Essentially Sharp Estimates on the Entropy Regularization Error in Discrete Discounted Markov Decision Processes

Johannes Müller [1] **Semih Çaycı** [1]

Abstract

We study the error introduced by entropy regularization of infinite-horizon discrete discounted Markov decision processes. We show that this error decreases exponentially in the inverse regularization strength both in a weighted KL-divergence and in value with a problem-specific exponent. We provide a lower bound matching our upper bound up to a polynomial factor. Our proof relies on the correspondence of the solutions of entropy-regularized Markov decision processes with gradient flows of the unregularized reward with respect to a Riemannian metric common in natural policy gradient methods. Further, this correspondence allows us to identify the limit of the gradient flow as the generalized maximum entropy optimal policy, thereby characterizing the implicit bias of the Kakade gradient flow which corresponds to a time-continuous version of the natural policy gradient method. We use this to show that for entropy-regularized natural policy gradient methods the overall error decays exponentially in the square root of the number of iterations.

1. Introduction

Entropy regularization plays an important role in reinforcement learning and is usually employed to encourage exploration thereby improving sample complexity and improving convergence of policy optimization techniques Ahmed et al. (2019). One benefit of entropy regularization is that it corresponds to a strictly convex regularizer in the space of state-action distribution, where the reward optimization problem is equivalent to a linear program. Employing this hidden strong convexity, entropy-regularized vanilla and natural policy gradient methods have been shown to converge exponentially for discrete and continuous problems, for gradient descent and gradient flows, and for tabular methods as well as under function approximation Mei et al. (2020); Cen et al. (2021); Çaycı et al. (2021); Müller & Montúfar (2024); Kerimkulov et al. (2023).

Adding entropy regularization changes the optimization problem thereby introducing an error for which a conclusive analysis remains elusive. For general bounded regularizers and regularization strength $\tau \geq 0$ an $O(\tau)$ estimate on the regularization error can be established Geist et al. (2019), which was subsequently used in the overall error analysis of entropy-regularized natural policy gradients Dai et al. (2018); Lee et al. (2018). This result covers general regularizers but neither uses the structure of Markov decision processes nor the entropic regularizer. Together with the $O(e^{-\tau k \eta})$ convergence of entropy-regularized natural policy gradients this implies $O(\frac{\log k}{\eta k})$ convergence of the overall error with regularization strength $\tau = \frac{\log k}{\eta k}$, where k denotes the number of iterations, see Cen et al. (2021).

1.1. Contributions

The main contribution of this article is a sharp analysis of the entropy regularization error in infinite-horizon discrete discounted Markov decision processes. We summarize our contributions as the following:

- *Sharp analysis of entropy regularization.* In Theorem 2.4 and Theorem 2.5 we give essentially matching upper and lower bounds on the entropy regularization error and show exponential convergence $\tilde{O}(e^{-\Delta \tau^{-1}})$ with a problem-dependent exponent $\Delta > 0$, where $\tilde{O}$ hides polynomial factors.

- *Implicit bias.* We show that the gradient flows corresponding to natural policy gradients converge towards the generalized maximum entropy optimal policy characterizing their implicit bias, see Theorem 2.5.

- *Overall error analysis.* We show that for regularized natural policy gradient methods the overall error decreases exponentially in the square-root of the number of iterations $\tilde{O}(e^{-\sqrt{\Delta \eta k/2}})$, see Theorem 2.6.

[1]Department of Mathematics, RWTH Aachen University, Aachen, 52062, Germany. Correspondence to: Johannes Müller <mueller@mathc.rwth-aachen.de>.

Workshop on Foundations of Reinforcement Learning and Control at the 41st International Conference on Machine Learning, Vienna, Austria. Copyright 2024 by the author(s).

1.2. Related works

At the core of our argument lies the observation that the optimal regularized policies solve the gradient flow of the unregularized reward with respect to the Kakade metric, which can be seen as the continuous time limit of unregularized natural policy gradient methods for tabular softmax policies. This correspondence uses the isometry between the Kakade metric and the conditional Fisher-Rao metric on the state-action distributions Müller & Montúfar (2024), which allows us to use the theory of Hessian gradient flows in convex optimization Alvarez et al. (2004).

Natural policy gradients are known to converge at a $O(\frac{1}{k})$ rate Agarwal et al. (2021) which was used in Khodadadian et al. (2022) to show assymptotic $O(e^{-ck})$ convergence for all $c < \Delta$. Our continuous-time analysis uses similar arguments, but we provide an anytime analysis, an essentially matching lower bound, and show convergence towards the generalized maximum entropy optimal policy.

A continuous-time analysis of gradient flows with respect to the Kakade metric has been conducted for Markov decision processes with discrete and continuous state and action spaces in Müller & Montúfar (2024); Kerimkulov et al. (2023). For unregularized reward exponential convergence was established in Müller & Montúfar (2024) without control over the exponent or the coefficient. Under the presence of entropy regularization with strength $\tau > 0$ exponential $O(e^{-\tau t})$ convergence was established in Müller & Montúfar (2024); Kerimkulov et al. (2023). For gradient flows with respect to the Fisher metric in state-action space exponential convergence $O(e^{-\delta t})$ with a problem specific exponent was established, however, a lower bound is missing there, and the exponent $\delta \leq \Delta$ is dominated by the one for Kakade's gradient flows Müller et al. (2024).

For unregularized natural actor-critic methods a $O(\log k)$ bound on the distance of the policies to the maximum entropy optimal policy was established by Hu et al. (2021). Working in continuous time allows us to show exponential convergence towards the generalized maximum entropy optimal policy. A similar algorithmic bias showing convergence to the maximum entropy optimal policy was established for natural policy gradient methods that decrease regularization and increase step sizes during optimization Li et al. (2023). In contrast to our implicit bias result, this approach considers an asymptotically vanishing but explicit regularization.

1.3. Notation

The *probability simplex* $\Delta_{\mathcal{X}} := \{\mu \in \mathbb{R}^{\mathcal{X}}_{\geq 0} : \sum_{x\in\mathcal{X}} \mu(x) = 1\}$ over a finite set $\mathcal{X}$ denotes all probability vectors. Given a second finite set $\mathcal{Y}$ we can identify the Cartesian product $\Delta^{\mathcal{Y}}_{\mathcal{X}} = \Delta_{\mathcal{X}} \times \cdots \Delta_{\mathcal{X}}$ with the set of stochastic matrices or the set of Markov kernels. We refer to $\Delta^{\mathcal{Y}}_{\mathcal{X}}$ as the *conditional probability polytope* and for $P \in \Delta^{\mathcal{Y}}_{\mathcal{X}}$ we write $P(x|y)$ for the entries of the Markov kernel. Finally, for two vectors $\mu, \nu \in \mathbb{R}^{\mathcal{X}}$ we denote the *Hadamard product* by $\mu \odot \nu \in \mathbb{R}^{\mathcal{X}}$ with entries $(\mu \odot \nu)(x) := \mu(x)\nu(x)$.

2. Main results

We consider a state and action spaces $\mathcal{S}$ and $\mathcal{A}$. The transition dynamics are described by a fixed Markov kernel $P \in \Delta^{\mathcal{S}\times\mathcal{A}}_{\mathcal{S}}$, where $P(s'|s,a)$ describes the probability of transitioning from s to s' under the action a. We consider stochastic memoryless policies $\pi \in \Delta^{\mathcal{S}}_{\mathcal{A}}$ and the discounted infinite-horizon entropy-regularized reward, given by

$$R_\tau(\pi) := (1-\gamma)\mathbb{E}_\pi\Bigg[\sum_{t\in\mathbb{N}} \gamma^t \Big(r(S_t, A_t) - \tau D_{\mathrm{KL}}(\pi(\cdot|S_t), \pi_0(\cdot|S_t)\Big)\Bigg] \tag{1}$$

where $\mu \in \Delta_{\mathcal{S}}$ is an initial distribution over the states, $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, and $\gamma \in [0,1)$ is the *discount factor*. Here, the states and actions are produced according to $S_0 \sim \mu$, $A_t \sim \pi(\cdot|S_t)$, $S_{t+1} \sim P(\cdot|S_t, A_t)$ and we refer to $\pi_0 \in \Delta^{\mathcal{S}}_{\mathcal{A}}$ *reference policy* and to $\tau \geq 0$ as the *regularization strength*. It is the goal of this article to understand the error induced by entropy regularization.

