below, weakly lower semi-continuous with respect to the weak topology of $W^{1,p}(\Omega)$ and continuous and equi-coercive with respect to the norm topology of $W^{1,p}(\Omega)$.

(A3) For every $f \in W^{1,p}(\Omega)^*$, there is a unique minimizer $u_f \in W_0^{1,p}(\Omega)$ of F^f and the solution map

$$S : W^{1,p}(\Omega)^* \to W_0^{1,p}(\Omega) \quad \text{with } f \mapsto u_f$$

is demi-continuous, i.e. maps strongly convergent sequences to weakly convergent ones.

Now we can state the main result, which is a special case of Theorem 5.21. The proof is based on the tool of Γ-convergence and can be found in Section 5.3

Theorem 5.1 (Uniform convergence of the deep Ritz method). *For $f \in W^{1,p}(\Omega)$ and $\delta_n > 0$ we define the approximate solution set*

$$S_n(f) := \left\{ \theta \in \Theta_n : L_n(\theta) \le \inf_{\theta' \in \Theta_n} L_n(\theta') + \delta_n \right\}.$$

Furthermore, denote the unique minimizer of E^f in $W_0^{1,p}(\Omega)$ by u_f and fix $R > 0$ and let $\lambda_n \to \infty$ and $\delta_n \to 0$ for $n \to \infty$. Then it holds that

$$\sup \left\{ \|u_\theta - u_f\|_{L^p(\Omega)} : \theta \in S_n(f),\ \|f\|_{W^{1,p}(\Omega)^*} \le R \right\} \to 0 \quad \textit{for } n \to \infty.$$

The theorem above can be formulated for abstract variational problems, however, for the sake of readability we presented here in the context of the Sobolev spaces $W^{1,p}(\Omega)$ for expository reasons, see Section 5.3 for a more general statement. The result ensures that under certain conditions the approximated solutions found by the deep Ritz method converge uniformly towards the true solution of the problem. For instance, we require the optimization to be successful, i.e., $\delta_n \to 0$, which is a non trivial assumption that we make since otherwise there is no reason why we could hope for convergence. Further, we assume that the penalization strength increases towards $+\infty$, which is intuitive and can be assured by the practitioner. The second requirement on the penalization strength is formulated in Assumption (A1) that we would like to comment on in more detail here.

Universal approximation. The penalization strength is not allowed to grow arbitrarily fast since otherwise the universal approximation Assumption (A1) might not be satisfied. It is well known by the classical works on neural network approximation that for most architectures of growing size are able to approximate any given function in $W^{1,p}(\Omega)$ [141] for smooth activation functions and more recent works have also established approximation rates and treated ReLU networks [127, 306, 262, 138, 106, 88]. Hence, the existence of $\theta_n \in \Theta_n$ such that $\|u_{\theta_n} - u\|_{W^{1,p}(\Omega)} \to 0$ for $n \to \infty$ is satisfied for virtually all architectures and activation functions. Since by the continuity of the trace operator we have $\|u\|_{L^p(\partial\Omega)} \leq c\|u\|_{W^{1,p}(\Omega)}$ for all $u \in W^{1,p}(\Omega)$ and therefore $\|u_{\theta_n}\|_{L^p(\partial\Omega)} \to 0$ for $n \to \infty$. If the rate of decay is independent of the target function, which it typically is, see [96], we can choose a growing sequence $(\lambda_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$ growing to $+\infty$ such that $\lambda_n \|u_{\theta_n}\|^p_{L^p(\partial\Omega)} \to 0$. Therefore, we have seen that whenever universal approximation holds uniformly, we can find a suitable sequence of penalization strengths, which depends on the boundary values required for universal approximation, such that Assumption (A1) is satisfied. Therefore, the admissible choice of penalization strengths depends on the specific choice of network architecture. We discuss the specific choice in more depth for linear problems in Subsection 5.1.2 and Subsection 5.4.3. For ReLU activation we have the special case that universal approximation of a function with zero boundary values is possible with neural network functions with exact zero boundary values.

Theorem 5.2 (Universal approximation with zero boundary values, [102]). *Consider an open set $\Omega \subseteq \mathbb{R}^d$ and let $u \in W^{1,p}_0(\Omega)$ with $p \in [1, \infty)$. Then for all $\varepsilon > 0$ there exists a function $u_\varepsilon \in W^{1,p}_0(\Omega)$ that can be expressed by a ReLU network of depth $\lceil \log_2(d + 1)\rceil + 1$ such that*

$$\|u - u_\varepsilon\|_{W^{1,p}(\Omega)} \leq \varepsilon.$$

The proof is a simple consequence that ReLU networks of unrestricted size are able to exactly compute arbitrary piecewise linear functions [20] and is provided in Section 5.3. An important consequence of this result is that for the specific choice of ReLU networks of depth $\lceil \log_2(d + 1)\rceil + 1$ and growing width arbitrary strong penalization is admissible is covered by Theorem 5.1.

The p-Laplace operator as an example. As an example of a nonlinear PDE that is covered by Theorem 5.1 we discuss the p-Laplacian. To this end, consider the p-Dirichlet energy for $p \in (1, \infty)$ given by

$$E \colon W^{1,p}(\Omega) \to \mathbb{R}, \quad u \mapsto \frac{1}{p}\int_\Omega |\nabla u|^p \,\mathrm{d}x.$$

Note that for $p \neq 2$ the associated Euler-Lagrange equation – the p-Laplace equation – is nonlinear. In strong formulation it is given by

$$\begin{aligned} -\operatorname{div}(|\nabla u|^{p-2}\nabla u) &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

see for example [274] or [246]. As a neural network architecture one can choose ReLU networks of depth $\lceil \log_2(d+1)\rceil + 1$ and width n and hence by Theorem 5.2 Assumption (A1) is satisfied and arbitrary penalization strengths $\lambda_n \to \infty$ can be chosen. For the technical Assumption (A3) we refer to Subsection 5.3.4. Finally, to provide the demi-continuity we must consider the operator $S\colon W_0^{1,p}(\Omega)^* \to W_0^{1,p}(\Omega)$ mapping f to the unique minimizer u_f of $E - f$ on $W_0^{1,p}(\Omega)$. By the Euler-Lagrange formalism, u minimizes F^f if and only if

$$\int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla v \mathrm{d}x = f(v) \quad \text{for all } v \in W_0^{1,p}(\Omega).$$

Hence, the solution map S is precisely the inverse of the mapping

$$W_0^{1,p}(\Omega) \to W_0^{1,p}(\Omega)^*, \quad u \mapsto \left(v \mapsto \int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla v \mathrm{d}x \right)$$

and this map is demi-continuous, see for example [246].

Related works. The uniform convergence result Theorem 5.1 provides the first convergence analysis of the deep Ritz method for nonlinear problems. Later, error estimates for the specific case of the p-Laplace where established in [155]. For linear elliptic problems a larger body of works establishing error estimates exist, which we discuss in more detail in Subsection 5.1.2.

5.1.2. An error estimate for the deep Ritz method with boundary penalty. When neural networks are used for the approximate minimization of a suitably convex variational energy the error will scale essentially like the approximation error of the neural networks, see Subsection 5.4.1. However, the relaxation of the exact Dirichlet boundary conditions to approximate zero boundary values together with a boundary penalty introduces an additional error. In the previous subsection we have shown that successful optimization together with a penalization strength increasing according to the approximation properties of the neural networks yields a uniform convergence guarantee for nonlinear problems. In this subsection focus on linear problems and provide quantitative results on the error made by the deep Ritz method and the admissible choices of penalization strengths.

For ease of presentation, we discuss our approach for the concrete equation

$$\begin{aligned} -\operatorname{div}(A\nabla u) &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{5.12}$$

where $A \in L^\infty(\Omega, \mathbb{R}^{d\times d})$ is a symmetric and elliptic coefficient matrix. The weak formulation of this equation gives rise to the bilinear form

$$a\colon H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}, \quad a(u,v) = \int_\Omega A\nabla u \cdot \nabla v \mathrm{d}x$$

and the energy

$$E\colon H^1(\Omega) \to \mathbb{R}, \quad E(u) = \frac{1}{2}a(u,u) - f(u)$$

where $f \in H^1(\Omega)^*$. Using the boundary penalty method as an approximation for (5.12) leads to the bilinear form

$$a_\lambda\colon H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}, \quad a_\lambda(u,v) = \int_\Omega A\nabla u \nabla v \mathrm{d}x + \lambda \int_{\partial\Omega} uv \mathrm{d}s$$

for a penalty parameter $\lambda > 0$ and the energy

$$E_\lambda\colon H^1(\Omega) \to \mathbb{R}, \quad E_\lambda(u) = \frac{1}{2}a_\lambda(u,u) - f(u).$$

The central error estimation is collected in the following Theorem. Note that we require $H^2(\Omega)$ regularity of the solution to equation (5.12).

Theorem 5.3. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with $C^{1,1}$ boundary, $f \in L^2(\Omega)$ and assume $A \in C^{0,1}(\Omega, \mathbb{R}^{d\times d})$ is symmetric, uniformly elliptic with ellipticity constant $\alpha > 0$[2]. By $u_f \in H^1_0(\Omega)$ we denote the solution of* (5.12) *and by u_λ the minimizer of the penalized energy E_λ over $H^1(\Omega)$. Fix an arbitrary subset $V \subset H^1(\Omega)$ and denote the coercivity constants of a_λ by $\alpha_\lambda > 0$ and set $\delta := E_\lambda(v) - \inf_{\tilde{v}\in V} E_\lambda(\tilde{v})$. Then there is a constant $c > 0$, only depending on A and Ω, such that for every $v \in V$ and $\lambda > 0$ it holds that*

$$\|v - u_f\|_{H^1(\Omega)} \leq \sqrt{\frac{2\delta}{\alpha_\lambda} + \frac{1}{\alpha_\lambda}\inf_{\tilde{v}\in V}\|\tilde{v} - u_\lambda\|^2_{a_\lambda}} + c\lambda^{-1}\|f\|_{L^2(\Omega)}, \tag{5.13}$$

where $\|u\|^2_{a_\lambda} := a_\lambda(u,u)$ is the norm induced by a_λ. Further, the constant c can be bounded in terms of domain Ω and the coefficient matrix A, see Theorem 5.28.

Proof sketch. The proof relies on the error decomposition

$$\|v - u_f\|_{H^1(\Omega)} \leq \|v - u_\lambda\|_{H^1(\Omega)} + \|u_\lambda - u_f\|_{H^1(\Omega)},$$

where $u_\lambda \in H^1(\Omega)$ is the (unique) minimizer of the energy E_λ. The first term can be estimated using standard arguments, see Subsection 5.4.1, where the second term is more delicate. The difference $v_\lambda := u_\lambda - u_f$ (weakly) solves the Robin boundary value problem

$$\begin{aligned} -\operatorname{div}(A\nabla v_\lambda) &= 0 \quad \text{in } \Omega \\ \partial_\nu v_\lambda + \lambda v_\lambda &= u_f \quad \text{on } \partial\Omega, \end{aligned}$$

where ∂_ν denotes the normal derivative, see Subsection 5.4.2. This allows to expand v_λ in an eigenbasis of weakly A-harmonic functions, which provides an explicit formula for v_λ and subsequently the desired estimate $\|u_\lambda - u_f\|_{H^1(\Omega)} \leq c\lambda^{-1}$. □

The strategy of the proof of Theorem 5.28 holds for a broader class of elliptic zero boundary value problems. The essential requirement is that the bilinear form a of the differential operator is coercive on $H^1_0(\Omega)$ and that a_λ is coercive on all of $H^1(\Omega)$. Then, regularity of the solution u_f of the zero boundary value problem is required to identify the equation u_f satisfies when tested with functions in $H^1(\Omega)$ and not only $H^1_0(\Omega)$, see (5.22).

Theorem 5.3 bounds the distance of a function in terms of the optimization error, the approximation power of the ansatz class and the penalization strength. Now we discuss the trade off of choosing the penalization strength λ too large or too small and discuss the implications of different scalings of λ in dependecy of the approximation capabilities of the ansatz classes. To do so, we consider a sequence $(V_n)_{n\in\mathbb{N}} \subseteq H^1(\Omega)$ of ansatz classes and penalization strengths $\lambda_n \sim n^\sigma$. Further, we denote the minimizers of the energies E_{λ_n} over V_n by $v_n^* \in V_n$. It is our goal to choose $\sigma \in \mathbb{R}$ in such a way that the upper bound of $\|v_n^* - u_f\|_{H^1(\Omega)}$ in (5.13) decays with the fastest possible rate. Neglecting constants, the bound evaluates to

$$\|v_n^* - u_f\|_{H^1(\Omega)} \lesssim \sqrt{\frac{1}{\alpha_{\lambda_n}}\inf_{v\in V_n}\|v - u_{\lambda_n}\|^2_{a_{\lambda_n}}} + \lambda_n^{-1}. \tag{5.14}$$

[2], i.e., $v^\top A(x)v \geq \alpha\|v\|^2$ for all $x \in \Omega, v \in \mathbb{R}^d$

We can assume without loss of generality that $\sigma > 0$ and hence $\lambda_n \geq 1$, because otherwise the upper bound will not decrase to zero. Note that in this case we have $\alpha_{\lambda_n} \geq \alpha_1 > 0$ and hence the we obtain

$$\|v_n^* - u_f\|_{H^1(\Omega)} \lesssim \sqrt{\inf_{v \in V_n} \left(\|\nabla(v - u_{\lambda_n})\|^2_{L^2(\Omega)} + n^\sigma \|v - u_{\lambda_n}\|^2_{L^2(\partial\Omega)} \right)} + n^{-\sigma}. \tag{5.15}$$

Here, the trade off in choosing σ and therefore λ_n too large or small is evident. The implications of this trade off are captured in the following result.

Theorem 5.4 (Rates for NN training with boundary penalty). *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with $C^{r+1,1}$ boundary for some $r \in \mathbb{N}$, $f \in H^r(\Omega)$ and assume $A \in C^{r,1}(\Omega, \mathbb{R}^{d\times d})$ is symmetric, uniformly elliptic with ellipticity constant $\alpha > 0$ and denote the solution to* (5.12) *by $u_f \in H^1_0(\Omega)$. For every $n \in \mathbb{N}$, there is a ReLU network with parameter space Θ_n of dimension $O(\log_2^2(n^{(r+2)/d}) \cdot n)$ such that if $\lambda_n \sim n^\sigma$ for $\sigma = \frac{2r+3}{2d}$ one has for any $\rho < \frac{2r+3}{4d}$ that*

$$\|u_{\theta_n} - u_f\|_{H^1(\Omega)} \lesssim \sqrt{\delta_n + n^{-2\rho}} + n^{-\rho} \quad \textit{for all } \theta_n \in \Theta_n, \tag{5.16}$$

where $\delta_n = E_{\lambda_n}(u_{\theta_n}) - \inf_{\tilde\theta \in \Theta_n} E_{\lambda_n}(u_{\tilde\theta})$.

Note that the solution $u_f \in H^{r+2}(\Omega)$ can be approximated at a rate of $O(n^{-(r+1)/d})$ (see Theorem 5.11), which is a faster rate than the rate $O(n^{-\rho})$ obtained by the deep Ritz method with successful training. In Subsection 5.1.2 we discuss approaches how faster rates can be achieved under stronger a priori estimates for Robin boundary value problems or with approximation results that offer a finer analysis of the required boundary values for approximation.

The error of the deep Ritz method decays at a rate increasing with the smoothness of the problem. This fact can be especially useful in high spatial dimensions, which is consistent with the empirical findings that the deep Ritz method can be effective in the numerical solution of high dimensional problems [110]. Note that also finite element methods can achieve rates increasing with the smoothness of data, however they require the delicate construction of higher order elements.

Combination with different approximation results. In Theorem 5.4 we focus on the ReLU activation in this section, whereas in practice often other architectures and activation functions are used, see [110, 132]. However, our results from Section 5.4 can handle arbitrary function classes and hence reduce the computation of error estimates to the computation of approximation bounds. Therefore, they can be combined with other approximation results for neural networks in Sobolev norm including the works of [127, 306, 262, 138, 106, 88].

The boundary penalty method for FEM. The boundary penalty method has been applied in the context of finite element approximations [27] and studied in terms of its convergence rates in [27, 258, 34]. The idea of the finite element approach is analogue to the idea of using neural networks for the approximate solution of variational problems. However, one constructs a nested sequence of finite dimensional vectorspaces $V_h \subseteq H^1(\Omega)$ arising from some triangulation with fineness $h > 0$ and computes the minimizer u_h of the penalized energy E_λ over V_h. Choosing a suitable triangulation and piecewise affine linear elements and setting $\lambda \sim h^{-1}$ one obtains the error estimate

$$\|u_h - u_f\|_{H^1(\Omega)} \lesssim h,$$

see [258]. At the core of those estimates lies a linear version of Céa's Lemma, which already incorporates boundary values. However, the proof of this lemma heavily relies on the fact that the class of ansatz functions is linear and that its minimizer solves a linear equation. This is not the case for non linear function classes like neural networks. Therefore, our estimates require a different strategy. However, the optimal rate of convergence for the boundary penalty method with finite elements can be recovered as a special case of our results, see Subsection 5.4.

Related works. There exist various results that estimate the error of the deep Ritz method for elliptic equations in a similar setting to the one considered by us [306, 149, 106]. However, all these works consider the case where the variational energy of the original PDE is coercive on the entire space $H^1(\Omega)$, e.g., for example $-\Delta u + \mu u = f$ for some constant $\mu > 0$. Thus, their analysis of the boundary penalty in the manuscripts can immediately be deduced from classic results [194] and more importantly it fails to capture the trade-off in the penalization strength. For Neumann boundary conditions and under the assumption that the solution of the Poisson problem lies in the Barron space the generalization error of the deep Ritz method has been studied in [174]. More recently, error estimates for the deep Ritz method in the context of the p-Laplace and the fractional Laplace operator have been established [155, 125]. Further, error estimates for a generalization of the deep Ritz method dealing with mixed boundary values is due to [177]. Our main contribution (see Theorem 5.3) is to relax the coercivity assumption on the operator in the boundary penalty method, allowing operators that are coercive only on the space $H^1_0(\Omega)$ and hence providing a result that covers the Poisson equation $-\Delta u = f$ with Dirichlet boundary conditions $u|_{\partial\Omega} = 0$.

5.1.3. Implications of exact boundary values in residual minimization.

For the deep Ritz method we obtained an error decay, which might be slower than the approximation rate, which is due to a trade of in the penalization strength. Here, we study the effect of exact boundary values for the ansatz of residual minimization. The situation is different here as both function space problems corresponding to the case of exact boundary values (5.3) and to penalized boundary values (5.11) have the unique solution to the original PDE (PE) as their minimizer. Hence, for both cases error estimates can be obtained by standard arguments and under the assumption of successful training the error will scale like the approximation error. However, there is a qualitative difference between exact and penalized boundary values as for exact boundary values the H^2 error scales like the H^2 approximation error where for penalized boundary values the $H^{1/2}$ error scales like the H^2 approximation error. Whereas in for the deep Ritz method penalized boundary values resulted in a decrease in the approximation rate for residual minimization penalized boundary values lead to estimates with respect to a weaker norm.

We consider again (PE), in particular, we assume that the problem is H^2 regular meaning that there is a constant $C_{\text{reg}} > 0$, satisfying

$$\|u\|_{H^2(\Omega)} \leq C_{\text{reg}} \|\Delta u\|_{L^2(\Omega)} \quad \text{for all } u \in H^2(\Omega) \cap H^1_0(\Omega).$$

Furthermore, we assume that Θ is a parameter set of a neural network type ansatz class, such that for every $\theta \in \Theta$ we have $u_\theta \in H^2(\Omega)$ and $(u_\theta)|_{\partial\Omega} = g$. As our strategy is to minimize the residual we define the loss function

$$\mathcal{L}\colon \Theta \to \mathbb{R}, \quad \mathcal{L}(\theta) = \|\Delta u_\theta + f\|^2_{L^2(\Omega)}.$$

This is for example satisfied when $\partial\Omega \in C^{1,1}$, $f \in L^2(\Omega)$. Alternatively, one can replace the assumption $\partial\Omega \in C^{1,1}$ by requiring that the domain Ω is convex. We refer to [124] for a detailed discussion of the regularity properties of elliptic equations.

The following result is a direct consequence of the H^2 regularity we assumed and a similar result is due to [283], although not exploiting the benefits of exact boundary conditions. Albeit being of simple nature, we believe it can be of practical relevance due to its easy and explicit error control.

Theorem 5.5 (H^2 estimate with exact boundary values). *It holds for every $\theta \in \Theta$ that*

$$\|u_\theta - u_f\|_{H^2(\Omega)} \leq C_{reg}\sqrt{\mathcal{L}(\theta)}.$$

For convex domains, we may estimate the regularity constant explicitely. It holds

$$C_{reg} \leq \sqrt{1 + C_P} \leq \sqrt{1 + \left(\frac{|\Omega|}{\omega_d}\right)^{\frac{1}{d}}},$$

where d is the dimension of Ω, ω_d denotes the volume of the unit ball in $\mathbb{R}^d$ and C_P is the Poincaré constant for functions in $H^1_0(\Omega)$.

Note that in contrast to the estimate (5.13) for the deep Ritz method Theorem 5.5 establishes an a posteriori estimate, i.e., the right hand side can be evaluated during or at the end of the optimization process.

Let us now turn towards the case of penalized boundary values. Again, by Θ we denote the parameter set of a neural network type ansatz class, such that for every $\theta \in \Theta$ we have $u_\theta \in H^2(\Omega)$, but make no assumptions on its boundary values. As our strategy is to minimize the residual we define the loss function with boundary penalty

$$\mathcal{L}_\tau\colon \Theta \to \mathbb{R}, \quad \mathcal{L}_\tau(\theta) = \|\Delta u_\theta + f\|^2_{L^2(\Omega)} + \tau\|u_\theta - g\|^2_{L^2(\partial\Omega)},$$

where $\tau \in (0, \infty)$ is a positive penalization parameter.

Theorem 5.6 (H^s estimates with penalized boundary values). *Assume that the domain $\Omega \subseteq \mathbb{R}^d$ has a smooth boundary $\partial\Omega \in C^\infty$. Then for $s \in \mathbb{R}$ there is a constant $c > 0$ such that*

$$\|u_\theta - u_f\|_{H^s(\Omega)} \leq c\sqrt{\mathcal{L}_\tau(\theta)} \quad \textit{for all } \theta \in \Theta \tag{5.17}$$

and all parametric classes and data $f \in L^2(\Omega), g \in H^{3/2}(\partial\Omega)$ if and only if $s \leq 1/2$.

Comparing this to Theorem 5.5 we see that (5.17) also provides an a posteriori estimate, however with respect to the weaker $H^{1/2}$ norm.

Stronger estimates through stronger penalty. We have seen that the $L^2(\partial\Omega)$ penalization can not lead to estimates in a stronger Sobolev norm than $H^{1/2}(\Omega)$. However, inspecting inequality (5.37) one could – at least in theory – penalize the boundary values in the $H^{3/2}(\partial\Omega)$ norm and would then obtain $H^2(\Omega)$ estimates. Not that the $H^{3/2}(\partial\Omega)$ norm is difficult to approximate in practice.

Stronger estimates through interpolation. It is possible to bound the H^s error for $s \geq 1/2$ of residual minimization with L^2 boundary penalty for the expense of worse rates and under the cost of an additional factor for which it is not clear whether it is bounded. Similar to [53] one can use an interpolation inequality for $s \in [1/2, 2]$ to obtain

$$\|u\|_{H^s(\Omega)} \leq \|u\|_{H^{1/2}(\Omega)}^{2(2-s)/3} \cdot \|u\|_{H^2(\Omega)}^{(2s-1)/3} \quad \text{for all } u \in H^2(\Omega).$$

Together with the a posteriori estimate on the $H^{1/2}$ norm, this yields

$$\begin{aligned} \|u_f - u_\theta\|_{H^s(\Omega)} &\leq \|u_f - u_\theta\|_{H^{1/2}(\Omega)}^{2(2-s)/3} \cdot \|u_f - u_\theta\|_{H^2(\Omega)}^{(2s-1)/3} \lesssim \|u_f - u_\theta\|_{H^2(\Omega)}^{(2s-1)/3} \cdot L(\theta)^{(2-s)/3} \\ &\leq \left(\|u_f\|_{H^2(\Omega)} + \|u_\theta\|_{H^2(\Omega)}\right)^{(2s-1)/3} \cdot L(\theta)^{(2-s)/3}. \end{aligned}$$

Hence, if it is possible to control the H^2 norm of the neural network functions, one obtains an a posteriori estimate on the H^s error. Note however, that the H^2 norm of the neural networks functions is not controlled through the loss function L and hence, this estimates requires an additional explicit or implicit control on the H^2 norm in order to be informative. Note, however, that the exponent of the a posteriori estimate decreases towards zero for $s \to 2$ and the estimate collapses to a trivial bound for $s = 2$.

Related works. Various theoretical analysis for physics informed neural networks exist, however, none of these study the influence of exact boundary values. A very general result showing the existence of uniform estimates without quantifying the uniformity is due to [197]. Quantitative estimates on the L^2 error made by physics informed networks for Kolmogorov PDEs and the Navier-Stokes equations are established in [89, 87]. Error estimates for physics informed networks with respect to the $H^{1/2}$ norm of the form of Theorem 5.5 have been shown in [261] and the generalization error of PINNs has been studied in [201, 29] and the convergence for growing data has been shown in [260]. In practice, ansatz functions with exact boundary values have become increasingly popular as it has been observed to simplify the training process and produce more accurate solutions, see for instance [47, 243, 185, 74]. The works of [74, 77] explicitly compare penalized boundary conditions to exactly enforced ones in numerical studies and found improved accuracy and a faster training process. This is in accordance with [161] that illustrates the difficulties in the training process stemming from soft penalties in residual minimization. It is also possible to encode Neumann or Robin boundary conditions in a similar way, we refer the reader to [185]. However, we mention that the approximation capabilities of such ansatz classes have not been studied so far.

5.1.4. Outlook.

Where have described our contributions above we highlight the following directions for future research:

- *Approximation theory for boundary values:* When working with the boundary penalty method, the boundary values required for the approximation of a function play an important role, see Assumption (A1) and Theorem 5.3 as well as the discussion in Subsection 5.4.3 and in particular (5.15). This asserts that under the assumption of perfect optimization the error can be estimated according to

$$\|v_n^* - u_f\|_{H^1(\Omega)} \lesssim \sqrt{\inf_{v \in V_n} \left(\|\nabla(v - u_{\lambda_n})\|_{L^2(\Omega)}^2 + \lambda_n \|v - u_{\lambda_n}\|_{L^2(\partial\Omega)}^2\right)} + \lambda_n^{-1}$$

and hence it is natural to study the approximation error

$$\inf_{v\in V_n}\left(\|\nabla(v-u_{\lambda_n})\|^2_{L^2(\Omega)}+\lambda_n\|v-u_{\lambda_n}\|^2_{L^2(\partial\Omega)}\right)$$

for different neural network based classes. For example, although we have shown in Theorem 5.2 that ReLU networks can approximate functions $u\in H^1_0(\Omega)$ while maintaining exact zero boundary values the rate at which they are able to do this is not known. Further, the approximation properties of function classes with exact zero boundary values constructed from neural networks according to (5.6) are unclear.

- *Penalization strategies:* Based on our theoretical guarantee in Theorem 5.3 we have deduced suggestions for scaling of the penalization strengths, see also the discussion in Subsection 5.4.3. It would be interesting to investigate whether these suggestions can be used as a general guidance in real world problems. Note that the benefits of successively increasing the penalization strength have been demonstrated [77].
- *Theoretical analysis of optimizers:* Our results decompose the error of the methods into terms depending on the approximation error, the optimization error and the penalization error in the case of the deep Ritz method. Hence, it is natural to study the behavior of different optimizers for this problem. A first analysis has been carried out in [184] using an NTK argument to show global convergence of gradient descent for shallow networks under an overparametrization assumption when working with a PINN approach. However, in contrast to supervised learning problems global convergence is not observed in practice. We believe the problems encountered in neural network based PDE solvers to be fundamentally different to supervised learning problems as samples are easy to generate. We believe that the theoretical understanding of the optimization process requires the development of new theoretical tools rather than an application of existing paradigms from the analysis of supervised learning problems. Further, we believe that the development of efficient optimizers are required for the advancement of neural network based approaches for the numerical analysis of PDEs. Chapter 6 is devoted to the development of a natural gradient method that achieves high accuracy for PINNs and the deep Ritz method.

5.2 Preliminaries regarding Sobolev spaces and neural networks

5.2.1. Notation of Sobolev spaces and Friedrich's inequality. We denote the space of functions on $\Omega\subseteq\mathbb{R}^d$ that are integrable in p-th power by $L^p(\Omega)$, where we assume that $p\in[1,\infty)$. Endowed with

$$\|u\|^p_{L^p(\Omega)}:=\int_\Omega|u|^p\mathrm{d}x$$

this is a Banach space, i.e., a complete normed space. If u is a multivariate function with values in $\mathbb{R}^m$ we interpret $|\cdot|$ as the Euclidean norm. We denote the subspace of $L^p(\Omega)$ of functions with weak derivatives up to order k in $L^p(\Omega)$ by $W^{k,p}(\Omega)$, which is a Banach space with the norm

$$\|u\|^p_{W^{k,p}(\Omega)} := \sum_{l=0}^{k} \|D^l u\|^p_{L^p(\Omega)}.$$

This space is called a *Sobolev space* and we denote its dual space, i.e., the space consisting of all bounded and linear functionals on $W^{k,p}(\Omega)$ by $W^{k,p}(\Omega)^*$. The closure of all compactly supported smooth functions $C_c^\infty(\Omega)$ in $W^{k,p}(\Omega)$ is denoted by $W_0^{k,p}(\Omega)$. It is well known that if Ω has a Lipschitz continuous boundary the operator that restricts a Lipschitz continuous function on $\overline{\Omega}$ to the boundary admits a linear and bounded extension $\operatorname{tr}\colon W^{1,p}(\Omega) \to L^p(\partial\Omega)$. This operator is called the *trace operator* and its kernel is precisely $W_0^{1,p}(\Omega)$. Further, we write $\|u\|_{L^p(\partial\Omega)}$ whenever we mean $\|\operatorname{tr}(u)\|_{L^p(\partial\Omega)}$. In the following we mostly work with the case $p = 2$ and write $H^k_{(0)}(\Omega)$ instead of $W^{k,2}_{(0)}(\Omega)$.

