Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Geometry and convergence of natural policy gradient methods

by

Guido Montúfar and Johannes Müller

Preprint no.: 31 2022

Geometry and convergence of natural policy gradient methods

Johannes Müller
Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
jmueller@mis.mpg.de
Guido Montúfar
Departments of Mathematics and Statistics, UCLA, CA, USA
Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
montufar@math.ucla.edu

November 4, 2022

Abstract

We study the convergence of several natural policy gradient (NPG) methods in infinite-horizon discounted Markov decision processes with regular policy parametrizations. 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. 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 penalization strength.
Keywords Markov decision process, Natural policy gradient, State-action frequency, Hessian geometry, stochastic policy

1 Introduction

Markov decision processes (MDPs) are an important model for sequential decision making in interaction with an environment and constitute a theoretical framework for modern reinforcement learning (RL). This framework has been successfully applied in recent years to solve increasingly complex tasks from robotics to board and video games [62, 63, 56, 45, 60]. In MDPs the goal is to identify a *policy* π, i.e., a procedure to select actions at every time step, which maximizes an expected time-aggregated reward $R(\pi)$. We will assume that the set of possible states $\mathcal{S}$ and the set of possible actions $\mathcal{A}$ are finite, and 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. We will study gradient-based policy optimization methods and more specifically *natural policy gradient (NPG)* methods. Inspired by the seminal works of Amari [5, 8],

various NPG methods have been proposed [29, 47, 49]. In general, they take the form

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

where $G(\theta)^+$ 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. Most of our analysis does not actually depend on the specific choice of the pseudo inverse, but in Section 6 we will use the Moore-Penrose pseudo 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 [29] proposed using $P(\theta) = \pi_\theta$ and taking for g a product of Fisher metrics weighted by the state probabilities 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. [56] offered reasons to regard Kakade's metric as the true Fisher metric in this case, yet other choices of the weights can be motivated by axiomatic approaches to define a Fisher metric of conditional probabilities [35, 46]. 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. [47] 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.

- 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 and 4.3 and Proposition 16).
- 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 (Theorems 25 and 26 and Corollaries 30 and 31).
- 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 26 and Corollary 31, 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 regularization strength. This yields a relatively short proof for the local quadratic convergence of regularized NPG methods with Newton step sizes (Theorem 33). This recovers as a special case the local quadratic convergence of Kakade's NPG under state-wise entropy regularization previously shown in [19].

Related work The application of natural gradients to optimization in MDPs was first proposed by Kakade [29], taking as a metric on $\Delta_{\mathcal{A}}^{\mathcal{S}}$ the product of Fisher metrics on $\Delta_{\mathcal{A}}^{s}$, $s \in \mathcal{S}$, weighted by the stationary state distribution. The relation of this metric to finite-horizon Fisher information matrices was studied by Bagnell and Schneider [12] as well as by Peters et al. [56]. Later, Morimura et al. [47] 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 [15]. 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 [15, 73, 74].

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 [44, 43]. For continuous state and action spaces, vanilla PG converges linearly for entropic regularization and shallow policy networks in the mean-field regime [34].

For Kakade's NPG, [1] 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 [16], constant step sizes [31, 3, 70], and for geometrically increasing step sizes [69]. For regularized problems, the method converges linearly for small step sizes and locally quadratically for Newton-like step size [19]. These results have been extended to more general frameworks using state-mixtures of Bregman divergences on the policy polytope [33, 72, 37], which however do not include NPG methods defined in state-action space such as Morimuras NPG. For projected PGs, [1] shows sublinear convergence at a rate $O(t^{-1/2})$, and the result has been improved to a sublinear rate $O(t^{-1})$ [69], and to a linear rate for step sizes chosen by exact line search [16].

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 [24, 25]. For partially observable systems, a gradient domination property has been established in [11]. NPG methods with dimension-free global convergence guarantees have been studied for multi-agent MDPs and potential games [2]. The sample complexity of a Bregman policy gradient arising from a strongly convex function in parameter space has been studied in [27]. For the linear quadratic regulator, global linear convergence guarantees for vanilla, Gauss-Newton and Kakades natural policy gradient methods are provided in [26]; note that 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 $O(\eta^{-1}|\mathcal{S}|^{2^{\Omega((1-\gamma)^{-1})}})$ on the iteration complexity for softmax PG method with step size η has been established in [36].

Notation We denote the simplex of probability distributions over a finite set $\mathcal{X}$ by $\Delta_{\mathcal{X}}$. An element $\mu \in \Delta_{\mathcal{X}}$ is a vector with non-negative entries $\mu_x = \mu(x)$, $x \in \mathcal{X}$ adding to one, $\sum_x \mu_x = 1$. We denote the set of Markov kernels from a finite set $\mathcal{X}$ to another finite set $\mathcal{Y}$ by $\Delta_{\mathcal{Y}}^{\mathcal{X}}$. An element $Q \in \Delta_{\mathcal{Y}}^{\mathcal{X}}$ is a $|\mathcal{X}| \times |\mathcal{Y}|$ row stochastic matrix with entries $Q_{xy} = Q(y|x)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$. Given $Q^{(1)} \in \Delta_{\mathcal{Y}}^{\mathcal{X}}$ and $Q^{(2)} \in \Delta_{\mathcal{Z}}^{\mathcal{Y}}$ we denote their composition into a kernel from $\mathcal{X}$ to $\mathcal{Z}$ by $Q^{(2)} \circ Q^{(1)} \in \Delta_{\mathcal{Z}}^{\mathcal{X}}$. Given $p \in \Delta_{\mathcal{X}}$ and $Q \in \Delta_{\mathcal{Y}}^{\mathcal{X}}$ we denote their composition into a joint probability distribution by $p * Q \in \Delta_{\mathcal{X}\times\mathcal{Y}}$, $(p * Q)(x, y) \coloneqq p(x)Q(y|x)$. 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 $\mu \in \mathbb{R}^{\mathcal{X}}_{\geq 0}$ we denote its *Shannon entropy* by

$$H(\mu) \coloneqq -\sum_x \mu(x)\log(\mu(x)),$$

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

$$H(\mu|\mu_X) \coloneqq -\sum_{x,y} \mu(x,y)\log\frac{\mu(x,y)}{\mu_X(x)} = H(\mu) - H(\mu_X). \tag{1}$$

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

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 $df_p\colon T_p\mathcal{M} \to T_{f(p)}\mathcal{N}$. In the Euclidean case, we also write $Df(p)$ for the Jacobian matrix with entries $Df(p)_{ij} = \partial_j f_i(p)$. 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.

For $A\in\mathbb{R}^{n\times m}$, we denote its Moore-Penrose inverse by $A^+ \in \mathbb{R}^{m\times n}$. Note that AA^+ is the orthogonal (Euclidean) projection onto range(A) and A^+A is the orthogonal (Euclidean) projection onto $\ker(A)$. Finally, 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 $f(t)\leq cg(t)$ for $t\to t_0$, where we allow $t_0 = +\infty$.

2 Markov decision processes

A *Markov decision process* or shortly *MDP* is a tuple $(\mathcal{S}, \mathcal{A}, \alpha, r)$. We assume that $\mathcal{S}$ and $\mathcal{A}$ are finite sets which we call the *state* and *action space* respectively. We fix a Markov kernel $\alpha \in \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$ which we call the *transition mechanism*. Further, we consider an *instantaneous reward vector* $r\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. In the case of partially observable MDPs (POMDPs) one also has a fixed kernel $\beta\in\Delta_{\mathcal{O}}^{\mathcal{S}}$ called the *observation mechanism*. The system is *fully observable* if $\beta = \mathrm{id}$,[1] in which case the POMDP simplifies to an MDP.

As *policies* we consider elements $\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}$. More generally, in POMDPs we would consider *effective policies* $\pi = \pi'\circ\beta \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ with $\pi'\in\Delta_{\mathcal{A}}^{\mathcal{O}}$. We will focus on the MDP case, however. A policy induces 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

$$P_\pi(s',a'|s,a) \coloneqq \alpha(s'|s,a)(\pi\circ\beta)(a'|s') \quad \text{and} \quad p_\pi(s'|s) \coloneqq \sum_{a\in\mathcal{A}} (\pi\circ\beta)(a|s)\alpha(s'|s,a). \tag{2}$$

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 which we denote by $\mathbb{P}^{\pi,\mu}$. For a *discount rate* $\gamma\in(0,1)$ we define

$$R(\pi) = R_\gamma^\mu(\pi) \coloneqq \mathbb{E}_{\mathbb{P}^{\pi,\mu}}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right],$$

called the *expected discounted reward*. The *expected mean reward* is obtained as the limit with $\gamma\to 1$ when this exists. We will focus on the discounted case, however. The goal is to maximize R over the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{O}}$. For a policy π we define the *value function* $V^\pi = V_\gamma^\pi \in \mathbb{R}^{\mathcal{S}}$ via $V^\pi(s) \coloneqq R_\gamma^{\delta_s}(\pi)$, $s\in\mathcal{S}$, where δ_s is the Dirac distribution concentrated at s.

[1] More generally, the system is fully observable if the supports of $\{\beta(\cdot|s)\}_{s\in\mathcal{S}}$ are disjoint subsets of $\mathcal{O}$.

A short calculation shows that $R(\pi) = \sum_{s,a} r(s,a)\eta^\pi(s,a) = \langle r, \eta^\pi \rangle_{\mathcal{S}\times\mathcal{A}}$ [71], where

$$\eta^\pi(s,a) \coloneqq (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}^{\pi,\mu}(s_t = s, a_t = a). \tag{3}$$

The vector η^π is an element of $\Delta_{\mathcal{S}\times\mathcal{A}}$ called the *expected (discounted) state-action frequency* [22], or (discounted) visitation/occupancy measure, or on-policy distribution [64]. Denoting the state marginal of η^π by $\rho^\pi \in \Delta_{\mathcal{S}}$ we have $\eta^\pi(s,a) = \rho^\pi(s)(\pi \circ \beta)(a|s)$. We denote the set of all state-action frequencies in the fully and in the partially observable cases by

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

Note that the expected cumulative reward function $R\colon \Delta_{\mathcal{A}}^{\mathcal{O}} \to \mathbb{R}$ factorizes according to

$$\Delta_{\mathcal{A}}^{\mathcal{O}} \to \Delta_{\mathcal{A}}^{\mathcal{S}} \to \mathcal{N}^\mu \to \mathbb{R}, \quad \pi \mapsto \pi \circ \beta \mapsto \eta^\pi \mapsto \langle r, \eta^\pi \rangle_{\mathcal{S}\times\mathcal{A}}.$$

We recall the following well-known facts.