Important concepts in Markov decision processes are the *state* and *state-action distributions*

$$d^\pi(s) := (1-\gamma)\sum_{t\in\mathbb{N}} \gamma^t \mathbb{P}^{\pi,\mu}(S_t = s) \quad \text{and} \tag{2}$$

$$\nu^\pi(s,a) := (1-\gamma)\sum_{t\in\mathbb{N}} \gamma^t \mathbb{P}^{\pi,\mu}(S_t = s, A_t = a). \tag{3}$$

Definition 2.1 (Kakade divergence)**.** For two policies $\pi_1, \pi_2 \in \Delta^{\mathcal{S}}_{\mathcal{A}}$ we call

$$D_{\mathrm{K}}(\pi_1, \pi_2) := \sum_{s\in\mathcal{S}} d^{\pi_1}(s) D_{\mathrm{KL}}(\pi_1(\cdot|s), \pi_2(\cdot|s)) \tag{4}$$

the *Kakade divergence* between π_1 and π_2.[1]

A direct computation shows $R_\tau(\pi) = R(\pi) + \tau D_{\mathrm{K}}(\pi, \pi_0)$. Note that the Kakade divergence is not a Bregman divergence, see Remark A.2.

Assumption 2.2 (State exploration). For any $\pi \in \Delta^{\mathcal{S}}_{\mathcal{A}}$ it holds that $d^\pi(s) > 0$ for all $s \in \mathcal{S}$.

Setting 2.3. *Consider a finite discounted Markov decision process $(\mathcal{S}, \mathcal{A}, P, \gamma, r)$, an initial distribution $\mu \in \Delta_{\mathcal{S}}$ and fix a policy $\pi_0 \in \mathrm{int}(\Delta^{\mathcal{S}}_{\mathcal{A}})$ and let Assumption 2.2 hold and denote the optimal reward by $R^\star := \max\{R(\pi) : \pi \in \Delta^{\mathcal{S}}_{\mathcal{A}}\}$.*

[1] Note that D_{K} depends both on the initial distribution $\mu \in \Delta_{\mathcal{S}}$ as well as on the discount factor $\gamma \in [0,1)$.

Further, let $(\pi_t)_{t\geq 0}$ denote the solutions of the entropy-regularized problems

$$\pi_t = \operatorname*{arg\,max}_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}} \left\{ R(\pi) - t^{-1} D_{\mathrm{K}}(\pi,\pi_0) \right\}. \tag{5}$$

We denote the projection onto the optimal policies by

$$\pi^\star = \operatorname*{arg\,min}_{\pi\in\Pi^\star} D_{\mathrm{K}}(\pi,\pi_0), \tag{6}$$

where $\Pi^\star := \{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}} : R(\pi) = R^\star\}$ denotes the set of optimal policies and set $c := D_{\mathrm{K}}(\pi^\star,\pi_0)$. Finally, we set

$$\Delta := -\max\left\{A^\star(s,a) \neq 0 : s\in\mathcal{S}, a\in\mathcal{A}\right\}, \tag{7}$$

where $A^\star$ is the advantage function under an optimal policy.

Now we give essentially matching upper and lower bounds on the entropy regularization error measured in value.

Theorem 2.4 (Convergence in value)**.** *Consider Setting 2.3. There exist $c_1, c_2 > 0$ such that for $t \geq 1$ it holds that*

$$R^\star - R(\pi_t) \leq c_1 e^{-\Delta(t-1)+\gamma c\log t} \quad \textit{and} \tag{8}$$

$$R^\star - R(\pi_t) \geq c_2 e^{-\Delta(t-1)-c\log t - 2\|r\|_\infty}. \tag{9}$$

We can give the constants in closed form, see Theorem C.2 The distance of the optimal entropy regularized policies decays at the same rate as the suboptimality gap.

Theorem 2.5 (Convergence of policies)**.** *Consider Setting 2.3. It holds that*

$$D_{\mathrm{K}}(\pi^\star,\pi_t) \leq \frac{e^{-\Delta(t-1)+\gamma c\log t}}{1-e^{-\Delta(t-1)+\gamma c\log t}}, \tag{10}$$

if $e^{-\Delta(t-1)+\gamma c\log t} < 1$ which is satisfied for $t > 0$ large enough. Further, for a suitable constant $c_3 > 0$ we have

$$D_{\mathrm{K}}(\pi^\star,\pi_t) \geq c_3 e^{-\Delta(t-1)-c\log t - 2\|r\|_\infty}. \tag{11}$$

As $(\pi_t)_{t\geq 0}$ solve a continuous version of the natural policy gradient method, Theorem 2.5 describes the *implicit bias* of these methods towards a generalized maximum entropy optimal policy, see Remark B.9.

We can combine our improved estimate on the regularization error with any guarantee for a regularized policy optimization technique. In particular, for the popular natural policy gradient method we can use the results of Cen et al. (2021) and obtain the following guarantee.

Theorem 2.6 (Overall error analysis)**.** *Consider Setting 2.3 with $t \geq 1$ and assume that $(\pi_k)_{k\in\mathbb{N}}$ are iterates produced by natural policy ascent with a tabular softmax parametrization with stepsize $\eta > 0$. For $t^{-1} = \sqrt{2\Delta/\eta k}$ it holds that*

$$R^\star - R(\pi_k) = O\left(\left(\eta k \Delta^{-1}\right)^{c/2} \cdot e^{-\sqrt{\Delta\eta k/2}}\right). \tag{12}$$

In contrast, unregularized natural policy gradient methods converge at a rate $\tilde{O}(e^{-\Delta\eta k})$ Khodadadian et al. (2022).

3. Proof structure

Our anlysis relies on the observation that the optimal regularized policies $(\pi_t)_{t\geq 0}$ solve a gradient flow with respect to a Riemannian metric proposed in the context of natural policy gradients. The gradient of the reward with respect to this metric admits an expression via the advantage function which allows for an explicit convergence analysis.

3.1. Geometry of Kakade's metric

The following Riemannian metric was proposed in the context of natural gradients Kakade (2001).

The following expression of the gradient of the reward with respect to the Kakade metric allows for an explicit convergence analysis of the corresponding gradient flow later.

Theorem 3.1 (Gradient with respect to the Kakade metric)**.** *Let Assumption 2.2 hold. Then for all $\pi \in \operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$ and $a \in \mathcal{A}$, $s \in \mathcal{S}$ it holds that*

$$\nabla^{\mathrm{K}} R(\pi)(s,a) = A^\pi(s,a)\pi(a|s), \tag{13}$$

where $\mathcal{A}^\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ denotes the advantage function of π.

Based on the expression of the Kakade gradient, we call

$$\partial_t \pi_t(a|s) = A^{\pi_t}(s,a)\pi_t(a|s) \tag{14}$$

the *Kakade gradient flow*.

The Fisher-Rao metric arises from the negative entropy Ay et al. (2017), similarly, one can define a conditional version. For $\nu \in \mathbb{R}_{>0}^{\mathcal{S}\times\mathcal{A}}$ we define the *conditional entropy* via

$$H_{A|S}(\nu) := \sum_{s\in\mathcal{S},a\in\mathcal{A}} \nu(s,a)\log\frac{\nu(s,a)}{\sum_{a'\in\mathcal{A}}\nu(s,a')}. \tag{15}$$

Definition 3.2 (Conditional Fisher-Rao metric)**.** We call the metric $g^{A|S}$ on $\operatorname{int}(\mathscr{D})$ given by

$$g_\nu^{A|S}(v,w) := v^\top \nabla^2 H_{A|S}(\nu) w \quad \text{for } v,w \in T\mathscr{D} \tag{16}$$

the *conditional Fisher-Rao metric* and the denote the corresponding gradient by $\nabla^{A|S}$. We call the corresponding Bregman divergence $D^{A|S}$ the *conditional KL-divergence*.