In order to study the boundary penalty method we use the Friedrich inequality, which states that the $L^p(\Omega)$ norm of a function can be estimated by the norm of its gradient and boundary values. We refer to [121] for a proof.

Proposition 5.7 (Friedrich's inequality). *Let $\Omega \subseteq \mathbb{R}^d$ be a bounded and open set with Lipschitz boundary $\partial\Omega$ and $p \in (1, \infty)$. Then there exists a constant $c > 0$ such that*

$$\|u\|^p_{W^{1,p}(\Omega)} \le c^p \cdot \left(\|\nabla u\|^p_{L^p(\Omega)} + \|u\|^p_{L^p(\partial\Omega)}\right) \quad \text{for all } u \in W^{1,p}(\Omega). \tag{5.18}$$

5.2.2. Neural networks. Here we introduce our notation for the functions represented by a feedforward neural network. Consider natural numbers $d, m, L, N_0, \dots, N_L \in \mathbb{N}$ and let

$$\theta = ((A_1, b_1), \dots, (A_L, b_L))$$

be a tuple of matrix-vector pairs where $A_l \in \mathbb{R}^{N_l \times N_{l-1}}, b_l \in \mathbb{R}^{N_l}$ and $N_0 = d, N_L = m$. Every matrix vector pair (A_l, b_l) induces an affine linear map $T_l\colon \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}$. The *neural network function with parameters* θ and with respect to some *activation function* $\rho\colon \mathbb{R} \to \mathbb{R}$ is the function

$$u_\theta^\rho\colon \mathbb{R}^d \to \mathbb{R}^m, \quad x \mapsto T_L(\rho(T_{L-1}(\rho(\cdots \rho(T_1(x)))))).$$

The set of all neural network functions of a certain architecture is given by $\{u_\theta^\rho : \theta \in \Theta\}$, where Θ collects all parameters of the above form with respect to fixed natural numbers $d, m, L, N_0, \dots, N_L$. If we have $f = u_\theta^\rho$ for some $\theta \in \Theta$ we say the function f can be *realized* by the neural network $\mathcal{F}_\Theta^\rho$. Note that we often drop the superscript ρ if it is clear from the context.

A particular activation function often used in practice and relevant for our results is the *rectified linear unit* or *ReLU activation function*, which is defined via $x \mapsto \max\{0, x\}$. [20] showed that the class of ReLU networks coincides with the class of continuous and piecewise linear functions. In particular they are weakly differentiable. Since piecewise linear functions are dense in $H_0^1(\Omega)$ we obtain the following universal approximation result.

Theorem 5.2 (Universal approximation with zero boundary values, [102]). *Consider an open set $\Omega \subseteq \mathbb{R}^d$ and let $u \in W_0^{1,p}(\Omega)$ with $p \in [1, \infty)$. Then for all $\varepsilon > 0$ there exists a function $u_\varepsilon \in W_0^{1,p}(\Omega)$ that can be expressed by a ReLU network of depth $\lceil \log_2(d+1) \rceil + 1$ such that*

$$\|u - u_\varepsilon\|_{W^{1,p}(\Omega)} \le \varepsilon.$$

Our proof uses that every continuous, piecewise linear function can be represented by a neural network with ReLU activation function and then shows how to approximate Sobolev functions with zero boundary conditions by such functions. The precise definition of a piecewise linear function is the following.

Definition 5.8 (Continuous piecewise linear function). We say a function $f\colon \mathbb{R}^d \longrightarrow \mathbb{R}$ is *continuous piecewise linear* or shorter *piecewise linear* if there exists a finite set of closed polyhedra whose union is $\mathbb{R}^d$, and f is affine linear over each polyhedron. Note every piecewise linear functions is continuous by definition since the polyhedra are closed and cover the whole space $\mathbb{R}^d$, and affine functions are continuous.

Theorem 5.9 (Universal expression). *Every ReLU neural network function $u_\theta\colon \mathbb{R}^d \longrightarrow \mathbb{R}$ is a piecewise linear function. Conversely, every piecewise linear function $f\colon \mathbb{R}^d \longrightarrow \mathbb{R}$ can be expressed by a ReLU network of depth at most $\lceil \log_2(d+1) \rceil + 1$.*

For the proof of this statement we refer to [20]. We turn now to the approximation capabilities of piecewise linear functions.

Lemma 5.10. *Let $\varphi \in C_c^\infty(\mathbb{R}^d)$ be a smooth function with compact support. Then for every $\varepsilon > 0$ there is a piecewise linear function s_ε such that for all $p \in [1, \infty]$ it holds*

$$\|s_\varepsilon - \varphi\|_{W^{1,p}(\mathbb{R}^d)} \leq \varepsilon \quad \textit{and} \quad \operatorname{supp}(s_\varepsilon) \subseteq \operatorname{supp}(\varphi) + B_\varepsilon(0).$$

Here, we set $B_\varepsilon(0)$ to be the ε-ball around zero, i.e. $B_\varepsilon(0) = \{x \in \mathbb{R} : |x| < \varepsilon\}$.

Proof. In the following we will denote by $\|\cdot\|_\infty$ the uniform norm on $\mathbb{R}^d$. To show the assertion choose a triangulation $\mathcal{T}$ of $\mathbb{R}^d$ of width $\delta = \delta(\varepsilon) > 0$, consisting of rotations and translations of one non-degenerate simplex K. We choose s_ε to agree with φ on all vertices of elements in $\mathcal{T}$. Since φ is compactly supported it is uniformly continuous and hence it is clear that $\|\varphi - s_\varepsilon\|_\infty < \varepsilon$ if δ is chosen small enough.

To show convergence of the gradients we show that also $\|\nabla\varphi - \nabla s_\varepsilon\|_\infty < \varepsilon$, which will be shown on one element $K \in \mathcal{T}$ and as the estimate is independent of K is understood to hold on all of $\mathbb{R}^d$. So let $K \in \mathcal{T}$ be given and denote its vertices by $x_1, \dots, x_{d+1}$. We set $v_i = x_{i+1} - x_1$, $i = 1, \dots, d$ to be the vectors spanning K. By the one dimensional mean value theorem we find ξ_i on the line segment joining x_1 and x_i such that

$$\partial_{v_i} s_\varepsilon(v_1) = \partial_{v_i}\varphi(\xi_i).$$

Note that $\partial_{v_i} s_\varepsilon$ is constant on all of K where it is defined. Now for arbitrary $x \in K$ we compute with setting $w = \sum_{i=1}^d \alpha_i v_i$ for $w \in \mathbb{R}^d$ with $|w| \leq 1$. Note that the α_i are bounded uniformly in w, where we use that all elements are the same up to rotations and translations.

$$\begin{aligned} |\nabla\varphi(x) - \nabla s_\varepsilon(x)| &= \sup_{|w|\leq 1} |\nabla\varphi(x)w - \nabla s_\varepsilon(x)w| \\ &\leq \sup_{|w|\leq 1} \sum_{i=1}^d |\alpha_i| \cdot \underbrace{|\partial_{v_i}\varphi(x) - \partial_{v_i} s_\varepsilon(x)|}_{=(*)} \end{aligned}$$

where again $(*)$ is uniformly small due to the uniform continuity of $\nabla\varphi$. Noting that the $W^{1,\infty}$-case implies the claim for all $p \in [1, \infty)$ finishes the proof. □

We turn to the proof of Theorem 5.2, which we state again for the convenience of the reader.

Proof of Theorem 5.2. Let $u \in W_0^{1,p}(\Omega)$ and $\varepsilon > 0$. By the density of $C_c^\infty(\Omega)$ in $W_0^{1,p}(\Omega)$, see for instance [65], we choose a smooth function $\varphi_\varepsilon \in C_c^\infty(\Omega)$ such that $\|u - \varphi_\varepsilon\|_{W^{1,p}(\Omega)} \leq \varepsilon/2$. Furthermore we use Lemma 5.10 and choose a piecewise linear function u_ε such that $\|\varphi_\varepsilon - u_\varepsilon\|_{W^{1,p}(\Omega)} \leq \varepsilon/2$ and such that u_ε has compact support in Ω. This yields

$$\|u - u_\varepsilon\|_{W^{1,p}(\Omega)} \leq \|u - \varphi_\varepsilon\|_{W^{1,p}(\Omega)} + \|\varphi_\varepsilon - u_\varepsilon\|_{W^{1,p}(\Omega)} \leq \varepsilon$$

and by Theorem 5.9 we know that u_ε is in fact a realization of a neural network with depth at most $\lceil \log_2(d+1) \rceil + 1$. □

To the best of our knowledge this is the only available universal approximation results where the approximating neural network functions are guaranteed to have zero boundary values. This relies on the special properties of the ReLU activation function and it is unclear for which classes of activation functions universal approximation with zero boundary values hold.

The difference of this result to other universal approximation results [141, 79] is the approximating neural network function are guaranteed to have zero boundary values. This is a special property of the ReLU activation function and implies the consistency of the boundary penalty method for arbitrary penalization strengths as we will see later. In order to quantify the error that is being made by the variational training of ReLU networks with boundary penalty, we use the following result from [126], where other results on approximation bounds in Sobolev spaces have been obtained in [127, 306, 262, 263, 138, 107, 88].

Theorem 5.11 (Quantitative universal approximation, [126]). *Let $\Omega \subseteq \mathbb{R}^d$ be a bounded and open set with Lipschitz regular boundary*[3]*, let $k \in (1, \infty)$, $p \in [1, \infty]$ and fix an arbitrary function $u \in W^{k,p}(\Omega)$. Then, for every $n \in \mathbb{N}$, there is a ReLU network u_n with $O(\log_2^2(n^{k/d}) \cdot n)$ many parameters and neurons such that*

$$\|u - u_n\|_{W^{s,p}(\Omega)} \leq c(s) \cdot \|u\|_{W^{k,p}(\Omega)} \cdot n^{-(k-s)/d}$$

for every $s \in [0, 1]$.

Proof. The approximation results in [126] are stated for functions with the unit cube $[0,1]^d$ as a domain. However, by scaling and possibly extending functions to the whole of $\mathbb{R}^d$ this implies analogue results for functions defined on bounded Sobolev extension domains Ω.

We examine the proof of [126] in order to see that the approximating network architectures do not depend on s. Let us fix a function $u \in W^{k,p}([0,1]^d)$. In their notation, for $M \in \mathbb{N}$, there are functions $(\phi_m)_{m=1,\dots,M^d}$ and polynomials $(p_m)_{m=1,\dots,M^d}$, such that

$$\left\| u - \sum \phi_m p_m \right\|_{W^{s,p}([0,1]^d)} \lesssim \|u\|_{W^{k,p}([0,1]^d)} M^{-(k-s)}$$

and a ReLU network function u_M with $N \lesssim M^d \log(M^k)$ parameters such that

$$\left\| \sum \phi_m p_m - u_M \right\|_{W^{s,p}([0,1]^d)} \leq c(s) \|u\|_{W^{k,p}(\Omega)} M^{-(k-s)}.$$

This follows from the Lemma C.3, C.4 and Lemma C.6 in [126] with $\varepsilon = M^{-k}$. Note that the functions and networks provided by those lemmata do not depend on s, which is evident

[3] or more generally, that Ω is a Sobolev extension domain

as the estimates are first shown for $s = 0, 1$ and then generalized through interpolation. Now, by the triangle inequality, we have

$$\|u - u_M\|_{W^{s,p}([0,1]^d)} \leq \tilde{c}(s)\|u\|_{W^{k,p}([0,1]^d)} M^{-(k-s)}.$$

Now the claim follows by choosing $M \sim n^{1/d}$. The additional square of the logarithm appears since they are considering networks with skip connections and those are then expressed as networks without skip connections, see also Corollary 4.2 in [126]. □

5.3 Proofs regarding convergence guarantees for the deep Ritz method for nonlinear problems

5.3.1. Prior on Γ-convergence. We recall the definition of Γ-convergence with respect to the weak topology of reflexive Banach spaces. For further reading we point the reader towards [83].

Definition 5.12 (Γ-convergence). Let X be a reflexive Banach space as well as $F_n, F\colon X \to (-\infty, \infty]$. Then $(F_n)_{n\in\mathbb{N}}$ is said to be Γ-*convergent* to F if the following two properties are satisfied.

(*i*) *Liminf inequality:* For every $x \in X$ and $(x_n)_{n\in\mathbb{N}}$ with $x_n \rightharpoonup x$ we have

$$F(x) \leq \liminf_{n\to\infty} F_n(x_n).$$

(*ii*) *Recovery sequence:* For every $x \in X$ there is $(x_n)_{n\in\mathbb{N}}$ with $x_n \rightharpoonup x$ such that

$$F(x) = \lim_{n\to\infty} F_n(x_n).$$

The sequence $(F_n)_{n\in\mathbb{N}}$ is called *equi-coercive* if the set

$$\bigcup_{n\in\mathbb{N}} \left\{x \in X : F_n(x) \leq r\right\}$$

is bounded in X (or equivalently relatively compact with respect to the weak topology) for all $r \in \mathbb{R}$. We say that a sequence $(x_n)_{n\in\mathbb{N}}$ are *quasi minimizers* of the functionals $(F_n)_{n\in\mathbb{N}}$ if we have

$$F_n(x_n) \leq \inf_{x\in X} F_n(x) + \delta_n$$

where $\delta_n \to 0$.

We need the following property of Γ-convergent sequences. We want to emphasize the fact that there are no requirements regarding the continuity of any of the functionals and that the functionals $(F_n)_{n\in\mathbb{N}}$ are not assumed to admit minimizers.

Theorem 5.13 (Convergence of quasi-minimizers). *Let X be a reflexive Banach space and $(F_n)_{n\in\mathbb{N}}$ be an equi-coercive sequence of functionals that* Γ*-converges to F. Then, any sequence $(x_n)_{n\in\mathbb{N}}$ of quasi-minimizers of $(F_n)_{n\in\mathbb{N}}$ is relatively compact with respect to the weak topology of X and every weak accumulation point of $(x_n)_{n\in\mathbb{N}}$ is a global minimizer of F. Consequently, if F possesses a unique minimizer x, then $(x_n)_{n\in\mathbb{N}}$ converges weakly to x.*

5.3.2. Abstract Γ-convergence result for the deep Ritz method. For the abstract results we work with an abstract energy $E\colon X \to \mathbb{R}$. This reduces technicalities in the proofs and separates abstract functional analytic considerations from applications.

Setting 5.14. *Let $(X, \|\cdot\|_X)$ and $(B, \|\cdot\|_B)$ be reflexive Banach spaces and $\gamma \in \mathcal{L}(X, B)$ be a continuous linear map. We set X_0 to be the kernel of γ, i.e., $X_0 = \gamma^{-1}(\{0\})$. Let $\rho\colon \mathbb{R} \to \mathbb{R}$ be some activation function and denote by $(\Theta_n)_{n\in\mathbb{N}}$ a sequence of neural network parameters. We assume that any function represented by such a neural network is a member of X and we define*

$$A_n := \{x_\theta : \theta \in \Theta_n\} \subseteq X. \tag{5.19}$$

Here, x_θ denotes the function represented by the neural network with the parameters θ. Let $E\colon X \to (-\infty, \infty]$ be a functional and $(\lambda_n)_{n\in\mathbb{N}}$ a sequence of real numbers with $\lambda_n \to \infty$. Furthermore, let $p \in (1, \infty)$ and $f \in X^*$ be fixed and define the functional $F_n^f\colon X \to (-\infty, \infty]$ by

$$F_n^f(x) = \begin{cases} E(x) + \lambda_n \|\gamma(x)\|_B^p - f(x) & \text{for } x \in A_n, \\ \infty & \text{otherwise}, \end{cases}$$

as well as $F^f\colon X \to (-\infty, \infty]$ by

$$F^f(x) = \begin{cases} E(x) - f(x) & \text{for } x \in X_0, \\ \infty & \text{otherwise}. \end{cases}$$

Then assume the following holds:

(B.A1) *For every $x \in X_0$ there is $x_n \in A_n$ such that $x_n \to x$ and $\lambda_n \|\gamma(x_n)\|_B^p \to 0$ for $n \to \infty$.*

(B.A2) *The functional E is bounded from below, weakly lower semi-continuous with respect to the weak topology of $(X, \|\cdot\|_X)$ and continuous with respect to the norm topology of $(X, \|\cdot\|_X)$.*

(B.A3) *The sequence $(F_n^f)_{n\in\mathbb{N}}$ is equi-coercive with respect to the norm $\|\cdot\|_X$.*

Remark 5.15. We discuss the Assumptions (B.A1) to (B.A3) in view of their applicability to concrete problems.

(*i*) In applications, $(X, \|\cdot\|_X)$ will usually be a Sobolev space with its natural norm, the space B contains boundary values of functions in X and the operator γ is a boundary value operator, e.g. the trace map. However, if the energy E is coercive on all of X, i.e. without adding boundary terms to it, we might choose $\gamma = 0$ and obtain $X_0 = X$. This is the case for non-essential boundary value problems.

(*ii*) The Assumption (B.A1) compensates that in general, we cannot penalize with arbitrary strength. However, if we can approximate any member of X_0 by a sequence $x_{\theta_n} \in A_n \cap X_0$ then any divergent sequence $(\lambda_n)_{n\in\mathbb{N}}$ can be chosen. This is for example the case for the ReLU activation function and the space $X_0 = H_0^1(\Omega)$. More precisely, we can choose A_n to be the class of functions expressed by a (fully connected) ReLU network of depth $\lceil \log_2(d+1) \rceil + 1$ and width n, see Theorem 5.2.

Theorem 5.16 (Γ-convergence). *Assume we are in Setting 5.14. Then the sequence $(F_n^f)_{n\in\mathbb{N}}$ of functionals Γ-converges towards F^f. In particular, if $(\delta_n)_{n\in\mathbb{N}}$ is a sequence of non-negative real* *numbers converging to zero, any sequence of δ_n-quasi minimizers of F_n^f is bounded and all its weak accumulation points are minimizers of F^f. If additionally F^f possesses a unique minimizer $x^f \in X_0$, any sequence of δ_n-quasi minimizers converges to x^f in the weak topology of X.*

Proof. We begin with the limes inferior inequality. Let $x_n \rightharpoonup x$ in X and assume that $x \notin X_0$. Then $f(x_n)$ converges to $f(x)$ as real numbers and $\gamma(x_n)$ converges weakly to $\gamma(x) \neq 0$ in B. Combining this with the weak lower semi-continuity of $\|\cdot\|_B^p$ we get, using the boundedness from below, that

$$\liminf_{n\to\infty} F_n^f(x_n) \geq \inf_{x\in X} E(x) + \liminf_{n\to\infty} \lambda_n \|\gamma(x_n)\|_B^p - \lim_{n\to\infty} f(x_n) = \infty.$$

Now let $x \in X_0$. Then by the weak lower semi-continuity of E we find

$$\liminf_{n\to\infty} F_n^f(x_n) \geq \liminf_{n\to\infty} E(x_n) - f(x) \geq E(x) - f(x) = F^f(x).$$

Now let us have a look at the construction of the recovery sequence. For $x \notin X_0$ we can choose the constant sequence and estimate

$$F_n^f(x_n) \geq E(x) + \lambda_n \|\gamma(x)\|_B^p - f(x).$$

Hence we find that $F_n^{f_n}(x) \to \infty = F^f(x)$. If $x \in X_0$ we approximate it with a sequence $(x_n) \subseteq X$ according to Assumption (B.A1), such that $x_n \in A_n$ and $x_n \to x$ in $\|\cdot\|_X$ and $\lambda_n \|\gamma(x_n)\|_B^p \to 0$. It follows that

$$F_n^f(x_n) = E(x_n) + \lambda_n \|x_n\|_B^p - f(x_n) \to E(x) - f(x) = F^f(x).$$

□

A sufficient criterion for equi-coercivity of the sequence $(F_n^f)_{n\in\mathbb{N}}$ from Assumption (B.A3) in terms of the functional E is given by the following lemma.

Lemma 5.17 (Criterion for Equi-Coercivity). *Assume we are in Setting 5.14. If there is a constant $c > 0$ such that it holds for all $x \in X$ that*

$$E(x) + \|\gamma(x)\|_B^p \geq c \cdot \left(\|x\|_X^p - \|x\|_X - 1\right),$$

then the sequence $(F_n^f)_{n\in\mathbb{N}}$ is equi-coercive.

Proof. It suffices to show that the sequence

$$G_n^f \colon X \to \mathbb{R} \quad \text{with} \quad G_n^f(x) = E(x) + \lambda_n \|\gamma(x)\|_B^p - f(x)$$

is equi-coercive, as $G_n^f \leq F_n^f$. So let $r \in \mathbb{R}$ be given and assume that $r \geq G_n^f(x)$. We estimate assuming without loss of generality that $\lambda_n \geq 1$

$$\begin{aligned} r &\geq E(x) + \lambda_n \|\gamma(x)\|_B^p - f(x) \\ &\geq c \cdot \left(\|x\|_X^p - \|x\|_X - 1\right) - \|f\|_{X^*} \cdot \|x\|_X \\ &\geq \tilde{c} \cdot \left(\|x\|_X^p - \|x\|_X - 1\right). \end{aligned}$$

As $p > 1$, a scaled version of Young's inequality clearly implies a bound on the set

$$\bigcup_{n\in\mathbb{N}} \left\{x \in X : G_n^f(x) \leq r\right\}$$

and hence the sequence $(F_n^f)_{n\in\mathbb{N}}$ is seen to be equi-coercive. □

5.3.3. Abstract uniform convergence result for the deep Ritz method. In this section we present an extension of Setting 5.14 that allows to prove uniform convergence results over certain bounded families of right-hand sides.

Setting 5.18. *Assume we are in Setting 5.14. Furthermore, let there be an additional norm $|\cdot|$ on X such that the dual space $(X, |\cdot|)^*$ is reflexive. However, we do not require $(X, |\cdot|)$ to be complete. Then, let the following assumptions hold*

(B.A4) *The identity* $\operatorname{Id}\colon (X, \|\cdot\|_X) \to (X, |\cdot|)$ *is completely continuous, i.e., maps weakly convergent sequences to strongly convergent ones.*

(B.A5) *For every $f \in X^*$, there is a unique minimizer $x_f \in X_0$ of F^f and the solution map*

$$S\colon X_0^* \to X_0 \quad \textit{with } f \mapsto x^f$$

is demi-continuous, i.e. maps strongly convergent sequences to weakly convergent ones.

Remark 5.19. As mentioned earlier, $(X, \|\cdot\|_X)$ is usually a Sobolev space with its natural norm. The norm $|\cdot|$ may then chosen to be an $L^p(\Omega)$ or $W^{s,p}(\Omega)$ norm, where s is strictly smaller than the differentiability order of X. In this case, Rellich's compactness theorem [65] provides Assumption (B.A4).

Lemma 5.20 (Compactness). *Assume we are in Setting 5.18. Then the solution operator $S\colon (X, |\cdot|)^* \to (X_0, |\cdot|)$ is completely continuous, i.e., maps weakly convergent sequences to strongly convergent ones.*

Proof. We begin by clarifying what we mean with S being defined on $(X, |\cdot|)^*$. Denote by i the inclusion map $i\colon X_0 \to X$ and consider

$$(X, |\cdot|)^* \xrightarrow{\operatorname{Id}^*} (X, \|\cdot\|_X)^* \xrightarrow{i^*} (X_0, \|\cdot\|_X)^* \xrightarrow{S} (X_0, \|\cdot\|_X) \xrightarrow{\operatorname{Id}} (X_0, |\cdot|).$$

By abusing notation, always when we refer to S as defined on $(X, |\cdot|)^*$ we mean the above composition, i.e., $\operatorname{Id} \circ S \circ i^* \circ \operatorname{Id}^*$. Having explained this, it is clear that it suffices to show that Id^* maps weakly convergent sequences to strongly convergent ones since i^* is continuous, S demi-continuous and Id strongly continuous. This, however, is a consequence of Schauder's theorem, see for instance [9], which states that a linear map $L \in \mathcal{L}(X, Y)$ between Banach spaces is compact if and only if $L^* \in \mathcal{L}(Y^*, X^*)$ is. Here, compact means that L maps bounded sets to relatively compact ones. Let X_c denote the completion of $(X, |\cdot|)$. Then, using the reflexivity of $(X, \|\cdot\|_X)$ it is easily seen that $\operatorname{Id}\colon (X, \|\cdot\|_X) \to X_c$ is compact. Finally, using that $(X, |\cdot|)^* = X_c^*$ the desired compactness of Id^* is established. □

The following theorem is the main result of this section. It shows that the convergence of the Deep Ritz method is uniform on bounded sets in the space $(X, |\cdot|)^*$. The proof of the uniformity follows an idea from [76], where in a different setting a compactness result was used to amplify pointwise convergence to uniform convergence across bounded sets, compare to Theorem 4.1 and Corollary 4.2 in [76].

Theorem 5.21 (Uniform Convergence of the Deep Ritz Method). *Assume that we are in Setting 5.18 and let $\delta_n \searrow 0$ be a sequence of real numbers. For $f \in X^*$ we set*

$$S_n(f) := \left\{ x \in X : F_n^f(x) \leq \inf_{z \in X} F_n^f(z) + \delta_n \right\},$$

which is the approximate solution set corresponding to f and δ_n. Furthermore, denote the unique minimizer of F^f in X_0 by x^f and fix $R > 0$. Then we have

$$\sup\left\{|x_n^f - x^f| : x_n^f \in S_n(f),\ \|f\|_{(X,\,|\cdot|)^*} \leq R\right\} \to 0 \quad \textit{for } n \to \infty.$$

In the definition of this supremum, f is measured in the norm of the space $(X, |\cdot|)^*$. This means that $f : (X, |\cdot|) \to \mathbb{R}$ is continuous, which is a more restrictive requirement than the continuity with respect to $\|\cdot\|_X$. Also the computation of this norm takes place in the unit ball of $(X, |\cdot|)$, i.e.

$$\|f\|_{(X,\,|\cdot|)^*} = \sup_{|x|\leq 1} f(x).$$

Before we prove Theorem 5.21 we need a Γ-convergence result similar to Theorem 5.16. The only difference is, that now also the right-hand side may vary along the sequence.

Proposition 5.22. *Assume that we are in Setting 5.18, however, we do not need Assumption* (B.A5) *for this result. Let $f_n, f \in (X, |\cdot|)^*$ such that $f_n \rightharpoonup f$ in the weak topology of the reflexive space $(X, |\cdot|)^*$. Then the sequence $(F_n^{f_n})_{n\in\mathbb{N}}$ of functionals Γ-converges to F^f in the weak topology of $(X, \|\cdot\|_X)$. Furthermore, the sequence $(F_n^{f_n})_{n\in\mathbb{N}}$ is equi-coercive.*

Proof. The proof is almost identical to the one of Theorem 5.16 but since it is brief, we include it for the reader's convenience. We begin with the limes inferior inequality. Let $x_n \rightharpoonup x$ in X and $x \notin X_0$. Then $x_n \to x$ with respect to $|\cdot|$, which implies that $f_n(x_n)$ converges to $f(x)$. Using that $\gamma(x_n) \rightharpoonup \gamma(x)$ in B combined with the weak lower semi-continuity of $\|\cdot\|_B^p$ we get

$$\liminf_{n\to\infty} F_n^{f_n}(x_n) \geq \inf_{x\in X} E(x) + \liminf_{n\to\infty} \lambda_n \|\gamma(x_n)\|_B^p - \lim_{n\to\infty} f_n(x_n) = \infty.$$

Now let $x \in X_0$. Then by the weak lower semi-continuity of E we find

$$\liminf_{n\to\infty} F_n^{f_n}(x_n) \geq \liminf_{n\to\infty} E(x_n) - f(x) \geq E(x) - f(x) = F^f(x).$$

Now let us have a look at the construction of the recovery sequence. For $x \notin X_0$ we can choose the constant sequence and estimate

$$F_n^{f_n}(x) \geq \inf_{x\in X} E(x) + \lambda_n \|\gamma(x)\|_B^p - \|f_n\|_{(X,|\cdot|)'} \cdot |x|.$$

As $\|f_n\|_{(X,|\cdot|)^*}$ is bounded we find $F_n^{f_n}(x) \to \infty = F^f(x)$. If $x \in X_0$ we approximate it with a sequence $(x_n) \subseteq X$ according to Assumption (B.A1), such that $x_n \in A_n$ and $x_n \to x$ in $\|\cdot\|_X$ and $\lambda_n \|\gamma(x_n)\|_B^p \to 0$. It follows that

$$F_n^{f_n}(x_n) = E(x_n) + \lambda_n \|x_n\|_B^p - f_n(x_n) \to E(x) - f(x) = F^f(x).$$

The equi-coercivity was already assumed in (B.A3) so it does not need to be shown. □

Proof of Theorem 5.21. We can choose $(f_n) \subseteq (X, |\cdot|)^*$ and $\|f_n\|_{(X,|\cdot|)^*} \leq R$ and $x_n^{f_n} \in S_n(f_n)$ such that

$$\sup_{\substack{\|f\|_{(X,\,|\cdot|)^*}\leq R \\ x_n^f \in S_n(f)}} \left|x_n^f - x^f\right| \leq \left|x_n^{f_n} - x^{f_n}\right| + \frac{1}{n}.$$

Now it suffices to show that $|x_n^{f_n} - x^{f_n}|$ converges to zero. Since $(f_n)_{n\in\mathbb{N}}$ is bounded in $(X, |\cdot|)^*$ and this space is reflexive we can without loss of generality assume that $f_n \rightharpoonup f$

in $(X, |\cdot|)^*$. This implies by Lemma 5.20 that $x^{f_n} \to x^f$ in $(X, |\cdot|)$. The Γ-convergence result of the previous proposition yields $x_n^{f_n} \rightharpoonup x^f$ in X and hence $x_n^{f_n} \to x^f$ with respect to $|\cdot|$, which concludes the proof. □

5.3.4. A nonlinear PDE: The p-Laplace. As an example for the uniform convergence of the Deep Ritz method we discuss the p-Laplacian. To this end, consider the p-Dirichlet energy for $p \in (1, \infty)$ given by

$$E \colon W^{1,p}(\Omega) \to \mathbb{R}, \quad u \mapsto \frac{1}{p} \int_\Omega |\nabla u|^p \, \mathrm{d}x.$$

Note that for $p \neq 2$ the associated Euler-Lagrange equation – the p-Laplace equation – is nonlinear. In strong formulation it is given by

$$\begin{aligned} -\operatorname{div}(|\nabla u|^{p-2} \nabla u) &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

see for example [274] or [246]. Choosing the ReLU activation function, the abstract setting is applicable as we will describe now. For the Banach spaces we choose

$$X = W^{1,p}(\Omega), \quad B = L^p(\partial\Omega), \quad |u| = \|u\|_{L^p(\Omega)}$$

where the norms $\|\cdot\|_X$ and $\|\cdot\|_B$ are chosen to be the natural ones. Clearly, $W^{1,p}(\Omega)$ endowed with the norm $\|\cdot\|_{W^{1,p}(\Omega)}$ is reflexive by our assumption $p \in (1, \infty)$. Note that it holds

$$\left(W^{1,p}(\Omega), \|\cdot\|_{L^p(\Omega)} \right)^* = L^p(\Omega)^* \cong L^{p'}(\Omega),$$

which is also reflexive. We set $\gamma = \operatorname{tr}$, i.e.

$$\operatorname{tr} \colon W^{1,p}(\Omega) \to L^p(\partial\Omega) \quad \text{with} \quad u \mapsto u|_{\partial\Omega}$$

We use the same ansatz sets $(A_n)_{n \in \mathbb{N}}$ as in the previous example, hence Assumption (B.A1) holds. Rellich's theorem provides the complete continuity of the embedding

$$\left(W^{1,p}(\Omega), \|\cdot\|_{W^{1,p}(\Omega)} \right) \to \left(W^{1,p}(\Omega), \|\cdot\|_{L^p(\Omega)} \right)$$

which shows Assumption (B.A4). As for Assumption (B.A3), Friedrich's inequality provides the assumptions of Lemma 5.17. Furthermore, E is continuous with respect to $\|\cdot\|_{W^{1,p}(\Omega)}$ and convex, hence also weakly lower semi-continuous. By Poincaré's and Young's inequality we find for all $u \in W_0^{1,p}(\Omega)$ that

$$\begin{aligned} F^f(u) &= \frac{1}{p} \int_\Omega |\nabla u|^p \mathrm{d}x - f(u) \\ &\geq C \|u\|_{W^{1,p}(\Omega)}^p - \|f\|_{W^{1,p}(\Omega)'} \|u\|_{W^{1,p}(\Omega)} \\ &\geq C \|u\|_{W^{1,p}(\Omega)}^p - \tilde{C}. \end{aligned}$$

Hence, a minimizing sequence in $W_0^{1,p}(\Omega)$ for F^f is bounded and as F^f is strictly convex on $W_0^{1,p}(\Omega)$ it possesses a unique minimizer. Finally, to provide the demi-continuity we must consider the operator $S\colon W_0^{1,p}(\Omega)^* \to W_0^{1,p}(\Omega)$ mapping f to the unique minimizer u_f of $E - f$ on $W_0^{1,p}(\Omega)$. By the Euler-Lagrange formalism, u minimizes F^f if and only if

$$\int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla v \mathrm{d}x = f(v) \quad \text{for all } v \in W_0^{1,p}(\Omega).$$

Hence, the solution map S is precisely the inverse of the mapping

$$W_0^{1,p}(\Omega) \to W_0^{1,p}(\Omega)^*, \quad u \mapsto \left(v \mapsto \int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla v \mathrm{d}x \right)$$

and this map is demi-continuous, see for example [246].