Proposition 1 (State-action polytope of MDPs, [22])**.** *The set $\mathcal{N}$ of state-action frequencies is a polytope given by $\mathcal{N} = \Delta_{\mathcal{S}\times\mathcal{A}} \cap \mathcal{L} = \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}} \cap \mathcal{L}$, where*

$$\mathcal{L} = \left\{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \ell_s(\eta) = 0 \text{ for all } s \in \mathcal{S}, \eta \geq 0\right\}, \tag{4}$$

and $\ell_s(\eta) \coloneqq \sum_a \eta_{sa} - \gamma \sum_{s',a'} \eta_{s'a'}\alpha(s|s',a') - (1-\gamma)\mu_s$.

The state-action polytope for a two-state MDP is shown in Figure 3. We note that in in the case of partially observable MDPs, the set of state-action frequencies $\mathcal{N}^\beta$ does not form a polytope, but rather a polynomially constrained set involving polynomials of higher degree depending on the properties of the observation kernel [50].

The result above shows that a (fully observable) Markov decision process can be solved by means of linear programming. Indeed, if η^* is a solution of the linear program $\langle r, \eta \rangle_{\mathcal{S}\times\mathcal{A}}$ over $\mathcal{N}$, one can compute the maximizing policy over $\Delta_{\mathcal{A}}^{\mathcal{S}}$ by conditioning, $\pi^*(a|s) = \eta^*(s,a)/\sum_{a'} \eta^*(s,a)$. We propose to study the evolution of natural policy gradient methods in state-action space $\mathcal{N} \subseteq \Delta_{\mathcal{S}\times\mathcal{A}}$. Indeed, we show that the evolution of diverse natural policy gradient algorithms in the state-action polytope solves the gradient flow of a (regularized) linear objective with respect to a Hessian geometry in state-action space. This perspective facilitates relatively short proofs for the global convergence of natural policy gradient methods and can also provide rates. In order to relate Riemannian geometries in the policy space $\Delta_{\mathcal{A}}^{\mathcal{S}}$ to Riemannian geometries in the state-action polytope $\mathcal{N}$ we need the following assumption.

Assumption 2 (Positivity)**.** For every $s \in \mathcal{S}$ and $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$, we assume that $\sum_a \eta_{sa}^\pi > 0$.

Assumption 2 holds in particular if either $\alpha > 0$ and $\gamma > 0$ or $\gamma < 1$ and $\mu > 0$ entrywise [50]. This assumption is standard in linear programming approaches and necessary for the convergence of policy gradient methods in MDPs [30, 44]. With this assumption in place we have the following.

Proposition 3 (Inverse of state-action map, [50])**.** *Under Assumption 2, the mapping $\Delta_{\mathcal{A}}^{\mathcal{S}} \to \mathcal{N}, \omega \mapsto \eta$ is rational and bijective with rational inverse given by conditioning $\mathcal{N} \to \Delta_{\mathcal{A}}^{\mathcal{S}}, \eta \mapsto \omega$, where $\omega_{as} = \frac{\eta_{sa}}{\sum_{a'} \eta_{sa'}}$.*

This result shows that the (interior of the) set of policies and the (interior of the) state-action polytope are diffeomorphic. Hence, we can port the Riemannian geometry on any of the two sets to the other by using the pull back along $\pi \mapsto \eta$ or the conditioning map $\eta \mapsto \pi$.

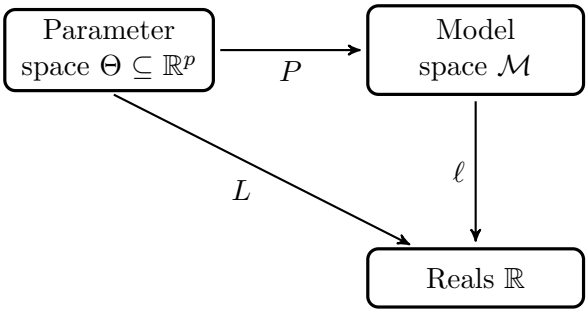

Figure 1: Schematic drawing of parametric models with an objective function ℓ and resulting parameter objective function L; note that neither the choice of geometry in the model space nor the factorization or the model space itself is uniquely determined by the objective function L.

3 Natural gradients

In this section we provide some background on the notion of natural gradients.

3.1 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 1. This general setup, with an objective function that factorizes as $L(\theta) = \ell(P(\theta))$, covers several usual 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 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 [5] 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 [55, 42, 23, 28], policy optimization [29, 56, 47] and inverse problems [54]. 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 [38, 39, 9]. 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 (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 1. For $\theta \in \Theta$ we define the Gram matrix $G(\theta)_{ij} \coloneqq g_{P(\theta)}(dP_\theta e_i, dP_\theta e_j)$ and call $\nabla^N L(\theta) \coloneqq G(\theta)^+ \nabla L(\theta)$ 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 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 $p \coloneqq P(\theta)$. Moving in the direction $v \in T_\theta\Theta$ in 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 $T_\theta\mathcal{M}_\Theta \coloneqq \operatorname{range}(d_\theta P) \subseteq T_p\mathcal{M}$. Hence, the best direction one can take on $\mathcal{M}_\Theta$ by infinitesimally varying the parameter θ is given by

$$\operatorname*{arg\,max}_{w \in T_\theta\mathcal{M}_\Theta, g_p(w,w)=1} \partial_w \ell(p),$$

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 5 (Natural gradient leads to steepest descent in model space)**.** *Consider the settings of Definition 4, where $\mathcal{M}$ is a Riemannian manifold with metric g. Let $\nabla^N L(\theta) \coloneqq G(\theta)^+ \nabla_\theta L(\theta)$ denote the natural gradient with respect to this factorization. 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 [6, Subsection 12.1.2]; for singular Gram matrices we refer to [66, Theorem 1].

3.2 Choice of a geometry in model space

Invariance axiomatic geometries A celebrated theorem by Chentsov [20] 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. [10].

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

$$g_p(u, v) = (f^* g')_p(u, v) \coloneqq g'_{f(p)}(f_* u, f_* v), \quad \text{for all } p \in \mathcal{E} \text{ and } u, v \in T_p\mathcal{E},$$

where $f_*\colon T_p\mathcal{E} \to T_{f(p)}\mathcal{E}'$ is the pushforward of f. And 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 any such mapping 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 [18] 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 [35] characterized a class of invariant metrics of positive matrices that restrict to products of Fisher metrics over stochastic matrices.[2] The maps considered by Lebanon do not map stochastic matrices to stochastic matrices, which motivated [46] 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 [7, 61].

Example 6 (Hessian geometries)**.** The following Riemannian geometries are induced by strictly convex functions.

1. *Euclidean geometry:* The Euclidean geometry on $\mathbb{R}^d$ is induced by the squared Euclidean norm $x \mapsto \sum_i x_i^2$.
2. *Fisher geometry:* The Fisher metric on $\mathbb{R}^d_{>0}$ is induced by the negative entropy $x \mapsto \sum_i x_i \log(x_i)$.
3. *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).

[2] 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 $g(u_1+u_2, v_1+v_2) = g_1(u_1,v_1) + g_2(u_2,v_2)$.

4. *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

$$\phi_\sigma(x) \coloneqq \begin{cases} \sum_i x_i \log(x_i) & \text{if } \sigma = 1 \\ -\sum_i \log(x_i) & \text{if } \sigma = 2 \\ \frac{1}{(2-\sigma)(1-\sigma)} \sum x_i^{2-\sigma} & \text{otherwise,} \end{cases} \tag{5}$$

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

$$g^\sigma_x(v, w) = \sum_i \frac{v_i w_i}{x_i^\sigma}. \tag{6}$$

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 [7], 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.

5. *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 $\phi_C(\mu) \coloneqq H(\mu|\mu_X) = H(\mu) - H(\mu_X)$ from (1). This is a convex function on the simplex $\Delta_{\mathcal{X}\times\mathcal{Y}}$ [53]. The Hessian of the conditional entropy is given by

$$\partial_{(s,a)}\partial_{(s',a')}\phi_C(\mu) = \delta_{xx'}\left(\delta_{yy'}\mu(x,y)^{-1} - \mu_X(x)^{-1}\right), \tag{7}$$

as can be verified by explicit computation or the chain rule for Hessian matrices (see also proof of Theorem 11). This Hessian does not induce a Riemannian geometry on the entire simplex since is not positive definite on the tangent space $T\Delta_{\mathcal{X}\times\mathcal{Y}}$, as can be seen by considering 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 $\text{int}(P) \to \text{int}(\Delta_{\mathcal{Y}}^{\mathcal{X}}), \mu \mapsto \mu_{Y|X}$. It can be shown by explicit computation (analogous to the proof of Theorem 11) that the Hessian $\nabla^2\phi_C(\mu)$ is the metric tensor of the pull back of the Riemmanian metric

$$g\colon T\Delta_{\mathcal{Y}}^{\mathcal{X}} \times T\Delta_{\mathcal{Y}}^{\mathcal{X}} \to \mathbb{R}, \quad g_{\mu(\cdot|\cdot)}(v, w) \coloneqq \sum_x \nu(x) \sum_y \frac{v(x,y)w(x,y)}{\mu(y|x)}.$$

This argument can be adapted to sets $\{\mu \in \Delta_{\mathcal{X}\times\mathcal{X}} : \mu_X = \nu(\mu_{Y|X})\}$, where $\nu\colon \text{int}(\Delta_{\mathcal{Y}}^{\mathcal{X}}) \to \text{int}(\Delta_{\mathcal{X}})$ depends smoothly on the conditional $\mu_{Y|X} \in \Delta_{\mathcal{Y}}^{\mathcal{X}}$.

We note that the Bregman divergence induced by the conditional entropy is the conditional relative entropy [53],

$$\begin{aligned} D_{\phi_C}(\mu^{(1)}, \mu^{(2)}) &= D_{KL}(\mu^{(1)}, \mu^{(2)}) - D_{KL}(\mu^{(1)}_X, \mu^{(2)}_X) \\ &= \sum_x \mu^{(1)}_X(x) D_{KL}(\mu^{(1)}(\cdot|x), \mu^{(2)}(\cdot|x)). \end{aligned}$$

Local Hessian of Bregman divergences Let ϕ be a twice differentiable strictly 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

$$\nabla_y^2 D_\phi(x,y)|_{y=x} = \nabla_y^2 D_\phi(y,x)|_{y=x} = \nabla^2\phi(x). \tag{8}$$

To see this, we set $f(y) \coloneqq D_\phi(x,y)$. Then it is straight forward to see 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 Gauss-Newton method 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 Gauss-Newton preconditioner of the objective function $\phi \circ P$ [41]. 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 preconditioner.