The maximization of the regularized reward R_τ subject to $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ is equivalent to the regularized linear program

$$\max r^\top \nu - \tau D_{A|S}(\nu,\nu_0) \quad \text{subject to } \nu \in \mathscr{D}, \tag{17}$$

where $\nu_0 = \nu^{\pi_0}$, see Neu et al. (2017).

The Kakade metric has been characterized as the pullback of the conditional Fisher-Rao metric.

Theorem 3.3 (Müller & Montúfar (2024))**.** *Consider a finite Markov decision process and let Assumption 2.2 hold. Then the following mapping is an isometry*

$$(\operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{S}}), g^{\mathrm{K}}) \to (\operatorname{int}(\mathscr{D}), g^{A|S}), \quad \pi \mapsto \nu^\pi. \tag{18}$$

As the reward is a linear function of the state-action distributions we can apply results from Hessian gradient flows of linear programs Alvarez et al. (2004).

Corollary 3.4. *The following statements hold:*

1. Well-posedness: *The Kakade gradient flow* (14) *admits a unique global solution* $(\pi_t)_{t\geq 0}$.
2. Central path property: *For all* $t \geq 0$ *it holds that*
$$\pi_t = \operatorname*{arg\,max}_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}} \Big\{ R(\pi) - t^{-1} D_{\mathrm{K}}(\pi, \pi_0) \Big\}. \tag{19}$$
3. Sublinear convergence: *For all* $t \geq 0$ *we have*
$$0 \leq R^\star - R(\pi_t) \leq t^{-1} \cdot \min_{\pi^\star\in\Pi^\star} D_{\mathrm{K}}(\pi^\star, \pi_0). \tag{20}$$
4. Implicit bias: *It holds that*
$$\lim_{t\to+\infty} \pi_t = \pi^\star = \operatorname*{arg\,min}_{\pi\in\Pi^\star} D_{\mathrm{K}}(\pi, \pi_0). \tag{21}$$

Importantly, (19) provides an equivalence between Kakade gradient flows and the solutions of the regularized problems.

3.2. Tight analysis

A central step in our proof is to show that the selection probabilities of suboptimal actions decay exponentially fast in t. This can be seen as a continuous time analogon to (Khodadadian et al., 2022, Lemma 3.4).

Lemma 3.5. *Consider Setting 2.3. Then for all* $t \geq t_0 > 0$, $s \in \mathcal{S}$, *and* $a \in \mathcal{A}$ *it holds that*

$$\pi_t(a|s) \leq \pi_0(a|s) e^{A^\star(s,a)(t-1)+\gamma c \log t + 2\|r\|_\infty} \quad \textit{and}$$
$$\pi_t(a|s) \geq \pi_0(a|s) e^{A^\star(s,a)(t-1)-c \log t - 2\|r\|_\infty}.$$

Proof. We consider the Kakade gradient flow (14). As a consequence of the sublinear convergence (20) it holds that

$$\partial_t \pi_t(a|s) \leq \left(A^\star(s,a) + \frac{\gamma c}{t} \right) \pi_t(a|s) \quad \text{and}$$
$$\partial_t \pi_t(a|s) \geq \left(A^\star(s,a) - \frac{c}{t} \right) \pi_t(a|s)$$

where $c = D(\pi^\star, \pi_0)$. Now, Grönwall's inequality yields

$$\begin{aligned}\pi_T(a|s) &\leq \pi_{t_0}(a|s) e^{A^\star(s,a)(T-t_0)+\gamma c \int_{t_0}^T t^{-1} \mathrm{d}t} \\ &= \pi_{t_0}(a|s) e^{A^\star(s,a)(T-t_0)+\gamma c \log T - \gamma c \log t_0}\end{aligned}$$

and similarly for the lower bound. Further, note that $\|A^\pi\|_\infty \leq 2\|r\|_\infty$ and hence Grönwall yields

$$\pi_0(a|s) e^{-2\|r\|_\infty t_0} \leq \pi_{t_0}(a|s) \leq \pi_0(a|s) e^{2\|r\|_\infty t_0}$$

and choosing $t_0 = 1$ finishes the proof. □

The proofs of the main results presented in Section 2 rely on the following auxiliary results, which proofs we defer to Appendix C. The combination of Lemma 3.5 and the following lemma yields the proof of Theorem 2.4.

Lemma 3.6 (Sub-optimality gap)**.** *For any policy* $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ *it holds that*

$$R^\star - R(\pi) \leq \frac{2\|r\|_\infty}{1-\gamma} \sum_{s\in\mathcal{S}, a\notin A_s^\star} d^\pi(s)\pi(a|s) \quad \textit{and} \tag{22}$$
$$R^\star - R(\pi) \geq \frac{\Delta}{1-\gamma} \cdot \sum_{s\in\mathcal{S}, a\notin A_s^\star} d^\pi(s)\pi(a|s). \tag{23}$$

The following lemma underlies the proof of Theorem 2.5.

Lemma 3.7. *Let* $\pi^\star \in \Pi^\star$ *be the Kakade projection of* $\pi \in \operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$ *onto* $\Pi^\star$. *Then it holds that*

$$D_{\mathrm{K}}(\pi^\star, \pi) \leq \frac{\max_{s\in\mathcal{S}} \sum_{a\notin A_s^\star} \pi(a|s)}{1 - \max_{s\in\mathcal{S}} \sum_{a\notin A_s^\star} \pi(a|s)} \quad \textit{and} \tag{24}$$
$$D_{\mathrm{K}}(\pi^\star, \pi) \geq \sum_{s\in\mathcal{S}} d^\star(s) \sum_{a\notin A_s^\star} \pi(a|s). \tag{25}$$

To estimate the unregularized reward of a policy close to the optimal regularized policy we use the following result.

Lemma 3.8. *With* $c = \frac{\sqrt{2}|\mathcal{S}|\|r\|_\infty}{1-\gamma}$ *for any policy* $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ *we have*

$$R^\star - R(\pi) \leq R^\star - R(\pi_\tau^\star) + c\|\log \pi_\tau^\star - \log \pi\|_\infty^{\frac{1}{2}}, \tag{26}$$

where $\pi_\tau^\star$ *denotes the optimal regularized policy.*

Entropy-regularized natural policy gradient methods with step size $\eta > 0$ and regularization strength $\tau > 0$ satisfy $\|\log \pi_\tau^\star - \log \pi\|_\infty = O(e^{-\tau\eta k})$ Cen et al. (2021). This together with Lemma 3.8 yields Theorem 2.6.

4. Conclusion

We provide essentially sharp estimates on the error introduced by entropy regularization in discrete discounted Markov decision processes showing $\tilde{O}(e^{-\Delta\tau^{-1}})$ convergence with a problem-dependent exponent $\Delta > 0$. At the heart lies the observation that the optimal regularized policies solve a gradient flow with respect to a Riemannian metric playing a central role in natural policy gradient methods. The tight estimate on the entropy regularization error leads to an improved $\tilde{O}(e^{-\sqrt{\Delta\eta k/2}})$ overall error analysis of natural policy gradient methods. However, this leaves a gap to the $\tilde{O}(e^{-\Delta\eta k})$ convergence of unregularized natural policy methods that can be addressed in future works.

Acknowledgments

The authors acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the project number 442047500 through the Collaborative Research Center *Sparsity and Singular Structures* (SFB 1481).