5.4 Proofs regarding error estimates for the deep Ritz method with boundary penalty

5.4.1. A Céa lemma. The following proof of Céa's Lemma is based on the curvature properties of a quadratic, coercive energy defined on a Hilbert space. Note that in the following proposition, V does not need to be a vector space.

Proposition 5.23 (Céa's Lemma). *Let X be a Hilbert space, $V \subseteq X$ any subset and $a\colon X \times X \to \mathbb{R}$ a symmetric, continuous and α-coercive bilinear form. For $f \in X^*$ define the quadratic energy $E(u) := \frac{1}{2}a(u,u) - f(u)$ and denote its unique minimizer by u^*. Then for every $v \in V$ it holds that*

$$\|v - u^*\|_X \leq \sqrt{\frac{2\delta}{\alpha} + \frac{1}{\alpha}\inf_{\tilde{v}\in V}\|\tilde{v} - u^*\|_a^2},$$

where $\delta = E(v) - \inf_{\tilde{v}\in V} E(\tilde{v})$ and $\|u\|_a^2 := a(u,u)$ is the norm induced by a.

Proof. As E is quadratic it can be exactly expanded using Taylor's formula. Hence, for every $h \in X$ it holds that

$$E(u + h) = E(u^*) + \frac{1}{2}D^2E(u^*)(h,h) = E(u^*) + \frac{1}{2}a(h,h) = E(u^*) + \frac{1}{2}\|h\|_a^2,$$

where we used $DE(u^*) = 0$. Inserting $v - u^*$ for h we obtain

$$E(v) - E(u^*) = \frac{1}{2}a(v - u^*, v - u^*) \geq \frac{\alpha}{2}\|v - u^*\|_X^2.$$

On the other hand we compute

$$\begin{aligned} E(v) - E(u^*) &= E(v) - \inf_{\tilde{v}\in V} E(\tilde{v}) + \inf_{\tilde{v}\in V}(E(\tilde{v}) - E(u^*)) \\ &= \delta + \frac{1}{2}\inf_{\tilde{v}\in V}\|\tilde{v} - u^*\|_a^2. \end{aligned}$$

Combining the two estimates and rearranging terms yields the assertion. □

Proposition 5.23 is all we need to derive error estimates for coercive problems with non-essential boundary conditions. We give an example.

Corollary 5.24 (Neumann Problem). *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain and let f be a fixed member of $H^1(\Omega)^*$. Denote by $u \in H^1(\Omega)$ the weak solution to*

$$-\Delta u + u = f \quad \text{in } H^1(\Omega)^*.$$

Let Θ be the parameter set of a neural network architecture such that $u_\theta \in H^1(\Omega)$ for every $\theta \in \Theta$. Then for every $\theta \in \Theta$ it holds

$$\|u_\theta - u\|_{H^1(\Omega)} \leq \sqrt{2\delta + \inf_{\eta\in\Theta} \|u_\eta - u\|^2_{H^1(\Omega)}}$$

where

$$\delta = \|u_\theta\|^2_{H^1(\Omega)} - f(u_\theta) - \inf_{\eta\in\Theta}\left[\|u_\eta\|^2_{H^1(\Omega)} - f(u_\eta)\right].$$

Proof. The bilinear form corresponding to the above Neumann problem is

$$a\colon H^1(\Omega)\times H^1(\Omega) \to \mathbb{R}, \quad a(u,v) = \int_\Omega \nabla u \nabla v + uv \mathrm{d}x$$

and therefore its coercivity constant is $\alpha = 1$ and the associated norm $\|\cdot\|_a$ is the natural one on $H^1(\Omega)$. Employing Proposition 5.23 yields the assertion. □

Remark 5.25. Corollary 5.24 yields $H^1(\Omega)$ convergence of the Deep Ritz Method provided the ansatz class possesses universal approximation properties with respect to the $H^1(\Omega)$ norm. This is of course also a necessary requirement and fulfilled by a wide class of network architectures and activation functions, see [141, 79] or for approximation rates [126]. We stress that any (quantitative) universal approximation theorem for Sobolev topologies can be combined with the above result, such as Theorem 5.11 for ReLU neural networks.

Furthermore, the form of the differential equation in the above corollary can easily be generalized. One can for example consider general second order elliptic PDEs in divergence form with non-essential boundary conditions as long as these are coercive and can be derived from a minimization principle.

Remark 5.26 (Dimension Dependence and Adaptation to Smoothness). Assume the solution u to the Neumann problem is a member of $H^k(\Omega)$ for some $k > 1$. Then applying the quantitative universal approximation Theorem 5.11 we estimate

$$\|u_\theta - u\|_{H^1(\Omega)} \lesssim \sqrt{2\delta + c\cdot \|u\|_{H^k(\Omega)} n^{-(k-1)/d}},$$

where $d \in \mathbb{N}$ is the spatial dimension. While this estimate is not dimension independent, it indicates how smoothness mitigates the deterioration of error decay rates for high dimensions. We see that the merit of neural networks to achieve approximation rates increasing with the smoothness of the target function carries over to the error decay in the deep Ritz method. In contrast, to achieve the approximation rate and an error decay rate of $(k-1)/d$ with finite elements one needs to for example use P^{k-1} elements [111], which complicates the ansatz class and therefore the approach.

Remark 5.27 (Practical Realization of Rates). There is a gap between the theory and the practice of neural network based methods for the solution of PDEs. Error decay rates, as predicted by our results cannot be observed in practice due to the difficulties of computing minima of non-convex functions. In practice, one observes moderate errors that don't decrease beyond a certain accuracy when the numbers of the parameters of the neural network ansatz architecture are increased. We refer to [170] and the references therein for a more detailed description of this phenomenon.

However, what one observes is that neural network based methods in general and the Deep Ritz Method in particular are well suited for problems in high spatial dimensions or a high dimensional parameter space. Practical evidence can already be found in the paper introducing the Deep Ritz Method, see [110] and further high dimensional examples – even of industrial scale – can be found in [132]. We propose to view our results as a qualitative explanation of these observations.

5.4.2. An error estimate for the boundary penalty method. The treatment of Dirichlet boundary conditions corresponds to a constrained optimization problem, as in standard neural network architectures zero boundary values cannot be directly encoded. We use the boundary penalty method as a way to enforce Dirichlet boundary conditions. For ease of presentation, we discuss our approach for the concrete equation

$$\begin{aligned} -\operatorname{div}(A\nabla u) &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{5.20}$$

where $A \in L^\infty(\Omega, \mathbb{R}^{d\times d})$ is a symmetric and elliptic coefficient matrix. The weak formulation of this equation gives rise to the bilinear form

$$a\colon H^1(\Omega)\times H^1(\Omega) \to \mathbb{R}, \quad a(u,v) = \int_\Omega A\nabla u\cdot\nabla v\mathrm{d}x$$

and the energy

$$E\colon H^1(\Omega) \to \mathbb{R}, \quad E(u) = \frac{1}{2}a(u,u) - f(u)$$

where $f \in H^1(\Omega)^*$. Using the boundary penalty method as an approximation for (5.20) leads to the bilinear form

$$a_\lambda\colon H^1(\Omega)\times H^1(\Omega) \to \mathbb{R}, \quad a_\lambda(u,v) = \int_\Omega A\nabla u\nabla v\mathrm{d}x + \lambda\int_{\partial\Omega} uv\mathrm{d}s$$

for a penalty parameter $\lambda > 0$ and the energy

$$E_\lambda\colon H^1(\Omega) \to \mathbb{R}, \quad E_\lambda(u) = \frac{1}{2}a_\lambda(u,u) - f(u).$$

The central error estimation is collected in the following Theorem. Note that we require $H^2(\Omega)$ regularity of the solution to equation (5.20).

Theorem 5.28. *Let $\Omega\subset\mathbb{R}^d$ be a bounded domain with $C^{1,1}$ boundary, $f\in L^2(\Omega)$ and assume $A\in C^{0,1}(\Omega,\mathbb{R}^{d\times d})$ is symmetric, uniformly elliptic with ellipticity constant $\alpha > 0$. By $u^*\in H^1_0(\Omega)$ we denote the solution of (5.20) and by u_λ the minimizer of the penalized energy E_λ over $H^1(\Omega)$. Fix an arbitrary subset $V\subset H^1(\Omega)$ and denote the coercivity constants of a_λ by $\alpha_\lambda > 0$ and set $\delta := E_\lambda(v) - \inf_{\tilde v\in V} E_\lambda(\tilde v)$. Then there is a constant $c > 0$, only depending on A and Ω, such that for every $v\in V$ and $\lambda > 0$ it holds that*

$$\|v-u^*\|_{H^1(\Omega)} \leq \sqrt{\frac{2\delta}{\alpha_\lambda} + \frac{1}{\alpha_\lambda}\inf_{\tilde v\in V}\|\tilde v - u_\lambda\|^2_{a_\lambda}} + c\lambda^{-1}\|f\|_{L^2(\Omega)}, \tag{5.21}$$

where $\|u\|^2_{a_\lambda} := a_\lambda(u,u)$ is the norm induced by a_λ. Further, we can choose

$$c := c_F c_{reg}\sqrt{\|a_1\|}\|T\|_{\mathcal{L}(H^2(\Omega);\mathcal{H}(\Omega))},$$

where $T\colon H^2(\Omega) \to H^1(\Omega)$ maps a function u to the A-harmonic extension of[4] $\partial_A u$, c_F *denotes the Friedrich constant (see Proposition 5.7) and c_{reg} is the operator norm of*

$$(-\operatorname{div}(A\nabla\cdot))^{-1}\colon L^2(\Omega) \to H^2(\Omega) \cap H^1_0(\Omega).$$

Proof. The central idea of the proof consists of the error decomposition

$$\|v - u^*\|_{H^1(\Omega)} \le \|v - u_\lambda\|_{H^1(\Omega)} + \|u_\lambda - u^*\|_{H^1(\Omega)}.$$

The first error can be treated using Céa's Lemma. Note that a_λ is in fact coercive on $H^1(\Omega)$, which is a consequence of Friedrich's inequality, see Proposition 5.7. For the second term one uses a Fourier series expansion in a Steklov basis. The latter is useful for weakly A-harmonic functions, hence we investigate the equation satisfied by $v_\lambda := u^* - u_\lambda$. Due to the regularity assumption on Ω and A we have $\operatorname{div}(A\nabla u^*) \in L^2(\Omega)$ and may integrate by parts to obtain for all $\varphi \in H^1(\Omega)$

$$\int_\Omega f\varphi \mathrm{d}x = -\int_\Omega \operatorname{div}(A\nabla u^*)\varphi \mathrm{d}x = \int_\Omega A\nabla u^* \nabla\varphi \mathrm{d}x - \int_{\partial\Omega} \partial_A u^* \varphi \mathrm{d}s. \tag{5.22}$$

Using the optimality condition of u_λ yields

$$\int_\Omega (A\nabla u_\lambda)\cdot\nabla\varphi \mathrm{d}x + \lambda\int_{\partial\Omega} u_\lambda\varphi \mathrm{d}s = \int_\Omega f\varphi \mathrm{d}x \quad \forall\varphi \in H^1(\Omega).$$

Subtracting these two equations we obtain that v_λ satisfies

$$\int_\Omega (A\nabla v_\lambda)\cdot\nabla\varphi \mathrm{d}x + \int_{\partial\Omega} (\lambda v_\lambda - \partial_A u^*)\varphi \mathrm{d}s = 0 \quad \forall\varphi \in H^1(\Omega).$$

This implies that v_λ is weakly A-harmonic, i.e.,

$$\int_\Omega (A\nabla v_\lambda)\cdot\nabla\varphi \mathrm{d}x = 0 \quad \forall\varphi \in H^1_0(\Omega),$$

We claim that there exists a basis $(e_j)_{j\in\mathbb{N}}$ of the space of weakly A-harmonic functions and that v_λ can be written in terms of this basis as

$$v_\lambda = \frac{1}{\lambda}\sum_{j=0}^{\infty} c(\lambda)_j e_j \tag{5.23}$$

for suitable coefficients $c(\lambda)_j \in \mathbb{R}$. Further, we claim that this Fourier expansion leads to the estimate

$$\|v_\lambda\|_{H^1(\Omega)} \le \frac{c}{\lambda}\|f\|_{L^2(\Omega)} \tag{5.24}$$

with c as specified in the statement of the Theorem, which finishes the proof. The remaining details are provided in the following Section. □

We presented the proof in its above form to draw attention to its key elements and to discuss possible limitations and generalizations.

Remark 5.29 (Limitations). Our proof requires crucially the $H^2(\Omega)$ regularity of the solution u^* to the Dirichlet problem. This is in contrast with the error estimates for non-essential boundary value problems that do not require additional regularity.

[4] here $\partial_A u = \nu \cdot A\nabla u$ and ν denotes the outer normal of Ω

Remark 5.30 (Generalizations). The strategy of the proof of Theorem 5.28 holds for a broader class of elliptic zero boundary value problems. The essential requirement is that the bilinear form a of the differential operator is coercive on $H_0^1(\Omega)$ and that a_λ is coercive on all of $H^1(\Omega)$. Then, regularity of the solution u^* of the zero boundary value problem is required to identify the equation u^* satisfies when tested with functions in $H^1(\Omega)$ and not only $H_0^1(\Omega)$, see (5.22).

Remark 5.31 (Optimality of the rate λ^{-1}). We demonstrate that the rate[5]

$$\|u_\lambda - u^*\|_{H^1(\Omega)} \lesssim \lambda^{-1}$$

can not in general be improved. To this end we consider the concrete example

$$a_\lambda \colon H^1(0,1)^2 \longrightarrow \mathbb{R}, \quad (u,v) \mapsto \int_0^1 u'v'\mathrm{d}x + \lambda(u(0)v(0) + u(1)v(1)).$$

The minimizer of E_λ with $f \equiv 1$ solves the ODE

$$-u'' = 1 \quad \text{in } (0,1)$$

with Robin boundary conditions

$$\begin{aligned} -u'(0) + \lambda u(0) &= 0 \\ u'(1) + \lambda u(1) &= 0. \end{aligned}$$

Its solution is given by

$$u_\lambda(x) = -\frac{1}{2}x^2 + \frac{1}{2}x + \frac{1}{2\lambda}.$$

On the other hand the associated Dirichlet problem is solving the same ODE subject to $u(0) = u(1) = 0$ and has the solution

$$u^*(x) = -\frac{1}{2}x^2 + \frac{1}{2}x.$$

Consequently the difference $u_\lambda - u^*$ measured in $H^1(0,1)$ norm is precisely $\frac{1}{2\lambda}$.

A solution formula based on Steklov eigenfunctions. The Steklov theory yields the existence of an orthonormal eigenbasis of the space

$$\mathcal{H}(\Omega) := \left\{ w \in H^1(\Omega) : a(w,v) = 0 \text{ for all } v \in H_0^1(\Omega) \right\}.$$

of weakly a-harmonic functions, which we can use for a Fourier expansion of v_λ in order to obtain the desired estimate. For a recent and more general discussion of Steklov theory we refer to [23]. The Steklov eigenvalue problem consists of finding $(\mu, w) \in \mathbb{R} \times H^1(\Omega)$ such that

$$a(w,\varphi) = \mu \int_{\partial\Omega} w\varphi \mathrm{d}s \quad \text{for all } \varphi \in H^1(\Omega). \tag{5.25}$$

We call μ a *Steklov eigenvalue* and w a corresponding *Steklov eigenfunction*.

Lemma 5.32 (Orthogonal decomposition). *We can decompose the space $H^1(\Omega)$ into*

$$H^1(\Omega) = \mathcal{H}(\Omega) \oplus_{a_1} H_0^1(\Omega)$$

with the decomposition being a_1-orthogonal.

[5] We write $\lesssim$ and $\gtrsim$ if the inequality $\leq$ or $\geq$ holds up to a constant; if both $\lesssim$ and $\gtrsim$ hold, we write $\sim$.

Proof. By the definition of a_1 and $\mathcal{H}(\Omega)$ the two spaces are a_1 orthogonal. To see that it spans all of $H^1(\Omega)$ let $u \in H^1(\Omega)$ be given. Let u_a be the unique solution of $a(u_a, \cdot) = 0$ in $H^1_0(\Omega)^*$ subject to $\mathrm{tr}(u_a) = \mathrm{tr}(u)$. Then the decomposition is given as $u = u_a + (u - u_a) = u_a + u^*$. □

Theorem 5.33 (Steklov spectral theorem). *Let $\Omega \subseteq \mathbb{R}^d$ be open and a be a positive semi-definite bilinear form on $H^1(\Omega)$ and $\mathcal{H}(\Omega) \hookrightarrow L^2(\partial\Omega)$ be compact. Then there exists a non decreasing sequence $(\mu_j)_{j\in\mathbb{N}} \subseteq [0, \infty)$ with $\mu_j \to \infty$ and a sequence $(e_j)_{j\in\mathbb{N}} \subseteq \mathcal{H}(\Omega)$ such that μ_j is a Steklov eigenvalue with eigenfunction e_j. Further, $(e_j)_{j\in\mathbb{N}}$ is a complete orthonormal system in $\mathcal{H}(\Omega)$ with respect to a_1.*

Proof. This can be derived from the spectral theory for compact operators as for example described in [101, Section 8.10]. In the notation of [101], set $X = \mathcal{H}(\Omega)$ with inner product a_1 and let $Y = L^2(\partial\Omega)$ equipped with its natural inner product. Then this yields a divergent sequence $0 < \tilde{\mu}_1 \leq \tilde{\mu}_2 \leq \dots$ growing to ∞ and $(e_j)_{j\in\mathbb{N}} \subseteq \mathcal{H}(\Omega)$ with $a_1(e_i, e_j) = \delta_{ij}$ and

$$a_1(e_j, w) = \tilde{\mu}_j \int_{\partial\Omega} e_j w \mathrm{d}s \quad \text{for all } w \in \mathcal{H}(\Omega). \tag{5.26}$$

For $\varphi \in H^1(\Omega)$, let $\varphi = \varphi_a + \varphi_0$ be the orthogonal decomposition of the preceding lemma and compute

$$a_1(e_j, \varphi) = a_1(e_j, \varphi_0) + a_1(e_j, \varphi_a) = a_1(e_j, \varphi_a) = \tilde{\mu}_j \int_{\partial\Omega} e_j \varphi_a \mathrm{d}s = \tilde{\mu}_j \int_{\partial\Omega} e_j \varphi \mathrm{d}s.$$

Using the definition of a_1 we obtain

$$a(e_j, \varphi) = (\tilde{\mu}_j - 1) \int_{\partial\Omega} e_j \varphi \mathrm{d}s \quad \text{for all } \varphi \in H^1(\Omega).$$

Setting $\mu_j := \tilde{\mu}_j - 1$ and noting that the above equality implies $\mu_j \geq 0$ concludes the proof. □

As a direct consequence we obtain the following representation formula.

Corollary 5.34 (Fourier expansion in the Steklov eigenbasis). *Let $w \in \mathcal{H}(\Omega)$. Then we have*

$$w = \sum_{j=0}^{\infty} c_j e_j,$$

where

$$c_j = (1 + \mu_j) \int_{\partial\Omega} w e_j \mathrm{d}s. \tag{5.27}$$

Proof. Using that e_j is a Steklov eigenvector, we can compute the Fourier coefficients

$$c_j = a_1(w, e_j) = a(w, e_j) + \int_{\partial\Omega} w e_j \mathrm{d}s = (1 + \mu_j) \int_{\partial\Omega} w e_j \mathrm{d}s.$$

□

Lemma 5.35 (Solution formula). *Let $v_\lambda \in H^1(\Omega)$ be the unique solution of*

$$a(v_\lambda, \varphi) + \int_{\partial\Omega} (\lambda v_\lambda - \partial_A u^*) \varphi \, ds = 0 \quad \textit{for all } \varphi \in H^1(\Omega). \tag{5.28}$$

Then we have

$$v_\lambda = \frac{1}{\lambda}\sum_{j=0}^{\infty} c(\lambda)_j e_j,$$

where

$$c(\lambda)_j = \frac{1+\mu_j}{1+\frac{\mu_j}{\lambda}} \cdot \int_{\partial\Omega} (\partial_A u^*) e_j \mathrm{d}s.$$

Proof. Note that v_λ is weakly harmonic and hence, we can apply the previous corollary to compute the Fourier coefficients of v_λ. Using that v_λ solves (5.28) and that e_j is a Steklov eigenfunction we compute

$$\int_{\partial\Omega} v_\lambda e_j \mathrm{d}s = \frac{1}{\lambda}\int_{\partial\Omega}(\partial_A u^*)e_j \mathrm{d}s - \frac{1}{\lambda}a(v_\lambda, e_j) = \frac{1}{\lambda}\int_{\partial\Omega}(\partial_A u^*)e_j \mathrm{d}s - \frac{\mu_j}{\lambda}\int_{\partial\Omega} v_\lambda e_j \mathrm{d}s.$$

Rearranging this yields the following equation, which completes the proof

$$\int_{\partial\Omega} v_\lambda e_j \mathrm{d}s = \frac{1}{\lambda}\cdot\frac{1}{1+\frac{\mu_j}{\lambda}}\int_{\partial\Omega}(\partial_A u^*)e_j \mathrm{d}s.$$

□

Proof of Theorem 5.28. We use the explicit solution formula from Lemma 5.35 for $v_\lambda = u^* - u_\lambda$ to provide the missing claims in the proof of Theorem 5.28.

Completing the proof of Theorem 5.28. We have already convinced ourselves that the difference $v_\lambda := u_\lambda - u^*$ indeed solves (5.28). By the means of Lemma 5.35 it suffices to bound

$$\left\| \sum_{j=0}^{\infty} c(\lambda)_j e_j \right\|_{H^1(\Omega)}$$

independently of $\lambda > 0$. Let us denote the A-harmonic extension of $\partial_A u^*$ with w, we obtain

$$c(\lambda)_j^2 \leq (1+\mu_j)^2 \left(\int_{\partial\Omega}(\partial_A u^*)e_j \mathrm{d}s\right)^2 = a_1(w, e_j)^2.$$

Now we can estimate

$$\begin{aligned}\sum_{j=0}^{\infty}(1+\mu_j)^2\left(\int_{\partial\Omega}(\partial_A u^*)e_j \mathrm{d}s\right)^2 &= \sum_{j=0}^{\infty} a_1(w,e_j)^2 = a_1(w,w)\\ &\leq \|a_1\|\,\|w\|_{H^1(\Omega)}^2\\ &\leq \|a_1\|\,\|u^*\|_{H^2(\Omega)}^2\|T\|_{\mathcal{L}(H^2(\Omega);H^1(\Omega))}^2\\ &\leq \|a_1\|\,c_{reg}^2\|f\|_{L^2(\Omega)}^2\|T\|_{\mathcal{L}(H^2(\Omega);H^1(\Omega))}^2,\end{aligned}$$

where $T\colon H^2 \longrightarrow H^1$ is the mapping that assigns a function u the harmonic extension of $\partial_A u$. Consequently, we obtain by Parseval's identity

$$\|v_\lambda\|_{H^1(\Omega)} \leq c_F\|v_\lambda\|_{a_1} = \frac{c_F}{\lambda}\sqrt{\sum_{j=1}^{\infty} c(\lambda)_j^2} \leq \frac{c_F}{\lambda}\sqrt{\|a_1\|}\, c_{reg}\|f\|_{L^2(\Omega)}\|T\|_{\mathcal{L}(H^2(\Omega);H^1(\Omega))},$$

which finishes the proof. □

Estimates under lower regularity of the right-hand side. If $f \notin L^2(\Omega)$, Theorem 5.28 cannot be applied as the estimation of the term $\|u_\lambda - u^*\|_{H^1(\Omega)}$ requires that u^* is a member of $H^2(\Omega)$. The next Lemma shows that at the expense of a worse rate and norm, we can still estimate this term for distributional right-hand sides $f \in H^1(\Omega)^*$.

Lemma 5.36. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with $C^{1,1}$ boundary, assume that $A \in C^{0,1}(\Omega, \mathbb{R}^{d\times d})$ is symmetric and uniformly elliptic, $f \in H^1(\Omega)^*$ and let u^* and u_λ be as in Theorem 5.28. Then it holds*

$$\|u_\lambda - u^*\|_{L^2(\Omega)} \leq c \cdot \|f\|_{H^1(\Omega)^*}\lambda^{-1/2}.$$

Proof. We set $v_\lambda = u_\lambda - u^*$ and denote by $w \in H^1_0(\Omega)$ the solution to the equation $-\operatorname{div}(A\nabla w) = v_\lambda$ in $H^1_0(\Omega)^*$. Then, by our assumptions on Ω and A, the function w is a member of $H^2(\Omega)$. This yields upon integration by parts and the fact that v_λ is weakly A-harmonic that it holds

$$\begin{aligned}
\|v_\lambda\|^2_{L^2(\Omega)} &= \int_\Omega A\nabla w \cdot \nabla v_\lambda \mathrm{d}x - \int_{\partial\Omega} \partial_A w v_\lambda \mathrm{d}s \\
&= -\int_{\partial\Omega} \partial_A w u_\lambda \mathrm{d}s \\
&\leq c\|w\|_{H^2(\Omega)}\|u_\lambda\|_{L^2(\partial\Omega)} \\
&\leq c\|v_\lambda\|_{L^2(\Omega)}\|u_\lambda\|_{L^2(\partial\Omega)}.
\end{aligned}$$

It remains to estimate $\|u_\lambda\|_{L^2(\partial\Omega)}$. Note that u_λ satisfies

$$0 = \int_\Omega A\nabla u_\lambda \cdot \nabla u_\lambda \mathrm{d}x + \lambda \int_{\partial\Omega} u_\lambda^2 \mathrm{d}s - f(u_\lambda).$$

We get after rearranging and using Young's inequality

$$\begin{aligned}
\|u_\lambda\|^2_{L^2(\partial\Omega)} &= \frac{2}{\lambda}\left(f(u_\lambda) - \left(\int_\Omega A\nabla u_\lambda \cdot \nabla u_\lambda \mathrm{d}x + \frac{\lambda}{2}\|u_\lambda\|_{L^2(\partial\Omega)}\right)\right) \\
&\leq \frac{2}{\lambda}\left(\|f\|_{H^1(\Omega)^*}\|u_\lambda\|_{H^1(\Omega)} - \alpha_{\lambda/2}\|u_\lambda\|^2_{H^1(\Omega)}\right) \\
&\leq \frac{\|f\|^2_{H^1(\Omega)^*}}{2\alpha_{\lambda/2}\lambda},
\end{aligned}$$

which completes the proof. □

5.4.3. Penalization strength and error decay.

We have seen that the distance of an ansatz function can be bounded in terms of the optimization error, the approximation power of the ansatz class and the penalization strength. In this section we discuss the trade off of choosing the penalization strength λ too large or too small and discuss the implications of different scalings of λ in dependecy of the approximation capabilities of the ansatz classes. We combine our general discussion with Theorem 5.11 to obtain Theorem 5.38, however, our discussion can be combined with any result guaranteeing approximation rates of a sequence of ansatz classes.