4 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 [68, 32, 65, 40, 14] offer a flexible approach to reward optimization. They have been used in robotics [56] and have been combined with deep neural networks [62, 63, 60]. 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 [49]. 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 [29]. 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. [47], is the state-action space $\mathcal{N} \subseteq \Delta_{\mathcal{S}\times\mathcal{A}}$, for which the objective becomes a rather simple, linear function. By Proposition 3 the two model spaces $\Delta_{\mathcal{A}}^{\mathcal{S}}$ and $\mathcal{N}$ are diffeomorphic under mild assumptions, which allows us to study any NPG method defined with respect to the policy space in state-action space. 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.

4.1 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

$$\theta_{k+1} \coloneqq \theta_k + \Delta t \nabla R(\theta_k), \tag{9}$$

where $\Delta t > 0$ is the step size.

For $\gamma \in (0,1)$, the reward function $\pi \mapsto R(\pi)$ is a rational function. Hence, in principle it can be differentiated using any automatic differentiation method. One can use the celebrated policy gradient theorem and use matrix inversion to compute the parameter update.

Theorem 7 (Policy gradient theorem)**.** *Consider an MDP $(\mathcal{S}, \mathcal{A}, \alpha, r), \gamma \in [0,1)$ and a parametrized policy class. It holds that*

$$\partial_{\theta_i} R(\theta) = \sum_s \rho_\theta(s) \sum_a \partial_{\theta_i} \pi_\theta(a|s) Q^{\pi_\theta}(s,a) = \sum_{s,a} \eta_\theta(s,a) \partial_{\theta_i} \log(\pi_\theta(a|s)) Q^{\pi_\theta}(s,a),$$

where $Q^\pi \coloneqq (I - \gamma P_\pi)^{-1} r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is the state-action value function.

In a reinforcement learning setup, one does not have direct access to the transition α and hence to P_π (2) nor Q^π, and sometimes even $\mathcal{S}$ is not known a priori. In this case, one has to estimate the gradient from interactions with the environment [14, 13, 48, 64]. In this work, however, we study the planning problem in MDPs, i.e., we assume that we have access to exact gradient evaluations.

Policy parametrizations Many results on the convergence of policy gradient methods have been providede for *tabular softmax policies*. The tabular softmax parametrization is given by

$$\pi_\theta(a|s) \coloneqq \frac{e^{\theta_{sa}}}{\sum_{a'} e^{\theta_{sa'}}} \quad \text{for all } a \in \mathcal{A}, s \in \mathcal{S}, \tag{10}$$

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 8 (Regular policy parametrization)**.** We call a policy parametrization $\mathbb{R}^p \to \operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{S}}), \theta \mapsto \pi_\theta$ *regular* if it is differentiable and satisfies

$$\operatorname{span}\{\partial_{\theta_i} \pi_\theta : i = 1, \dots, p\} = T_{\pi_\theta} \Delta_{\mathcal{A}}^{\mathcal{S}} \quad \text{for every } \theta \in \mathbb{R}^p. \tag{11}$$

We will focus on regular policy parametrizations. Nonetheless, we observe that policy optimization with constrained search variables can also be an attractive option and refer to [51] for a discussion in context of POMDPs.

Regularization in MDPs In practice, the reward function is often regularized as $R_\lambda = R - \lambda\psi$. This is often motivated to encourage exploration [68] and has also been shown to lead to fast convergence for strictly convex regularizers ψ [44, 19]. One popular regularizer is the conditional entropy in state-action space, see [53, 44, 19],

$$\psi_C(\theta) = \sum_s \rho_\theta(s) \sum_a \pi_\theta(a|s) \log(\pi_\theta(a|s)) = H(\eta_\theta) - H(\rho_\theta), \tag{12}$$

which has also been used to successfully design trust region and proximal methods for reward optimization [58, 59]. It is also possible to take the functions ϕ_σ defined in (5) as regularizers. This includes the entropy function, which is studied in state-action space in [53] and logarithmic barriers, which are studied in policy space in [1]. Introducing a regularizer changes the optimization problem and usually also the optimizer. The difference introduced by this regularization can be estimated in terms of the regularization strength λ. For logarithmic barriers in state-action space, this follows from standard estimates for interior point methods [17]. For entropic regularization in state-action space, this is elaborated in [67], and for the conditional entropy this is done in [44, 19]. We will see later that several of these regularizers lead to Hessian geometries in state-action space that correspond to different natural gradients that have been proposed in the context of policy optimization.

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 $\{\pi_\theta \circ \beta : \theta \in \Theta\}$. 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 5 and Section 6 do not carry over to POMDPs since they assume tabular softmax (state) policies.

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

$$\pi_{k+1} \coloneqq \Pi_{\Delta_{\mathcal{A}}^{\mathcal{S}}}(\pi_k + \Delta t G(\pi_k)^{+} \nabla R(\pi)),$$

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 [1, 69] for convergence rates of these methods.

4.2 Kakade's natural policy gradient

Kakade [29] proposed 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:

$$\begin{aligned} G_K(\theta)_{ij} &= \sum_s \rho_\theta(s) \sum_a \pi_\theta(a|s) \partial_{\theta_i} \log(\pi_\theta(a|s)) \partial_{\theta_j} \log(\pi_\theta(a|s)) \\ &= \sum_{s,a} \eta_\theta(s,a) \partial_{\theta_i} \log(\pi_\theta(a|s)) \partial_{\theta_j} \log(\pi_\theta(a|s)) \\ &= \sum_s \rho_\theta(s) \sum_a \frac{\partial_{\theta_i} \pi_\theta(a|s) \partial_{\theta_j} \pi_\theta(a|s)}{\pi_\theta(a|s)}. \end{aligned} \tag{13}$$

Definition 9 (Kakade's NPG and geometry in policy space)**.** We refer to the natural gradient $\nabla^K R(\theta) \coloneqq G_K(\theta)^{+} \nabla_\theta R(\pi_\theta)$ as *Kakade's natural policy gradient (K-NPG)*, where G_K is defined

in (13). Hence, Kakade's NPG is the NPG induced by the factorization $\theta \mapsto \pi_\theta \mapsto R(\theta)$ and the Riemannian metric on $\mathrm{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$ given by

$$g_\pi^K(v,w) \coloneqq \sum_s \rho^\pi(s) \sum_a \frac{v(s,a)w(s,a)}{\pi(a|s)} \quad \text{for all } v, w \in T_\pi \Delta_{\mathcal{A}}^{\mathcal{S}}. \tag{14}$$

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 10. In [29] the definition of G_K was heuristically motivated by the fact that the reward is also a mix of instantaneous rewards according to the state frequencies, $R(\pi) = \sum_s \rho^\pi(s) \sum_a \pi(a|s) r(s,a)$. The invariance axiomatic approaches discussed in [35, 46] 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 [56, 12, 52] argued that the Gram matrix G_K corresponds to the limit of the Fisher information matrices of finite-path probability measures as the path length tends to infinity.

Interpration as Hessian geometry of conditional entropy regularization 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 [12, 56, 46]. 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 11 (Kakade's geometry as conditional entropy Hessian geometry)**.** *Consider an MDP $(\mathcal{S}, \mathcal{A}, \alpha)$ and fix $\mu \in \Delta_{\mathcal{S}}$ and $\gamma \in (0,1)$ such that Assumption 2 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 $\pi \mapsto \eta^\pi$. In particular, 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.,*

$$\begin{aligned} G_K(\theta)_{ij} &= \sum_{s,a} \frac{\partial_{\theta_i}\eta_\theta(s,a)\partial_{\theta_j}\eta_\theta(s,a)}{\eta_\theta(s,a)} - \sum_s \frac{\partial_{\theta_i}\rho_\theta(s)\partial_{\theta_j}\rho_\theta(s)}{\rho_\theta(s)} \\ &= \sum_{s,a} \partial_{\theta_i} \log(\eta_\theta(s,a))\partial_{\theta_j}\log(\eta_\theta(s,a))\eta_\theta(s,a) \\ &\quad - \sum_s \partial_{\theta_i}\log(\rho_\theta(s))\partial_{\theta_j}\log(\rho_\theta(s))\rho_\theta(s). \end{aligned} \tag{15}$$

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

$$\begin{aligned} G(\eta)_{(s,a),(s',a')} &= g_\pi^K(\partial_{(s,a)}\eta(\cdot|\cdot), \partial_{(s',a')}\eta(\cdot|\cdot)) \\ &= \sum_{\tilde{s},\tilde{a}} \rho(\tilde{s}) \frac{\partial_{(s,a)}\eta(\tilde{a}|\tilde{s})\partial_{(s',a')}\eta(\tilde{a}|\tilde{s})}{\eta(\tilde{a}|\tilde{s})} \\ &= \sum_{\tilde{s},\tilde{a}} \rho(\tilde{s})^2 \frac{\partial_{(s,a)}\eta(\tilde{a}|\tilde{s})\partial_{(s',a')}\eta(\tilde{a}|\tilde{s})}{\eta(\tilde{s},\tilde{a})}. \end{aligned} \tag{16}$$

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

$$G(\eta)_{(s,a),(s',a')} = \delta_{ss'}\left(\delta_{aa'}\eta(s,a)^{-1} - \rho(s)^{-1}\right). \tag{17}$$

We aim to show that $G(\eta) = \nabla^2\phi_C(\eta)$, where $\phi_C(\eta) = H(\eta) - H(\rho)$, where $\rho(s) = \sum_a \eta(s,a)$ denotes the state-marginal. Note that $\nabla^2 H(\eta) = \mathrm{diag}(\eta)$, which is the first term appearing in (17). 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 (17). Overall this implies $G(\eta) = \nabla^2\phi_C(\eta)$. □

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 [53].

Remark 12. Kakade's NPG is known to converge at a locally quadratic rate under conditional entropy regularization [19], 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 (14) is not the Hessian geometry induced by ψ in policy space, which would take the form

$$\begin{aligned}\nabla^2\psi(\pi) = &\sum_s \rho^\pi(s)\nabla^2 H(\pi(\cdot|s)) + \sum_s (\nabla H(\cdot|s)^\top \nabla\rho^\pi(s) + \nabla H(\cdot|s)\nabla\rho^\pi(s)^\top) \\ &+ \sum_s H(\pi(\cdot|s))\nabla^2\rho^\pi(s).\end{aligned}$$

Instead, the metric proposed by Kakade only considers the contribution of the first term, see (14). As we will see in Sections 5 and 6, 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.