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

Ahmed, Z., Le Roux, N., Norouzi, M., and Schuurmans, D. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pp. 151–160. PMLR, 2019.

Alvarez, F., Bolte, J., and Brahic, O. Hessian Riemannian gradient flows in convex programming. *SIAM journal on control and optimization*, 43(2):477–501, 2004.

Amari, S.-i. *Information geometry and its applications*, volume 194. Springer, 2016.

Ay, N., Jost, J., Vân Lê, H., and Schwachhöfer, L. *Information geometry*, volume 64. Springer, 2017.

Bagnell, J. A. and Schneider, J. Covariant policy search. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pp. 1019–1024, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

Boll, B., Cassel, J., Albers, P., Petra, S., and Schnörr, C. A geometric embedding approach to multiple games and multiple populations. *arXiv preprint arXiv:2401.05918*, 2024.

Bubeck, S. et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Çaycı, S., He, N., and Srikant, R. Linear convergence of entropy-regularized natural policy gradient with linear function approximation. *arXiv preprint arXiv:2106.04096*, 2021.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021.

Csiszár, I. and Matúš, F. Generalized minimizers of convex integral functionals, bregman distance, pythagorean identities. *Kybernetika*, 48(4):637–689, 2012. URL http://eudml.org/doc/247058.

Dai, B., Shaw, A., Li, L., Xiao, L., He, N., Liu, Z., Chen, J., and Song, L. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International conference on machine learning*, pp. 1125–1134. PMLR, 2018.

Derman, C. *Finite state Markovian decision processes*. Academic Press, Inc., 1970.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

Hu, Y., Ji, Z., and Telgarsky, M. Actor-critic is implicitly biased towards high entropy optimal policies. *arXiv preprint arXiv:2110.11280*, 2021.

Johnson, E., Pike-Burke, C., and Rebeschini, P. Optimal Convergence Rate for Exact Policy Mirror Descent in Discounted Markov Decision Processes. *Advances in Neural Information Processing Systems*, 36, 2024.

Kakade, S. M. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Kallenberg, L. C. Survey of linear programming for standard and nonstandard Markovian control problems. Part I: Theory. *Zeitschrift für Operations Research*, 40:1–42, 1994.

Kerimkulov, B., Leahy, J.-M., Šiška, D., Szpruch, Ł., and Zhang, Y. A Fisher-Rao gradient flow for entropy-regularised Markov decision processes in Polish spaces. *arXiv preprint arXiv:2310.02951*, 2023.

Khodadadian, S., Jhunjhunwala, P. R., Varma, S. M., and Maguluri, S. T. On linear and super-linear convergence of natural policy gradient algorithm. *Systems & Control Letters*, 164:105214, 2022.

Lan, G. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pp. 1–48, 2022.

Laroche, R. and Des Combes, R. T. On the occupancy measure of non-markovian policies in continuous mdps. In *International Conference on Machine Learning*, pp. 18548–18562. PMLR, 2023.

Lee, K., Choi, S., and Oh, S. Sparse Markov decision processes with causal sparse Tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018.

Li, Y., Lan, G., and Zhao, T. Homotopic policy mirror descent: policy convergence, algorithmic regularization,

and improved sample complexity. *Mathematical Programming*, pp. 1–57, 2023.

Liu, J., Li, W., and Wei, K. Elementary analysis of policy gradient methods. *arXiv preprint arXiv:2404.03372*, 2024.

Mei, J., Xiao, C., Szepesvári, C., and Schuurmans, D. On the Global Convergence Rates of Softmax Policy Gradient Methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020.

Montúfar, G., Rauh, J., and Ay, N. On the Fisher metric of conditional probability polytopes. *Entropy*, 16(6):3207–3233, 2014.

Müller, J. *Geometry of Optimization in Markov Decision Processes and Neural Network Based PDE Solvers*. PhD thesis, University of Leipzig, 2023.

Müller, J. and Montufar, G. The Geometry of Memoryless Stochastic Policy Optimization in Infinite-Horizon POMDPs. In *International Conference on Learning Representations*, 2022.

Müller, J. and Montúfar, G. Geometry and convergence of natural policy gradient methods. *Information Geometry*, 7(Suppl 1):485–523, 2024.

Müller, J., Çaycı, S., and Montúfar, G. Fisher-Rao Gradient Flows of Linear Programs and State-Action Natural Policy Gradients. *arXiv preprint arXiv:2403.19448*, 2024.

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Neu, G., Jonsson, A., and Gómez, V. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Peters, J., Vijayakumar, S., and Schaal, S. Reinforcement learning for humanoid robotics. In *Proceedings of the third IEEE-RAS international conference on humanoid robots*, pp. 1–20, 2003.

Sethi, D., Šiška, D., and Zhang, Y. Entropy annealing for policy mirror descent in continuous time and space. *arXiv preprint arXiv:2405.20250*, 2024.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Wolfer, G. and Watanabe, S. Information geometry of Markov kernels: a survey. *Frontiers in Physics*, 11: 1195562, 2023.

A. Preliminaries

In this section, we provide background material from the theory of Markov decision processes. We put an emphasis on regularization and see that entropy-regularized Markov decision processes are equivalent to a regularized linear programming formulation of the Markov decision process, where the regularizer is given by a conditional entropy term. We conclude with a general discussion of regularized linear programs and revisit the central path property. This states that the solutions of regularized linear programs with regularization strength t^{-1} solve the gradient flow of the linear program with respect to the Riemannian metric induced by the convex regularizer. Our explicit analysis of the entropy regularization error is built on this dynamic interpretation of the solutions of the optimizers of the regularized problems.

A.1. Markov decision processes and entropy regularization

We consider a finite set $\mathcal{S}$ of states of some system and a finite set $\mathcal{A}$ of actions that can used to control the state $s \in \mathcal{S}$. The transition dynamics are described by a fixed Markov kernel $P \in \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$, where $P(s'|s,a)$ describes the probability of transitioning from s to s' under the action a. It is the goal in Markov decision processes and reinforcement learning to design a *policy* $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, where the entries $\pi(a|s)$ of the stochastic matrix describe the probability of selecting an action a when in state s. A common optimality criterion is the *discounted infinite horizon reward* given by

$$R(\pi) \coloneqq (1-\gamma)\mathbb{E}_{\mathbb{P}^{\pi,\mu}}\left[\sum_{t\in\mathbb{N}} \gamma^t r(S_t, A_t)\right], \tag{27}$$

where $\mu \in \Delta_{\mathcal{S}}$ is an initial distribution over the states and $\mathbb{P}^{\pi,\mu}$ denotes the law of the Markov process on $\mathcal{S}\times\mathcal{A}$ induced by the iteration

$$S_0 \sim \mu, A_t \sim \pi(\cdot|S_t), S_{t+1} \sim P(\cdot|S_t, A_t), \tag{28}$$

$r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is the *instantaneous reward* vector, and $\gamma \in [0,1)$ is the *discount factor*. To encourage exploration it is common to regularize the reward with an entropy term which yields the *entropy-regularized* or *KL-regularized* reward

$$R_\tau^\mu(\pi) \coloneqq (1-\gamma)\mathbb{E}_{\mathbb{P}^{\pi,\mu}}\left[\sum_{t\in\mathbb{N}} \gamma^t (r(S_t, A_t) - \tau D_{\mathrm{KL}}(\pi(\cdot|S_t), \pi_0(\cdot|S_t))\right] \tag{29}$$

for some *reference policy* $\pi_0 \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ and *regularization strength* $\tau \geq 0$. The entropy or KL-regularized reward optimization problem is given by

$$\max R_\tau(\pi) \quad \text{subject to } \pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}. \tag{30}$$

The regularization can be interpreted as a convex regularization and consequently was used to show exponential convergence (Mei et al., 2020; Cen et al., 2021; Çaycı et al., 2021; Lan, 2022), however, it introduces an error and leads to a new optimal policy $\pi_\tau^\star$, which might not be optimal with respect to the unregularized reward R. It is our goal to understand how well the entropy-regularized reward optimization problem (30) approximates the unregularized reward optimization problem. For this we provide an explicit analysis of the *entropy regularization error* $R^\star - R(\pi_\tau^\star)$ and $\min_{\pi^\star\in\Pi^\star} D(\pi^\star, \pi_\tau^\star)$, where $R^\star = \max_\pi R(\pi)$ denotes the optimal reward, $\Pi^\star$ the set of optimal policies, and $D(\cdot,\cdot)$ is some notion of distance.