We consider a sequence $(V_n)_{n\in\mathbb{N}} \subseteq H^1(\Omega)$ of ansatz classes and penalization strengths $\lambda_n \sim n^\sigma$. Further, we denote the minimizers of the energies E_{λ_n} over V_n by $v_n^* \in V_n$. It is our goal to choose $\sigma \in \mathbb{R}$ in such a way that the upper bound of $\|v_n^* - u^*\|_{H^1(\Omega)}$ in (5.21) decays with the fastest possible rate. Neglecting constants, the bound evaluates to

$$\|v_n^* - u^*\|_{H^1(\Omega)} \lesssim \sqrt{\frac{1}{\alpha_{\lambda_n}} \inf_{v\in V_n} \|v - u_{\lambda_n}\|_{a_{\lambda_n}}^2} + \lambda_n^{-1}. \tag{5.29}$$

We can assume without loss of generality that $\sigma > 0$ and hence $\lambda_n \geq 1$, because otherwise the upper bound will not decrase to zero. Note that in this case we have $\alpha_{\lambda_n} \geq \alpha_1 > 0$ and hence the we obtain

$$\|v_n^* - u^*\|_{H^1(\Omega)} \lesssim \sqrt{\inf_{v\in V_n} \left(\|\nabla(v - u_{\lambda_n})\|_{L^2(\Omega)}^2 + n^\sigma \|v - u_{\lambda_n}\|_{L^2(\partial\Omega)}^2 \right)} + n^{-\sigma}. \tag{5.30}$$

Here, the trade off in choosing σ and therefore λ_n too large or small is evident. We discuss the implications of this upper bound in three different scenarios.

Approximation rates with zero boundary values. Consider the case where there is an element $v_n \in V_n \cap H_0^1(\Omega)$ such that

$$\|v_n - u^*\|_{H^1(\Omega)} \lesssim n^{-r}.$$

Using the Euler-Lagrange equations $a_\lambda(u_\lambda, \cdot) = f(\cdot)$ and $a(u^*, \cdot) = f(\cdot)$ we can estimate

$$\begin{aligned} \frac{1}{2} \inf_{v\in V_n} \|v - u_\lambda\|_{a_\lambda}^2 = \inf_{v\in V_n} E_\lambda(v) - E_\lambda(u_\lambda) &\leq \inf_{v\in V_n\cap H_0^1(\Omega)} E_\lambda(v) - E(u_\lambda) \\ &\leq \inf_{v\in V_n\cap H_0^1(\Omega)} E(v) - E(u^*) = \frac{1}{2} \inf_{v\in V_n\cap H_0^1(\Omega)} \|v - u^*\|_a^2 \lesssim n^{-2r} \end{aligned}$$

independently of λ. Hence, the estimate (5.29) yields

$$\|v_n^* - u^*\|_{H^1(\Omega)} \lesssim \sqrt{n^{-2r}} + \lambda_n^{-1} \lesssim n^{-r},$$

whenever $\lambda_n \gtrsim n^r$. Note that in this case, no trade off in λ exists and the approximation rate with zero boundary values can always be achieved up to optimization. However, the curvature α_{λ_n} of E_{λ_n} increases with λ_n. Thus, it seems reasonable to choose λ_n as small as possible, i.e., $\lambda_n \sim n^r$. Approximation rates with zero boundary values have not been established so far for neural networks to the best of our knowledge.

Approximation error of u_λ independent of λ. Now, we consider the case, without an approximation rate with exact zero boundary values, but where the sequence $(V_n)_{n\in\mathbb{N}}$ of ansatz classes admits approximation rates in both $H^1(\Omega)$ and $L^2(\partial\Omega)$. More precisely, we assume that there are real numbers $s \geq r > 0$ such that for every (sufficiently big) λ and every $n \in \mathbb{N}$ there is an element $v_n \in V_n$ satisfying

$$\|v_n - u_\lambda\|_{H^1(\Omega)} \leq cn^{-r} \quad \text{and} \quad \|v_n - u_\lambda\|_{L^2(\partial\Omega)} \leq cn^{-s},$$

for some $c > 0$ independent on λ. Then the estimate in (5.30) yields

$$\|v_n^* - u^*\|_{H^1(\Omega)} \lesssim \sqrt{n^{-2r} + n^{\sigma-2s}} + n^{-\sigma}.$$

The resulting rate of decay of the upper bound of the error is then

$$\rho(\sigma) = \min\left(\frac{1}{2}\min(2r, 2s - \sigma), \sigma\right) = \min\left(r, s - \frac{\sigma}{2}, \sigma\right).$$

which is maximized at $\sigma^* = 2s/3$ with a value of

$$\rho^* = \min\left(\frac{2}{3}s, r\right). \tag{5.31}$$

In this case, the upper bound does not necessarily decay at the same rate as the approximation error, which decays with rate r. Note that because $H^1(\Omega)$ embeds into $L^2(\partial\Omega)$ we can assume without loss of generality that $s \geq r$.

We made the assumption that the approximation rates of r and s holds with the same constant independently of λ. This is for example the case, if the solutions u_λ are uniformly in λ bounded in $H^s(\Omega)$ for some $s > 1$.

Approximation rates for u^*. Now we want to discuss the case, where we weaken the approximation assumption from above, which is uniformly in λ. More precisely, we assume that there is a constant $c > 0$ and elements $v_n \in V_n$ satisfying

$$\|v_n - u^*\|_{H^1(\Omega)} \leq cn^{-r} \quad \text{and} \quad \|v_n - u^*\|_{L^2(\partial\Omega)} \leq cn^{-s}.$$

By (5.24) (or equivalently Theorem 5.28 with $V = H^1(\Omega)$ and $v = u_\lambda$) and the triangle inequality we have

$$\|v_n - u_\lambda\|_{H^1(\Omega)} \leq \|v_n - u^*\|_{H^1(\Omega)} + \|u^* - u_\lambda\|_{H^1(\Omega)} \leq cn^{-r} + c'n^{-\sigma} \leq \tilde{c}n^{-\tilde{r}}$$

and similarly

$$\|v_n - u_\lambda\|_{L^2(\partial\Omega)} \leq cn^{-s} + c'n^{-\sigma} \leq \tilde{c}n^{-\tilde{s}},$$

where $\tilde{r} = \min(r, \sigma)$ and $\tilde{s} = \min(s, \sigma)$. Hence, we have reduced this case to the previous case and find that the right hand side of (5.30) decays at a rate of

$$\begin{aligned} \rho(\sigma) &= \min(\min(r, \sigma), \min(s, \sigma) - \sigma/2, \sigma) = \min(r, \min(s, \sigma) - \sigma/2) \\ &= \min(r, \min(s - \sigma/2, \sigma/2)) = \min(r, s - \sigma/2, \sigma/2). \end{aligned} \tag{5.32}$$

This function is maximized at $\sigma^* = s$ with a value of $\rho^* = \min(s/2, r)$. Like before, we can without loss of generality assume that $s \geq r$.

Remark 5.37. Note that in general the decay rate $\rho^* = \min(s/2, r)$ of the upper bound (5.29) can be smaller than the approximation rate r. This is in contrast to problems with non-essential boundary values for which the error decays proportional to the approximation error by Cea's lemma. We stress that the defect in the decay rate of the right hand side of (5.29) is not an artefact of our computations but in fact sharp.

Let us now come back to the original problem of the Dirichlet problem (5.20). For a right hand side $f \in H^r(\Omega)$, standard regularity results yield $u^* \in H^{r+2}(\Omega)$. Theorem 5.11 provides rates for the approximation in $H^s(\Omega)$ for $s \in [0, 1]$, which lead to the following result.

Theorem 5.38 (Rates for NN training with boundary penalty). *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with $C^{r+1,1}$ boundary for some $r \in \mathbb{N}$, $f \in H^r(\Omega)$ and assume $A \in C^{r,1}(\Omega, \mathbb{R}^{d\times d})$ is symmetric, uniformly elliptic with ellipticity constant $\alpha > 0$ and denote the solution to* (5.20) *by $u^* \in H^1_0(\Omega)$. For every $n \in \mathbb{N}$, there is a ReLU network with parameter space Θ_n of dimension $\mathcal{O}(\log_2^2(n^{(r+2)/d}) \cdot n)$ such that if $\lambda_n \sim n^\sigma$ for $\sigma = \frac{2r+3}{2d}$ one has for any $\rho < \frac{2r+3}{4d}$ that*

$$\|u_{\theta_n} - u^*\|_{H^1(\Omega)} \lesssim \sqrt{\delta_n + n^{-2\rho}} + n^{-\rho} \quad \textit{for all } \theta_n \in \Theta_n, \tag{5.33}$$

where $\delta_n = E_{\lambda_n}(u_{\theta_n}) - \inf_{\tilde{\theta}\in\Theta_n} E_{\lambda_n}(u_{\tilde{\theta}})$.

Proof. For $\varepsilon > 0$ it holds that $H^{1/2+\varepsilon}(\Omega) \hookrightarrow L^2(\partial\Omega)$. Thus, Theorem 5.11 guarantees the existence of u_{θ_n} with $\theta_n \in \Theta_n$ and the claimed number of parameters such that $\|u_{\theta_n} - u^*\|_{H^1(\Omega)} \lesssim n^{-(r+1)/d} = n^{-\tilde{r}}$ and

$$\|u_{\theta_n} - u^*\|_{L^2(\partial\Omega)} \lesssim \|u_{\theta_n} - u^*\|_{H^{1/2+\varepsilon}(\Omega)} \lesssim n^{-(r+2-(1/2+\varepsilon))/d} = n^{-\tilde{s}},$$

where $\tilde{s} = (2r + 3 - 2\varepsilon)/(2d)$. By (5.32), the estimate (5.33) holds for

$$\rho = \min(\tilde{r}, \tilde{s} - \sigma/2, \sigma/2) = \min\left(\frac{r+1}{d}, \frac{2r+3-4\varepsilon}{4d}, \frac{2r+3}{4d}\right) = \frac{2r+3-4\varepsilon}{4d}.$$

□

Remark 5.39 (Adaptation to Smoothness). The discussion from Remark 5.26 carries over to the case of Dirichlet boundary values and boundary penalties, i.e., the error of the deep Ritz method decays at a rate increasing with the smoothness of the problem. This fact can be especially useful in high spatial dimensions, which is consistent with the empirical findings that the deep Ritz method can be effective in the numerical solution of high dimensional problems [110]. Note that also finite element methods can achieve rates increasing with the smoothness of data, however they require the delicate construction of higher order elements.

Remark 5.40 (Combination with different approximation results). We focus on the ReLU activation in this section, whereas in practice often other architectures and activation functions are used, see [110, 132]. However, our results from Section 3 can handle arbitrary function classes and hence reduce the computation of error estimates to the computation of approximation bounds. Therefore, they can be combined with other approximation results for neural networks in Sobolev norm including the works of [127, 306, 262, 138, 106, 88].

Remark 5.41 (The boundary penalty method for FEM). The boundary penalty method has been applied in the context of finite element approximations [27] and studied in terms of its convergence rates in [27, 258, 34]. The idea of the finite element approach is analogue to the idea of using neural networks for the approximate solution of variational problems. However, one constructs a nested sequence of finite dimensional vector spaces $V_h \subseteq H^1(\Omega)$ arising from some triangulation with fineness $h > 0$ and computes the minimizer u_h of the penalized energy E_λ over V_h. Choosing a suitable triangulation and piecewise affine linear elements and setting $\lambda \sim h^{-1}$ one obtains the error estimate

$$\|u_h - u^*\|_{H^1(\Omega)} \lesssim h,$$

see [258]. At the core of those estimates lies a linear version of Céa's Lemma, which already incorporates boundary values. However, the proof of this lemma heavily relies on the fact that the class of ansatz functions is linear and that its minimizer solves a linear equation. This is not the case for non linear function classes like neural networks. Therefore, our estimates require a different strategy. However, the optimal rate of convergence for the boundary penalty method with finite elements can be deduced from our results. In fact, one can choose a suitable triangulation and an operator $r_h \colon H^2(\Omega) \to V_h$ such that

$$\|r_h u - u\|_{H^1(\Omega)} \lesssim h\|u\|_{H^2(\Omega)},$$

where the approximating functions u_h have zero boundary values as they arise from interpolation. By the general discussion from above for the ansatz classes with approximation rates with exact zero boundary values, choosing $\lambda \gtrsim h^{-1}$ yields an error bound decaying like the approximation error $\|u_h - u^*\|_{H^1(\Omega)} \lesssim h$.

5.5 Proofs regarding the implications of exact boundary values in residual minimization

In this section we show the theoretical benefits of using neural network type ansatz functions that satisfy Dirichlet boundary conditions exactly in the residual minimization method for the Poisson problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= g \quad \text{on } \partial\Omega, \end{aligned} \tag{5.34}$$

where $f \in L^2(\Omega)$ and $g \in H^{3/2}(\partial\Omega)$. We see that the exact boundary conditions improve the mode of convergence from $H^{1/2}$ to H^2. Although being formulated for the Laplace operator, those results hold for any elliptic operator, which is H^2 regular.

5.5.1. H^2 estimates for residual minimization with exact boundary values.

We start by considering the case of exact boundary conditions and present two main results, one that allows to quantify the H^2 error using the value of the loss function and the other, an estimate based on Céa's Lemma that allows to link the approximation capabilities of the network class to the error made by residual minimization.

Setting 5.42. *We consider again* (5.34), *in particular, we assume that the problem is H^2 regular meaning that there is a constant $C_{reg} > 0$, satisfying*

$$\|u\|_{H^2(\Omega)} \le C_{reg}\|\Delta u\|_{L^2(\Omega)} \quad \textit{for all } u \in H^2(\Omega) \cap H^1_0(\Omega).$$

Furthermore, we assume that Θ is a parameter set of a neural network type ansatz class, such that for every $\theta \in \Theta$ we have $u_\theta \in H^2(\Omega)$ and $(u_\theta)|_{\partial\Omega} = g$. As our strategy is to minimize the residual we define the loss function

$$\mathcal{L}\colon \Theta \to \mathbb{R}, \quad \mathcal{L}(\theta) = \|\Delta u_\theta + f\|^2_{L^2(\Omega)}.$$

Remark 5.43. Setting 5.42 is for example satisfied when $\partial\Omega \in C^{1,1}$, $f \in L^2(\Omega)$. Alternatively, one can replace the assumption $\partial\Omega \in C^{1,1}$ by requiring that the domain Ω is convex. We refer to [124] for a detailed discussion of the regularity properties of elliptic equations.

The following result is a trivial corollary of the H^2 regularity we assumed and a similar result is due to [283], although not exploiting the benefits of exact boundary conditions. Albeit being of simple nature, we believe it can be of practical relevance due to its easy and explicit error control.

Theorem 5.44. *Assume we are in the situation of Setting 5.42, then it holds for every $\theta \in \Theta$ that*

$$\|u_\theta - u_f\|_{H^2(\Omega)} \le C_{reg}\sqrt{\mathcal{L}(\theta)}.$$

For convex domains, we may estimate the regularity constant explicitely. It holds

$$C_{reg} \le \sqrt{1 + C_P} \le \sqrt{1 + \left(\frac{|\Omega|}{\omega_d}\right)^{\frac{1}{d}}},$$

where d is the dimension of Ω, ω_d denotes the volume of the unit ball in $\mathbb{R}^d$ and C_P is the Poincaré constant for functions in $H^1_0(\Omega)$.

Proof. The difference $u_\theta - u_f$ lies in $H^2(\Omega) \cap H^1_0(\Omega)$ and solves $-\Delta(u_f - u_\theta) = \Delta u_\theta + f$. The $H^2(\Omega)$ regularity theory then implies the desired estimate. Let us now assume that Ω is convex and derive the explicit estimate on C_{reg}. We expand the $H^2(\Omega)$ norm of $u_f \in H^2(\Omega) \cap H^1_0(\Omega)$

$$\|u_f\|^2_{H^2(\Omega)} = \|u_f\|^2_{L^2(\Omega)} + \|\nabla u_f\|^2_{L^2(\Omega)} + \|D^2 u_f\|^2_{L^2(\Omega)}$$

Due to the zero boundary values and the convexity of Ω we have

$$\|D^2 u\|^2_{L^2(\Omega)} = \|\Delta u\|^2_{L^2(\Omega)} = \|f\|^2_{L^2(\Omega)}$$

and we refer the reader to [124] for details. The first two terms can be estimated jointly using the a priori estimates of the Lax-Milgram Theorem, this yields

$$\|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2(\Omega)} \leq C_P^2 \|f\|^2_{H^1_0(\Omega)^*} \leq C_P^2 \|f\|^2_{L^2(\Omega)}.$$

This is due to the fact that C_P^{-1} is the coercivity constant of the Dirichlet Laplacian bilinear form, see [112]. The explicit estimate of the Poincaré constant C_P can be found in [151]. □

Remark 5.45. Some remarks are in order.

(i) The zero boundary conditions are essential. If one instead resorts to a $L^2(\partial\Omega)$ penalty of the boundary values the best convergence one can hope for is $H^{1/2}(\Omega)$. We elaborate this in Section 5.5.2.

(ii) The theorem allows to compute an explicit upper bound on the error made by residual minimization, once the training returns a parameter θ via computing the (continuous) loss. In particular, no access to the solution u_f is required. This means that if boundary conditions are encoded in the ansatz functions, the loss itself is a consistent a posteriori error estimator for the residual minimization method.

(iii) The root in the estimate above does not indicate a slow convergence. In fact, the loss itself is a squared $L^2(\Omega)$ norm and the root accounts for that.

The next theorem allows to quantify the error made by the residual minimization method using the optimization quality and the expressiveness of the ansatz class. It is an application of the non-linear Céa Lemma as formulated by [213].

Theorem 5.46. *Assume we are in Setting 5.42, then for any $\theta \in \Theta$ it holds*

(5.35)
$$\|u_f - u_\theta\|_{H^2(\Omega)} \leq \sqrt{C^2_{reg}\delta + C^2_{reg} \inf_{\tilde\theta \in \Theta} \|\Delta(u_{\tilde\theta} - u_f)\|^2_{L^2(\Omega)}} \leq \sqrt{C^2_{reg}\delta + C^2_{reg} \inf_{\tilde\theta \in \Theta} \|u_{\tilde\theta} - u_f\|^2_{H^2(\Omega)}},$$

where $\delta = \mathcal{L}(\theta) - \inf_{\tilde\theta \in \Theta} \mathcal{L}(\tilde\theta)$.

Proof. We define the energy

$$E \colon H^2(\Omega) \to \mathbb{R}, \quad E(u) = \|\Delta u + f\|^2_{L^2(\Omega)}.$$

Note that E is defined on a different domain than the loss function $\mathcal{L}$, which is why we reserve an own symbol for it. The energy E is a quadratic energy

$$\begin{aligned}\|\Delta u + f\|^2_{L^2(\Omega)} &= \int_\Omega (\Delta u)^2 \mathrm{d}x + 2\int_\Omega f\Delta u \mathrm{d}x + \int_\Omega f^2 \mathrm{d}x\\ &= \frac{1}{2}a(u,u) - F(u) + c,\end{aligned}$$

where the bilinear form $a : H^2(\Omega) \times H^2(\Omega) \to \mathbb{R}$, the functional $F \in H^2(\Omega)^*$ and the constant c are given by

$$a(u,v) = 2\int_\Omega \Delta u \Delta v \mathrm{d}x, \quad F(u) = 2\int_\Omega f(-\Delta u)\mathrm{d}x, \quad c = \int_\Omega f^2 \mathrm{d}x.$$

The unique minimizer of E in the affine subspace $H^2(\Omega) \cap H^1_g(\Omega)$ is precisely the solution u_f to the Poisson problem (5.34). The bilinear form a is coercive on the subspace $H^2(\Omega) \cap H^1_0(\Omega)$, which follows from elliptic regularity theory, see for instance [124]. This allows to exploit a Céa Lemma for non-linear ansatz spaces, as described in [213, in Proposition 3.1]. To transfer this to the affine space $H^2(\Omega) \cap H^1_g(\Omega)$ we choose $u_g \in H^2(\Omega)$ such that $-\Delta u_g = 0$ and $(u_g)|_{\partial\Omega} = g$. For an arbitrary u_θ we then expand

$$\|u_\theta - u_f\|_{H^2(\Omega)} = \|(u_\theta - u_g) - (u_f - u_g)\|_{H^2(\Omega)}.$$

Now note that $u_f - u_g$ solves $-\Delta(u_f - u_g) = f$ with zero boundary values, hence $u_f - u_g$ is the unique minimizer of E over the subspace $H^2(\Omega) \cap H^1_0(\Omega)$ and we can apply Céa's Lemma with the ansatz set $\{u_\theta - u_g : \theta \in \Theta\}$

$$\|u_\theta - u_f\|_{H^2(\Omega)} \leq \sqrt{\frac{2\delta}{\alpha} + \frac{1}{\alpha}\inf_{\tilde\theta\in\Theta}\|u_{\tilde\theta} - u_f\|^2_a},$$

where $\|\cdot\|_a$ denotes the norm induced by a. Using that the coercivity constant α of a is $2/C^2_{\mathrm{reg}}$ and the norm $\|\cdot\|_a = 2\|\Delta\cdot\|_{L^2(\Omega)}$ we conclude. □

Remark 5.47 (General Elliptic Equations). The discussion of this chapter can be extended to more general elliptic equations. For coefficients $A \in C^{0,1}(\Omega, \mathbb{R}^{d\times d})$, a right-hand side $f \in L^2(\Omega)$ and boundary values $g \in H^{3/2}(\partial\Omega)$ consider the equation

$$\begin{aligned}-\operatorname{div}(A\nabla u) &= f \quad \text{in } \Omega,\\ u &= g \quad \text{on } \partial\Omega.\end{aligned}$$

If we assume that $\partial\Omega \in C^{1,1}$ (or that Ω is convex) and the coefficients are uniformly elliptic, i.e., for a constant $c_A > 0$ satisfy $A(x)\xi\cdot\xi \geq c_A|\xi|^2$ uniformly in $x \in \Omega$ and $\xi \in \mathbb{R}^d$, the problem admits a unique solution $u_f \in H^2(\Omega)$ and we can estimate

$$\|u_f\|_{H^2(\Omega)} \leq c_{\mathrm{reg}}\left(\|f\|_{L^2(\Omega)} + \|g\|_{H^{3/2}(\partial\Omega)}\right).$$

Arguing as in the proof of Theorem 5.44 we obtain

$$\|u_\theta - u_f\|_{H^2(\Omega)} \leq c_{\mathrm{reg}}\sqrt{\mathcal{L}(\theta)}.$$

Similarly, Theorem 5.46 can be transferred to this setting.

5.5.2. Failure without exact boundary values. In this section we show that *not* enforcing exact boundary values in the neural network ansatz functions leads to considerably weaker error estimates. Throughout this subsection, we work under the following assumptions.

Setting 5.48. *We consider again* (5.34)*. We assume that Θ is a parameter set of a neural network type ansatz class, such that for every $\theta \in \Theta$ we have $u_\theta \in H^2(\Omega)$, but make no assumptions on its boundary values. As our strategy is to minimize the residual we define the loss function with boundary penalty*

$$\mathcal{L}_\tau \colon \Theta \to \mathbb{R}, \quad \mathcal{L}_\tau(\theta) = \|\Delta u_\theta + f\|^2_{L^2(\Omega)} + \tau \|u_\theta - g\|^2_{L^2(\partial\Omega)},$$

where $\tau \in (0, \infty)$ is a positive penalization parameter.

Without exact boundary values, the penalization of the deviations of the boundary values is required in order to enforce them approximately. Note that if u_θ has exact boundary values, it holds that $\mathcal{L}_\tau(\theta) = \mathcal{L}(\theta)$. With the penalization introduced above, we obtain a similar result to Theorem 5.44 but only with respect to the weaker $H^{1/2}$-norm, which is to the estimate by [261] is sharp. However, we sharpen this result by showing that $1/2$ is the largest exponent for which such an estimate can hold in general.

Theorem 5.49. *Assume that we are in Setting 5.48 and that the domain $\Omega \subseteq \mathbb{R}^d$ has a smooth boundary $\partial\Omega \in C^\infty$. Then for $s \in \mathbb{R}$ there is a constant $c > 0$ such that*

$$\|u_\theta - u_f\|_{H^s(\Omega)} \le c\sqrt{\mathcal{L}_\tau(\theta)} \quad \textit{for all } \theta \in \Theta \tag{5.36}$$

and all parametric classes and data $f \in L^2(\Omega)$, $g \in H^{3/2}(\partial\Omega)$ if and only if $s \le 1/2$.

Proof. First, we show that the estimate holds for $s \le 1/2$, where it suffices to show it for $s = 1/2$. For this, we use the estimate

$$\|u\|_{H^s(\Omega)} \le c\left(\|-\Delta u\|_{H^{s-2}(\Omega)} + \|u\|_{H^{s-1/2}(\partial\Omega)}\right), \tag{5.37}$$

for all $u \in C^\infty(\overline{\Omega})$ and $s \in \mathbb{R}$, see Theorem 2.1 in [247] or Lemma 6.2 in [261]. Setting $s = 1/2$ and noting that it extends to functions $u \in H^2(\Omega)$ yields

$$\|u\|_{H^{1/2}(\Omega)} \le c\left(\|\Delta u\|_{H^{-3/2}(\Omega)} + \|u\|_{L^2(\partial\Omega)}\right) \le \tilde{c}\left(\|\Delta u\|_{L^2(\Omega)} + \|u\|_{L^2(\partial\Omega)}\right).$$

Setting $u := u_\theta - u_f$ yields

$$\|u_\theta - u_f\|_{H^{1/2}(\Omega)} \le (1 + \tau^{-1/2})\tilde{c}\sqrt{\mathcal{L}_\tau(\theta)}.$$

To show that the estimate (5.36) can not in general be established for any stronger norms, we assume that it holds for some $s \in \mathbb{R}$. As in the proof of Theorem 5.46 we define the energy, this time penalising boundary values

$$E_\tau \colon H^2(\Omega) \to \mathbb{R}, \quad E_\tau(u) := \|\Delta u + f\|^2_{L^2(\Omega)} + \tau\|u - g\|^2_{L^2(\partial\Omega)}.$$

If the estimate (5.36) holds for general parametric classes, this yields

$$\|v - u_f\|^2_{H^s(\Omega)} \le c \cdot E_\tau(v) \quad \text{for all } v \in H^2(\Omega), f \in L^2(\Omega), g \in H^{3/2}(\partial\Omega).$$

Choosing $f = 0$ and $g = 0$ yields

$$\|v\|^2_{H^s(\Omega)} \le c \cdot E_\tau(v) = c \cdot \left(\|\Delta v\|^2_{L^2(\Omega)} + \tau\|v\|^2_{L^2(\partial\Omega)}\right) \quad \text{for all } v \in H^2(\Omega).$$

For $h \in H^{3/2}(\partial\Omega)$ let $u_h \in H^2(\Omega)$ denote the unique harmonic extension, i.e., the solution of

$$\begin{aligned} -\Delta u_h &= 0 \quad \text{in } \Omega \\ u_h &= h \quad \text{on } \partial\Omega. \end{aligned}$$

Now we have

$$\|h\|^2_{H^{s-1/2}(\partial\Omega)} \le c\|u_h\|^2_{H^s(\Omega)} \le \tilde{c}\left(\|\Delta u_h\|^2_{L^2(\Omega)} + \tau\|u_h\|^2_{L^2(\partial\Omega)}\right) = \tilde{c}\tau \cdot \|h\|^2_{L^2(\partial\Omega)} \tag{5.38}$$

for all $h \in H^{3/2}(\partial\Omega)$. In order to see that this implies $s \le 1/2$ we assume the contrary and set $\varepsilon := s - 1/2 > 0$. Then, the embedding $H^{3/2}(\partial\Omega) \hookrightarrow H^{\varepsilon}(\partial\Omega)$ is dense and hence (5.38) extends to $h \in H^{\varepsilon}(\partial\Omega)$. This yields that all norms $\|\cdot\|_{H^{\delta}(\partial\Omega)}$ for $\delta \in (0, \varepsilon)$ are equivalent to $\|\cdot\|_{L^2(\partial\Omega)}$, which implies that all spaces $H^{\delta}(\partial\Omega)$ agree, which constitutes a contradiction. □

Remark 5.50 (Stronger estimates through stronger penalty). We have seen that the $L^2(\partial\Omega)$ penalization can not lead to estimates in a stronger Sobolev norm than $H^{1/2}(\Omega)$. However, inspecting inequality (5.37) one could – at least in theory – penalize the boundary values in the $H^{3/2}(\partial\Omega)$ norm and would then obtain $H^2(\Omega)$ estimates. As the $H^{3/2}(\partial\Omega)$ norm is difficult to approximate in practice, this is no feasible numerical approach.

Remark 5.51 (Stronger estimates through interpolation). It is possible to bound the H^s error for $s \ge 1/2$ of residual minimization with L^2 boundary penalty for the expense of worse rates and under the cost of an additional factor for which it is not clear whether it is bounded. Similar to [53] one can use an interpolation inequality for $s \in [1/2, 2]$ to obtain

$$\|u\|_{H^s(\Omega)} \le \|u\|^{2(2-s)/3}_{H^{1/2}(\Omega)} \cdot \|u\|^{(2s-1)/3}_{H^2(\Omega)} \quad \text{for all } u \in H^2(\Omega).$$

Together with the a posteriori estimate on the $H^{1/2}$ norm, this yields

$$\begin{aligned} \|u_f - u_\theta\|_{H^s(\Omega)} &\le \|u_f - u_\theta\|^{2(2-s)/3}_{H^{1/2}(\Omega)} \cdot \|u_f - u_\theta\|^{(2s-1)/3}_{H^2(\Omega)} \lesssim \|u_f - u_\theta\|^{(2s-1)/3}_{H^2(\Omega)} \cdot L(\theta)^{(2-s)/3} \\ &\le \left(\|u_f\|_{H^2(\Omega)} + \|u_\theta\|_{H^2(\Omega)}\right)^{(2s-1)/3} \cdot L(\theta)^{(2-s)/3}. \end{aligned}$$

Hence, if it is possible to control the H^2 norm of the neural network functions, one obtains an a posteriori estimate on the H^s error. Note however, that the H^2 norm of the neural networks functions is not controlled through the loss function L and hence, this estimates requires an additional explicit or implicit control on the H^2 norm in order to be informative. Note, however, that the power of the a posteriori estimate decreases towards zero for $s \to 2$ and the estimate collapses to a trivial bound for $s = 2$.

5.5.3. Higher order Sobolev norms as a residual measurement.

We discuss the potential benefit of using (higher order) Sobolev norms to measure the residual, as was already proposed by [271]. We are again supposing the exact enforcement of boundary conditions. Our precise setting is the following.