4.3 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 [47] 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

$$G_M(\theta)_{ij} = \sum_{s,a} \partial_{\theta_i}\log(\eta_\theta(s,a))\partial_{\theta_j}\log(\eta_\theta(s,a))\eta_\theta(s,a). \tag{18}$$

Definition 13 (Morimura's NPG)**.** We refer to the natural gradient $\nabla^M R(\theta) := G_M(\theta)^+\nabla_\theta R(\pi_\theta)$ as *Morimura's natural policy gradient (M-NPG)*, where G_M is defined in (18). 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 (15) 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 [47]. 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 [7]. Hence, we can interpret the Fisher metric as the Hessian geometry of the entropy regularized reward $\eta \mapsto \langle r, \eta\rangle - H(\eta)$.

4.4 General Hessian natural policy gradient

Generalizing the above definitions, we define general state-action space Hessian NPGs as follows. Consider a twice differentiable function $\phi\colon \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0} \to \mathbb{R}$ 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

$$G_\phi(\theta)_{ij} \coloneqq \sum_{s,s',a,a'} \partial_{\theta_i}\eta_\theta(s,a)\partial_{(s,a)}\partial_{(s',a')}\phi(\eta_\theta)\partial_{\theta_j}\eta_\theta(s',a'),$$

which is the Gram matrix with respect to the Hessian geometry in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$.

Definition 14 (Hessian NPG)**.** We refer to the natural gradient $\nabla^\phi R(\theta) \coloneqq G_\phi(\theta)^+\nabla_\theta R(\pi_\theta)$ 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 (5). 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 15 (σ-NPG)**.** We refer to the natural gradient $\nabla^\sigma R(\theta) \coloneqq G_\sigma(\theta)^+\nabla_\theta R(\pi_\theta)$ 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 (6).

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.

5 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 tabular softmax policies. Although we focus on this case, we observe that our results directly extend to regular parametrizations of the interior of the policy polytope $\Delta^{\mathcal{S}}_{\mathcal{A}}$. Leveraging tools from the theory of gradient flows in Hessian geometries established in [4] 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 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 in order to study natural policy gradient flows for tabular softmax policies.

Proposition 16 (Evolution in state-action space)**.** *Consider an MDP $(\mathcal{S}, \mathcal{A}, \alpha)$, a Riemannian metric g on $\operatorname{int}(\mathcal{N}) = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$ and an differentiable objective function $\mathfrak{R}\colon \operatorname{int}(\Delta_{\mathcal{S}\times\mathcal{A}}) \to \mathbb{R}$. Consider a regular policy parametrization and the objective $R(\theta) \coloneqq \mathfrak{R}(\eta_\theta)$ and a solution $\theta\colon [0,T] \to \Theta = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ of the natural policy gradient flow*

$$\partial_t \theta(t) = \nabla^N R(\theta(t)) = G(\theta(t))^+ \nabla R(\theta(t)), \tag{19}$$

where $G(\theta)_{ij} = g_\eta(\partial_{\theta_i}\eta_\theta, \partial_{\theta_j}\eta_\theta)$ and $G(\theta)^+$ denotes some pseudo inverse of $G(\theta)$. Then, setting $\eta(t) \coloneqq \eta_{\theta(t)}$ we have that $\eta\colon [0,T] \to \Delta_{\mathcal{S}\times\mathcal{A}}$ is the gradient flow with respect to the metric $g|_{\mathcal{N}}$ and the objective $\mathfrak{R}$, i.e., solves

$$\partial_t \eta(t) = \nabla^{g|_{\mathcal{N}}} \mathfrak{R}(\eta(t)) = \Pi_{T\mathcal{L}}(\nabla^g \mathfrak{R}(\eta(t))), \tag{20}$$

where $\Pi^g_{T\mathcal{L}}$ is the g-orthogonal projection onto the tangent space $T\mathcal{L}$ with $\mathcal{L}$ defined in (4)*.*

Proof. This is a direct consequence of Theorem 5. □

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.

By Proposition 16 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 [29] and Morimura et al. [47]. We study the evolution of these flows in state-action space and leverage results on Hessian gradient flows of convex problems in [4] to obtain global convergence rates for different NPG methods.

5.1 Convergence of unregularized Hessian natural policy gradient flows

First, we study the case of unregularized reward, i.e., where the state-action objective is linear and given by $\mathfrak{R}(\eta) = \langle r, \eta\rangle$. In this case we obtain global convergence guarantees including rates. In particular, our general result covers the σ-NPGs and thus Morimura's NPGs as well as Kakade's NPGs. For the remainder of this section we work under the following assumptions.

Setting 17. Let $(\mathcal{S}, \mathcal{A}, \alpha)$ be an MDP, $\mu \in \Delta_{\mathcal{S}}$ and $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and let the positivity Assumption 2 hold. We denote the state-action polytope by $\mathcal{N} = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geq 0} \cap \mathcal{L}$, see Proposition 1, and its (relative) interior and boundary by $\mathrm{int}(\mathcal{N}) = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0} \cap \mathcal{L}$ and $\partial\mathcal{N} = \partial\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geq 0} \cap \mathcal{L}$ respectively. We consider an objective function $\mathfrak{R}\colon \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R} \cup \{-\infty\}$ that is finite, differentiable and concave on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$ and continuous on its domain $\mathrm{dom}(\mathfrak{R}) = \{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \mathfrak{R}(\eta) \in \mathbb{R}\}$. Further, 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 $\eta \in \mathrm{int}(\mathcal{N})$. Further, with $\eta\colon [0,T) \to \mathcal{N}$ we denote a solution of the Hessian gradient flow

$$\partial_t \eta(t) = \Pi_{T\mathcal{L}}(\nabla^2\phi(\eta(t))^{-1}\nabla\mathfrak{R}(\eta(t))), \tag{21}$$

which is the gradient flow with respect to the Hessian geometry induced by ϕ on $\mathcal{N}$. We denote[3] $R^* \coloneqq \sup_{\eta\in\mathcal{N}} \mathfrak{R}(\eta) < \infty$ 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 η_0 and η^* by π_0 and π^*, see Proposition 3.

We observe that the Hessian of the conditional entropy only defines a Riemannian metric on $\mathrm{int}(\mathcal{N})$, even if not over all of $\Delta_{\mathcal{S}\times\mathcal{A}}$. Note that in general η^* might lie on the boundary and for linear $\mathfrak{R}$ corresponding to unregularized reward it necessarily lies on the boundary.

Sublinear rates for general case We begin by providing a sublinear rate of convergence for general NPG flows, which we the specialize to Kakade and σ-NPGs.

Lemma 18 (Convergence of Hessian natural policy gradient flows)**.** *Consider Setting 17 and assume that there exists a solution $\eta\colon [0,T) \to \mathrm{int}(\mathcal{N})$ of the NPG flow (21) with initial condition $\eta(0) = \eta_0$. Then for any $\eta' \in \mathcal{N}$ and $t \in [0,T)$ it holds that*

$$\mathfrak{R}(\eta') - \mathfrak{R}(\eta(t)) \leq D_\phi(\eta', \eta_0)t^{-1}, \tag{22}$$

where D_ϕ denotes the Bregman divergence of ϕ. In particular it holds that $\mathfrak{R}(\eta(t)) \to R^*$ as $T \to \infty$. Further, this convergence happens at a rate $O(t^{-1})$ if there is a maximizer $\eta^* \in \mathcal{N}$ of $\mathfrak{R}$ with $\phi(\eta^*) < \infty$.

Proof. This is precisely the statement of Proposition 4.4 in [4]; 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. Note that

$$\partial_t D_\phi(\eta, \eta(t)) = \langle \nabla\mathfrak{R}(\eta(t)), \eta - \eta(t)\rangle \leq \mathfrak{R}(\eta(t)) - \mathfrak{R}(\eta), \tag{23}$$

which can either be seen by inspecting the proof of equation (4.4) in [4] and noting that the proof does not require the stronger assumption made there or by explicit computation. Integration and the monotonicity of $t \mapsto \mathfrak{R}(\eta(t))$ yields the claim. □

The previous result is very general 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 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.

[3] Note that $\mathfrak{R}$ is bounded over the bounded set $\mathcal{N}$ as a concave function.

Definition 19 (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:

1. *Domain:* It holds that $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0} \subseteq \mathrm{dom}(\phi) \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geq 0}$, where $\mathrm{dom}(\phi) = \{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \phi(\eta) < \infty\}$.

2. *Smoothness and convexity:* We assume ϕ to be continuous on $\mathrm{dom}(\phi)$ and twice continuous differentiable on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$ and such that $\nabla^2\phi(\eta)$ is positiv definite on $T_\eta\mathcal{N} = T\mathcal{L} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ for every $\eta \in \mathrm{int}(\mathcal{N})$.

3. *Gradient blowup at boundary:* For any $(\eta_k) \subseteq \mathrm{int}(\mathcal{N})$ 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 [4], 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 $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geq 0}$ 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 20. 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:

1. The functions ϕ_σ defined in (5) 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 $\sigma = 0$.

2. The conditional entropy ϕ_C defined in (12) 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 $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geq 0}$.

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.

Let us now turn towards Kakade's natural policy gradient, which is the Hessian NPG induced by the conditional entropy ϕ_C defined in (1). The Bregman divergence of the conditional entropy (see [57]) is given by

$$\begin{aligned} D_{\phi_C}(\eta_1,\eta_2) &= \sum_{s,a}\eta_1(s,a)\log\left(\frac{\eta_1(s,a)}{\eta_2(s,a)}\right) - \sum_{s,a}\eta_1(s,a)\log\left(\frac{\sum_{a'}\eta_1(s,a')}{\sum_{a'}\eta_2(s,a')}\right) \\ &= D_{KL}(\eta_1,\eta_2) - D_{KL}(\rho_1,\rho_2) = \sum_s \rho_1(s) D_{KL}(\eta_1(\cdot|s),\eta_2(\cdot|s)), \end{aligned}$$

which has been studied in the context of mirror descent algorithms of the linear programming formulation of MDPs in [53].