A central role in theoretical and algorithmic approaches to Markov decision processes and reinforcement learning plays the *value function* $V^\pi \in \mathbb{R}^{\mathcal{S}}$ given by

$$V_\tau^\pi(s) \coloneqq R_\tau^{\delta_s}(\pi) \quad \text{for all } s \in \mathcal{S}, \tag{31}$$

which stores the reward obtained when starting in a deterministic state $s \in \mathcal{S}$. The unregularized and regularized *state-action* or *Q-value functions* are given by

$$Q_\tau(s,a) \coloneqq (1-\gamma)r(s,a) + \gamma \sum_{s'\in\mathcal{S}} V_\tau^\pi(s')P(s'|s,a) \quad \text{for all } s \in \mathcal{S}, a \in \mathcal{A}. \tag{32}$$

The *advantage function* of a policy π given by

$$A_\tau^\pi(s,a) \coloneqq Q_\tau^\pi(s,a) - V_\tau^\pi(s), \tag{33}$$

in words, $A_\tau^\pi(s,a)$ describes how much better is it to select action a and follow the policy π afterwards compared to following π. Finally, we define the *optimal reward* and *optimal value functions* via $R^\star \coloneqq \max_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}} R(\pi)$ and

$$V_\tau^\star(s) \coloneqq \max_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}} V_\tau^\pi(s) \quad \text{and } Q_\tau^\star(s,a) \coloneqq \max_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}} Q_\tau^\pi(s,a) \quad \text{for } s\in\mathcal{S}, a\in\mathcal{A} \tag{34}$$

and the *optimal advantage function* via

$$A_\tau^\star(s,a) \coloneqq Q_\tau^\star(s,a) - V^\star(s) \quad \text{for } s\in\mathcal{S}, a\in\mathcal{A}. \tag{35}$$

It is well known that there are optimal policies $\pi_\tau^\star \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ satisfying $V_\tau^\star = V^{\pi_\tau^\star}$ and $Q_\tau^\star = Q^{\pi_\tau^\star}$, which is known as the Bellman principle, see Geist et al. (2019). In the unregularized case, the optimal advantage function $A^\star$ satisfies $A^\star(s,a)\le 0$ for all $s\in\mathcal{S}, a\in\mathcal{A}$, and an action $a\in\mathcal{A}$ is optimal in a state $s\in\mathcal{S}$ if and only if $A^\star(s,a)=0$. For $s\in\mathcal{S}$, we denote the set of optimal actions by $A_s^\star \coloneqq \{a\in\mathcal{A} : A^\star(s,a)=0\}$ and the set of optimal policies is given by

$$\Pi^\star \coloneqq \left\{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}} : R(\pi)=R^\star\right\} = \left\{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}} : \operatorname{supp}(\pi(\cdot|s))\subseteq A_s^\star \text{ for all } s\in\mathcal{S}\right\}. \tag{36}$$

We denote the unregularized value and advantage functions by $V^\pi = V_0^\pi$, $Q^\pi = Q_0^\pi$, $q^\pi = q_0^\pi$, $A^\pi = A_0^\pi$, and $B^\pi = B_0^\pi$. Note that $Q^\pi = q^\pi$ and $A^\pi = B^\pi$.

Of central importance in the theory of Markov decision processes are the *state distributions*, which measure how much time the process spends at the individual states for a given policy $\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}$. This is formalized by

$$d^\pi(s) = d_\gamma^{\pi,\mu}(s) \coloneqq (1-\gamma)\sum_{t\in\mathbb{N}} \gamma^t \mathbb{P}^{\pi,\mu}(S_t=s). \tag{37}$$

Note that indeed $d^\pi\in\Delta_{\mathcal{S}}$ by the geometric series. Sometimes, the state distributions are called state frequencies, state occupancy measures or (state) visitation distributions. We work with the following notion of distance between policies.

Definition A.1 (Kakade divergence)**.** For two policies $\pi_1,\pi_2\in\Delta_{\mathcal{A}}^{\mathcal{S}}$ we call

$$D_{\mathrm{K}}(\pi_1,\pi_2) = D_{\mathrm{K}}^\mu(\pi_1,\pi_2) \coloneqq \sum_{s\in\mathcal{S}} d^{\pi_1}(s) D_{\mathrm{KL}}(\pi_1(\cdot|s),\pi_2(\cdot|s)) \tag{38}$$

the *Kakade divergence* between π_1 and π_2. Note that D_{K} depends on the initial distribution $\mu\in\Delta_{\mathcal{S}}$ and the discount factor $\gamma\in[0,1)$.

The Kakade divergence arises naturally when studying entropy regularization since

$$R_\tau(\pi) = R(\pi) - \tau D_{\mathrm{K}}(\pi,\pi_0) \quad \text{for all } \pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}. \tag{39}$$

Note that although D_{KL} is a Bregman divergence, the Kakade divergence is not, which hinders the direct use of mirror descent tools developed in the context of convex optimization (Bubeck et al., 2015).

Remark A.2 (Kakade divergence is not Bregman). Bregman divergences are well-studied in convex optimization (Alvarez et al., 2004; Bubeck et al., 2015), however, the Kakade divergence does not fall into this class. Indeed, Bregman divergences are convex in their first argument, which is not generally true for the Kakade divergence. To construct a specific example, where $D_{\mathrm{K}}(\cdot,\pi)$ is not convex, we consider the Markov decision process with two states and actions shown in Figure 1. Further, we choose the reference policy π to be the uniform policy and if we consider $\pi_p(a_1|s_i)=p$ for $i=1,2$, then we

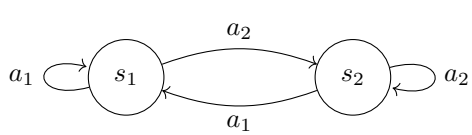

Figure 1. Transition graph of the Markov decision process.

obtain

$$g(p) \coloneqq D_{\mathrm{K}}(\pi_p,\pi) = (1-\gamma)\phi(p) + \gamma^2 p^2 \phi(p), \tag{40}$$

where $\phi(p) := -p\log p - (1-p)\log(1-p)$. Taking the second derivative yields

$$\partial_p^2 g(p) = -(1-\gamma)\left(\frac{1}{1-p} - \frac{1}{p^2}\right) - \gamma^2\left(\frac{1}{1-p} + \frac{1}{p}\right)p - 2\gamma^2(\log p - \log(1-p)). \tag{41}$$

Taking $p \to 1$ yields $\partial_p^2 g(p) \to -\infty$ showing the non-convexity of g and therefore $D_{\mathrm{K}}(\cdot, \pi)$.

As we discuss later, the Kakade divergence is the pullback of the conditional KL divergence on the space of state-action distributions, which renders the entropy-regularized reward optimization problem equivalent to a linear program with a Bregman regularizer, see Proposition A.6.

The following sublinear estimate on the regularization error is well known and commonly applied when approximating unregularized Markov decision processes via regularization, see Geist et al. (2019); Cen et al. (2021). Here, we follow the terminology common in optimization theory, where linear convergence refers to exponential convergence $O(e^{-ct})$ and sublinear to an algebraic convergence rate $O(t^{-\kappa})$ (Nesterov et al., 2018).