Setting 5.52. *Let $p \in (1, \infty)$ and $k \ge 0$ be fixed. Assume that $\Omega \subseteq \mathbb{R}^d$ is a bounded, open domain with $C^{k+1,1}$ boundary and let $f \in W^{k,p}(\Omega)$ and $g \in W^{2+k-1/p,p}(\partial\Omega)$. Denote by u_f the solution to* (5.34)*. Furthermore, let Θ be a parameter set of a neural network class, such that for every $\theta \in \Theta$ we have $u_\theta \in W^{k+2,p}(\Omega)$ and $u|_{\partial\Omega} = g$. We define the loss function*

$$\mathcal{L}\colon \Theta \to \mathbb{R}, \quad \mathcal{L}(\theta) = \|\Delta u_\theta + f\|^p_{W^{k,p}(\Omega)}. \tag{5.39}$$

In total analogy to Theorem 5.44 we obtain the following result.

Theorem 5.53. *Assume we are in the situation of Setting 5.52, then it holds for every $\theta \in \Theta$ that*

$$\|u_\theta - u_f\|_{W^{k+2,p}(\Omega)} \leq C_{reg}(p,k)\sqrt[p]{\mathcal{L}(\theta)}.$$

Proof. The essential ingredient is the L^p regularity theory that holds under the assumptions made in Setting 5.52, see for instance chapter 2.5 in [124]. The relevant result is that

$$-\Delta\colon W^{k+2,p}(\Omega) \cap W_0^{1,p}(\Omega) \to W^{k,p}(\Omega)$$

is a linear homeomorphism, where $C_{\text{reg}}(p,k)$ denotes the operator norm of its inverse. □

Remark 5.54. The above result might be interesting if approximation of higher derivatives is desired. Furthermore, the empirical findings of [271] suggest that measuring the residual in a Sobolev norm might lead to fewer iterations in a gradient based optimization routine.

5.5.4. Estimates for parabolic equations. The same observation made for the Poisson equation can be exploited for linear parabolic equations when both initial and boundary values are satisfied exactly by the ansatz class. Here, the key is maximal parabolic L^2 regularity theory. We begin by describing our setting.

Setting 5.55. *We consider again a domain $\Omega \subseteq \mathbb{R}^d$ that is H^2 regular for the Laplacian and a finite time interval $I = [0,T]$. For $f \in L^2(I, L^2(\Omega))$, $g \in H^{3/2}(\partial\Omega)$ and $u_0 \in H_0^1(\Omega)$ we consider the parabolic problem*

$$\begin{aligned} d_t u - \Delta u &= f \quad \text{in } I \times \Omega \\ u(t)|_{\partial\Omega} &= g \quad \text{for all } t \in I \\ u(0) &= u_0. \end{aligned} \tag{5.40}$$

Let Θ be a parameter set of a neural network class such that for every $\theta \in \Theta$ the function u_θ is a member of the space

$$\mathcal{X} = H^1(I, L^2(\Omega)) \cap L^2(I, H^2(\Omega) \cap H_g^1(\Omega)), \quad \|u\|_{\mathcal{X}} = \|d_t u\|_{L^2(I,L^2(\Omega))} + \|u\|_{L^2(I,H^2(\Omega))}$$

with $u_\theta(0) = u_0$. This means that both initial and boundary conditions are satisfied exactly. For an introduction to vector-valued Sobolev spaces we refer the reader to [59]. Then we define the loss function

$$\mathcal{L}(\theta) = \|d_t u_\theta - \Delta u_\theta - f\|^2_{L^2(I,L^2(\Omega))}$$

The following theorem is analogue to the case of the Laplacian and relies on a parabolic regularity result.

Theorem 5.56. *Assume xwe are in Setting 5.55. Then it holds for all $\theta \in \Theta$ that*

$$\|u_\theta - u_f\|_{\mathcal{X}} \leq C\sqrt{\mathcal{L}(\theta)}$$

Proof. We denote by $H_0^1(I, L^2(\Omega))$ the vector-valued Sobolev space with vanishing initial values. Maximal parabolic $L^2(\Omega)$ regularity theory tells us that

$$d_t - \Delta\colon H_0^1(I, L^2(\Omega)) \cap L^2(I, H^2(\Omega) \cap H_0^1(\Omega)) \longrightarrow L^2(I, L^2(\Omega))$$

is a linear homeomorphism and this implies the assertion, see for instance [19] for more information on maximal parabolic regularity. The constant C is then the operator norm of $(d_t - \Delta)^{-1}$. □

Remark 5.57. Of course this result is not limited to the heat equation. Indeed one can replace $-\Delta$ by a self-adjoint, coercive operator that satisfies $H^2(\Omega)$ regularity, we refer the reader again to [19] for the corresponding regularity theory. For information on the dependency of the constant C on data, we refer to [12], especially Theorem 4.10.8.

Remark 5.58. [202] report error estimates for parabolic equations not enforcing initial and boundary conditions in the ansatz architecture. We stress that even though the solutions there are assumed to be classical, smooth solutions the error is only estimated in the $L^2(I \times \Omega)$ norm, which is weaker than the estimates presented here. This is again due to advantage of exact boundary and initial conditions.

CHAPTER 6

Energy natural gradients for neural network based PDE solvers

Neural network based PDE solvers have recently experienced an enormous growth in popularity and attention within the scientific community following the works of [109, 129, 269, 110, 237, 176]. Like in Chapter 5 we focus on methods, which parametrize the solution of the PDE by a neural network and use a formulation of the PDE in terms of a minimization problem to construct a loss function used to train the network. The works following this ansatz can be divided into the two approaches: (a) residual minimization of the PDEs residual in strong form, this is known under the name *physics informed neural networks* or *deep Galerkin method*, see for example [100, 164, 269, 237]; (b) if existent, leveraging the variational formulation to obtain a loss function, this is known as the *deep Ritz method* [110], see also [43, 295] for in depth reviews of these methods.

One central reason for the rapid development of these methods is their mesh free nature, which allows easy incorporation of data and their promise to be effective in high-dimensional and parametric problems, that render mesh-based approaches infeasible. Nevertheless, in practice when these approaches are tackled directly with well established optimizers like GD, SGD, Adam or (quasi-)Newton methods, they often fail to produce accurate solutions even for problems of small size. This phenomenon is increasingly well documented in the literature where it is attributed to an insufficient optimization leading to a variety of optimization procedures being suggested, where accuracy better than in the order of 10^{-3} relative L^2 error can rarely be achieved [264, 292, 293, 161, 84, 314]. The only exceptions are ansatzes, which are conceptionally different from direct gradient based optimization, more precisely greedy algorithms and a reformulation as a min-max game [264, 314].

Contributions. We provide a simple, yet effective optimization method that achieves high accuracy for a range of PDEs when combined with the PINN ansatz. Although we evaluate the approach on PDE related tasks, it can be applied to a wide variety of training problems. Our main contributions can be summarized as follows:

- We introduce the notion of *energy natural gradients*. This natural gradient is defined via the Hessian of the training objective in function space, see Definition 6.1. When the same discretization of the function space is used for both the computation of the objective and the natural gradient then the energy natural gradient can be interpreted as a generalized Gauss-Newton method.

- We show that an energy natural gradient update in parameter space corresponds to a Newton update in function space. In particular, for quadratic energies the function space update approximately moves into the direction of the error u^*-u_θ, see Theorem 6.2.

- We demonstrate the capabilities of the energy natural gradient combined with a simple line search to achieve an accuracy, which is several orders of magnitude higher compared to standard optimizers like GD, Adam and Newton's method. These examples include PINN formulations of stationary and evolutionary PDEs as well as the deep Ritz formulation of a nonlinear ODE. The numerical evaluation is contained in Section 6.2.

Related works. Here, we focus on improving the training process and thereby the accuracy of PINNs. It has been documented in various works that direction optimization of the parameters rarely achieves high accuracy [264, 292, 293, 161, 84, 314] with quasi-Newton methods regarded as being among the most efficient optimizers [191, 52]. After our work on energy natural gradients a Gauss-Newton method has been suggested and theoretically analyzed for the deep Ritz method, however, without providing accuracy greater than 10^{-3} [130].

It has been observed that the magnitude of the gradient contributions from the PDE residuum, the boundary terms and the initial conditions often possess imbalanced magnitudes. To address this, different weighting strategies for the individual components of the loss have been developed [292, 197, 293]. Albeit improving PINN training, non of the mentioned works reports relative L^2 errors below 10^{-4}.

The choice of the collocation points in the discretization of PINN losses has been investigated in a variety of works [183, 215, 85, 313, 291, 303]. Common in all these studies is the observation that collocation points should be concentrated in regions of high PDE residual and we refer to [85, 303] for an extensive comparisons of the different proposed sampling strategies in the literature. Further, for time dependent problems curriculum learning is reported to mitigate training pathologies associated with solving evolution problems with a long time horizon [291, 161]. Again, while all aforementioned works considerably improve PINN training, in non of the contributions errors below 10^{-4} could be achieved.

Different optimization strategies, which are conceptionally different to a direct gradient based optimization of the objective, have been proposed in the context of PINNs. For instance, greedy algorithms where used to incrementally build a shallow neural neuron by neuron, which led to high accuracy, up to relative errors of 10^{-8}, for a wide range of PDEs [264]. However, the proposed greedy algorithms are only computationally tractable for shallow neural networks. Another ansatz is to reformulate the quadratic PINN loss as a saddle-point problem involving a network for the approximation of the solution and a discriminator network that penalizes a non-zero residual. The resulting saddle-point formulation cab be solved with competitive gradient descent [314] and the authors report highly accurate – up to 10^{-8} relative L^2 error – PINN solutions for a number of example problems. This approach however comes at the price of training two neural networks and exchanging a minimization problem for a saddle-point problem. Finally, particle swarm optimization methods have been proposed in the context of PINNs, where they improve over the accuracy of standard optimizers, but fail to achieve accuracy better than 10^{-3} despite their computation burden [84].

Natural gradient methods are an established optimization algorithm and we give an overview in Section 6.1 and discuss here only works related to the numerical solution of PDEs. In fact, without explicitly referring to the natural gradient literature and terminology, natural gradients are used in the PDE constrained optimization community in the context of finite elements. For example, in certain situations the mass or stiffness matrices can be interpreted as Gramians, showing that this ansatz is indeed a natural gradient method. For explicit examples we refer to [252, 253]. In the context of neural network based approaches, a variety of natural gradients induced by Sobolev, Fisher-Rao and Wasserstein geometries have been proposed and tested for PINNs [223]. This work focuses on the efficient implementation of these methods and does not consider energy based natural gradients, which we find to be necessary in order to achieve high accuracy.

Notation. To keep this chapter self contained we present all notation used here.

We denote the space of functions on $\Omega \subseteq \mathbb{R}^d$ that are integrable in p-th power by $L^p(\Omega)$ and endow it with its canonical norm.

For a sufficiently smooth function u we denote its partial derivatives by $\partial_i u = \partial u / \partial x_i$ and denote the tensor associated by the l-th derivative by $(D^l u)_{i_1,\dots,i_l} := \partial_{i_1} \dots \partial_{i_l} u$. We denote the gradient of a sufficiently smooth function u by $\nabla u = (\partial_1 u, \dots, \partial_d u)^\top$ and the Laplace operator Δ is defined by $\Delta u := \sum_{i=1}^d \partial_i^2 u$.

We denote the *Sobolev space* of functions with weak derivatives up to order k in $L^p(\Omega)$ by $W^{k,p}(\Omega)$, which is a Banach space with the norm

$$\|u\|^p_{W^{k,p}(\Omega)} := \sum_{l=0}^{k} \|D^l u\|^p_{L^p(\Omega)}.$$

In the following we mostly work with the case $p = 2$ and write $H^k(\Omega)$ instead of $W^{k,2}(\Omega)$.

Consider natural numbers $d, m, L, N_0, \dots, N_L$ and let $\theta = ((A_1, b_1), \dots, (A_L, b_L))$ be a tuple of matrix-vector pairs where $A_l \in \mathbb{R}^{N_l \times N_{l-1}}, b_l \in \mathbb{R}^{N_l}$ and $N_0 = d, N_L = m$. Every matrix vector pair (A_l, b_l) induces an affine linear map $T_l \colon \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}$. The *neural network function with parameters* θ and with respect to some *activation function* $\rho \colon \mathbb{R} \to \mathbb{R}$ is the function

$$u_\theta \colon \mathbb{R}^d \to \mathbb{R}^m, \quad x \mapsto T_L(\rho(T_{L-1}(\rho(\cdots \rho(T_1(x)))))).$$

The *number of parameters* and the *number of neurons* of such a network is given by $\sum_{l=0}^{L-1}(n_l + 1)n_{l+1}$. We call a network *shallow* if it has depth 2 and *deep* otherwise. In the remainder, we restrict ourselves to the case $m = 1$ since we only consider real valued functions. Further, in our experiments we choose tanh as an activation function in order to assume the required notion of smoothness of the network functions u_θ and the parametrization $\theta \mapsto u_\theta$.

For $A \in \mathbb{R}^{n \times m}$ we denote any pseudo inverse of A by A^+.

6.1 Energy natural gradients

Before we introduce natural gradients and in particular energy natural gradients that arise from the Hessian geometry in function space we provide a general setup for variational problems that covers PINNs and the deep Ritz method.

6.1.1. Notation and setup. Various neural network based approaches for the approximate solution of PDEs that cast the solution of the PDE as the minimizer of a typically convex energy over some function space and use this energy to optimize the networks parameters have been suggested [43, 295, 160]. We present two prominent approaches and introduce the unified setup that we use to treat both of these approaches later.

Physics-informed neural networks (PINNs). Consider a general partial differential equation of the form

$$\begin{aligned} \mathcal{L}u &= f \quad \text{in } \Omega \\ \mathcal{B}u &= g \quad \text{on } \partial\Omega, \end{aligned} \tag{6.1}$$

where $\Omega \subseteq \mathbb{R}^d$ is an open set, $\mathcal{L}$ is a – possibly non-linear – partial differential operator and $\mathcal{B}$ is a boundary value operator. We assume that the solution u is sought in a Hilbert space X and that the right-hand side f and the boundary values g are square integrable functions on Ω and $\partial\Omega$ respectively. In this situation, we can reformulate (6.1) as an minimization problem with objective function

$$E(u) = \int_\Omega (\mathcal{L}u - f)^2 \mathrm{d}x + \tau \int_{\partial\Omega} (\mathcal{B}u - g)^2 \mathrm{d}s, \tag{6.2}$$

for a penalization parameter $\tau > 0$. A function $u \in X$ solves (6.1) if and only if $E(u) = 0$. In order to obtain an approximate solution, one can parametrize the function u_θ by a neural network and minimize the network parameters $\theta \in \mathbb{R}^p$ according to the loss function

$$L(\theta) := \int_\Omega (\mathcal{L}u_\theta - f)^2 \mathrm{d}x + \tau \int_{\partial\Omega} (\mathcal{B}u_\theta - g)^2 \mathrm{d}s. \tag{6.3}$$

This general approach to formulate equations as minimization problems is known as *residual minimization* and in the context of neural networks for PDEs can be traced back to [100, 164]. More recently, this ansatz was popularized under the names *deep Galerkin method* or *physics-informed neural networks*, where the loss can also be augmented to encorporate a regression term steming from real world measurements of the solution [269, 237]. In practice, the integrals in the objective function have to be discretized in a suitable way.

The deep Ritz method. When working with weak formulations of PDEs it is standard to consider the variational formulation, i.e., to consider an energy functional such that the Euler-Lagrange equations are the weak formulation of the PDE. This idea was already exploited by Walter Ritz [242] to compute the coefficients of polynomial approximations to solutions of PDEs and popularized in the context of neural networks in [110] who coined the name *deep Ritz method* for this approach. For example for the Poisson equation $-\Delta u = f$ the variational energy is given by

$$u \mapsto \frac{1}{2}\int_\Omega |\nabla u|^2 \mathrm{d}x - \int_\Omega f u \mathrm{d}x$$

compared to the residual energy (6.2). In particular, the energies require different smoothness of the functions and are hence defined on different Sobolev spaces.

Incorporating essential boundary values in the Deep Ritz Method differs from the PINN approach. Whereas in PINNs for any $\tau > 0$ the unique minimizer of the energy is the solution of the PDE, in the deep Ritz method the minimizer of the penalized energy solves a Robin boundary value problem, which can be interpreted as a perturbed problem.

In order to achieve a good approximation of the original problem the penalty parameters need to be large, which leads to ill conditioned problems [213, 77].

General Setup. Both, physics informed neural networks as well as the deep Ritz method fall in the general framework of minimizing an energy $E\colon X \to \mathbb{R}$ xor more precisely the associated objective function $L(\theta) := E(u_\theta)$ over the parameter space of a neural network. Here, we assume X to be a Hilbert space of functions and the functions u_θ computed by the neural network with parameters θ to lie in X and assume that E admits a unique minimizer $u^\star \in X$. Further, we assume that the parametrization $P\colon \mathbb{R}^p \to X, \theta \mapsto u_\theta$ is differentiable and denote its range by $\mathcal{F}_\Theta = \{u_\theta : \theta \in \mathbb{R}^p\}$. We denote the generalized tangent space on this parametric model by

$$T_\theta \mathcal{F}_\Theta := \operatorname{span}\{\partial_{\theta_i} u_\theta : i = 1, \dots, p\}. \tag{6.4}$$

6.1.2. Energy natural gradients and Newton's method in function space. The concept of *natural gradients* was popularized by Amari in the context of parameter estimation in supervised learning and blind source separation [13]. The idea here is to modify the update direction in a gradient based optimization scheme to emulate gradient in a suitable representation space of the parameters. Whereas, this ansatz was already formulated for general metrics it is usually attributed to the use of the Fisher metric on the representation space, but also products of Fisher metrics, Wasserstein and Sobolev geometries have been successfully used [153, 175, 223]. After the initial applications in supervised learning and blind source separation, it was successfully adopted in reinforcement learning [153, 232, 28, 206], inverse problems [223], neural network training [249, 230, 192] and generative models [257, 178]. One sublety in the natural gradients is the definition of a geometry in the function space. This can either be done axiomatically or through the Hessian of a potential function [15, 14, 290, 209]. We follow the idea to work with the natural gradient induced by the Hessian of the convex function space objective in which the natural gradient can be interpreted as a generalized Gauss-Newton method, which has been suggested for neural network training for supervised learning tasks [240, 66, 119, 192, 128]. Contrary to existing works we encounter infinite dimensional and not strongly convex objective in our applications. Further, we develop a function space perspective rescribing the function space update directions as projections of the Newton update direction in function space onto the tangent space of the model.

Here, we consider the setting of the minimization of a convex energy $E\colon X \to \mathbb{R}$ defined on a Hilbert space X, which covers both physics informed neural networks and the deep Ritz method. As an objective function for the optimization of the networks parameters we use $L(\theta) = E(u_\theta)$ like before. We define the *energy Gram matrices* by

$$G_E(\theta)_{ij} := D^2E(u_\theta)(\partial_{\theta_i} u_\theta, \partial_{\theta_j} u_\theta). \tag{6.5}$$

Definition 6.1 (Energy Natural Gradient). Consider the problem $\min_{\theta \in \mathbb{R}^p} L(\theta)$, where $L(\theta) = E(u_\theta)$ and denote the Euclidean gradient by $\nabla L(\theta)$. Then we call

$$\nabla^E L(\theta) := G_E^+(\theta) \nabla L(\theta), \tag{6.6}$$

the *energy natural gradient (E-NG)*[1].

[1] Note that this is different from the *energetic natural gradients* proposed in [279], which defines natural gradients based on the energy distance rather than the Fisher metric.

It is possible to choose other inner products in the function space X for the definition of the Gram matrix and hence the natural gradient. For example if $X = H^s(\Omega)$ one can use the Sobolev inner product to obtain a Gram matrix

$$G_S(\theta)_{ij} := \langle \partial_{\theta_i} u_\theta, \partial_{\theta_j} u_\theta \langle_{H^s(\Omega)}$$

as it was proposed in [223]. We refer to this approach as *Hilbert* or *Sobolev* natural gradients.

For a linear PDE of the form (6.1) the residual yields a quadratic energy and the energy Gram matrix takes the form

$$G_E(\theta)_{ij} = \int_\Omega \mathcal{L}(\partial_{\theta_i} u_\theta)\mathcal{L}(\partial_{\theta_j} u_\theta)\mathrm{d}x + \tau \int_{\partial\Omega} \mathcal{B}(\partial_{\theta_i} u_\theta)\mathcal{B}(\partial_{\theta_j} u_\theta)\mathrm{d}s \tag{6.7}$$

On the other hand, the deep Ritz method for a quadratic energy $E(u) = \frac{1}{2}a(u,u) - f(u)$, where a is a symmetric and coercive bilinear form and $f \in X^*$ yields

$$G_E(\theta)_{ij} = a(\partial_{\theta_i} u_\theta, \partial_{\theta_j} u_\theta). \tag{6.8}$$

For the energy natural gradient we have the following result relating energy natural gradients to Newton updates.

Theorem 6.2 (Energy natural gradient in function space). *If we assume that D^2E is coercive everywhere, then we have*[2]

$$DP_\theta \nabla^E L(\theta) = \Pi^{D^2E(u_\theta)}_{T_\theta \mathcal{F}_\Theta}(D^2E(u_\theta)^{-1}\nabla E(u_\theta)). \tag{6.9}$$

Assume now that E is a quadratic function with bounded and positive definite second derivative $D^2E = a$ that admits a minimizer $u^* \in X$. Then it holds that

$$DP_\theta \nabla^E L(\theta) = \Pi^{a}_{T_\theta \mathcal{F}_\Theta}(u_\theta - u^*). \tag{6.10}$$

For the proof of this result we follow an analogue approach to [226], which addresses finite dimensional spaces.

Lemma 6.3. *Let X be a vector space with a scalar product $\langle \cdot, \cdot \rangle \colon X \times X \to \mathbb{R}$ and consider a linear map $A \colon \mathbb{R}^p \to X$ for some $p \in \mathbb{N}$. Let $G \in \mathbb{R}^{p \times p}$ be given by $G_{ij} := \langle Ae_i, Ae_j \rangle$ and consider the adjoint operator $A^* \colon X \to \mathbb{R}^p$ given by*

$$A^* y := \sum_{i=1}^{p} \langle y, Ae_i \rangle e_i. \tag{6.11}$$

Then it holds that

$$AG^+A^*x = \Pi_{R(A)}(x), \tag{6.12}$$

where $\Pi_{R(A)}(x)$ denotes the projection of x onto the range $R(A) = \{Av : v \in \mathbb{R}^p\}$ of A, which is the unique element satisfying

$$\langle \Pi_{R(A)}(x), z \rangle = \langle x, z \rangle \quad \textit{for all } z \in R(A). \tag{6.13}$$

[2] Here, we interpret the bilinear form $D^2E(u_\theta) \colon H \times H \to \mathbb{R}$ as an operator $D^2E(u_\theta) \colon H \to H$; further $\Pi^{D^2E(u_\theta)}_{T_\theta \mathcal{F}_\Theta}$ denotes the projection with respect to the inner product defined by $D^2E(u_\theta)$.

Proof. It is elementary to check that the adjoint satisfies $\langle A^*x, v\rangle = \langle x, Av\rangle$. Picking some orthonormal basis $(b_i)_{i=1,\dots,d}$ of $R(A)$, the orthogonal projection of $x \in X$ to $R(A)$ exists and is given by $\sum_i \langle x, b_i\rangle b_i$. Without loss of generality we can assume $x \in R(A)$ and otherwise replace x by its projection onto $R(A)$ since A^* vanishes on $R(A)^\perp$.

Let us use the notation $v_i := Ae_i$. Note that clearly $AG^+A^*x \in R(A)$. Hence, it remains to show that $\langle AG^+A^*x, v_i\rangle = \langle x, v_i\rangle$ for all $i = 1, \dots, p$. It holds that $A^*v_i = \sum_j \langle Ae_i, Ae_j\rangle e_j = Ge_i$ and we can express $x = \sum_i a_i v_i$. Using the symmetry of G we can compute

$$\begin{aligned} \langle AG^+A^*x, v_i\rangle &= \langle G^+A^*x, A^*v_i\rangle \\ &= \sum_j a_j \langle G^+A^*v_j, Ge_i\rangle \\ &= \sum_j a_j \langle GG^+Ge_j, e_i\rangle \\ &= \sum_j a_j \langle Ge_j, e_i\rangle \\ &= \sum_j a_j \langle A^*v_j, e_i\rangle \\ &= \sum_j a_j \langle v_j, Ae_i\rangle \\ &= \langle x, v_i\rangle, \end{aligned} \tag{6.14}$$

which completes the proof. □

Theorem 6.4. *Let $(\mathcal{M}, g)$ be a Riemannian Hilbert manifold with model space X, where for any $x \in \mathcal{M}$ the Riemannian metric g_x defines a scalar product on the tangent space $T_x\mathcal{M} \cong X$ rendering $T_x\mathcal{M}$ complete. Consider a differentiable objective function $E\colon \mathcal{M} \to \mathbb{R}$ and a differentiable parametrization $P\colon \mathbb{R}^p \to \mathcal{M}$ and define the Gram matrix in the usual way $G(\theta)_{ij} := g_{P(\theta)}(\partial_{\theta_i}P(\theta), \partial_{\theta_j}P(\theta))$ and consider the objective function $L\colon \mathbb{R}^p \to \mathbb{R}$ given by $\theta \mapsto E(u_\theta)$. Then it holds that*

$$DP_\theta G(\theta)^+ \nabla L(\theta) = \Pi_{T_\theta P(\mathbb{R}^p)} \nabla E(P(\theta)). \tag{6.15}$$

Proof. This follows directly from Lemma 6.3 by setting $X := T_{P(\theta)}\mathcal{M}$ and $A = DP_\theta$, where by the gradient chain rule it holds that $\nabla L(\theta) = DP(\theta)^*\nabla E(u_\theta)$. □

Proof of Theorem 6.2. The case of strongly convex energy E is a falls into the setting of Theorem 6.4 by defining the Riemannian metric via $g_u := DE^2(u)$. It remains to show that the Riemannian gradient with respect to the metric induced by the second derivative D^2E is given by $D^2E(u)^{-1}\nabla E(u)$. This follows from

$$D^2E(u)(D^2E(u)^{-1}\nabla E(u), v) = \langle \nabla E(u), v\rangle = DE(u)v. \tag{6.16}$$

Consider now the case of a symmetric quadratic function E with positiv definite second derivative D^2E and assume that E admits a unique minimizer $u^* \in X$. Lemma 6.3 with $A = DP_\theta$ implies

$$DP_\theta G(\theta)^+ DP_\theta^{*,a}(u - u^*) = \Pi_{T_\theta \mathcal{F}_\Theta}(u - u^*), \tag{6.17}$$

where $DP_\theta^{*,a}$ denotes the adjoint of DP_θ with respect to the inner product a. Hence, it remains to show $\nabla L(\theta) = DP_\theta^{*,a}(u - u^*)$. Note that $E(u) = \frac{1}{2}a(u - u^*, u - u^*) + c$ for a suitable constant $c \in \mathbb{R}$. This follows from the computation

$$\begin{aligned} \langle DP_\theta^{*,a}(u - u^*), e_i\rangle_{\mathbb{R}^p} &= a(u_\theta - u^*, DP_\theta e_i) \\ &= a(u_\theta - u^*, \partial_{\theta_i} u_\theta) \\ &= DE(u_\theta)\partial_{\theta_i} u_\theta \\ &= \partial_{\theta_i} L(\theta), \end{aligned} \tag{6.18}$$

where we used the chain rule in the last step. □

In particular, we see from (6.9) and (6.10) that using the energy natural gradient in parameter space is closely related to a Newton update in function space, where for quadratic energies the Newton direction is given by the error $u_\theta - u^\star$.

Interpretation as a generalized Gauss-Newton method. For an objective function $L(\theta) = \frac{1}{2}\|f(\theta)\|_2^2$ for $f\colon \mathbb{R}^p \to \mathbb{R}^n$ the entries of the Gauss-Newton matrix are given by $A_{GN}(\theta)_{ij} = \partial_{\theta_i} f(\theta)^\top \partial_{\theta_j} f(\theta)$. Typically, as a motivation for this choice the decomposition

$$\partial_{\theta_i}\partial_{\theta_j} L(\theta) = \partial_{\theta_i} f(\theta)^\top \partial_{\theta_j} f(\theta) + \sum_{k=1}^n f_k(\theta)\partial_{\theta_i}\partial_{\theta_j} f_k(\theta)$$

of the Hessian of the objective is used to argue that the Gauss-Newton matrix approximates the Hessian. For a general loss $L(\theta) = E(P(\theta))$ entries of the Hessian of the objective function is give by

$$\partial_{\theta_i}\partial_{\theta_j} L(\theta) = \partial_{\theta_i} DE(u_\theta)\partial_{\theta_j} u_\theta = D^2E(u_\theta)(\partial_{\theta_i} u_\theta, \partial_{\theta_j} u_\theta) + DE(u_\theta)\partial_{\theta_i}\partial_{\theta_j} u_\theta, \tag{6.19}$$

where the first term equals the entry $G_E(\theta)_{ij}$ of the energy gram matrix. In analogy to the classical Gauss-Newton method the first term can be interpreted as a generalized Gauss-Newton matrix [249, 192, 230]. Therefore, the energy natural gradient – and in general any natural gradient induced by the Hessian of the function space objective – can be interpreted as a generalized Gauss-Newton method. Recently, a generalized Gauss-Newton method that coincides with the energy natural gradient has been proposed and analyzed for the special case of a quadratic deep Ritz energy [130].

However, we do not refer to the proposed method as a generalized Gauss-Newton method since we do not see it as an approximation of Newton's method. Much rather, by Theorem 6.2 we see it as a better choice of the update direction as in function space it corresponds to a Newton update.