Theorem 21 (Convergence of Kakade's NPG flow for unregularized reward)**.** *Consider Setting 17 with $\phi = \phi_C$ being the conditional entropy defined in* (12) *and let $\mathfrak{R}(\eta) = \langle r,\eta\rangle$ denote the unregularized reward and fix an element $\eta_0 \in \mathrm{int}(\mathcal{N})$. Then there exists a unique global solution $\eta\colon [0,\infty) \to \mathrm{int}(\mathcal{N})$ of Kakade's NPG flow with initial condition $\eta(0) = \eta_0$, i.e., of* (21) *with $\phi = \phi_C$, and it holds that*

$$R^* - \mathfrak{R}(\eta(t)) \leq t^{-1} D_{\phi_C}(\eta^*,\eta_0) = t^{-1}\sum_s \rho^*(s) D_{KL}(\pi^*(\cdot|s),\pi_0(\cdot|s)),$$

where D_{ϕ_C} denotes the conditional relative entropy. In particular, we have $\operatorname{dist}(\eta(t), S) \in O(t^{-1})$*, where* $S = \{\eta \in \mathcal{N} : \langle r, \eta\rangle = R^*\}$ *denotes the solution set and* dist *denotes the Euclidean distance.*

Proof. The well posedness follows by a similar reasoning as in [4, Theorem 4.1]. Now the result follows directly from Lemma 18. □

Now we elaborate the consequences of the general convergence result Lemma 18 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 22 (Convergence of σ-NPG flow for unregularized reward)**.** *Consider Setting 17 with $\phi = \phi_\sigma$ for some $\sigma \in [1, \infty)$ being defined in (5). 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 (21) with inital condition $\eta(0) = \eta_0$ and it holds that $R^* - \mathfrak{R}(\eta(t)) = O(f_\sigma(t))$ as $t \to \infty$, where*

$$f_\sigma(t) \coloneqq \begin{cases} t^{-1} & \text{for } \sigma \in [1, 2) \\ \log(t)t^{-1} & \text{for } \sigma = 2 \\ t^{\sigma-3} & \text{for } \sigma \in (2, \infty). \end{cases}$$

In particular, we have $\operatorname{dist}(\eta(t), S) \in O(f_\sigma(t))$*, 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. By the preceding Lemma 18 it suffices to show the well posedness of the σ-NPG flow. The result [4, Theorem 4.1] guarantees the well posedness for Hessian gradient flows for smooth Legendre type functions. Note however that they work with slightly stronger assumptions, which are that the gradient blowup of the Legendre type functions occurs not only on the boundary of $\mathcal{N}$ but on the boundary of $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geq 0}$ and that the objective $\mathfrak{R}$ is globally defined. Consolidating the proof in [4] reveals that both of these relaxations do not change the validity or proof of the statement.

It is easy to see that for $\sigma \geq 1$ the functions ϕ_σ are of Legendre type and smooth and hence we can apply the preceding Lemma 18. Let η^* be a maximizer, which necessarily lies at the boundary of $\mathcal{N}$ (except for constant reward) and therefore has at least one zero entry. For $\sigma \in [1, 2)$ we have that $\phi_\sigma(\eta^*) < \infty$ and hence we obtain $R^* - \mathfrak{R}(\eta(t)) \in D_{\phi_\sigma}(\eta^*, \eta_0)t^{-1}$. 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

$$\begin{aligned} R^* - \mathfrak{R}(\eta(t)) &= R^* - \langle r, \eta_\delta\rangle + \langle r, \eta_\delta\rangle - \mathfrak{R}(\eta(t)) = O(\delta) + D_{\phi_\sigma}(\eta_\delta, \eta_0)t^{-1} \\ &= O(\delta) + \left(\phi_\sigma(\eta_\delta) - \phi_\sigma(\eta_0) - \langle \nabla\phi_\sigma(\eta_0), \eta_\delta - \eta_0\rangle\right) t^{-1} \\ &= O(\delta) + O(\log(\delta) + 1)t^{-1}. \end{aligned}$$

Setting $\delta = t^{-1}$ we obtain $R^* - \mathfrak{R}(\eta(t)) = O(t^{-1}) + O((\log(t^{-1}) + 1)t^{-1}) = O(\log(t)t^{-1})$ for $t \to \infty$. For $\sigma \in (2, \infty)$ the calculation follows in analogue fashion. Noting that $\operatorname{dist}(\eta(t), S) \sim R^* - \mathfrak{R}(\eta(t))$ finishes the proof. □

Remark 23. Theorem 22 and Theorem 21 show global convergence of σ-NPG and Kakade's NPG flows to a maximizer of the unregularized problem. 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 3. For $\sigma \geq 3$ Theorem 22 is uninformative since $\mathfrak{R}(\eta(t))$ is non increasing. However, in our experiments we observed that (discretizations of) σ-NPG flows still converge for $\sigma \geq 3$, although the plateau problem becomes more pronounced, as can be seen in Figure 3.

Furthermore, 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 [4].

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 24 (Convergence rates for gradient flow trajectories)**.** *Consider Setting 17 and assume that there is a global solution $\eta\colon [0,\infty) \to \operatorname{int}(\mathcal{N})$ of the Hessian gradient flow (21). Assume that there is $\eta^* \in \mathcal{N}$ such that $\phi(\eta^*) < +\infty$ as well as a neighborhood N of η^* in $\mathcal{N}$ and $\omega \in (0,\infty)$ and $\tau \in [1,\infty)$ such that*

$$\mathfrak{R}(\eta^*) - \mathfrak{R}(\eta) \geq \omega D_\phi(\eta^*, \eta)^\tau \quad \text{for all } \eta \in N. \tag{24}$$

Then there is a constant $c > 0$ such that

1. *if $\tau = 1$, then $D_\phi(\eta^*, \eta(t)) \leq ce^{-\omega t}$,*
2. *if $\tau > 1$, then $D_\phi(\eta^*, \eta(t)) \leq ct^{-1/(\tau-1)}$.*

The lower bound (24) 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 state-action trajectories in the sense of the Bregman divergence.

Proof. The statement of this result can be found in [4, Proposition 4.9], where however stronger assumptions are made and hence we provide a short proof. First, note that our assumption implies that η^* is the unique global maximizer of $\mathfrak{R}$ over $\mathcal{N}$. By (23) it holds that $u(t) \coloneqq \partial_t D_\phi(\eta^*, \eta(t))$ is strictly decreasing as long as $\eta(t) \neq \eta^*$. Note that if $\eta(t) = \eta^*$ for some $t \in [0,\infty)$, we have $u(t') = 0$ for all $t' \geq t$ and hence the statement becomes trivial. Therefore, we can assume $u(t) > 0$ for all $t > 0$. By Lemma 18 it holds that $\mathfrak{R}(\eta(t)) \to \mathfrak{R}(\eta^*)$ and hence $\eta(t) \to \eta^*$; this is due to the compactness of $\mathcal{N}$ and because the continuity of $\mathfrak{R}$ implies that every accumulation point of $\eta(t)$ is a maximizer and thus equal to η^*. Hence, $\eta(t) \in N$ for $t \geq t_0$. For the statement about the asymptotic behavior we may therefore assume without loss of generality that $\eta(t) \in N$ for all $t \geq 0$. Combining (23) and (24) we obtain $u'(t) \leq -\omega u(t)^\tau$. Dividing by the right hand side and integrating the inequality we obtain $u(t) \leq u(0)e^{-\omega t}$ for $\tau = 1$ and $u(t) \leq \omega^{1/(1-\tau)}(\tau-1)^{1/(1-\tau)}t^{1/(1-\tau)}$. □

Theorem 25 (Linear convergence of unregularized Kakade's NPG flow)**.** *Consider Setting 17, where $\phi = \phi_C$ is the conditional entropy defined in (12) and assume that there is a unique maximizer η^* of the unregularized reward $\mathfrak{R}$. Then $R^* - \mathfrak{R}(\eta(t)) = O(e^{-ct})$ for some $c > 0$.*

Proof. Let ϕ_C denote the conditional entropy, so that $D_{\phi_C}(\eta^*, \eta) = D_{KL}(\eta^*, \eta) - D_{KL}(\rho^*, \rho) \leq D_{KL}(\eta^*, \eta)$. Hence, we obtain just like in the case of σ-NPG flows for $\sigma = 1$ that $D_{\phi_C}(\eta^*, \eta) = O(\mathfrak{R}(\eta^*) - \mathfrak{R}(\eta))$ for $\eta \to \eta^*$ and hence $D_{\phi_C}(\eta^*, \eta(t)) = O(e^{-ct})$ for some $c > 0$ by Lemma 24. Hence, it remains to estimate $\mathfrak{R}(\eta^*) - \mathfrak{R}(\eta) = O(\|\eta^* - \eta\|_1)$ 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

$$\begin{aligned} D_{\phi_C}(\eta^*, \eta) &= \sum_s \rho^*(s) D_{KL}(\pi^*(\cdot|s), \pi^*(\cdot|s)) = -\sum_s \rho^*(s) \log(\pi(a_s^*|s)) \\ &\geq \sum_s \rho^*(s)(1 - \pi(a_s^*|s)) = 2^{-1} \sum_s \rho^*(s) \|\pi^*(\cdot|s) - \pi(\cdot|s)\|_1 \\ &\geq 2^{-1} \left(\min_s \rho^*(s) \right) \cdot \|\pi^* - \pi\|_1. \end{aligned}$$

Here, we have used $\log(t) \leq t - 1$ as well as

$$\begin{aligned} \|\pi^*(\cdot|s) - \pi(\cdot|s)\|_1 &= \sum_{a \neq a_s^*} |\pi^*(a|s) - \pi(a|s)| + |\pi^*(a|s) - \pi(a|s)| \\ &= \sum_{a \neq a_s^*} \pi(a|s) + (1 - \pi(a_s^*|s)) = 2(1 - \pi(a_s^*|s)). \end{aligned}$$

Now we observe that the mapping $\pi \mapsto \eta$ is L-Lipschitz with constant $L = O((1-\gamma)^{-1})$. The fact that $L = O((1-\gamma)^{-1})$ follows from the policy gradient theorem as $\partial_{\pi(a|s)} \eta^\pi = \rho^\pi(s)(I - \gamma P_\pi^\top)^{-1} e_{(s,a)}$, see also [50, Proposition 48]. In turn, it holds that

$$\|\eta^* - \eta(t)\|_1 \leq L \|\pi^* - \pi(t)\|_1 \leq \frac{2L}{\min_s \rho^*(s)} \cdot D_{\phi_C}(\eta^*, \eta(t)) = O(e^{-ct}).$$

Altogether this implies $R^* - \mathfrak{R}(\eta(t)) = O(e^{-ct})$, which concludes the proof. The O notation hides constants that scale with the norm of the instantaneous reward vector r, inversely with the minimum state probability, and inversely with $(1-\gamma)$ where γ is the discount rate. □

Theorem 26 (Improved convergence rates for σ-NPG flow)**.** *Consider Setting 17, where $\phi = \phi_\sigma$ for some $\sigma \in [1, 2)$ as defined in (5), and assume that there is a unique maximizer η^* of the unregularized reward $\mathfrak{R}$. Then $R^* - \mathfrak{R}(\eta(t)) \in O(g_\sigma(t))$, where*

$$g_\sigma(t) = \begin{cases} e^{-ct} & \text{if } \sigma = 1 \\ t^{-1/(\sigma-1)} & \text{if } \sigma \in (1, 2), \end{cases}$$

for some $c > 0$, where $\eta \colon [0, \infty) \to \operatorname{int}(\mathcal{N})$ denotes the solution of the σ-NPG flow.