Proposition A.3 (Sublinear estimate on the regularization error)**.** *Let $\pi_\tau^\star$ denote an optimal policy of the regularized reward $R_\tau(\pi) := R(\pi) - \tau D_{\mathrm{K}}(\pi, \pi_0)$ for some $\pi_0 \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ and denote the set of optimal policies by $\Pi^\star$. Then we have*

$$0 \le R^\star - R(\pi_\tau^\star) \le \tau \inf_{\pi^\star \in \Pi^\star} D_{\mathrm{K}}(\pi^\star, \pi_0), \tag{42}$$

where $\inf_{\pi\in\Pi^\star} D_{\mathrm{K}}(\pi, \pi_0) < +\infty$ for $\pi_0 \in \mathrm{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$.

Proof. For any $\pi^\star \in \Pi^\star$ we have

$$R^\star - \tau D_{\mathrm{K}}(\pi^\star, \pi_0) = R_\tau(\pi^\star) \le R_\tau(\pi_\tau^\star) \le R(\pi_\tau^\star).$$

Rearranging and taking the infimum over $\Pi^\star$ yields the claim. □

Note that Proposition A.3 holds for general regularizers and uses neither the reward nor the regularizer. In the remainder, we will improve this sublinear estimate $O(\tau)$ on the suboptimality of the entropy optimal regularized policy $\pi_\tau^\star$ to a linear estimate $\tilde{O}(e^{-\Delta\tau^{-1}})$, establish an essentially matching lower bound, and provide a similar estimate on $\min_{\pi^\star\in\Pi^\star} D_{\mathrm{K}}(\pi^\star, \pi_\tau^\star)$.

A.2. State-action geometry of entropy regularization

Similarly to the state-distributions, we define the *state-action distribution* of a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ via

$$\nu^\pi(s,a) = \nu_\gamma^{\pi,\mu}(s,a) := (1-\gamma)\sum_{t\in\mathbb{N}} \gamma^t \mathbb{P}^{\pi,\mu}(S_t = s, A_t = a). \tag{43}$$

Note that by the geometric series and stationarity of the policy, we have $\nu^\pi \in \Delta_{\mathcal{S}\times\mathcal{A}}$ and it holds that $\nu^\pi(s,a) = d^\pi(s)\pi(a|s)$. The state and state-action distributions are also known as *occupancy measures* or *state frequencies*. An important property of state-action distributions is given by

$$R(\pi) = r^\top \nu^\pi, \tag{44}$$

which can be seen using the dominated convergence theorem (Müller, 2023). Moreover, we have the classic characterization of the set of state-action distributions.

Proposition A.4 (State-action polytope, Derman (1970))**.** *The set $\mathscr{D} = \{\nu^\pi : \pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}\}$ of state-action distributions is a polytope given by*

$$\mathscr{D} = \Delta_{\mathcal{S}\times\mathcal{A}} \cap \left\{\nu \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \ell_s(\nu) = 0 \text{ for all } s \in \mathcal{S}\right\}, \tag{45}$$

where

$$\ell_s(\nu) := \sum_{a\in\mathcal{A}} \nu(s,a) - \gamma \sum_{a'\in\mathcal{A}, s'\in\mathcal{S}} P(s|s',a')\nu(s',a') - (1-\gamma)\mu(s). \tag{46}$$

In particular, the characterization of the state-action distributions of a Markov decision process as a polytope shows that the reward optimization problem is equivalent to the linear program

$$\max r^\top \nu \quad \text{subject to } \nu \in \mathscr{D}. \tag{47}$$

Further, it is well known that for a state-action distribution $\nu \in \mathscr{D}$ a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ with $\nu^\pi = \nu$ can be computed by conditioning, see for example (Kallenberg, 1994; Müller & Montufar, 2022; Laroche & Des Combes, 2023), and hence we have

$$\pi(a|s) = \nu(a|s) := \begin{cases} \frac{\nu(s,a)}{\sum_{a'} \nu(s,a')} & \text{if } \sum_{a'} \nu(s,a') > 0 \text{ and} \\ \frac{1}{|\mathcal{A}|} & \text{otherwise.} \end{cases} \tag{48}$$

The entropy-regularized reward optimization problem admits an interpretation as a regularized version of the linear program (47), where the regularizer is given by the conditional entropy.

Definition A.5 (Conditional entropy and KL)**.** For $\nu \in \mathbb{R}_{>0}^{\mathcal{S}\times\mathcal{A}}$ we define the *conditional entropy* via

$$H_{A|S}(\nu) := \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \nu(s,a) \log \frac{\nu(s,a)}{\sum_{a'} \nu(s,a')} = H(\nu) - H(d), \tag{49}$$

where $d(s) := \sum_{a\in\mathcal{A}} \nu(s,a)$. For $\nu_1, \nu_2 \in \mathbb{R}_{>0}^{\mathcal{S}\times\mathcal{A}}$ we call

$$D_{A|S}(\nu_1, \nu_2) := \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \nu_1(s,a) \log \frac{\nu_1(a|s)}{\nu_2(a|s)} = D_{\mathrm{KL}}(\nu_1, \nu_2) - D_{\mathrm{KL}}(d_1, d_2). \tag{50}$$

the *conditional KL-divergence* between ν_1 and ν_2, where $d_i(s) = \sum_{a\in\mathcal{A}} \nu_i(s,a)$.

Direct computation shows that $D_{A|S}$ is the Bregman divergence induced by $H_{A|S}$, see (Neu et al., 2017, Appendix A1).

Proposition A.6 (State-action geometry of entropic regularization)**.** *It holds that*

$$D_{\mathrm{K}}(\pi_1, \pi_2) = D_{A|S}(\nu^{\pi_1}, \nu^{\pi_2}) \quad \textit{for all } \pi_1, \pi_2 \in \Delta_{\mathcal{A}}^{\mathcal{S}} \tag{51}$$

showing that D_{K} is the pull back of the conditional KL-divergence $D_{A|S}$. In particular, the entropy-regularized reward optimization problem (30) *is equivalent to the regularized linear program*

$$\max r^\top \nu - \tau D_{A|S}(\nu, \nu_0) \quad \textit{subject to } \nu \in \mathscr{D}, \tag{52}$$

where $\nu_0 = \nu^{\pi_0}$, meaning that there is a unique solution $\nu_\tau^\star \in \mathrm{int}(\mathscr{D})$ and therefore a unique solution $\pi_\tau^\star \in \mathrm{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$ of the regularized problem and we have $\nu^{\pi_\tau^\star} = \nu_\tau^\star$.

Proof. This is a direct consequence of $R_\tau(\pi) = r^\top \nu^\pi - \tau D_{A|S}(\nu, \nu_0)$, see also Neu et al. (2017). □

The Pythagorean theorem can be generalized to Bregman divergences, which is well known in the field of information and convex optimization (Csiszár & Matúš, 2012; Bubeck et al., 2015; Amari, 2016; Ay et al., 2017). As the Kakade divergence is not a Bregman divergence, it is not included in those general results, but using the characterization of D_{K} as the pullback of a conditional KL-divergence $D_{A|S}$ allows us to provide a Pythagorean theorem for the Kakade divergence and s-rectangular policy classes.

Theorem A.7 (Pythagoras for Kakade divergence)**.** *Consider a set of policies $\Pi = \otimes_{s\in\mathcal{S}} \Pi_s \subseteq \Delta_{\mathcal{A}}^{\mathcal{S}}$ given by the cartesian product of polytopes $\Pi_s \subseteq \Delta_{\mathcal{A}}$. Further, fix $\pi_0 \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ and consider the Kakade projection*

$$\hat{\pi} = \arg\min_{\pi\in\Pi} D_{\mathrm{K}}(\pi, \pi_0) \tag{53}$$

of π_0 onto Π. Then for any $\pi \in \Pi$ we have

$$D_{\mathrm{K}}(\pi, \pi_0) \geq D_{\mathrm{K}}(\pi, \hat{\pi}) + D_{\mathrm{K}}(\hat{\pi}, \pi_0). \tag{54}$$

If further $\Pi_s = \Delta_{\mathcal{A}} \cap \mathcal{L}_s$ for affine spaces $\mathcal{L}_s \subseteq \mathbb{R}^{\mathcal{A}}$ for all $s \in \mathcal{S}$, then we have equality in (54).