6.1.3. Visualization of the function space update directions.

In order to demonstrate Theorem 6.2 we visualize the update directions in function space for the energy natural gradient as well as the update directions of Newton's method and the vanilla gradient and compare them to the actual error. Recall that when updating the parameter $\theta \in \mathbb{R}^p$ in direction $v \in \mathbb{R}^p$, i.e., $\theta' = \theta + \eta v$, the resulting update direction in function space is approximately $DP(\theta)v$ since by Taylor's theorem $u_{\theta'} = u_\theta + tDP(\theta)v + O(t^2\|v\|^2)$. Hence, we plot the push forwards

$$DP(\theta)G_E(\theta)^+\nabla L(\theta), \quad DP(\theta)H(\theta)^+\nabla L(\theta) \quad \text{and } DP(\theta)\nabla L(\theta)$$

of the energy natural gradient, the Newton update direction and the vanilla gradient for a random initialization of θ. Here, as a preconditioner for the Newton update direction we choose

$$H(\theta) = D^2L(\theta) - \min(0, \lambda_{min}(D^2L(\theta)))I,$$

where $\lambda_{min}(D^2L(\theta))$ denotes the smallest eigenvalue of the Hessian $D^2L(\theta)$. This is to ensure that preconditioner is positive semi-definite. When plotting the update directions in function space we normalize them to have values in $[-1, 1]$ to allow for a better visual comparison. As a reference for the update directions, we plot the error $u_\theta - u^*$ the model.

We first consider a two-dimensional Poisson equation

$$-\Delta u(x, y) = f(x, y) = 2\pi^2 \sin(\pi x) \sin(\pi y)$$

on the unit square $[0, 1]^2$ with zero boundary values and use a shallow network with 32 hidden neurons and the hyperbolic tangent tanh as an activation function. The solution is given by

$$u^*(x, y) = \sin(\pi x) \sin(\pi y).$$

We draw a parameter vector θ from a Gaussian distribution and in Figure 6.1 we plot the error $u_\theta - u^*$ as well as the push forwards of the energy natural gradient, the Newton update direction and the vanilla gradient. Visually, the energy natural gradient update direction matches the error $u_\theta - u^*$, which is in accordance with Theorem 6.2 that states that the update direction of energy natural gradients corresponds to the projection of the error $u_\theta - u^*$ onto the generalized tangent space of the neural network model. In contrast, the push forwards of the Newton update direction and the vanilla gradient are very different to the error $u_\theta - u^*$.

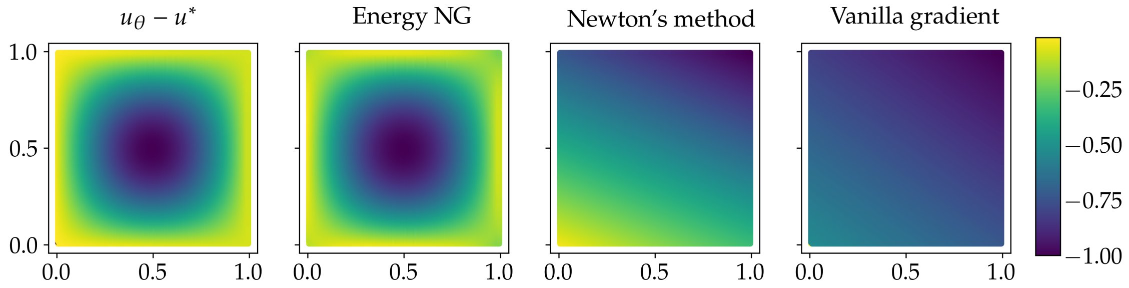

FIGURE 6.1. Shown are the error $u_\theta - u^*$ and the push forwards of the energy natural gradient, the Newton update direction and the vanilla gradient; all functions normed to lie in $[-1, 1]$ to allow for a visual comparison.

As a second example we consider the one-dimensional heat equation

$$\begin{aligned} \partial_t u(t, x) &= \frac{1}{4}\partial_x^2 u(t, x) && \text{for } (t, x) \in [0, 1]^2 \\ u(0, x) &= \sin(\pi x) && \text{for } x \in [0, 1] \\ u(t, x) &= 0 && \text{for } (t, x) \in [0, 1] \times \{0, 1\}. \end{aligned}$$

with solution

$$u^*(t, x) = \exp\left(-\frac{\pi^2 t}{4}\right) \sin(\pi x).$$

Again, we use a shallow network with tanh activation and 32 hidden neurons and draw a parameter θ according to a standard Gaussian distribution. Just like in the case of the Poisson equation, the push forward of the energy natural gradient matches the error $u_\theta - u^*$ very well. Again, Newton's direction and vanilla gradient descent fail to provide update directions in function space that lead to updates proportional to the error.

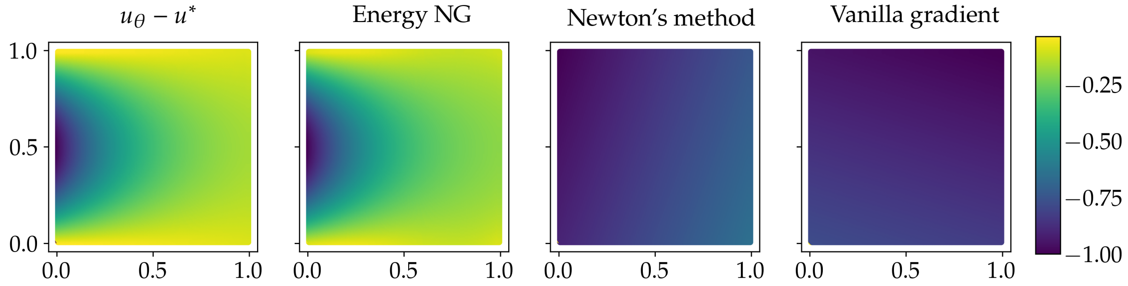

FIGURE 6.2. The first image shows $u_\theta - u^*$, the second image is the computed natural gradient and the last image is the pushforward of the standard parameter gradient. All gradients are pointwise normed to $[-1, 1]$ to allow visual comparison.

6.2 EXPERIMENTS

We test energy natural gradients on four problems: the PINN formulations of two two-dimensional Poisson equation, a PINN formulation of a one-dimensional heat equation and a deep Ritz formulation of a one-dimensional, nonlinear elliptic equation. We evaluate its performance against gradient descent, Adam and Newton's method.

6.2.1. DESCRIPTION OF THE EXPERIMENTS. For all our numerical experiments, we realize an energy natural gradient step with a line search as described in Algorithm 4. We choose the interval $[0, 1]$ for the line search determining the step size since a step size of 1 would correspond to an approximate Newton step in function space. However, since the parametrization of the model is non linear, it is beneficial to conduct the line search and can not simply choose the Newton step size. In our experiments, we use a grid search over a logarithmically spaced grid on $[0, 1]$ to determine the learning rate η^*. The assembly of

Algorithm 4 Energy Natural Gradient Descent with Line Search (E-NGD)

```
Input: initial parameters $\theta_0 \in \mathbb{R}^p$, maximum number of iterations $N_{max}$, functions $\lambda_k\colon \mathbb{R}^p \to \mathbb{R}_{\geq 0}$
for $k = 0, \dots, N_{max} - 1$ do
    Compute $\nabla L(\theta_k) \in \mathbb{R}^p$
    $G_E(\theta)_{ij} \leftarrow D^2E(\partial_{\theta_i} u_\theta, \partial_{\theta_j} u_\theta) + \lambda_k(\theta)\delta_{ij}$ for $i, j = 1, \dots, p$
    $\nabla^E L(\theta) \leftarrow G_E^+(\theta)\nabla L(\theta)$                ▷ Compute the natural gradient
    $\eta_k \leftarrow \arg\min_{\eta \in [0,1]} L(\theta - \eta\nabla^E L(\theta))$        ▷ Choose step size via line search
    $\theta_{k+1} = \theta_k - \eta_k \nabla^E L(\theta_k)$                      ▷ Update parameters
end for
```

the Gram matrix G_E can be done efficiently in parallel, avoiding a potentially costly loop over index pairs (i, j). Instead of computing the pseudo inverse of the Gram matrix $G_E(\theta)$ we solve the least square problem

$$\nabla^E L(\theta) \in \arg\min_{\psi\in\mathbb{R}^p} \|G_E(\theta)\psi - \nabla L(\theta)\|_2^2. \tag{6.20}$$

Although naive, this can easily be parallelized and performs fast and efficient in our experiments.

Further, we introduce a scaling parameter $\lambda_k(\theta) \geq 0$ for a Levenberg–Marquardt type modification of the energy natural gradient, i.e., we use the preconditioner $G_E(\theta)_{ij} \longleftarrow D^2E(\partial_{\theta_i}u_\theta, \partial_{\theta_j}u_\theta) + \lambda(\theta)\delta_{ij}$. In our experiments we test both the performance of energy natural gradients for $\lambda_k(\theta) = 0$ as well as for $\lambda_k(\theta) \propto L(\theta)$ as we observed this to be yield good results, see also [307]. For the numerical evaluation of the integrals in the loss function as well as in the entries of the Gram matrix we experiment both with fixed integration points on a regular grid and repeatedly and randomly drawn integration points.

Tested optimizers. We compare energy natural gradient to vanilla gradient descent, Adam and Newton's method. Here, we describe these baselines in more detail. First, we consider vanilla gradient descent (denoted as GD in our experiments) with a line search on a logarithmic grid. Then, we test the performance of Adam with an exponentially decreasing learning rate schedule to prevent oscillations, where we start with an initial learning rate of 10^{-3} that after $1.5 \cdot 10^4$ steps starts to decrease by a factor of 10^{-1} every 10^4 steps until a minimum learning rate of 10^{-7} is reached or the maximal amount of iterations is completed. Further, we chose not to include the Hilbert space natural gradient induced since we found it to not yield competitive results and sometimes even failing to reduce the error at all.

Further, we test Newton's method, which for a strongly convex produces the updates

$$\theta_{k+1} = \theta_k - \eta_k H(\theta_k)^+ \nabla^E L(\theta_k), \tag{6.21}$$

where $H(\theta_k) = D^2L(\theta_k)$ is the Hessian of the objective. Since the objective is non convex in our settings, we follow [222] and choose

$$H(\theta_k) = D^2L(\theta_k) - \min(0, \lambda_{min}(D^2L(\theta)))I, \tag{6.22}$$

where $\lambda_{min}(D^2L(\theta_k))$ denotes the smallest eigenvalue of the Hessian $D^2L(\theta_k)$. This ensures that $H(\theta_k)$ is positive semi-definite. Finally, we conduct a line search to choose the step size η_k.

Overall, we compare the following optimizers in our experiments:

- Gradient descent with line search (GD)
- Adam
- Newton's method (Newton)
- Energy natural gradient (E-NGD)
- Energy natural gradient with Levenberg–Marquardt modification (E-NGD-LM)

We chose not to include gradient descent with fixed step size as we found it to perform inferior to gradient descent with line search. We also did not include the Sobolev natural gradients proposed in [223] as we did not find them to yield competitive performance when compared to energy natural gradients or Newton's method. Further, we did not include Newton's method with a Levenberg-Marquardt modification as we found this to offer no benefit over the Newton method used here; note that we add a multiple of the identity matrix to the Hessian whenever the Hessian is has a negative eigenvalue (6.22), which was in practice almost always the case. We report the relative[3] L^2 errors during and after the optimization process.

Computation details. For our implementation we rely on the library JAX [60], where all required derivatives are computed using JAX' automatic differentiation module. The JAX implementation of the least square solve relies on a singular value decomposition. The code used in order to assemble the Gram matrices and compute the natural gradients can be found in the repository `https://github.com/MariusZeinhofer/Natural-Gradient-PINNs-ICML23`. For the implementation of Newton's method we use the function `jax.hessian` to compute the Hessian of the objective and employ `jax.numpy.linalg.eigvals` to compute its eigenvalues.

6.2.2. Two-dimensional Poisson equation.

We consider the two-dimensional Poisson equation

$$-\Delta u(x,y) = f(x,y) = 2\pi^2 \sin(\pi x)\sin(\pi y)$$

on the unit square $[0,1]^2$ with zero boundary values. The solution is given by

$$u^*(x,y) = \sin(\pi x)\sin(\pi y)$$

and the PINN loss of the problem is

$$L(\theta) = \frac{1}{N_\Omega}\sum_{i=1}^{N_\Omega}(\Delta u_\theta(x_i,y_i) + f(x_i,y_i))^2 + \frac{1}{N_{\partial\Omega}}\sum_{i=1}^{N_{\partial\Omega}} u_\theta(x_i^b,y_i^b)^2, \tag{6.23}$$

where $\{(x_i,y_i)\}_{i=1,\dots,N_\Omega}$ denote the interior collocation points and $\{(x_i^b,y_i^b)\}_{i=1,\dots,N_{\partial\Omega}}$ denote the collocation points on $\partial\Omega$. In this case the energy inner product on $H^2(\Omega)$ is given by

$$a(u,v) = \int_\Omega \Delta u \Delta v \mathrm{d}x + \int_{\partial\Omega} uv\mathrm{d}s. \tag{6.24}$$

Note that this inner product is not coercive[4] on $H^2(\Omega)$ and different from the $H^2(\Omega)$ inner product. The integrals in (6.24) are computed using the same collocation points as in the definition of the PINN loss function L in (6.23).

Algorithmic choices. To approximate the solution u^* we use a shallow neural network with the hyperbolic tangent as activation function and a width of 32, thus there are 129 trainable weights. We choose 900 equi-distantly spaced collocation points in the interior of Ω and 120 collocation points on the boundary. The energy natural gradient descent and the Hilbert natural gradient descent are applied for 10^3 iterations each, whereas we train for 10^5 iterations of GD and Adam. For all methods apart from Adam we conduct a line search evaluating the objective for the step sizes $\{2^{-k} : k = 0,\dots,30\}$. For the Levenberg-Marquardt type modification of the natural energy gradient descent we add $\lambda_k(\theta)I = 10^{-6-m}L(\theta)I$ if $k \in \{100m, 100m+99\}$ to the respective gram matrix. We initialize

[3] i.e., normalized by the norm of the solution

[4] the inner product is coercive with respect to the $H^{1/2}(\Omega)$ norm, see [214]

	Median	Minimum	Maximum
GD	$1.2 \cdot 10^{-2}$	$2.6 \cdot 10^{-3}$	$2.3 \cdot 10^{-2}$
Adam	$1.3 \cdot 10^{-3}$	$7.7 \cdot 10^{-4}$	$1.9 \cdot 10^{-3}$
E-NGD	$1.3 \cdot 10^{-7}$	$\mathbf{1.7 \cdot 10^{-8}}$	$\mathbf{2.5 \cdot 10^{-7}}$
E-NGD-LM	$\mathbf{9.3 \cdot 10^{-8}}$	$1.9 \cdot 10^{-8}$	$5.8 \cdot 10^{-7}$
Newton	$1.7 \cdot 10^{-6}$	$7.2 \cdot 10^{-7}$	$1.0 \cdot 10^{-5}$

TABLE 6.1. Median, minimum and maximum of the relative L^2 errors for the Poisson equation example achieved by different optimizers over 10 initializations. Here, the energy natural gradient methods and Newton's method are run for 10^3 and the other methods for 10^5 iterations.

the network's weights and biases according to a Gaussian with standard deviation 0.1 and vanishing mean.

Evaluation and discussion. In Table 6.1 we report the minimum, median and maximum error of the different optimized over ten random initializations. Further, in Table 6.2 we report the wallclock and CPU time required per iteration as well as the total wall clock time of the different optimizers in our experiments. In Figure 6.3 we present the evolution of the relative L^2 error for the different optimizers during the optimization.

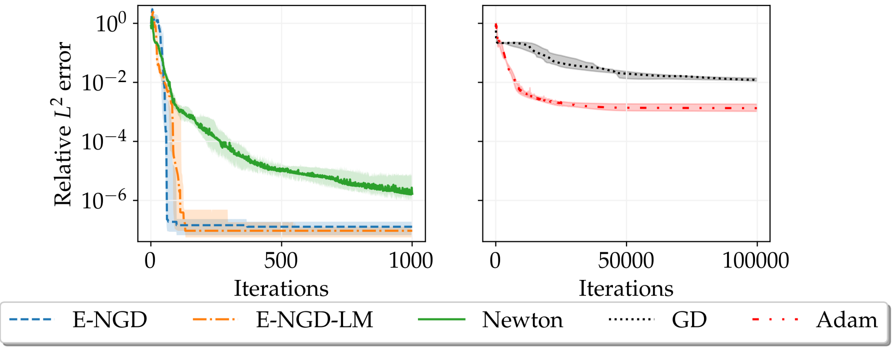

FIGURE 6.3. Median relative L^2 errors for the two dimensional Poisson equation example over 10 initializations for the five optimizers; the shaded area denotes the region between the first and third quartile; note that GD and Adam are run for 100 times more iterations and GD, Adam and Newton's method are given more computation time than NGD, see Table 6.2.

From Table 6.1 and Figure 6.3, we deduce that energy natural gradient with and without Levenberg-Marquardt type modification requires relatively few iterations to produce a highly accurate approximate solution of the Poisson equation. Newton's method achieves also solutions of high accuracy, which are around one order of magnitude less accurate than the ones obtained by the energy natural gradient methods while taking significantly more iterations and computation time. Further, the error decreases slower when compared to the two variants of energy natural gradient descent, see Figure 6.3. On the other hand vanilla gradient descent and Adam don't achieve high accuracy and

	Iteration CPU time	Iteration wall clock time	Full wall clock time
GD	$1.5 \cdot 10^{-2}$s	$9.7 \cdot 10^{-3}$s	16min 6s
Adam	$\mathbf{3.1 \cdot 10^{-3}}$**s**	$\mathbf{3.0 \cdot 10^{-3}}$**s**	5min 3s
E-NGD	$4.6 \cdot 10^{-1}$s	$5.9 \cdot 10^{-2}$s	**59s**
E-NGD-LM	$4.9 \cdot 10^{-1}$s	$6.8 \cdot 10^{-2}$s	1min 8s
Newton	1.8s	$2.9 \cdot 10^{-1}$s	4min 50s

Table 6.2. Median computation times for the optimizers for the two dimensional Poisson example. The experiments were conducted on a Apple M2 CPU with 16GB of RAM.

seem to saturate around a relative L^2 error of 10^{-2} and 10^{-3} respectively. Their computation time per iteration is faster compared to the natural gradient methods and Newton's method; here, an iteration of gradient descent takes more time compared to Adam because of the line search. Note however, that we run gradient descent and Adam for 100 times more iterations and hence allow them to take more absolute wall clock time, see Table 6.2. In this example one natural gradient update is around one order of magnitude more expensive as one iteration of gradient descent or Adam when compared in wall clock time, see Table 6.2. Overall, we find that energy natural gradient with and without a Levenberg-Marquardt modification yields converges in very few iterations for different initializations to approximate solutions of accuracy several orders of magnitude higher compared to the other methods.

6.2.3. A higher frequency Poisson equation. Next, we test the energy gradient method on a two-dimensional Poisson equation where the solution has high frequency parts. Usually such problems are regarded as being harder to solve with physics informed neural networks [191]. Here, we adapt the example from above and consider the Poisson equation

$$-\Delta u(x,y) = f(x,y) = 2k^2\pi^2 \sin(k\pi x)\sin(k\pi y)$$

on the unit square $[0,1]^2$ with zero boundary values where the solution is given by

$$u^*(x,y) = \sin(k\pi x)\sin(k\pi y),$$

where in our experiments we consider $k = 4$.

Algorithmic choices. We make the same choices as in Subsection 6.2.2 for the discretization of the integrals, the line search, the number of iterations and the Levenberg-Marquardt type modification. The only thing we change is the network architecture where we choose to work with a shallow network with 64 hidden neurons. We stick however to the choice of the hyperbolic tangent as an activation function and initialize the network's parameters according to a Gaussian with standard deviation 0.1 and vanishing mean.

Evaluation and discussion. In general, our evaluation is analogue to Subsection 6.2.2. In Table 6.3 we report the computation times, in Figure 6.4 we show the relative L^2 errors during training and we omit the table with the errors at the end of training.

When comparing the energy natural gradient methods and Newton's method we find that the two energy natural gradient methods yield higher accuracy. At the beginning

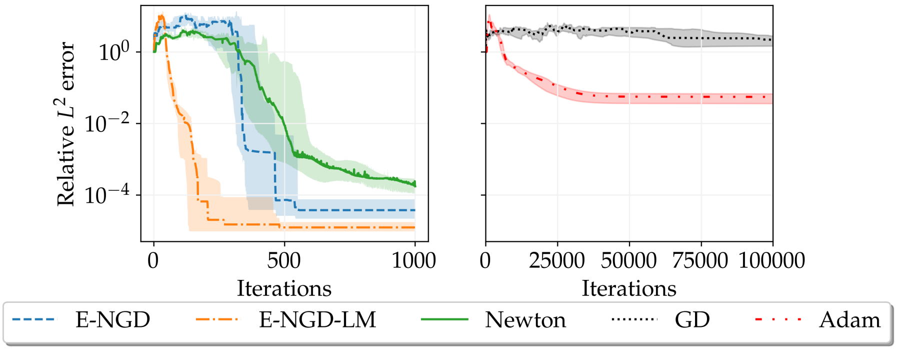

FIGURE 6.4. Median relative L^2 errors for the two dimensional Poisson equation example over 10 initializations for the five optimizers; the shaded area denotes the region between the first and third quartile; note that GD and Adam are run for 100 times more iterations and GD, Adam and Newtons method are given more computation time than NGD, see Table 6.2.

	Iteration CPU time	Iteration wall clock time	Full wall clock time
GD	$2.9 \cdot 10^{-2}$s	$1.5 \cdot 10^{-2}$s	25min 45s
Adam	$\mathbf{7.4 \cdot 10^{-3}}$**s**	$\mathbf{6.1 \cdot 10^{-3}}$**s**	10min 14s
E-NGD	1.4s	$2.5 \cdot 10^{-1}$s	**4min 10s**
E-NGD-LM	1.6s	$3.3 \cdot 10^{-1}$s	5min 28s
Newton	3.5s	1.1s	18min 57s

TABLE 6.3. Mean computation times for the optimizers for the two dimensional Poisson example. The experiments were conducted on a Apple M2 CPU with 16GB of RAM.

of training however energy natural gradient without Levenberg-Marquardt type modification and Newton's method both exhibit a plateau for the first few hundred iterations. This is not the case for the energy natural gradient with Levenberg-Marquardt type modification that achieves similar accuracy compared to the plain energy natural gradient and suffers from less plateaus. Again, we see that gradient descent with line search and Adam fail to produce high accuracy even when given significantly more computation time compared to the energy natural gradient methods.

6.2.4. Heat equation.

Let us consider the one-dimensional heat equation

$$\begin{aligned} \partial_t u(t,x) &= \frac{1}{4}\partial_x^2 u(t,x) && \text{for } (t,x) \in [0,1]^2 \\ u(0,x) &= \sin(\pi x) && \text{for } x \in [0,1] \\ u(t,x) &= 0 && \text{for } (t,x) \in [0,1] \times \{0,1\}. \end{aligned}$$

The solution is given by

$$u^*(t,x) = \exp\left(-\frac{\pi^2 t}{4}\right)\sin(\pi x)$$

and the PINN loss is

$$\begin{aligned} L(\theta) = \frac{1}{N_{\Omega_T}} \sum_{i=1}^{N_{\Omega_T}} &\left(\partial_t u_\theta(t_i, x_i) - \frac{1}{4} \partial_x^2 u_\theta(t_i, x_i) \right)^2 \\ &+ \frac{1}{N_{\text{in}}} \sum_{i=1}^{N_\Omega} \left(u_\theta(0, x_i^{\text{in}}) - \sin(\pi x_i^{\text{in}}) \right)^2 \\ &+ \frac{1}{N_{\partial\Omega}} \sum_{i=1}^{N_{\partial\Omega}} u_\theta(t_i^b, x_i^b)^2, \end{aligned}$$

where $\{(t_i, x_i)\}_{i=1,\dots,N_{\Omega_T}}$ denote collocation points in the interior of the space-time cylinder, $\{(t_i^b, x_i^b)\}_{i=1,\dots,N_{\partial\Omega}}$ denote collocation points on the spatial boundary and $\{(x_i^{\text{in}})\}_{i=1,\dots,N_{\text{in}}}$ denote collocation points for the initial condition. The energy inner product is defined on the space

$$a\colon \left(H^1(I, L^2(\Omega)) \cap L^2(I, H^2(\Omega)) \right)^2 \to \mathbb{R}$$

and given by

$$\begin{aligned} a(u, v) = \int_0^1 \int_\Omega &\left(\partial_t u - \frac{1}{4} \partial_x^2 u \right) \left(\partial_t v - \frac{1}{4} \partial_x^2 v \right) \mathrm{d}x \mathrm{d}t \\ &+ \int_\Omega u(0, x) v(0, x) \, \mathrm{d}x + \int_{I \times \partial\Omega} uv \, \mathrm{d}s \mathrm{d}t. \end{aligned}$$

In our implementation, the inner product is discretized by the same quadrature points as in the definition of the loss function.

Algorithmic choices. The network architecture and the training process are identical to the previous example of the Poisson problem, i.e., we use a shallow network with 64 neurons and the hyperbolic tangent as activation function. We run the energy natural gradient methods as well as Newton's method for $2 \cdot 10^3$ iterations and gradient descent and Adam for 10^5 iterations. We use the same choice for the strength of the Levenberg-Marquardt type modification as for the two Poisson equations. We initialize the network's parameters according to a Gaussian with standard deviation 0.1 and vanishing mean.

Evaluation and discussion. Again, we present the computation times in Table 6.4 and the relative L^2 error during training in Figure 6.5.

For the heat equation observe a very similar behavior of the different optimizers: both versions of the energy natural gradient descent achieve am accuracy of around one order of magnitude better compared to Newton's method while taking less time. We see that also in this example the Levenberg-Marquardt type modification removes the initial plateau of the energy natural gradient. Note again the saturation of gradient descent and Adam above 10^{-2} and 10^{-3} relative L^2 error respectively although they are given more computation time.

6.2.5. A nonlinear example with the deep Ritz method. Finally, we test the energy natural gradient method for a nonlinear problem utilizing the deep Ritz formulation.

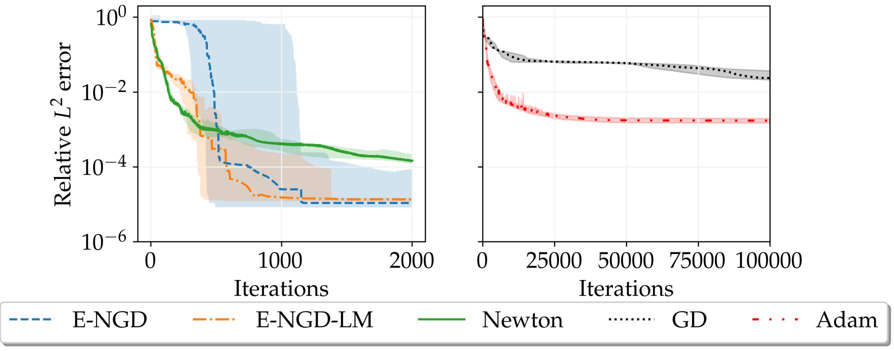

FIGURE 6.5. Median relative L^2 errors for the heat equation example throughout the training process for the five optimizers. The shaded area displays the region between the first and third quartile of 10 runs for different initializations. Note that GD and Adam are run for 100 times more iterations and GD, Adam and Newton's method are given more computation time than E-NGD, see Table 6.4.

	Iteration CPU time	Iteration wall clock time	Full wall clock time
GD	$7.6 \cdot 10^{-3}$s	$5.1 \cdot 10^{-3}$s	8min 33s
Adam	$\mathbf{4.9 \cdot 10^{-3}}$**s**	$\mathbf{4.9 \cdot 10^{-3}}$**s**	8min 12s
E-NGD	$7.2 \cdot 10^{-1}$s	$1.1 \cdot 10^{-1}$s	**3min 37s**
E-NGD-LM	1.3s	$2.3 \cdot 10^{-1}$s	7min 47s
Newton	1.3s	$3.0 \cdot 10^{-1}$s	10min 8s

TABLE 6.4. Mean computation times for the optimizers for the heat equation. For the time per iteration we averaged over 100 iterations. The experiments were conducted on a Apple M2 CPU with 16GB of RAM.

Consider the one-dimensional variational problem of finding the minimizer of the energy

$$E(u) := \frac{1}{2}\int_\Omega |u'|^2 \mathrm{d}x + \frac{1}{4}\int_\Omega u^4 \mathrm{d}x - \int_\Omega f u \,\mathrm{d}x \tag{6.25}$$

with $\Omega = [-1, 1]$ and $f(x) = \pi^2 \cos(\pi x) + \cos^3(\pi x)$. The associated Euler Lagrange equations yield the nonlinear differential equation

$$\begin{aligned} -u'' + u^3 &= f \quad \text{in } \Omega \\ \partial_n u &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{6.26}$$

and hence the minimizer is given by $u^*(x) = \cos(\pi x)$. Since the energy is not quadratic, the energy inner product depends on $u \in H^1(\Omega)$ and is given by

$$D^2E(u)(v, w) = \int_\Omega v'w' \,\mathrm{d}x + 3\int_\Omega u^2 vw \,\mathrm{d}x.$$

Algorithmic choices. To discretize the energy and the inner product we use trapezoidal integration with $2 \cdot 10^4$ equi-spaced quadrature points, which we found to be necessary in order to achieve high accuracy. We use a shallow neural network of width of 32 neurons and a hyperbolic tangent as an activation function. For all methods apart from Adam we conduct a line search evaluating the objective for the step sizes $\{2^{-k} : k = 0, \dots, 30\}$. For the Levenberg-Marquardt type modification of the natural energy gradient descent we choose $\lambda_k(\theta) = 10^{-8-m} L(\theta)$ if $k \in \{10m, 10m + 9\}$. Both versions of energy natural gradient descent and Newton's method are applied for 10^3 iterations, whereas we train GD and Adam for 10^5 iterations. Again, we initialize the network's parameters according to a Gaussian with standard deviation 0.1 and vanishing mean.

Evaluation and discussion. Once more, we observe that the energy NG updates efficiently lead to a very accurate approximation of the solution, see Figure 6.6 for a visualization of the training process and Table 6.5 for the computation times for the individual methods.

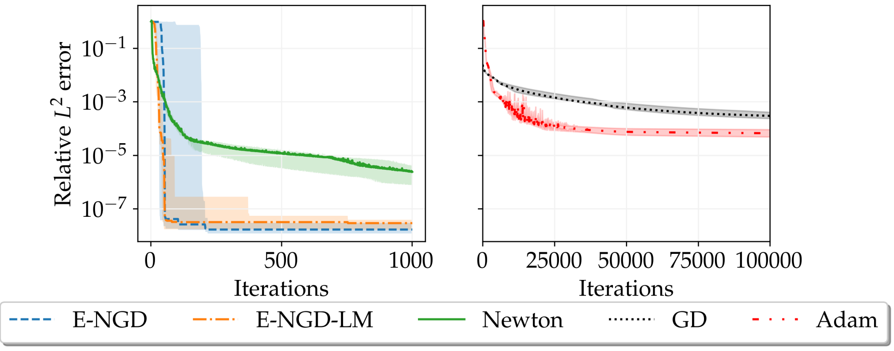

FIGURE 6.6. Relative L^2 errors for the nonlinear example throughout the training process for the five optimizers. The shaded area displays the region between the first and third quartile of 10 runs for different initializations. Note that GD and Adam are run for 100 times more iterations and GD, Adam and Newton's method are given significantly more computation time than E-NGD and E-NGD-LM, see Table 6.5.