Proof. The key is to show that (24) holds for $\tau = (2-\sigma)^{-1} \geq 1$. To see that this holds, we first consider the case $\sigma \in (1, 2)$, where we obtain

$$D_\sigma(\eta^*, \eta) = \sum_{s,a} \frac{\eta^*(s,a)^{2-\sigma}}{(1-\sigma)(2-\sigma)} - \sum_{s,a} \frac{\eta(s,a)^{2-\sigma}}{(1-\sigma)(2-\sigma)} - \sum_{s,a} \frac{\eta(s,a)^{1-\sigma}(\eta^*(s,a) - \eta(s,a))}{1-\sigma}.$$

We can bound every 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 24 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 $\mathfrak{R}(\eta^*) - \mathfrak{R}(\eta) = O(\|\eta^* - \eta\|_1)$ 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 $x \to y$. Now, Jensen's inequality yields

$$\|\eta^* - \eta\|_1 = O(D_\sigma(\eta^*, \eta)^{1/(2-\sigma)}).$$

Overall, we obtain

$$\mathfrak{R}(\eta^*) - \mathfrak{R}(\eta(t)) = O(\|\eta^* - \eta(t)\|_1^{1/(2-\sigma)}) = O(t^{-1/(1-\sigma)}).$$

The case $\sigma = 1$ can be treated similarly, where one obtains $D_\sigma(\eta^*, \eta) = O(\|\eta^* - \eta\|) = O(\mathfrak{R}(\eta^*) - \mathfrak{R}(\eta))$ for $\eta \to \eta^*$. To relate the L^1-norm to the Bregman divergence one can employ Pinsker's inequality $\|\eta^* - \eta\|_1 \leq \sqrt{2D_{KL}(\eta^*, \eta)} = \sqrt{2D_\sigma(\eta^*, \eta)}$. □

Compared to Theorem 22 the above Theorem 26 improves the $O(t^{-1})$ rates for $\sigma \in [1, 2)$. Later, we conduct numerical experiments that indicate that the rates $O(t^{-1/(\sigma-1)})$ also hold for $\sigma \geq 2$ and are tight.

Numerical examples We use the following example proposed by Kakade [29] and which was also used in [12, 47]. 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 2.

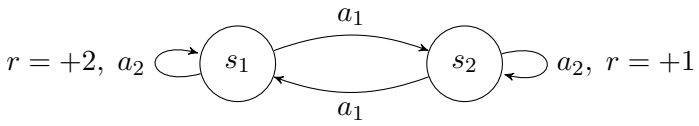

Figure 2: Transition graph and reward of the MDP example.

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 $\sigma \in \{-0.5, 0, 0.5, 1, 1.5, 2, 3, 4\}$ for 30 random (but the same for every method) initializations. We plot the trajectories in the state-action space (Figure 3) and in the policy polytope (Figure 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 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 $O(t^{-1})$ [44] and $O(t^{-1/(\sigma-1)})$ by Theorem 26. 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 Morimuras NPG we expect linear convergence by Theorem 25 and 26 respectively and hence use a semi-logarithmic plot.

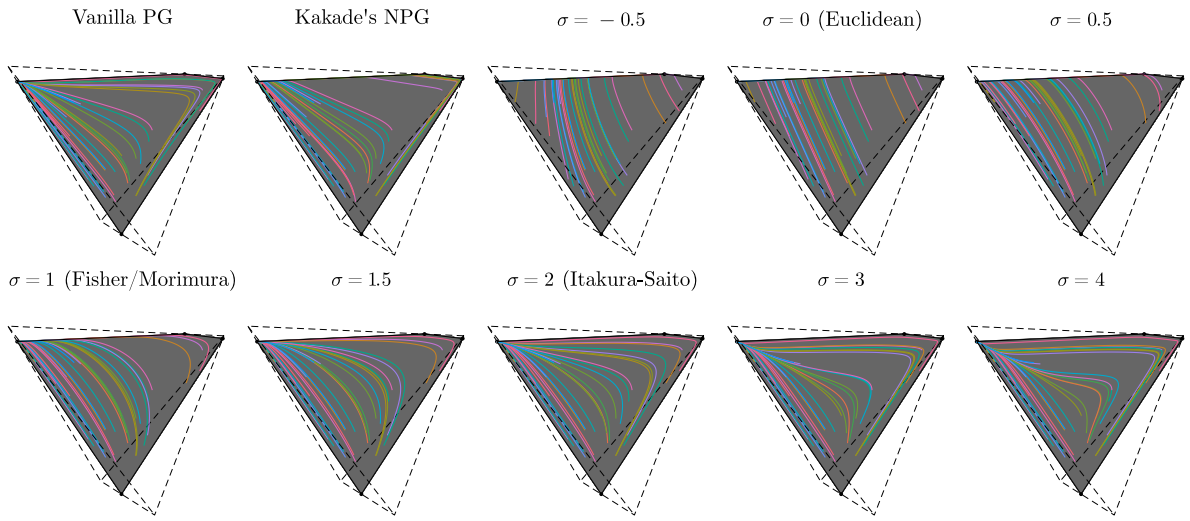

Figure 3: State-action trajectories for different PG methods, which are vanilla PG, Kakade's NPG and σ-NPG, where Morimura's NPG corresponds to $\sigma = 1$; the state-action polytope is shown in gray inside a three dimensional projection of the the simplex $\Delta_{\mathcal{S}\times\mathcal{A}}$; shown 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.

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 is the NPG flow is guaranteed to exhibit long time solutions. However, we observe finite-time convergence of the trajectories towards the global optimum (see Figure 5), which we suspect to be due to the discretization error.

For the other methods, namely vanilla PG, Kakade's NPG and σ-NPG with $\sigma \in [1, \infty)$, Theorem 22 and Theorem 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 3 and 4 respectively). When considering the convergence in objective value we observe that both Kakade's and Morimura's NPG exhibit a linear rate of convergence as asserted by Theorem 25 and Theorem 26, whereby Kakade's NPG appears to have more severe plateaus in some examples. 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 [44] and indicates that the rate $O(t^{-1/(\sigma-1)})$ shown for $\sigma \in (1, 2)$ is also valid in the regime $\sigma > 2$. 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.

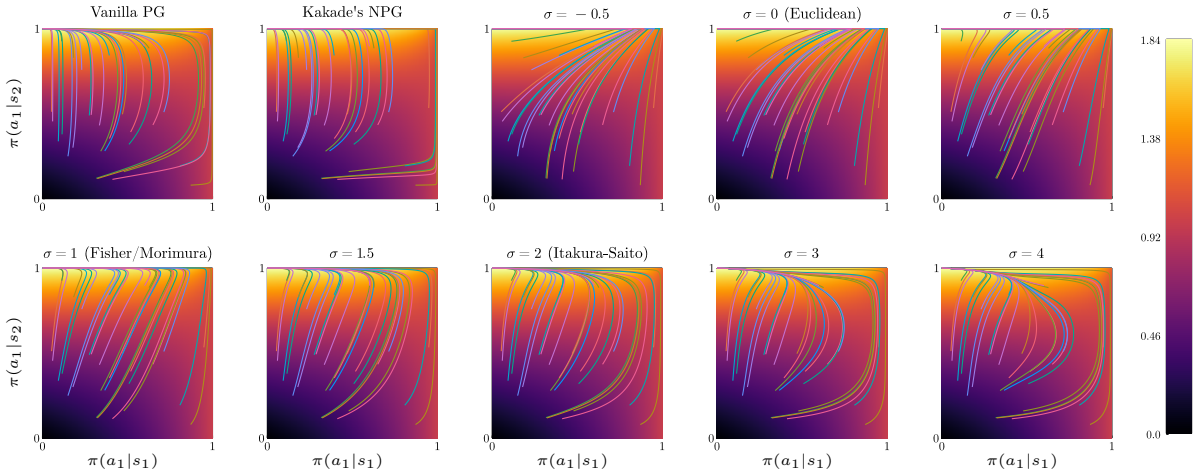

Figure 4: Plots of the trajectories of the individual methods inside the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S}} \cong [0,1]^2$; additionally, a heatmap of the reward function $\pi \mapsto R(\pi)$ is shown; the maximizer π^* is located at the upper left corner of the policy polytope.

5.2 Linear convergence of regularized Hessian natural policy gradient flows

It is known that strictly convex regularization in state-action space can yield linear convergence in reward optimization for vanilla and Kakade's natural policy gradients [44, 19]. Using Lemma 24 we generalize the result for Kakade's NPG and provide a result giving the linear convergence for general Hessian NPG.

Theorem 27 (Linear convergence for regularized problems)**.** *Consider Setting 17 and let ϕ be a Legendre type function and denote the regularized reward by $\mathfrak{R}_\lambda(\eta) = \langle r, \eta\rangle - \lambda\phi(\eta)$ 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 $\operatorname{int}(\mathcal{N})$. Assume that $\eta\colon [0,\infty) \to \operatorname{int}(\mathcal{N})$ solves the natural policy gradient flow with respect to the regularized reward $\mathfrak{R}_\lambda$ and the Hessian geometry induced by ϕ. For any $c \in (0,\lambda)$ there exists a constant $K > 0$ such that $D_\phi(\eta_\lambda^*, \eta(t)) \leq Ke^{-ct}$. In particular, for any $\kappa \in (\kappa_c, \infty)$ this implies $R_\lambda^* - \mathfrak{R}_\lambda(\eta(t)) \leq \kappa\lambda K e^{-ct}$, where κ_c denotes the condition number of $\nabla^2\phi(\eta^*)$.*

Proof. We first recall that by Lemma 18 it holds that $\mathfrak{R}(\eta(t)) \to \mathfrak{R}(\eta^*)$ and the uniqueness of the maximizer $\eta(t) \to \eta^* \in \operatorname{int}(\mathcal{N})$. By Lemma 24 it suffices to show that for any $\omega \in (0,1)$ it holds $\mathfrak{R}_\lambda(\eta^*) - \mathfrak{R}_\lambda(\eta) \geq \omega D_\phi(\eta^*, \eta)$ if η in a neighborhood of η^*. Note that

$$D_\phi(\eta^*, \eta) = \lambda^{-1} D_{\lambda\phi}(\eta^*, \eta) = D_{-\mathfrak{R}_\lambda}(\eta^*, \eta).$$

By Lemma 28 it follows that

$$\mathfrak{R}_\lambda(\eta^*) - \mathfrak{R}_\lambda(\eta) \geq \omega D_{-\mathfrak{R}_\lambda}(\eta^*, \eta) = \lambda\omega D_\phi(\eta^*, \eta), \tag{25}$$

which shows the linear convergence of the trajectory in the Bregman divergence. For arbitrary $m, M > 0$ such that $mI \prec \nabla^2\phi(\eta^*) \prec MI$ we can estimate

$$R_\lambda^* - \mathfrak{R}_\lambda(\eta(t)) = \mathfrak{R}_\lambda(\eta^*) - \mathfrak{R}_\lambda(\eta(t)) \leq \frac{\lambda M}{2} \cdot \|\eta^* - \eta(t)\|^2 \leq \frac{\lambda M}{m} \cdot D_\phi(\eta^*, \eta),$$

for $\eta(t)$ close to η^*, where we used that ϕ is m strongly convex in a neighborhood of η^*. □

In the proof of the previous theorem we used the following lemma.