Proof. Consider the set $D \coloneqq \{\nu^\pi : \pi \in \Pi\} \subseteq \mathscr{D}$ of state-action distributions arising from the policy class Π, which is again a polytope (Müller & Montufar, 2022, Remark 55). In particular, D is convex and we can pass to the corresponding state-action distributions $\nu_0, \hat{\nu}, \nu$ and apply the Pythagorean theorem for Bregman divergences, see Appendix E.2. If further $\Pi_s = \Delta_{\mathcal{A}} \cap \mathcal{L}_s$, then we have $D = \mathscr{D} \cap \mathcal{L}$ for an affine subspace $\mathcal{L}$ (Müller & Montufar, 2022, Proposition 14). Consequently, the Bregman projection $\hat{\nu}$ will always lie at the relative interior $\operatorname{int}(D)$ and we get equality in the Pythagorean theorem. □

A.3. Regularized linear programs and Hessian gradient flows

We have seen that the state-action distributions of the optimal regularized policies $\pi_\tau^\star$ solve the linear program associated with the Markov decision process with a conditional entropic regularization. Here, we see that the solutions of regularized linear programs solve the gradient flow of the unregularized linear objective with respect to the Riemannian metric induced by the convex regularizer. We refer to Alvarez et al. (2004) for a more general discussion of Hessian gradient flows, where slightly stronger assumptions on ϕ are made. We work in the following setting.

Setting A.8. *We consider the linear program*

$$\max c^\top x \quad \textit{subject to } x \in P, \tag{55}$$

with cost $c \in \mathbb{R}^d$ and feasible region given by a polytope $P \subseteq \mathbb{R}^d$. Further, we consider a twice continuously differentiable convex function ϕ defined on a neighborhood of $\operatorname{int}(P)$ and assume that $\nabla^2\phi(x)$ is strictly positive definite on TP for all $x \in \operatorname{int}(P)$, where TP denotes the tangent space of the polytope P, which is given by the affine span. We define the Riemannian Hessian metric $g_x^\phi(v, w) \coloneqq v^\top \nabla^2\phi(x) w$ on $\operatorname{int}(P)$ and denote the gradient of f with respect to g^ϕ by $\nabla^\phi f$. By $(x_t)_{t\in[0,T)} \subseteq \operatorname{int}(P)$ we denote a solution of the Hessian gradient flow

$$\partial_t x_t = \nabla^\phi f(x_t) \tag{56}$$

with initial condition $x_0 \in \operatorname{int}(P)$ and potential $f(x) = c^\top x$, where $T \in \mathbb{R}_{\geq 0} \cup \{+\infty\}$. Finally, we denote the Bregman divergence induced by ϕ by D_ϕ.

Note that $(x_t)_{t\in[0,T)} \subseteq \operatorname{int}(P)$ solves the Hessian gradient flow (56) if and only if we have

$$g_{x_t}^\phi(\partial_t x_t, v) = \langle \nabla^2\phi(x_t)\partial_t x_t, v\rangle = \langle \nabla f(x_t), v\rangle \quad \text{for all } v \in TP, t \in [0, T). \tag{57}$$

We can now formulate and prove the equivalence property between Hessian gradient flows of linear programs and solutions of Bregman regularized linear programs.

Proposition A.9 (Central path property)**.** *Consider Setting A.8. Then the Hessian gradient flow $(x_t)_{t\in[0,T)}$ of the linear program is characterized by*

$$x_t \in \arg\max \left\{c^\top x - t^{-1} D_\phi(x, x_0) : x \in P\right\} \quad \textit{for all } t \in (0, T). \tag{58}$$

Proof. Note that by the first-order stationarity conditions for equality-constrained optimization, a point $\hat{x} \in \operatorname{int}(P)$ maximizes $g(x) \coloneqq c^\top x - t^{-1} D_\phi(x, x_0)$ over the feasible region P of the linear program if and only if $\langle \nabla g(\hat{x}_t), v\rangle = 0$ for all $v \in TP$. Direct computation yields $\nabla g(x) = c - t^{-1}(\nabla\phi(x) - \nabla\phi(x_0))$ and hence the maximizers $\hat{x}$ of g over P are characterized by

$$t\langle c, v\rangle = \langle \nabla\phi(\hat{x}) - \nabla\phi(x_0), v\rangle \quad \text{for all } v \in TP.$$

On the other hand, for the gradient flow, we can use (57) and compute for $v \in TP$

$$\begin{aligned}
\langle \nabla\phi(x_t) - \nabla\phi(x_0), v\rangle &= \int_0^t \partial_s \langle \nabla\phi(x_s), v\rangle \mathrm{d}s \\
&= \int_0^t \langle \nabla^2\phi(x_s)\partial_s x_s, v\rangle \mathrm{d}s \\
&= \int_0^t \langle \nabla f(x_s), v\rangle \mathrm{d}s \\
&= \int_0^t \langle c, v\rangle \mathrm{d}s = t\langle c, v\rangle.
\end{aligned}$$

This shows x_t maximizes g over P as claimed. □

Corollary A.10 (Sublinear convergence)**.** *Consider Setting A.8 and denote the face of maximizers of the linear program (55) by $F^\star$ and fix $x^\star \in \arg\min_{x\in F^\star} D_\phi(x, x_0)$. Then it holds that*

$$c^\top x^\star - c^\top x_t \leq \frac{D_\phi(x^\star, x_0) - D_\phi(x_t, x_0)}{t} \leq \frac{D_\phi(x^\star, x_0)}{t} \quad \textit{for all } t \in [0, T). \tag{59}$$

Proof. By the central path property, we have

$$c^\top x_t - t^{-1} D_\phi(x_t, x_0) \geq c^\top x^\star - t^{-1} D_\phi(x^\star, x_0).$$

Rearranging yields the result. □

Corollary A.11 (Implicit bias of Hessian gradient flows of LPs)**.** *Consider Setting A.8 and denote the face of maximizers of the linear program (55) by $F^\star$, assume that ϕ is strictly convex and continuous on its domain, and assume that the Hessian gradient flow $(x_t)_{t\geq 0}$ exists for all times. Then it holds that*

$$\lim_{t\to+\infty} x_t = x^\star = \operatorname*{arg\,min}_{x\in F^\star} D_\phi(x, x_0). \tag{60}$$

In words, the Hessian gradient flow converges to the Bregman projection of x_0 to $F^\star$.

Proof. By compactness of P, the sequence $(x_{t_n})_{n\in\mathbb{N}}$ has at least one accumulation point for any sequence $t_n \to +\infty$. Hence, we can assume without loss of generality that $x_{t_n} \to \hat{x}$ and it remains to identify $\hat{x}$ as the information projection $x^\star \in F^\star$.

Surely, we have $\hat{x} \in F^\star$ as $c^\top \hat{x} = \lim_{n\to\infty} c^\top x_{t_n} = \max_{x\in P} c^\top x$ by Corollary A.10. Further, by the central path property we have for any optimizer $x' \in F^\star$ that

$$c^\top x_t - t^{-1} D_\phi(x_t, x_0) \geq c^\top x' - t^{-1} D_\phi(x', x_0)$$

and therefore

$$D_\phi(x', x_0) - D_\phi(x_t, x_0) \geq t c^\top (x' - x_t) \geq 0.$$

Hence, we have

$$D_\phi(\hat{x}, x_0) = \lim_{n\to\infty} D_\phi(x_{t_n}, x_0) \leq D_\phi(x', x_0)$$

and can conclude by minimizing over $x' \in F^\star$. □

B. Geometry and Sublinear Convergence of Kakade Gradient Flows

Our goal is to study the solutions $\pi_\tau^\star$ of the entropy-regularized reward R_τ. For linear programs, we have seen in Appendix A.3 that the solutions of the regularized problems solve the corresponding Hessian gradient flow. Recall, that the entropy-regularized reward optimization problem is equivalent to a linear program in state-action space with a conditional KL regularization. Hence, the state-action distributions solve a Hessian gradient flow and consequently, the optimal regularized policies solve a gradient flow with respect to some metric. In this section, we study this Riemannian metric on the space of policies and provide an explicit expression of the gradient dynamics.