	Iteration CPU time	Iteration wall clock time	Full wall clock time
GD	$2.4 \cdot 10^{-1}$s	$1.0 \cdot 10^{-1}$s	2h 49min 40s
Adam	$\mathbf{2.1 \cdot 10^{-2}s}$	$\mathbf{1.8 \cdot 10^{-2}s}$	30min 45s
E-NGD	1.3s	$3.0 \cdot 10^{-1}$s	**5min 1s**
E-NGD-LM	1.5s	$3.2 \cdot 10^{-1}$s	5min 19s
Newton	5.8s	3.0s	49min 50s

TABLE 6.5. Median computation times for the optimizers for the nonlinear example; experiments were conducted on a Apple M2 CPU with 16GB of RAM.

In this example both versions of the energy natural gradient achieve very high accuracy in very few iterations often converging in less than 50 iterations. Again, Adam and standard gradient descent saturate early with much higher errors than the natural gradient methods. In comparison, Newton's method does not seem to saturate but only achieves an accuracy two orders of magnitude larger compared to the energy natural gradients.

6.3 Conclusion and outlook

We propose to train physics informed neural networks with energy natural gradients, which is a natural gradient based on the geometric information of the Hessian in function space and can be interpreted as a generalized Gauss-Newton method. We show that the energy natural gradient update direction corresponds to the Newton direction in function space, modulo an orthogonal projection onto the tangent space of the model. We demonstrate experimentally that this optimization achieves highly accurate PINN solutions, well beyond the the accuracy that can be obtained with standard optimizers even if these methods are allowed several order of magnitude more computation time. The proposed method is compatible with arbitrary discretizations of the integrals appearing in the objective and the gram matrix as with arbitrary network architectures.

Important steps in the pursue of efficient neural network based PDE solvers that can be applied at an industrial scale include the following:

- *Efficient implementation:* An efficient implementation of energy natural gradients – possibly in matrix-free fashion – would vastly improve the applicability of physics informed neural networks to large scale and industrial problems.
- *Initialization schemes:* Since the convergence of energy natural gradient descent is sensitive to the initialization we believe that it is important to gain a better understand the behavior of different initialization schemes.
- *Levenberg-Marquardt schemes:* We observed Levenberg-Marquardt type modifications of energy natural gradient to reduce the plateaus of the energy natural gradient. Where our choices seemed to work well in practice a systematic procedure for the choice would improve the applicability of energy natural gradients.

Bibliography

[1] J. Abadie. "On the Kuhn-Tucker theorem". In: *Nonlinear Programming (NATO Summer School, Menton, (1964)*. North-Holland, Amsterdam, 1967, pp. 19–36.

[2] Alekh Agarwal et al. "On the theory of policy gradient methods: Optimality, approximation, and distribution shift". In: *Journal of Machine Learning Research* 22.98 (2021), pp. 1–76.

[3] Alekh Agarwal et al. *Reinforcement learning: Theory and algorithms*. 2022. URL: https://rltheorybook.github.io/.

[4] Antti Airola and Tapio Pahikkala. "Fast Kronecker product kernel methods via generalized vec trick". In: *IEEE transactions on neural networks and learning systems* 29.8 (2017), pp. 3374–3387.

[5] Carlo Alfano and Patrick Rebeschini. "Dimension-Free Rates for Natural Policy Gradient in Multi-Agent Reinforcement Learning". In: *arXiv preprint arXiv:2109.11692* (2021).

[6] Carlo Alfano and Patrick Rebeschini. "Linear Convergence for Natural Policy Gradient with Log-linear Policy Parametrization". In: *arXiv preprint arXiv:2209.15382* (2022).

[7] Carlo Alfano, Rui Yuan, and Patrick Rebeschini. "A Novel Framework for Policy Mirror Descent with General Parametrization and Linear Convergence". In: *Sixteenth European Workshop on Reinforcement Learning*. 2023. URL: https://openreview.net/forum?id=Rh429iw7d_C.

[8] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: a Hitchhiker's Guide*. Berlin; London: Springer, 2006. ISBN: 9783540326960 3540326960. DOI: 10.1007/3-540-29587-9.

[9] Hans Wilhelm Alt. "Linear Functional Analysis". In: *An application oriented introduction* (1992).

[10] Eitan Altman and Adam Shwartz. "Markov decision problems and state-action frequencies". In: *SIAM journal on control and optimization* 29.4 (1991), pp. 786–809.

[11] Felipe Alvarez, Jérôme Bolte, and Olivier Brahic. "Hessian Riemannian gradient flows in convex programming". In: *SIAM journal on control and optimization* 43.2 (2004), pp. 477–501.

[12] Herbert Amann. *Linear and Quasilinear Parabolic Problems: Volume I: Abstract Linear Theory*. Vol. 1. Springer Science & Business Media, 1995.

[13] Shun-Ichi Amari. "Natural gradient works efficiently in learning". In: *Neural computation* 10.2 (1998), pp. 251–276.

[14] Shun-ichi Amari. *Information geometry and its applications*. Vol. 194. Springer, 2016.

[15] Shun-ichi Amari and Andrzej Cichocki. “Information geometry of divergence functions”. In: *Bulletin of the polish academy of sciences. Technical sciences* 58.1 (2010), pp. 183–195.

[16] Shun-Ichi Amari and Scott C Douglas. “Why natural gradient?” In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP’98 (Cat. No. 98CH36181)*. Vol. 2. IEEE. 1998, pp. 1213–1216.

[17] Christopher Amato, Daniel S Bernstein, and Shlomo Zilberstein. “Solving POMDPs using quadratically constrained linear programs”. In: *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. 2006, pp. 341–343.

[18] Michael Arbel et al. “Kernelized Wasserstein Natural Gradient”. In: *International Conference on Learning Representations*. 2020.

[19] Wolfgang Arendt, Dominik Dier, and Stephan Fackler. “JL Lions’ Problem on Maximal Regularity”. In: *Archiv der Mathematik* 109.1 (2017), pp. 59–72.

[20] Raman Arora et al. “Understanding Deep Neural Networks with Rectified Linear Units”. In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=B1J_rgWRW.

[21] Kumar Ashutosh et al. “Lower Bounds for Policy Iteration on Multi-action MDPs”. In: *2020 59th ieee conference on decision and control (cdc)*. IEEE. 2020, pp. 1744–1749.

[22] Karl Johan Åström. “Optimal control of Markov processes with incomplete state information”. In: *Journal of mathematical analysis and applications* 10.1 (1965), pp. 174–205.

[23] Giles Auchmuty. “Bases and comparison results for linear elliptic eigenproblems”. In: *Journal of Mathematical Analysis and Applications* 390.1 (2012), pp. 394–406.

[24] Nihat Ay et al. *Information geometry*. Vol. 64. Springer, 2017.

[25] Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. “Open problem: Approximate planning of POMDPs in the class of memoryless policies”. In: *Conference on Learning Theory*. PMLR. 2016, pp. 1639–1642.

[26] Kamyar Azizzadenesheli, Yisong Yue, and Animashree Anandkumar. “Policy Gradient in Partially Observable Environments: Approximation and Convergence”. In: *arXiv:1810.07900* (2018).

[27] Ivo Babuška. “The finite element method with penalty”. In: *Mathematics of computation* 27.122 (1973), pp. 221–228.

[28] J. Andrew Bagnell and Jeff G. Schneider. “Covariant Policy Search”. In: *IJCAI*. 2003, pp. 1019–1024.

[29] Genming Bai et al. “Physics Informed Neural Networks (PINNs) For Approximating Nonlinear Dispersive PDEs”. In: *Journal of Computational Mathematics* 39.6 (2021), pp. 816–847.

[30] Leemon C Baird III. *Advantage updating*. Tech. rep. WRIGHT LAB WRIGHT-PATTERSON AFB OH, 1993.

[31] Chanderjit Bajaj. “The Algebraic Degree of Geometric Optimization Problems”. In: *Discrete & Computational Geometry* 3.2 (1988), pp. 177–191.

[32] Lorenzo Baldi and Bernard Mourrain. *Exact Moment Representation in Polynomial Optimization*. Preprint, arXiv:2012.14652. 2022.

[33] Serguei Barannikov et al. “Barcodes as summary of loss function’s topology”. In: *arXiv:1912.00043* (2019).

[34] John W Barrett and Charles M Elliott. “Finite element approximation of the Dirichlet problem using the boundary penalty method”. In: *Numerische Mathematik* 49.4 (1986), pp. 343–366.

[35] Saugata Basu. “Algorithms in Real Algebraic Geometry: A Survey”. In: *arXiv:1409.1534* (2014).

[36] Saugata Basu. “Different Bounds on the Different Betti Numbers of Semi-Algebraic Sets”. In: *Discrete and Computational Geometry* 30.1 (2003), pp. 65–85.

[37] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry (Algorithms and Computation in Mathematics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[38] Nicole Bäuerle and Ulrich Rieder. *Markov decision processes with applications to finance*. Springer Science & Business Media, 2011.

[39] Jonathan Baxter and Peter L Bartlett. “Infinite-Horizon Policy-Gradient Estimation”. In: *Journal of Artificial Intelligence Research* 15 (2001), pp. 319–350.

[40] Jonathan Baxter, Peter L Bartlett, et al. “Reinforcement Learning in POMDP’s via Direct Gradient Ascent”. In: *ICML*. Citeseer. 2000, pp. 41–48.

[41] M.S. Bazaraa, H.D. Sherali, and C.M. Shetty. *Nonlinear programming*. Third. Theory and algorithms. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006, pp. xvi+853. ISBN: 978-0-471-48600-8; 0-471-48600-0.

[42] James C Bean, John R Birge, and Robert L Smith. “Aggregation in dynamic programming”. In: *Operations Research* 35.2 (1987), pp. 215–220.

[43] Christian Beck et al. “An overview on deep learning-based approximation methods for partial differential equations”. In: *arXiv preprint arXiv:2012.12348* (2020).

[44] Marc Bellemare et al. “A geometric perspective on optimal representations for reinforcement learning”. In: *Advances in neural information processing systems* 32 (2019).

[45] Richard Bellman. “A Markovian decision process”. In: *Journal of mathematics and mechanics* 6.5 (1957), pp. 679–684.

[46] Richard Bellman. “The theory of dynamic programming”. In: *Bulletin of the American Mathematical Society* 60.6 (1954), pp. 503–515.

[47] Jens Berg and Kaj Nyström. “A unified deep artificial Neural Network Approach to Partial Differential Equations in complex Geometries”. In: *Neurocomputing* 317 (2018), pp. 28–41.

[48] D.P. Bertsekas. “Nonlinear programming”. In: *Journal of the Operational Research Society* 48.3 (1997), pp. 334–334.

[49] J. Bezanson et al. “Julia: A fresh approach to numerical computing”. In: *SIAM review* 59.1 (2017), pp. 65–98.

[50] Jalaj Bhandari and Daniel Russo. “Global optimality guarantees for policy gradient methods”. In: *arXiv preprint arXiv:1906.01786* (2019).

[51] Jalaj Bhandari and Daniel Russo. “On the linear convergence of policy gradient methods for finite MDPs”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 2386–2394.

[52] Rafael Bischof. *10 Useful Hints and Tricks for Improving Physics-Informed Neural Networks (PINNs)*. `https://towardsdatascience.com/10-useful-hints-and-tricks-for-improving-pinns-1a5dd7b86001`. Accessed: 2023-04-30.

[53] Animikh Biswas, Jing Tian, and Suleyman Ulusoy. "Error estimates for deep learning methods in fluid dynamics". In: *Numerische Mathematik* 151.3 (2022), pp. 753–777.

[54] David Blackwell. "Discounted dynamic programming". In: *The Annals of Mathematical Statistics* 36.1 (1965), pp. 226–235.

[55] David Blackwell. "Discrete dynamic programming". In: *The Annals of Mathematical Statistics* (1962), pp. 719–726.

[56] David Blackwell. "Positive dynamic programming". In: *Proceedings of the 5th Berkeley symposium on Mathematical Statistics and Probability*. Vol. 1. University of California Press Berkeley. 1967, pp. 415–418.

[57] Jan Blechschmidt and Oliver G Ernst. "Three ways to solve partial differential equations with neural networks—A review". In: *GAMM-Mitteilungen* 44.2 (2021), e202100006.

[58] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[59] Franck Boyer and Pierre Fabrie. *Mathematical Tools for the Study of the Incompressible Navier-Stokes Equations and related Models*. Vol. 183. Springer Science & Business Media, 2012.

[60] James Bradbury et al. *JAX: composable transformations of Python+NumPy programs*. Version 0.3.13. 2018. URL: `http://github.com/google/jax`.

[61] Dietrich Braess. *Finite elements: Theory, fast solvers, and applications in solid mechanics*. Cambridge University Press, 2007.

[62] Paul Breiding, Kemal Rose, and Sascha Timme. "Certifying zeros of polynomial systems using interval arithmetic". In: *ACM Transactions on Mathematical Software* 49.1 (2023), pp. 1–14.

[63] Paul Breiding and Sascha Timme. "HomotopyContinuation. jl: A package for homotopy continuation in Julia". In: *International Congress on Mathematical Software*. Springer. 2018, pp. 458–465.

[64] Paul Breiding et al. "Nonlinear Algebra and Applications". In: *arXiv:2103.16300* (2021).

[65] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer Science & Business Media, 2010.

[66] Tianle Cai et al. "Gram-Gauss-Newton method: Learning overparameterized neural networks for regression problems". In: *arXiv preprint arXiv:1905.11675* (2019).

[67] L. Campbell. "An extended Čencov characterization of the information metric". In: *Proceedings of the American Mathematical Society* 98 (1986), pp. 135–141.

[68] Fabrizio Catanese et al. "The maximum likelihood degree". In: *American Journal of Mathematics* 128.3 (2006), pp. 671–697.

[69] Michael J Catanzaro et al. "Moduli spaces of morse functions for persistence". In: *Journal of Applied and Computational Topology* 4.3 (2020), pp. 353–385.

[70] Türkü Özlüm Çelik et al. "Wasserstein distance to independence models". In: *Journal of Symbolic Computation* 104 (2021), pp. 855–873. URL: https://www.sciencedirect.com/science/article/pii/S0747717120301152.

[71] Shicong Cen et al. "Fast global convergence of natural policy gradient methods with entropy regularization". In: *Operations Research* (2021).

[72] N. N. Čencov. *Statistical decision rules and optimal inference*. Vol. 53. Translations of Mathematical Monographs. Translation from the Russian edited by Lev J. Leifman. Providence, R.I.: American Mathematical Society, 1982, pp. viii+499. ISBN: 0-8218-4502-0.

[73] Krishnendu Chatterjee, Martin Chmelík, and Mathieu Tracol. "What is decidable about partially observable Markov decision processes with ω-regular objectives". In: *Journal of Computer and System Sciences* 82.5 (2016), pp. 878–911. URL: https://www.sciencedirect.com/science/article/pii/S0022000016000246.

[74] Jingrun Chen, Rui Du, and Keke Wu. "A Comparison Study of Deep Galerkin Method and Deep Ritz Method for Elliptic Problems with Different Boundary Conditions". In: (2020).

[75] Yichen Chen and Mengdi Wang. "Lower Bound On the Computational Complexity of Discounted Markov Decision Problems". In: *arXiv preprint arXiv:1705.07312* (2017).

[76] Kirill Cherednichenko, Patrick Dondl, and Frank Rösler. "Norm-resolvent convergence in perforated domains". In: *Asymptotic Analysis* 110.3-4 (2018), pp. 163–184.

[77] Luca Courte and Marius Zeinhofer. "Robin Pre-Training for the Deep Ritz Method". In: *Northern Lights Deep Learning Conference* (2023).

[78] Salvatore Cuomo et al. "Scientific machine learning through physics–informed neural networks: where we are and what's next". In: *Journal of Scientific Computing* 92.3 (2022), p. 88.

[79] Wojciech Marian Czarnecki et al. "Sobolev training for neural networks". In: *arXiv preprint arXiv:1706.04859* (2017).

[80] Francois d'Epenoux. "A Probabilistic Production and Inventory Problem". In: *Management Science* 10.1 (1963), pp. 98–108.

[81] Robert Dadashi et al. "The value function polytope in reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 1486–1495.

[82] J. Dahl and E. D. Andersen. "A primal-dual interior-point algorithm for nonsymmetric exponential-cone optimization". In: *Mathematical Programming* (Mar. 2021).

[83] Gianni Dal Maso. *An Introduction to* Γ*-Convergence*. Vol. 8. Springer Science & Business Media, 2012.

[84] Caio Davi and Ulisses Braga-Neto. "PSO-PINN: Physics-Informed Neural Networks Trained with Particle Swarm Optimization". In: *arXiv preprint arXiv:2202.01943* (2022).

[85] Arka Daw et al. "Rethinking the Importance of Sampling in Physics-informed Neural Networks". In: *arXiv preprint arXiv:2207.02338* (2022).

[86] Guy De Ghellinck. "Les problemes de decisions sequentielles". In: *Cahiers du Centre d'Etudes de Recherche Opérationnelle* 2.2 (1960), pp. 161–179.

[87] Tim De Ryck, Ameya D Jagtap, and Siddhartha Mishra. "Error estimates for physics-informed neural networks approximating the Navier–Stokes equations".

In: *IMA Journal of Numerical Analysis* (Jan. 2023), drac085. ISSN: 0272-4979. DOI: 10.1093/imanum/drac085. eprint: https://academic.oup.com/imajna/advance-article-pdf/doi/10.1093/imanum/drac085/49005455/drac085.pdf. URL: https://doi.org/10.1093/imanum/drac085.

[88] Tim De Ryck, Samuel Lanthaler, and Siddhartha Mishra. "On the approximation of functions by tanh neural networks". In: *Neural Networks* 143 (2021), pp. 732–750.

[89] Tim De Ryck and Siddhartha Mishra. "Error analysis for physics-informed neural networks (PINNs) approximating Kolmogorov PDEs". In: *Advances in Computational Mathematics* 48.6 (2022), pp. 1–40.

[90] Ron S Dembo, Stanley C Eisenstat, and Trond Steihaug. "Inexact Newton methods". In: *SIAM Journal on Numerical analysis* 19.2 (1982), pp. 400–408.

[91] Eric V Denardo. "Contraction mappings in the theory underlying dynamic programming". In: *Siam Review* 9.2 (1967), pp. 165–177.

[92] Eric V Denardo. "On linear programming in a Markov decision problem". In: *Management Science* 16.5 (1970), pp. 281–288.

[93] Cyrus Derman. *Finite state Markovian decision processes*. Academic Press, 1970.

[94] Cyrus Derman. "On sequential decisions and Markov chains". In: *Management Science* 9.1 (1962), pp. 16–24.

[95] Guillaume Desjardins, Karen Simonyan, Razvan Pascanu, et al. "Natural neural networks". In: *Advances in neural information processing systems* 28 (2015).

[96] Ronald DeVore, Boris Hanin, and Guergana Petrova. "Neural network approximation". In: *Acta Numerica* 30 (2021), pp. 327–444.

[97] Dongsheng Ding et al. "Convergence and sample complexity of natural policy gradient primal-dual methods for constrained MDPs". In: *arXiv preprint arXiv:2206.02346* (2022).

[98] Dongsheng Ding et al. "Natural policy gradient primal-dual method for constrained Markov decision processes". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 8378–8390.

[99] Jiu Ding and Aihui Zhou. "Eigenvalues of rank-one updated matrices with some applications". In: *Applied Mathematics Letters* 20.12 (2007), pp. 1223–1226.

[100] MWMG Dissanayake and N Phan-Thien. "Neural-network-based approximations for solving partial differential equations". In: *communications in Numerical Methods in Engineering* 10.3 (1994), pp. 195–201.

[101] Manfred Dobrowolski. *Angewandte Funktionalanalysis: Funktionalanalysis, Sobolev-Räume und elliptische Differentialgleichungen*. Springer-Verlag, 2010.

[102] Patrick Dondl, Johannes Müller, and Marius Zeinhofer. "Uniform convergence guarantees for the deep Ritz method for nonlinear problems". In: *Advances in Continuous and Discrete Models* 2022.1 (2022), pp. 1–19.

[103] Joseph Leo Doob. *Stochastic processes*. Vol. 10. New York Wiley, 1953.

[104] Jan Draisma et al. "The Euclidean Distance Degree of an Algebraic Variety". In: *Foundations of Computational Mathematics* 16.1 (2016), pp. 99–149. URL: https://doi.org/10.1007/s10208-014-9240-x.

[105] Mareike Dressler et al. "Algebraic optimization of sequential decision problems". In: *Journal of Symbolic Computation* (2023), p. 102241.

[106] Chenguang Duan et al. "Analysis of Deep Ritz Methods for Laplace Equations with Dirichlet Boundary Conditions". In: *arXiv preprint arXiv:2111.02009* (2021).

[107] Chenguang Duan et al. "Convergence rate analysis for deep Ritz method". In: *arXiv preprint arXiv:2103.13330* (2021).

[108] Iain Dunning, Joey Huchette, and Miles Lubin. "JuMP: A modeling language for mathematical optimization". In: *SIAM review* 59.2 (2017), pp. 295–320.

[109] Weinan E, Jiequn Han, and Arnulf Jentzen. "Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations". In: *Communications in Mathematics and Statistics* 5.4 (2017), pp. 349–380.

[110] Weinan E and Bing Yu. "The Deep Ritz Method: A Deep Learning-Based Numerical Algorithm for Solving Variational Problems". In: *Communications in Mathematics and Statistics* 6.1 (2018), pp. 1–12.

[111] Alexandre Ern and Jean-Luc Guermond. *Theory and practice of finite elements*. Vol. 159. Springer, 2004.

[112] Lawrence C Evans. *Partial Differential Equations*. Vol. 19. Rhode Island, USA, 1998.

[113] Maryam Fazel et al. "Global convergence of policy gradient methods for the linear quadratic regulator". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1467–1476.

[114] Eugene A Feinberg and Gaojin He. "Complexity bounds for approximately solving discounted MDPs by value iterations". In: *Operations Research Letters* 48.5 (2020), pp. 543–548.

[115] Eugene A Feinberg and Jefferson Huang. "The value iteration algorithm is not strongly polynomial for discounted dynamic programming". In: *Operations Research Letters* 42.2 (2014), pp. 130–131.

[116] Eugene A Feinberg and Adam Shwartz. *Handbook of Markov decision processes: methods and applications*. Vol. 40. Springer Science & Business Media, 2012.

[117] Marshall Freimer. "Solving a Markovian decision problems by linear-programming". In: *Annals of Mathematical Statistics*. Vol. 33. 1. INST MATHEMATICAL STATISTICS IMS BUSINESS OFFICE-SUITE 7, 3401 INVESTMENT . . . 1962, p. 301.

[118] Roland W Freund and Ronald W Hoppe. *Stoer/Bulirsch: Numerische Mathematik 1*. Springer-Verlag, 2007.

[119] Matilde Gargiani et al. "On the promise of the stochastic generalized Gauss-Newton method for training DNNs". In: *arXiv preprint arXiv:2006.02409* (2020).

[120] Dean Gillette. "9. STOCHASTIC GAMES WITH ZERO STOP PROBABILITIES". In: *Contributions to the Theory of Games (AM-39), Volume III*. Princeton University Press, 1958, pp. 179–188.

[121] Carsten Gräser. "A note on Poincaré-and Friedrichs-type inequalities". In: *arXiv preprint arXiv:1512.02842* (2015).

[122] D Yu Grigor'ev and NN Vorobjov. "Counting connected components of a semialgebraic set in subexponential time". In: *Computational Complexity* 2.2 (1992), pp. 133–186.

[123] Yuri Grinberg and Doina Precup. "Average Reward Optimization Objective In Partially Observable Domains". In: *International Conference on Machine Learning*. PMLR. 2013, pp. 320–328.

[124] Pierre Grisvard. *Elliptic problems in nonsmooth domains*. SIAM, 2011.

[125] Yiqi Gu and Michael K Ng. "Deep Ritz Method for the Spectral Fractional Laplacian Equation Using the Caffarelli–Silvestre Extension". In: *SIAM Journal on Scientific Computing* 44.4 (2022), A2018–A2036.

[126] Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. "Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms". In: *Analysis and Applications* 18.05 (2020), pp. 803–859.

[127] Ingo Gühring and Mones Raslan. "Approximation rates for neural networks with encodable weights in smoothness spaces". In: *Neural Networks* 134 (2021), pp. 107–130.

[128] Tian-De Guo, Yan Liu, and Cong-Ying Han. "An Overview of Stochastic Quasi-Newton Methods for Large-Scale Machine Learning". In: *Journal of the Operations Research Society of China* 11.2 (2023), pp. 245–275.

[129] Jiequn Han, Arnulf Jentzen, and E Weinan. "Solving high-dimensional partial differential equations using deep learning". In: *Proceedings of the National Academy of Sciences* 115.34 (2018), pp. 8505–8510.

[130] Wenrui Hao et al. "Gauss Newton method for solving variational problems of PDEs with neural network discretizaitons". In: *arXiv preprint arXiv:2306.08727* (2023).

[131] J William Helton and Victor Vinnikov. "Linear matrix inequality representation of sets". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 60.5 (2007), pp. 654–674.

[132] Oliver Hennigh et al. "NVIDIA SimNet™: An AI-accelerated multi-physics simulation framework". In: *International Conference on Computational Science*. Springer. 2021, pp. 447–461.

[133] D. Henrion, J.B. Lasserre, and J. Löfberg. "GloptiPoly 3: moments, optimization and semidefinite programming". In: *Optimization Methods & Software* 24.4-5 (2009), pp. 761–779.

[134] Didier Henrion and Jean-Bernard Lasserre. "Detecting Global Optimality and Extracting Solutions in GloptiPoly". In: *Positive Polynomials in Control*. Ed. by Didier Henrion and Andrea Garulli. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 293–310.

[135] Onésimo Hernández-Lerma and Jean B Lasserre. *Discrete-time Markov control processes: basic optimality criteria*. Vol. 30. Springer Science & Business Media, 2012.

[136] Onésimo Hernández-Lerma et al. "Introduction: Optimal Control Problems". In: *An Introduction to Optimal Control Theory: The Dynamic Programming Approach*. Springer, 2023, pp. 1–12.

[137] Romain Hollanders, Jean-Charles Delvenne, and Raphaël M Jungers. "The complexity of policy iteration is exponential for discounted Markov decision processes". In: *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. IEEE. 2012, pp. 5997–6002.

[138] Sean Hon and Haizhao Yang. "Simultaneous neural network approximations in Sobolev spaces". In: *arXiv preprint arXiv:2109.00161* (2021).

[139] A Hordijk and LCM Kallenberg. "Linear Programming Methods for Solving Finite Markovian Decision Problems ". In: *DGOR*. Springer, 1981, pp. 468–482.

[140] Arie Hordijk and LCM Kallenberg. "Linear programming and Markov decision chains". In: *Management Science* 25.4 (1979), pp. 352–362.

[141] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks". In: *Neural networks* 4.2 (1991), pp. 251–257.

[142] Ronald A Howard. *Dynamic programming and Markov processes*. MIT Press, 1960.

[143] Feihu Huang, Shangqian Gao, and Heng Huang. "Bregman Gradient Policy Optimization". In: *International Conference on Learning Representations*. 2022. URL: `https://openreview.net/forum?id=ZU-zFnTum1N`.

[144] Jeffrey J. Hunter. "Chapter 2 - Generating Functions". In: *Mathematical Techniques of Applied Probability*. Ed. by Jeffrey J. Hunter. Academic Press, 1983, pp. 24–67. URL: `https://www.sciencedirect.com/science/article/pii/B9780123618016500083`.

[145] Rodrigo Toro Icarte et al. "The act of remembering: A study in partially observable reinforcement learning". In: *arXiv preprint arXiv:2010.01753* (2020).

[146] Mohammad Rasool Izadi et al. "Optimization of graph neural networks with natural gradient descent". In: *2020 IEEE international conference on big data (big data)*. IEEE. 2020, pp. 171–179.

[147] J. Nie and K. Ranestad. "Algebraic Degree of Polynomial Optimization". In: *SIAM J. Optim.* 20 (2009), pp. 485–502.

[148] J.B. Lasserre. "Global optimization with polynomials and the problem of moments". In: *SIAM J. Optim.* 11.3 (2000/01), pp. 796–817.

[149] Yuling Jiao et al. "Error analysis of deep Ritz methods for elliptic equations". In: *arXiv preprint arXiv:2107.14478* (2021).

[150] Colin Jones, E. C. Kerrigan, and Jan Maciejowski. *Equality Set Projection: A new algorithm for the projection of polytopes in halfspace representation*. Tech. rep. Cambridge, 2004, p. 45. URL: `http://publications.eng.cam.ac.uk/327023/`.

[151] Jürgen Jost. *Partial Differential Equations*. Springer, 2003.

[152] Sham Kakade and John Langford. "Approximately optimal approximate reinforcement learning". In: *In Proc. 19th International Conference on Machine Learning*. Citeseer. 2002.

[153] Sham M Kakade. "A natural policy gradient". In: *Advances in neural information processing systems* 14 (2001).

[154] Lodewijk CM Kallenberg. "Survey of linear programming for standard and nonstandard Markovian control problems. Part I: Theory". In: *Zeitschrift für Operations Research* 40.1 (1994), pp. 1–42.

[155] Alex Kaltenbach and Marius Zeinhofer. "The Deep Ritz Method for Parametric p-Dirichlet Problems". In: *arXiv preprint arXiv:2207.01894* (2022).

[156] Sajad Khodadadian et al. "On the linear convergence of natural policy gradient algorithm". In: *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE. 2021, pp. 3794–3799.

[157] Jens Kober, J Andrew Bagnell, and Jan Peters. "Reinforcement learning in robotics: A survey". In: *The International Journal of Robotics Research* 32.11 (2013), pp. 1238–1274.