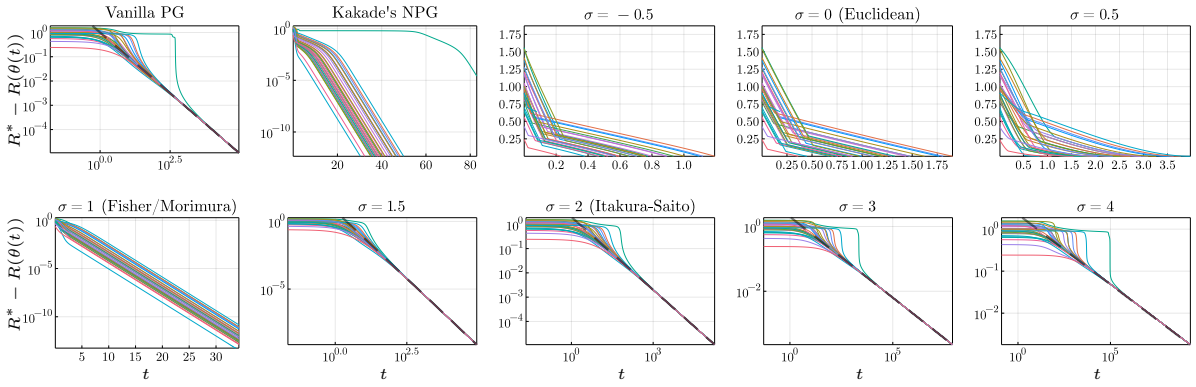

Figure 5: Plot of the optimality gaps $R^* - R(\theta(t))$ during optimization; note that for vanilla PG and $\sigma > 1$ these are log-log plots since we expect a decay like t^{-1} and $t^{-1/(\sigma-1)}$ respectively, which are shown as a dashed gray line; Kakade's and Morimura's NPG are at a log plot since we expect a linear convergence; finally, for $\sigma < 1$ we observe finite time convergence.

Lemma 28. *Let ϕ be a strictly convex function defined on an open convex set $\Omega \subseteq \mathbb{R}^d$ with unique minimizer η^*. Then for any $\omega \in (0,1)$ there is a neighborhood N_ω of x^* such that*

$$\phi(x) - \phi(x^*) \geq \omega D_\phi(x^*, x) \quad \text{for all } x \in N_\omega.$$

Proof. Set $f(x) \coloneqq D_\phi(x^*, x)$ and $g(x) \coloneqq D_\phi(x, x^*)$. It holds that $f(x^*) = g(x^*) = 0$ and since both functions are non-negative $\nabla f(x^*) = \nabla g(x^*) = 0$, which implies $g(x) = \phi(x) - \phi(x^*)$. By (8) 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

$$f(x) \leq \phi(x) - \phi(x^*) + \varepsilon \|x - x^*\|^2 \leq \left(1 + \frac{2\varepsilon}{m}\right)(\phi(x) - \phi(x^*))$$

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 $\omega \coloneqq (1 + 2\varepsilon m^{-1})^{-1}$ yields the claim. □

Remark 29 (Location of maximizers)**.** The condition that $\eta_\lambda^* \in \operatorname{int}(\mathcal{N})$ assumed in Theorem 27 is satisfied if the gradient blow-up condition from Definition 19 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 $\partial_v \mathfrak{R}_\lambda(\eta + t_0 v) > 0$. Then by the concavity of $\mathfrak{R}_\lambda$ and continuity of $\mathfrak{R}_\lambda$ we have

$$\mathfrak{R}_\lambda(\eta) \leq \mathfrak{R}_\lambda(\eta + t_0 v) - t_0 \partial_v \mathfrak{R}_\lambda(\eta + t_0 v) < \mathfrak{R}_\lambda(\eta + t_0 v),$$

and hence $\eta \neq \eta^*$.

Now we elaborate the consequences of this general convergence result given in Theorem 27 for Kakade and σ-NPG flows.

Corollary 30 (Linear convergence of regularized Kakade's NPG flow)**.** *Assume that* $\eta\colon [0,\infty) \to \operatorname{int}(\mathcal{N})$ *solves the natural policy gradient flow with respect to the regularized reward* $\mathfrak{R}_\lambda$ *and the Hessian geometry induced by* ϕ*. For any* $\omega \in (0,\lambda)$ *there exists a constant* $K > 0$ *such that* $D_\phi(\eta^*, \eta(t)) \leq Ke^{-\omega t}$*. In particular, for any* $\kappa \in (\kappa_c, \infty)$ *this implies* $R^*_\lambda - \mathfrak{R}_\lambda(\eta(t)) \leq \kappa K e^{-\omega t}$*, where* κ_c *denotes the condition number of* $\nabla^2\phi_C(\eta^*)$*.*

Proof. We want to use Remark 29. 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 2 it holds that $\rho(s) > 0$. Hence, if $\eta \in \partial\mathcal{N}$ we can take any $v \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ such that $\eta_\varepsilon \coloneqq \eta + \varepsilon v \in \operatorname{int}(\mathcal{N})$ 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 $\varepsilon \to 0$ since $\eta(s',a') = 0$ for some $s' \in \mathcal{S}, a' \in \mathcal{A}$ and $\rho_\varepsilon(s) \to \rho(s) > 0$ for all $s \in \mathcal{S}$. □

Corollary 31 (Linear convergence for regularized σ-NPG flow)**.** *Consider Setting 17 with* $\phi = \phi_\sigma$ *for some* $\sigma \in [1,\infty)$ *and denote the regularized reward by* $\mathfrak{R}_\lambda(\eta) = \langle r, \eta\rangle - \lambda\phi(\eta)$ *and fix an element* $\eta_0 \in \operatorname{int}(\mathcal{N})$*. Assume that* $\eta\colon [0,\infty) \to \operatorname{int}(\mathcal{N})$ *solves the natural policy gradient flow with respect to the regularized reward* $\mathfrak{R}_\lambda$ *and the Hessian geometry induced by* ϕ*. For any* $\omega \in (0,\lambda)$ *there exists a constant* $K > 0$ *such that* $D_\phi(\eta^*, \eta(t)) \leq Ke^{-\omega t}$*. In particular, for any* $\kappa \in (\kappa(\eta^*)^\sigma, \infty)$ *this implies* $R^*_\lambda - \mathfrak{R}_\lambda(\eta(t)) \leq \kappa K e^{-\omega t}$*, where* $\kappa(\eta^*) = \frac{\max \eta^*}{\min \eta^*}$

Proof. Again, we use Remark 29 it is straight forward to see that for the Legendre type functions ϕ_σ the unique maximizer η^* of $\mathfrak{R}_\lambda$ lies in the interior of $\mathcal{N}$. Hence, it remains to compute the condition number, for which we note that $\nabla^2\phi_\sigma(\eta^*) = \operatorname{diag}(\eta^*)^{-\sigma}$, which yields the result. □

Remark 32 (Extension to arbitrary regularizers)**.** The results above do not cover arbitrary combinations of Hessian geometries and regularizers. However, the proof of Theorem 27 can be adapted to this case, where the only part that requires adjustments is (25) 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.

6 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 [19, 37]. 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 thus formulate 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 penalization strength, which is reminiscent of Newton's method converging for step size 1.

Theorem 33 (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 on $T_\eta\mathcal{N} = T\mathcal{L} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ for every $\eta \in \operatorname{int}(\mathcal{N})$. Further, consider a regular policy parametrization and the regularized reward $R_\lambda(\theta) \coloneqq R(\theta) + \lambda\phi(\eta_\theta)$ 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, i.e.,*

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

with step size $\Delta t = \lambda$, where $G(\theta_k)^+$ denotes the Moore-Penrose inverse. Assume that $R_\lambda(\theta_k) \to R_\lambda^*$ for $k\to\infty$. Then $\theta_k \to \theta^*$ at a (locally) quadratic rate and hence $R_\lambda(\theta_k) \to R_\lambda^*$ at a (locally) quadratic rate.

The proof of this result relies on the following convergence result for inexact Newton methods.

Theorem 34 (Theorem 3.3 in [21])**.** *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 35 (Newton iteration in state-action space)**.** *The iterates of Newton's method in state-action space are given by*

$$\eta_{k+1} = \eta_k + \Pi^E_{T\mathcal{L}}(\nabla^2\mathfrak{R}_\lambda(\eta_k))^{-1}\Pi^E_{T\mathcal{L}}(\nabla\mathfrak{R}_\lambda(\eta_k)), \tag{26}$$

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 L defined in (4)*.*

Proof. The domain of the optimization problem is $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geq 0}\cap\mathcal{L}$ 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 $\iota\colon X \hookrightarrow L, x\mapsto x+\eta_0$ and set $f(x) \coloneqq \mathfrak{R}_\lambda(\iota x)$. 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

$$\begin{aligned}\eta_{k+1} - \eta_k &= \iota\nabla^2 f(x_k)^{-1}\nabla f(x_k) = \iota\iota^+\nabla^2\mathfrak{R}_\lambda(\eta_k)^{-1}(\iota^\top)^+\iota^\top\nabla\mathfrak{R}_\lambda(\eta_k)\\ &= \Pi^E_{T\mathcal{L}}(\nabla^2\mathfrak{R}_\lambda(\eta_k))^{-1}\Pi^E_{T\mathcal{L}}(\nabla\mathfrak{R}_\lambda(\eta_k)),\end{aligned}$$

where we used $AA^+ = \Pi_{\operatorname{range}(A)}$ and $(A^\top)^+A^\top = \Pi_{\ker(A^\top)} = \Pi_{\operatorname{range}(A)}$. □

Lemma 36. *Let (θ_k) be 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 $\eta_k \coloneqq \eta_{\theta_k}$ satisfy

$$\eta_{k+1} = \eta_k + \Delta t\cdot\Pi^E_{T\mathcal{L}}(\nabla^2\phi(\eta_k)^{-1}\Pi^E_{T\mathcal{L}}(\nabla\mathfrak{R}_\lambda(\eta_k))) + O(\Delta t^2\|G(\theta_k)^+\nabla R_\lambda(\theta_k)\|^2). \tag{27}$$

Proof. Writing P for the mapping $\theta \mapsto \eta_\theta$ and an application of Taylor's theorem implies that

$$\eta_{k+1} - \eta_k = \Delta t \cdot DP(\theta_k) G(\theta_k)^+ \nabla R_\lambda(\theta_k) + O(\Delta t^2 \|G(\theta_k)^+ \nabla R_\lambda(\theta_k)\|^2).$$

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 33. In our case, by the preceding two lemmata, we have

$$\|\varepsilon_k\| = O(\Delta t^2 \|G(\theta_k)^+ \nabla R_\lambda(\theta_k)\|^2) = O(\|\Pi_{T\mathcal{L}} \nabla \mathfrak{R}_\lambda(\eta_k)\|^2) = O(\|\nabla f(x_k)\|^2),$$

which proves the claim. □

Remark 37. 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 12). 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 (10) are overparametrized since in this case $\dim(\Theta) = |\mathcal{S}||\mathcal{A}|$.