[158] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[159] Vijay Konda and John Tsitsiklis. "Actor-critic algorithms". In: *Advances in neural information processing systems* 12 (1999).

[160] Nikola Kovachki et al. "Neural operator: Learning maps between function spaces". In: *arXiv preprint arXiv:2108.08481* (2021).

[161] Aditi Krishnapriyan et al. "Characterizing possible failure modes in physics-informed neural networks". In: *Advances in Neural Information Processing Systems* 34 (2021).

[162] H.W. Kuhn and A.W. Tucker. "Nonlinear Programming". In: *Second Berkeley Symposium on Mathematical Statistics and Probability*. 1951, pp. 481–492.

[163] HT Kung. "The computational complexity of algebraic numbers". In: *Proceedings of the fifth annual ACM symposium on Theory of computing*. 1973, pp. 152–159.

[164] Isaac E Lagaris, Aristidis Likas, and Dimitrios I Fotiadis. "Artificial neural networks for solving ordinary and partial differential equations". In: *IEEE transactions on neural networks* 9.5 (1998), pp. 987–1000.

[165] Guanghui Lan. "Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes". In: *Mathematical programming* (2022), pp. 1–48.

[166] Amy N. Langville and William J. Stewart. "The Kronecker product and stochastic automata networks". In: *Journal of Computational and Applied Mathematics* 167.2 (2004), pp. 429–447. URL: https://www.sciencedirect.com/science/article/pii/S0377042703009312.

[167] Romain Laroche and Remi Tachet Des Combes. "On the Occupancy Measure of Non-Markovian Policies in Continuous MDPs". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, 23–29 Jul 2023, pp. 18548–18562. URL: https://proceedings.mlr.press/v202/laroche23a.html.

[168] James-Michael Leahy et al. "Convergence of Policy Gradient for Entropy Regularized MDPs with Neural Network Approximation in the Mean-Field Regime". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 12222–12252.

[169] Guy Lebanon. "Axiomatic geometry of conditional models". In: *IEEE Transactions on Information Theory* 51 (2005), pp. 1283–1294.

[170] Kookjin Lee et al. "Partition of unity networks: Deep hp-approximation". In: *CEUR Workshop Proceedings*. Vol. 2964. CEUR-WS. 2021, p. 180.

[171] Gen Li et al. "Softmax policy gradient methods can take exponential time to converge". In: *Conference on Learning Theory*. PMLR. 2021, pp. 3107–3110.

[172] Haoya Li et al. "Quasi-Newton policy gradient algorithms". In: *arXiv preprint arXiv:2110.02398* (2021).

[173] Lihong Li, Thomas J Walsh, and Michael L Littman. "Towards a unified theory of state abstraction for MDPs." In: *AI&M*. 2006.

[174] Lingfeng Li et al. "Priori Error Estimate of Deep Mixed Residual Method for Elliptic PDEs". In: *arXiv preprint arXiv:2206.07474* (2022).

[175] Wuchen Li and Guido Montúfar. "Natural gradient via optimal transport". In: *Information Geometry* 1.2 (2018), pp. 181–214.

[176] Zongyi Li et al. “Fourier Neural Operator for Parametric Partial Differential Equations”. In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=c8P9NQVtmnO.

[177] Yulei Liao and Pingbing Ming. “Deep Nitsche Method: Deep Ritz Method with Essential Boundary Conditions”. In: *Communications in Computational Physics* 29.5 (2021), pp. 1365–1384. ISSN: 1991-7120. DOI: https://doi.org/10.4208/cicp.OA-2020-0219. URL: http://global-sci.org/intro/article_detail/cicp/18717.html.

[178] Alex Tong Lin et al. “Wasserstein proximal of GANs”. In: *International Conference on Geometric Science of Information*. Springer. 2021, pp. 524–533.

[179] Michael L. Littman. “An Optimization-based Categorization of Reinforcement Learning Environments”. In: *From Animals to Animats 2*. Ed. by Jean-Arcady Meyer, Herbert L. Roitblat, and Stewart W. Wilson. MIT Press, 1993, pp. 262–270. URL: http://www.cs.rutgers.edu/~mlittman/papers/sab92.giveout.ps.

[180] Michael L. Littman. “Memoryless Policies: Theoretical Limitations and Practical Results”. In: *Proceedings of the Third International Conference on Simulation of Adaptive Behavior: From Animals to Animats 3: From Animals to Animats 3*. SAB94. Brighton, United Kingdom: MIT Press, 1994, pp. 238–245.

[181] Michael L. Littman, Thomas L. Dean, and Leslie Pack Kaelbling. “On the Complexity of Solving Markov Decision Problems”. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. UAI’95. Montréal, Qué, Canada: Morgan Kaufmann Publishers Inc., 1995, pp. 394–402. ISBN: 1558603859.

[182] John Loch and Satinder P. Singh. “Using Eligibility Traces to Find the Best Memoryless Policy in Partially Observable Markov Decision Processes”. In: *Proceedings of the Fifteenth International Conference on Machine Learning*. ICML ’98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 323–331.

[183] Lu Lu et al. “DeepXDE: A deep learning library for solving differential equations”. In: *SIAM Review* 63.1 (2021), pp. 208–228.

[184] Tao Luo and Haizhao Yang. “Two-layer neural networks for partial differential equations: Optimization and generalization theory”. In: *arXiv preprint arXiv:2006.15733* (2020).

[185] Liyao Lyu et al. “Enforcing Exact Boundary and Initial Conditions in the Deep Mixed Residual Method”. In: *CSIAM Transactions on Applied Mathematics* 2.4 (2021), pp. 748–775. ISSN: 2708-0579. DOI: https://doi.org/10.4208/csiam-am.SO-2021-0011. URL: http://global-sci.org/intro/article_detail/csiam-am/19991.html.

[186] Omid Madani, Steve Hanks, and Anne Condon. “On the undecidability of probabilistic planning and related stochastic optimization problems”. In: *Artificial Intelligence* 147.1 (2003). Planning with Uncertainty and Incomplete Information, pp. 5–34. URL: https://www.sciencedirect.com/science/article/pii/S0004370202003788.

[187] Sridhar Mahadevan. “Average reward reinforcement learning: Foundations, algorithms, and empirical results”. In: *Recent advances in reinforcement Learning* (1996), pp. 159–195.

[188] Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. "Wasserstein Riemannian geometry of Gaussian densities". In: *Information Geometry* 1.2 (2018), pp. 137–179.

[189] Alan S Manne. "Linear Programming and Sequential Decisions". In: *Management Science* 6.3 (1960), pp. 259–267.

[190] Peter Marbach and John N Tsitsiklis. "Simulation-based optimization of Markov reward processes". In: *IEEE Transactions on Automatic Control* 46.2 (2001), pp. 191–209.

[191] Stefano Markidis. "The old and the new: Can physics-informed deep-learning replace traditional linear solvers?" In: *Frontiers in big Data* 4 (2021), p. 669097.

[192] James Martens. "New insights and perspectives on the natural gradient method". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 5776–5851.

[193] James Martens and Roger Grosse. "Optimizing neural networks with kronecker-factored approximate curvature". In: *International conference on machine learning*. PMLR. 2015, pp. 2408–2417.

[194] Bertrand Maury. "Numerical analysis of a finite element/volume penalty method". In: *SIAM Journal on Numerical Analysis* 47.2 (2009), pp. 1126–1148.

[195] Laurentiu G Maxim et al. "Linear optimization on varieties and Chern-Mather classes". In: *arXiv preprint arXiv:2208.09073* (2022).

[196] *Maze generation*. `https://rosettacode.org/wiki/Maze_generation`. Accessed: 2022-01-10.

[197] Remco van der Meer, Cornelis W Oosterlee, and Anastasia Borovykh. "Optimally weighted loss functions for solving pdes with neural networks". In: *Journal of Computational and Applied Mathematics* 405 (2022), p. 113887.

[198] Jincheng Mei et al. "Escaping the gravitational pull of softmax". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21130–21140.

[199] Jincheng Mei et al. "Leveraging non-uniformity in first-order non-convex optimization". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 7555–7564.

[200] Jincheng Mei et al. "On the global convergence rates of softmax policy gradient methods". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6820–6829.

[201] Siddhartha Mishra and Roberto Molinaro. "Estimates on the generalization error of physics-informed neural networks for approximating a class of inverse problems for PDEs". In: *IMA Journal of Numerical Analysis* 42.2 (June 2021), pp. 981–1022. ISSN: 0272-4979. DOI: `10.1093/imanum/drab032`. eprint: `https://academic.oup.com/imajna/article-pdf/42/2/981/43373937/drab032.pdf`. URL: `https://doi.org/10.1093/imanum/drab032`.

[202] Siddhartha Mishra and Roberto Molinaro. "Estimates on the generalization error of physics-informed neural networks for approximating PDEs". In: *IMA Journal of Numerical Analysis* (2022).

[203] Guido Montúfar, Keyan Ghazi-Zahedi, and Nihat Ay. "Geometry and determinism of optimal stationary control in partially observable Markov decision processes". In: *arXiv:1503.07206* (2015). URL: `http://arxiv.org/abs/1503.07206`.

[204] Guido Montúfar and Johannes Rauh. "Geometry of Policy Improvement". In: *International Conference on Geometric Science of Information*. Springer. 2017, pp. 282–290.

[205] Guido Montúfar, Johannes Rauh, and Nihat Ay. "On the Fisher metric of conditional probability polytopes". In: *Entropy* 16.6 (2014), pp. 3207–3233.

[206] Tetsuro Morimura et al. "A New Natural Policy Gradient by Stationary Distribution Metric". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2008, pp. 82–97.

[207] Tetsuro Morimura et al. "Derivatives of logarithmic stationary distributions for policy gradient reinforcement learning". In: *Neural computation* 22.2 (2010), pp. 342–376.

[208] Ted Moskovitz et al. "Efficient Wasserstein Natural Gradients for Reinforcement Learning". In: *International Conference on Learning Representations*. 2021. URL: `https://openreview.net/forum?id=OHgnfSrn2jv`.

[209] Johannes Müller and Guido Montúfar. "Geometry and convergence of natural policy gradient methods". In: *Information Geometry* (2023), pp. 1–39.

[210] Johannes Müller and Guido Montúfar. "Solving infinite-horizon POMDPs with memoryless stochastic policies in state-action space". In: *The 5th Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM 2022)*. 2022.

[211] Johannes Müller and Guido Montúfar. "The Geometry of Memoryless Stochastic Policy Optimization in Infinite-Horizon POMDPs". In: *International Conference on Learning Representations*. 2022. URL: `https://openreview.net/forum?id=A05I5IvrdL-`.

[212] Johannes Müller and Marius Zeinhofer. "Achieving High Accuracy with PINNs via Energy Natural Gradient Descent". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, 23–29 Jul 2023, pp. 25471–25485. URL: `https://proceedings.mlr.press/v202/muller23b.html`.

[213] Johannes Müller and Marius Zeinhofer. "Error Estimates for the Deep Ritz Method with Boundary Penalty". In: *Mathematical and Scientific Machine Learning*. PMLR. 2022, pp. 215–230.

[214] Johannes Müller and Marius Zeinhofer. "Notes on exact boundary values in residual minimisation". In: *Mathematical and Scientific Machine Learning*. PMLR. 2022, pp. 231–240.

[215] Mohammad Amin Nabian, Rini Jasmine Gladstone, and Hadi Meidani. "Efficient training of physics-informed neural networks via importance sampling". In: *Computer-Aided Civil and Infrastructure Engineering* 36.8 (2021), pp. 962–977.

[216] Hiroshi Nagaoka. "The exponential family of Markov chains and its information geometry". In: *arXiv preprint arXiv:1701.06119* (2017).

[217] Tim Netzer and Andreas Thom. "Polynomials with and without determinantal representations". In: *Linear algebra and its applications* 437.7 (2012), pp. 1579–1595.

[218] Gergely Neu, Anders Jonsson, and Vicenç Gómez. "A unified view of entropy-regularized Markov decision processes". In: *arXiv preprint arXiv:1705.07798* (2017).

[219] J. Nie. "Optimality conditions and finite convergence of Lasserre's hierarchy". In: *Mathematical programming* 146.1 (2014), pp. 97–121.

[220] Jiawang Nie. "Certifying convergence of Lasserre's hierarchy via flat truncation". In: *Mathematical Programming* 142 (2011), pp. 485–510.

[221] Jiawang Nie and Kristian Ranestad. "Algebraic Degree of Polynomial Optimization". In: *SIAM Journal on Optimization* 20.1 (2009), pp. 485–502. eprint: `https://doi.org/10.1137/080716670`. URL: `https://doi.org/10.1137/080716670`.

[222] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[223] Levon Nurbekyan, Wanzhou Lei, and Yunan Yang. "Efficient Natural Gradient Descent Methods for Large-Scale Optimization Problems". In: *arXiv:2202.06236* (2022).

[224] Frans A Oliehoek. "Decentralized pomdps". In: *Reinforcement Learning: State-of-the-Art*. Springer, 2012, pp. 471–503.

[225] Frans A Oliehoek, Christopher Amato, et al. *A concise introduction to decentralized POMDPs*. Vol. 1. Springer, 2016.

[226] Jesse van Oostrum, Johannes Müller, and Nihat Ay. "Invariance properties of the natural gradient in overparametrised systems". In: *Information Geometry* (2022), pp. 1–17.

[227] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.

[228] Christos H Papadimitriou and John N Tsitsiklis. "The complexity of Markov decision processes". In: *Mathematics of operations research* 12.3 (1987), pp. 441–450.

[229] Hyeyoung Park, S-I Amari, and Kenji Fukumizu. "Adaptive natural gradient learning algorithms for various stochastic models". In: *Neural Networks* 13.7 (2000), pp. 755–764.

[230] Razvan Pascanu and Yoshua Bengio. "Revisiting natural gradient for deep networks". In: *International Conference on Learning Representations*. 2014. URL: `https://openreview.net/forum?id=vz8AumxkAfz5U`.

[231] Leonid Peshkin, Nicolas Meuleau, and Leslie Pack Kaelbling. "Learning Policies with External Memory". In: *Proceedings of the 16th International Conference on Machine Learning*. Morgan Kaufmann, 1999, pp. 307–314.

[232] Jan Peters, Sethu Vijayakumar, and Stefan Schaal. "Reinforcement learning for humanoid robotics". In: *Proceedings of the third IEEE-RAS international conference on humanoid robots*. 2003, pp. 1–20.

[233] Yury Polyanskiy and Yihong Wu. "Lecture notes on information theory". In: *Lecture Notes for ECE563 (UIUC) and* 6.2012-2016 (2014), p. 7.

[234] Ian Post and Yinyu Ye. "The simplex method is strongly polynomial for deterministic Markov decision processes". In: *Mathematics of Operations Research* 40.4 (2015), pp. 859–868.

[235] Martin L Puterman. "Markov decision processes". In: *Handbooks in operations research and management science* 2 (1990), pp. 331–434.

[236] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[237] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations”. In: *Journal of Computational physics* 378 (2019), pp. 686–707.

[238] Vladimir Rakočević. “On continuity of the Moore-Penrose and Drazin inverses.” In: *Matematichki Vesnik* 49 (1997), pp. 163–172.

[239] Johannes Rauh, Nihat Ay, and Guido Montúfar. “A continuity result for optimal memoryless planning in POMDPs”. In: *The 4th Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM 2019)*. 2019.

[240] Yi Ren and Donald Goldfarb. “Efficient subsampled Gauss-Newton and natural gradient methods for training neural networks”. In: *arXiv preprint arXiv:1906.02353* (2019).

[241] Zhiyuan Ren and Bruce H Krogh. “State aggregation in Markov decision processes”. In: *Proceedings of the 41st IEEE Conference on Decision and Control, 2002.* Vol. 4. IEEE. 2002, pp. 3819–3824.

[242] Walter Ritz. “Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik.” In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 1909.135 (1909), pp. 1–61.

[243] Julian Roth, Max Schröder, and Thomas Wick. “Neural network guided adjoint computations in dual weighted residual error estimation”. In: *SN Applied Sciences* 4.2 (2022), p. 62.

[244] William E Roth. “On direct product matrices”. In: (1934).

[245] Jesus M Ruiz. “Semialgebraic and semianalytic sets”. In: *Cahiers du séminaire d'histoire des mathématiques* 1 (1991), pp. 59–70.

[246] Michael Růžička. *Nichtlineare Funktionalanalysis: Eine Einführung*. Springer-Verlag, 2006.

[247] Martin Schechter. “On L^p Estimates and Regularity II”. In: *Mathematica Scandinavica* 13.1 (1963), pp. 47–69.

[248] Bruno Scherrer. “Improved and generalized upper bounds on the complexity of policy iteration”. In: *Advances in Neural Information Processing Systems* 26 (2013).

[249] Nicol N Schraudolph. “Fast curvature matrix-vector products for second-order gradient descent”. In: *Neural computation* 14.7 (2002), pp. 1723–1738.

[250] John Schulman et al. “Proximal policy optimization algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017).

[251] John Schulman et al. “Trust region policy optimization”. In: *International conference on machine learning*. PMLR. 2015, pp. 1889–1897.

[252] Tobias Schwedes, Simon W Funke, and David A Ham. “An iteration count estimate for a mesh-dependent steepest descent method based on finite elements and Riesz inner product representation”. In: *arXiv preprint arXiv:1606.08069* (2016).

[253] Tobias Schwedes et al. “Mesh dependence in PDE-constrained optimisation”. In: *Mesh Dependence in PDE-Constrained Optimisation*. Springer, 2017, pp. 53–78.

[254] Guy Shani, Joelle Pineau, and Robert Kaplow. “A survey of point-based POMDP solvers”. In: *Autonomous Agents and Multi-Agent Systems* 27 (2013), pp. 1–51.

[255] Kun Shao et al. “A survey of deep reinforcement learning in video games”. In: *arXiv preprint arXiv:1912.10944* (2019).

[256] Lloyd S Shapley. “Stochastic games”. In: *Proceedings of the national academy of sciences* 39.10 (1953), pp. 1095–1100.

[257] Zebang Shen et al. “Sinkhorn natural gradient for generative models”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1646–1656.

[258] Zhong-Ci Shi. “On the convergence rate of the boundary penalty method”. In: *International journal for numerical methods in engineering* 20.11 (1984), pp. 2027–2032.

[259] Hirohiko Shima. *The geometry of Hessian structures*. World Scientific, 2007.

[260] Yeonjong Shin, Jérôme Darbon, and George Em Karniadakis. “On the Convergence of Physics Informed Neural Networks for Linear Second-Order Elliptic and Parabolic Type PDEs”. In: *Communications in Computational Physics* 28.5 (2020), pp. 2042–2074. ISSN: 1991-7120. DOI: `https://doi.org/10.4208/cicp.OA-2020-0193`. URL: `http://global-sci.org/intro/article_detail/cicp/18404.html`.

[261] Yeonjong Shin, Zhongqiang Zhang, and George Em Karniadakis. “Error Estimates of Residual Minimization using Neural Networks for linear PDEs”. In: *arXiv preprint arXiv:2010.08019* (2020).

[262] Jonathan W Siegel and Jinchao Xu. “Approximation rates for neural networks with general activation functions”. In: *Neural Networks* 128 (2020), pp. 313–321.

[263] Jonathan W Siegel and Jinchao Xu. “High-order approximation rates for shallow neural networks with cosine and ReLU^k activation functions”. In: *Applied and Computational Harmonic Analysis* 58 (2022), pp. 1–26.

[264] Jonathan W Siegel et al. “Greedy training algorithms for neural networks and applications to PDEs”. In: *Journal of Computational Physics* 484 (2023), p. 112084.

[265] David Silver et al. “Mastering the game of Go with deep neural networks and tree search”. In: *nature* 529.7587 (2016), pp. 484–489.

[266] David Silver et al. “Mastering the Game of Go without Human Knowledge”. In: *Nature* 550.7676 (2017), pp. 354–359.

[267] Satinder P Singh, Tommi Jaakkola, and Michael I Jordan. “Learning without state-estimation in partially observable Markovian decision processes”. In: *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 284–292.

[268] Satinder P Singh and Richard C Yee. “An upper bound on the loss from approximate optimal-value functions”. In: *Machine Learning* 16 (1994), pp. 227–233.

[269] Justin Sirignano and Konstantinos Spiliopoulos. “DGM: A deep learning algorithm for solving partial differential equations”. In: *Journal of computational physics* 375 (2018), pp. 1339–1364.

[270] Lucca Sodomaco. “The Distance Function from the Variety of partially symmetric rank-one Tensors”. PhD thesis. University of Florence, Department of Mathematics and Computer Science, 2020.

[271] Hwijae Son et al. “Sobolev Training for the Neural Network Solutions of PDEs”. In: *arXiv preprint arXiv:2101.08932* (2021).

[272] Matthijs TJ Spaan. “Partially observable Markov decision processes”. In: *Reinforcement learning: State-of-the-art* (2012), pp. 387–414.

[273] Ralph E Strauch. “Negative dynamic programming”. In: *The Annals of Mathematical Statistics* 37.4 (1966), pp. 871–890.

[274] Michael Struwe. *Variational Methods*. Vol. 31999. Springer, 1990.

[275] Seth Sullivant. *Algebraic statistics*. Vol. 194. American Mathematical Soc., 2018.

[276] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[277] Richard S Sutton et al. "Policy Gradient Methods for Reinforcement Learning with Function Approximation". In: *NIPs*. Vol. 99. Citeseer. 1999, pp. 1057–1063.

[278] Gerald Tesauro. "Temporal Difference Learning and TD-Gammon". In: *Commun. ACM* 38.3 (Mar. 1995), pp. 58–68. URL: https://doi.org/10.1145/203330.203343.

[279] Philip Thomas et al. "Energetic natural gradient descent". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 2887–2895.

[280] Sascha Timme. "Numerical Nonlinear Algebra". PhD thesis. Technische Universität Berlin (Germany), 2021.

[281] Matthew Trager, Kathlén Kohn, and Joan Bruna. "Pure and Spurious Critical Points: a Geometric Study of Linear Networks". In: *International Conference on Learning Representations*. 2019.

[282] Paul Tseng. "Solving H-horizon, stationary Markov decision problems in time proportional to $\log(H)$". In: *Operations Research Letters* 9.5 (1990), pp. 287–297.

[283] Remco van der Meer, Cornelis W. Oosterlee, and Anastasia Borovykh. "Optimally weighted loss functions for solving PDEs with Neural Networks". In: *Journal of Computational and Applied Mathematics* 405 (2022), p. 113887. ISSN: 0377-0427. DOI: https://doi.org/10.1016/j.cam.2021.113887. URL: https://www.sciencedirect.com/science/article/pii/S0377042721005100.

[284] Arthur F Veinott. "Discrete dynamic programming with sensitive discount optimality criteria". In: *The Annals of Mathematical Statistics* 40.5 (1969), pp. 1635–1660.

[285] Arthur F Veinott. "On finding optimal policies in discrete dynamic programming with no discounting". In: *The Annals of Mathematical Statistics* 37.5 (1966), pp. 1284–1294.

[286] Nikos Vlassis, Michael L Littman, and David Barber. "On the Computational Complexity of Stochastic Controller Optimization in POMDPs". In: *ACM Transactions on Computation Theory (TOCT)* 4.4 (2012), pp. 1–8.

[287] A. Wächter and L.T. Biegler. "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming". In: *Mathematical programming* 106.1 (2006), pp. 25–57.

[288] Paul Wagner. "A reinterpretation of the policy oscillation phenomenon in approximate policy iteration". In: *Advances in Neural Information Processing Systems* 24 (2011).

[289] Kaixin Wang et al. *The Geometry of Robust Value Functions*. 2022. DOI: 10.48550/ARXIV.2201.12929. URL: https://arxiv.org/abs/2201.12929.

[290] Li Wang and Ming Yan. "Hessian informed mirror descent". In: *Journal of Scientific Computing* 92.3 (2022), pp. 1–22.

[291] Sifan Wang, Shyam Sankaran, and Paris Perdikaris. "Respecting causality is all you need for training physics-informed neural networks". In: *arXiv preprint arXiv:2203.07404* (2022).

[292] Sifan Wang, Yujun Teng, and Paris Perdikaris. "Understanding and mitigating gradient flow pathologies in physics-informed neural networks". In: *SIAM Journal on Scientific Computing* 43.5 (2021), A3055–A3081.

[293] Sifan Wang, Xinling Yu, and Paris Perdikaris. "When and why PINNs fail to train: A neural tangent kernel perspective". In: *Journal of Computational Physics* 449 (2022), p. 110768.

[294] Jonathan Weed. "An explicit analysis of the entropic penalty in linear programming". In: *Conference On Learning Theory*. PMLR. 2018, pp. 1841–1855.

[295] E Weinan, Jiequn Han, and Arnulf Jentzen. "Algorithms for solving high dimensional PDEs: from nonlinear Monte Carlo to machine learning". In: *Nonlinearity* 35.1 (2021), p. 278.

[296] Stephan Wilhelm Weis. *Exponential Families with Incompatible Statistics and Their Entropy Distance*. Friedrich-Alexander-Universität Erlangen-Nürnberg (Germany), 2010.

[297] Douglas J White. "Dynamic programming, Markov chains, and the method of successive approximations". In: *Journal of Mathematical Analysis and Applications* 6.3 (1963), pp. 373–376.

[298] Chelsea C White III and Douglas J White. "Markov decision processes". In: *European Journal of Operational Research* 39.1 (1989), pp. 1–16.

[299] John Williams and Satinder Singh. "Experimental Results on Learning Stochastic Memoryless Policies for Partially Observable Markov Decision Processes". In: *Advances in Neural Information Processing Systems*. Ed. by M. Kearns, S. Solla, and D. Cohn. Vol. 11. MIT Press, 1999. URL: https://proceedings.neurips.cc/paper/1998/file/1cd3882394520876dc88d1472aa2a93f-Paper.pdf.

[300] Ronald J Williams. "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine learning* 8.3 (1992), pp. 229–256.

[301] Ronald J Williams and Leemon C Baird. *Tight performance bounds on greedy policies based on imperfect value functions*. Tech. rep. Technical report, College of Computer Science, Northeastern University, 1993.

[302] Philip Wolfe and George Bernard Dantzig. "Linear programming in a Markov chain". In: *Operations Research* 10.5 (1962), pp. 702–710.

[303] Chenxi Wu et al. "A comprehensive study of non-adaptive and residual-based adaptive sampling for physics-informed neural networks". In: *Computer Methods in Applied Mechanics and Engineering* 403 (2023), p. 115671.

[304] Yue Wu and Jesús A. De Loera. "Geometric Policy Iteration for Markov Decision Processes". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '22. Washington DC, USA: Association for Computing Machinery, 2022, pp. 2070–2078. ISBN: 9781450393850. DOI: 10.1145/3534678.3539478. URL: https://doi.org/10.1145/3534678.3539478.

[305] Lin Xiao. "On the Convergence Rates of Policy Gradient Methods". In: *Journal of Machine Learning Research* 23.282 (2022), pp. 1–36. URL: http://jmlr.org/papers/v23/22-0056.html.

[306] Jinchao Xu. "Finite Neuron Method and Convergence Analysis". In: *Communications in Computational Physics* 28.5 (2020), pp. 1707–1745. ISSN: 1991-7120. DOI: https://doi.org/10.4208/cicp.OA-2020-0191. URL: http://global-sci.org/intro/article_detail/cicp/18394.html.

[307] Nobuo Yamashita and Masao Fukushima. "On the rate of convergence of the Levenberg-Marquardt method". In: *Topics in Numerical Analysis: With Special Emphasis on Nonlinear Problems*. Springer. 2001, pp. 239–249.

[308] Yinyu Ye. "A new complexity result on solving the Markov decision problem". In: *Mathematics of Operations Research* 30.3 (2005), pp. 733–749.

[309] Yinyu Ye. "The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate". In: *Mathematics of Operations Research* 36.4 (2011), pp. 593–603.

[310] Lexing Ying and Yuhua Zhu. "A note on optimization formulations of Markov decision processes". In: *Communications in Mathematical Sciences* (2022). URL: `https://dx.doi.org/10.4310/CMS.2022.v20.n3.a5`.

[311] Rui Yuan et al. "Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: `https://openreview.net/forum?id=-z9hdsyUwVQ`.

[312] Tom Zahavy et al. "Reward is enough for convex MDPs". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 25746–25759.

[313] Bastian Zapf et al. "Investigating molecular transport in the human brain from MRI with physics-informed neural networks". In: *Scientific Reports* 12.1 (2022), pp. 1–12.

[314] Qi Zeng, Spencer H Bryngelson, and Florian Tobias Schaefer. "Competitive Physics Informed Networks". In: *ICLR 2022 Workshop on Gamification and Multiagent Solutions*. 2022. URL: `https://openreview.net/forum?id=rMz_scJ6lc`.

[315] Wenhao Zhan et al. "Policy Mirror Descent for Regularized Reinforcement Learning: A Generalized Framework with Linear Convergence". In: *SIAM Journal on Optimization* 33.2 (2023), pp. 1061–1091.

[316] Kaiqing Zhang et al. "Global Convergence of Policy Gradient Methods to (Almost) Locally Optimal Policies". In: *SIAM Journal on Control and Optimization* 58.6 (2020), pp. 3586–3612.

[317] Matthew S Zhang, Murat A Erdogdu, and Animesh Garg. "Convergence and Optimality of Policy Gradient Methods in Weakly Smooth Settings". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 8. 2022, pp. 9066–9073.

[318] Günter M Ziegler. *Lectures on Polytopes*. Vol. 152. Springer Science & Business Media, 2012.

Bibliographische Daten

Geometry of Optimization in Markov Decision Processes and Neural Network Based PDE Solvers (Geometrie der Optimierung in Markov-Entscheidungsprozessen und auf neuronalen Netzen basierender PDE-Löser)

Müller, Johannes Christoph

Universität Leipzig, Dissertation, 2023

x + 221 Seiten

22 Abbildungen

10 Tabellen

318 Referenzen

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 18ten September, 2023

. . . .
Johannes Christoph Müller

Daten zum Autor

Name	Johannes Chrisoph Müller
Geburtsdatum	03. Juni 1995
Geburtsort	Augsburg
10/2013-07/2016	B.Sc. Mathematik, Albert-Ludwigs-Universität Freiburg
10/2016-12/2019	M.Sc. Mathematik, Albert-Ludwigs-Universität Freiburg
10/2017-10/2018	M.Sc. Interdisziplinäre Mathematik, University of Warwick
Seit 01/2020	Doktorand der Mathematik