7 Discussion

We provide a study of 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 et al., 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 regular policy parametrizations 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 and with step size equal to the 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.

Our convergence analysis currently does not cover the case of general parametric policy classes nor the case of partially observable MDPs, which we consider important future directions. Further, we study only the planning problem, i.e., assume to have access to exact gradients, and hence a combination of our study of NPG methods in state-action space with estimation problems would be a natural extension.

Acknowledgments This project has been supported by ERC Starting Grant 757983 and DFG SPP 2298 Grant 464109215. GM has been supported by NSF CAREER Award DMS-2145630. JM acknowledges support from the International Max Planck Research School for Mathematics in the Sciences (IMPRS MiS) and the Evangelisches Studienwerk Villigst e.V..

Conflict of interest statement There is no conflict of interest.

Data availability statement A repository with computer code to reproduce the experiments will be made available.

References

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[2] Carlo Alfano and Patrick Rebeschini. Dimension-Free Rates for Natural Policy Gradient in Multi-Agent Reinforcement Learning. *arXiv:2109.11692*, 2021.

[3] Carlo Alfano and Patrick Rebeschini. Linear convergence for natural policy gradient with log-linear policy parametrization. *arXiv preprint arXiv:2209.15382*, 2022.

[4] Felipe Alvarez, Jérôme Bolte, and Olivier Brahic. Hessian Riemannian gradient flows in convex programming. *SIAM journal on control and optimization*, 43(2):477–501, 2004.

[5] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[6] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, Japan, 2016.

[7] Shun-ichi Amari and Andrzej Cichocki. Information geometry of divergence functions. *Bulletin of the polish academy of sciences. Technical sciences*, 58(1):183–195, 2010.

[8] 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)*, volume 2, pages 1213–1216. IEEE, 1998.

[9] Michael Arbel, Arthur Gretton, Wuchen Li, and G Montúfar. Kernelized Wasserstein natural gradient. In *International Conference on Learning Representations*, 2020.

[10] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information geometry*, volume 64. Springer, Cham, 2017.

[11] Kamyar Azizzadenesheli, Yisong Yue, and Animashree Anandkumar. Policy Gradient in Partially Observable Environments: Approximation and Convergence. *arXiv:1810.07900*, 2018.

[12] J. Andrew Bagnell and Jeff G. Schneider. Covariant policy search. In *IJCAI*, pages 1019–1024, 2003.

[13] Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

[14] Jonathan Baxter, Peter L Bartlett, et al. Reinforcement learning in POMDPs via direct gradient ascent. In *ICML*, pages 41–48. Citeseer, 2000.

[15] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv:1906.01786*, 2019.

[16] Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite MDPs. In *International Conference on Artificial Intelligence and Statistics*, pages 2386–2394. PMLR, 2021.

[17] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, Cambridge, 2004.

[18] L. Campbell. An extended Čencov characterization of the information metric. *Proceedings of the American Mathematical Society*, 98:135–141, 1986.

[19] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021.

[20] N. N. Čencov. *Statistical decision rules and optimal inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, R.I., 1982. Translation from the Russian edited by Lev J. Leifman.

[21] Ron S Dembo, Stanley C Eisenstat, and Trond Steihaug. Inexact Newton methods. *SIAM Journal on Numerical analysis*, 19(2):400–408, 1982.

[22] Cyrus Derman. *Finite state Markovian decision processes*. Academic Press, New York, 1970.

[23] Guillaume Desjardins, Karen Simonyan, Razvan Pascanu, et al. Natural neural networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[24] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained Markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

[25] Dongsheng Ding, Kaiqing Zhang, Jiali Duan, Tamer Başar, and Mihailo R Jovanović. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained MDPs. *arXiv:2206.02346*, 2022.

[26] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

[27] Feihu Huang, Shangqian Gao, and Heng Huang. Bregman gradient policy optimization. In *International Conference on Learning Representations*, 2022.

[28] Mohammad Rasool Izadi, Yihao Fang, Robert Stevenson, and Lizhen Lin. Optimization of graph neural networks with natural gradient descent. In *2020 IEEE international conference on big data (big data)*, pages 171–179. IEEE, 2020.

[29] Sham M Kakade. A natural policy gradient. *Advances in Neural Information Processing Systems*, 14, 2001.

[30] Lodewijk CM Kallenberg. Survey of linear programming for standard and nonstandard Markovian control problems. Part I: Theory. *Zeitschrift für Operations Research*, 40(1):1–42, 1994.

[31] Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3794–3799. IEEE, 2021.

[32] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in Neural Information Processing Systems*, 12, 1999.

[33] Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pages 1–48, 2022.

[34] James-Michael Leahy, Bekzhan Kerimkulov, David Siska, and Lukasz Szpruch. Convergence of policy gradient for entropy regularized MDPs with neural network approximation in the mean-field regime. In *International Conference on Machine Learning*, pages 12222–12252. PMLR, 2022.

[35] Guy Lebanon. Axiomatic geometry of conditional models. *IEEE Transactions on Information Theory*, 51:1283–1294, 2005.

[36] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. In *Conference on Learning Theory*, pages 3107–3110. PMLR, 2021.

[37] Haoya Li, Samarth Gupta, Hsiangfu Yu, Lexing Ying, and Inderjit Dhillon. Quasi-Newton policy gradient algorithms. *arXiv:2110.02398*, 2021.

[38] Wuchen Li and Guido Montúfar. Natural gradient via optimal transport. *Information Geometry*, 1(2):181–214, 2018.

[39] Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. Wasserstein Riemannian geometry of Gaussian densities. *Information Geometry*, 1(2):137–179, 2018.

[40] P. Marbach and J.N. Tsitsiklis. Simulation-based optimization of Markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, 2001.

[41] James Martens. New insights and perspectives on the natural gradient method. *The Journal of Machine Learning Research*, 21(1):5776–5851, 2020.

[42] James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.

[43] Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pages 7555–7564. PMLR, 2021.

[44] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

[45] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with deep reinforcement learning. *arXiv:1312.5602*, 2013.

[46] Guido Montúfar, Johannes Rauh, and Nihat Ay. On the Fisher metric of conditional probability polytopes. *Entropy*, 16(6):3207–3233, 2014.

[47] Tetsuro Morimura, Eiji Uchibe, Junichiro Yoshimoto, and Kenji Doya. A new natural policy gradient by stationary distribution metric. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 82–97. Springer, 2008.

[48] Tetsuro Morimura, Eiji Uchibe, Junichiro Yoshimoto, Jan Peters, and Kenji Doya. Derivatives of logarithmic stationary distributions for policy gradient reinforcement learning. *Neural computation*, 22(2):342–376, 2010.

[49] Ted Moskovitz, Michael Arbel, Ferenc Huszar, and Arthur Gretton. Efficient Wasserstein natural gradients for reinforcement learning. In *International Conference on Learning Representations*, 2021.

[50] Johannes Müller and Guido Montúfar. The geometry of memoryless stochastic policy optimization in infinite-horizon POMDPs. In *International Conference on Learning Representations*, 2022.

[51] 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.

[52] Hiroshi Nagaoka. The exponential family of Markov chains and its information geometry. In *28th Symposium on Information Theory and Its Applications (SITA2005)*, 2005.

[53] Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv:1705.07798*, 2017.

[54] Levon Nurbekyan, Wanzhou Lei, and Yunan Yang. Efficient natural gradient descent methods for large-scale optimization problems. *arXiv:2202.06236*, 2022.

[55] Hyeyoung Park, Shun-ichi Amari, and Kenji Fukumizu. Adaptive natural gradient learning algorithms for various stochastic models. *Neural Networks*, 13(7):755–764, 2000.

[56] Jan Peters, Sethu Vijayakumar, and Stefan Schaal. Reinforcement learning for humanoid robotics. In *Proceedings of the third IEEE-RAS international conference on humanoid robots*, pages 1–20, 2003.

[57] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

[58] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[59] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

[60] Kun Shao, Zhentao Tang, Yuanheng Zhu, Nannan Li, and Dongbin Zhao. A survey of deep reinforcement learning in video games. *arXiv:1912.10944*, 2019.

[61] Hirohiko Shima. *The geometry of Hessian structures*. World Scientific, Singapore, 2007.

[62] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[63] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the Game of Go without Human Knowledge. *Nature*, 550(7676):354–359, 2017.

[64] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, Cambridge, Massachusetts, 2018.

[65] Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *NIPS*, volume 99, pages 1057–1063. Citeseer, 1999.

[66] Jesse van Oostrum, Johannes Müller, and Nihat Ay. Invariance properties of the natural gradient in overparametrised systems. *Information Geometry*, pages 1–17, 2022.

[67] Jonathan Weed. An explicit analysis of the entropic penalty in linear programming. In *Conference On Learning Theory*, pages 1841–1855. PMLR, 2018.

[68] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[69] Lin Xiao. On the convergence rates of policy gradient methods. *arXiv:2201.07443*, 2022.

[70] Rui Yuan, Simon S Du, Robert M Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. *arXiv preprint arXiv:2210.01400*, 2022.

[71] Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex MDPs. *Advances in Neural Information Processing Systems*, 34:25746–25759, 2021.

[72] Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *arXiv:2105.11066*, 2021.

[73] Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global Convergence of Policy Gradient Methods to (Almost) Locally Optimal Policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020.

[74] 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*, volume 36, pages 9066–9073, 2022.