B.1. The Kakade metric and policy gradient theorems

The following metric on the policy space was proposed in the context of natural gradients by Kakade (2001).

Definition B.1 (Kakade metric)**.** We call the Riemannian metric g^{K} on $\operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$ defined by

$$g_\pi^{\mathrm{K}}(v, w) \coloneqq \sum_{s\in\mathcal{S}} d^\pi(s) \sum_{a\in\mathcal{A}} \frac{v(s,a) w(s,a)}{\pi(a|s)} \quad \text{for all } v, w \in T\Delta_{\mathcal{A}}^{\mathcal{S}} \tag{61}$$

the *Kakade metric*. For differentiable $f\colon \Delta_{\mathcal{A}}^{\mathcal{S}} \to \mathbb{R}$, we denote the Riemannian gradient by $\nabla^{\mathrm{K}} f(\pi)$.

The Kakade metric has been referred to as the Fisher-Rao metric on $\Delta_{\mathcal{A}}^{\mathcal{S}}$ as this reduces to the Fisher-Rao metric if $|\mathcal{S}| = 1$, see Kerimkulov et al. (2023). We choose the name Kakade metric here, as there exist multiple extensions of the Fisher-Rao metric to products of simplicies. For example in a game-theoretic context, when considering independently chosen strategies of the players, it might be more natural to work with the product metric, meaning, the sum of the Fisher-Rao metrics over the individual factors, which corresponds to the pullback of the Fisher-Rao metric on the simplex of joint distributions under the independence model (Montúfar et al., 2014; Boll et al., 2024). Other weightings of the Fisher-Rao metrics over the individual factors are also possible, see Montúfar et al. (2014) for an in-depth discussion of different choices and their invariance properties. Further, this specific Riemannian metric has been described as the limit of weighted Fisher-Rao metrics on the finite-horizon path spaces (Bagnell & Schneider, 2003; Peters et al., 2003; Wolfer & Watanabe, 2023). Note that although the Kakade metric is closely connected to the weighted entropy H_{K}, it is not the Hessian metric of H_{K}, in fact, it is not a Hessian metric at all (Müller & Montúfar, 2024, Remark 13).

In general, the Kakade metric is only a pseudo-metric and the following assumption ensures positive definiteness and we make it for the remainder of our analysis.

Assumption B.2 (State exploration). For any policy $\pi \in \mathrm{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$ the discounted state distribution is positive, meaning that $d^{\pi}(s) > 0$ for all $s \in \mathcal{S}$.

Kakade gradient flows are well-posed, meaning that they admit a unique solution $(\pi_t)_{t\in\mathbb{R}_{\geq 0}}$, both in the unregularized case (Alvarez et al., 2004; Müller & Montúfar, 2024) and the regularized case (Müller, 2023; Kerimkulov et al., 2023).

The *policy gradient theorem* states that

$$\partial_{\theta_i} R(\pi) = \frac{1}{1-\gamma} \sum_{s\in\mathcal{S}} d^{\pi_\theta}(s) \sum_{a\in\mathcal{A}} \partial_{\theta_i} \pi_\theta(a|s) A^{\pi_\theta}(s,a) \tag{62}$$

and connects the gradient of the reward to the advantage function (Sutton et al., 1999; Agarwal et al., 2021). Inspired by this, we provide an explicit formula for the gradient of the reward with respect to the Kakade metric.

Theorem B.3 (Gradient with respect to the Kakade metric)**.** *Let Assumption B.2 hold. Then for all $\pi \in \mathrm{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$ and $a \in \mathcal{A}$, $s \in \mathcal{S}$ it holds that*

$$\nabla^{\mathrm{K}} R(\pi)(s,a) = (1-\gamma)^{-1} A^{\pi}(s,a)\pi(a|s). \tag{63}$$

In the proof, we use the following auxiliary result.

Lemma B.4 (Derivatives of state-action distributions, Müller & Montufar (2022))**.** *It holds that*

$$\frac{\partial \nu^{\pi}}{\partial \pi(a|s)} = d^{\pi}(s)(I - \gamma P_{\pi}^{T})^{-1} e_{(s,a)}, \tag{64}$$

where $P_{\pi} \in \Delta_{\mathcal{S}\times\mathcal{A}}^{\mathcal{S}\times\mathcal{A}}$ is given by $P_{\pi}(s',a'|s,a) = \pi(a'|s')P(s'|s,a)$.

Proof of Theorem B.3. First, note that $A^{\pi} \odot \pi \in T\Delta_{\mathcal{A}}^{\mathcal{S}}$ where $(A^{\pi} \odot \pi)(s,a) = A^{\pi}(s,a)\pi(a|s)$ and that R and g^{K} can be extended to a neighborhood of $\Delta_{\mathcal{A}}^{\mathcal{S}}$. It suffices to show

$$(1-\gamma)^{-1} g_{\pi}^{\mathrm{K}}(A^{\pi} \odot \pi, v) = \partial_v R(\pi) \quad \text{for all } v \in T\Delta_{\mathcal{A}}^{\mathcal{S}}.$$

Since $R(\pi) = r^\top \nu^\pi$, we have $\frac{\partial R(\pi)}{\partial \pi(a|s)} = r^\top \frac{\partial \nu^\pi}{\partial \pi(a|s)}$. For any tangent vector $v \in T\Delta_{\mathcal{A}}^{\mathcal{S}}$ we can use Lemma B.4 to compute

$$\begin{aligned}
\partial_v R(\pi) &= \sum_{s\in\mathcal{S}, a\in\mathcal{A}} v(s,a) \cdot \frac{\partial R(\pi)}{\partial \pi(a|s)} \\
&= \sum_{s\in\mathcal{S}, a\in\mathcal{A}} d^\pi(s) v(s,a) \left\langle (I-\gamma P_\pi^T)^{-1} e_{(s,a)}, r \right\rangle \\
&= \sum_{s\in\mathcal{S}, a\in\mathcal{A}} d^\pi(s) v(s,a) \left\langle e_{(s,a)}, (I-\gamma P_\pi)^{-1} r \right\rangle \\
&= (1-\gamma)^{-1} \sum_{s\in\mathcal{S}, a\in\mathcal{A}} d^\pi(s) v(s,a) Q^\pi(s,a) \\
&= (1-\gamma)^{-1} \sum_{s\in\mathcal{S}, a\in\mathcal{A}} d^\pi(s) v(s,a) (Q^\pi(s,a) - V^\pi(s)) \\
&= (1-\gamma)^{-1} g_\pi^{\mathrm{K}}(A^\pi \odot \pi, v),
\end{aligned}$$

where we have used the Bellman equation $Q^\pi = (1-\gamma)(I-\gamma P_\pi)^{-1} r$ as well as $\sum_{a\in\mathcal{A}} v(s,a) = 0$, which holds for tangent vectors $v \in T\Delta_{\mathcal{A}}^{\mathcal{S}}$. □

Note that the gradient of the reward with respect to the Kakade metric is independent of the initial distribution $\mu \in \Delta_{\mathcal{S}}$ both for the regularized and the unregularized reward.

Definition B.5 (Kakade gradient flow)**.** We say that $(\pi_t)_{t\in[0,T)} \subseteq \mathrm{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$ solves the *Kakade gradient flow* if

$$\partial_t \pi_t = \nabla^{\mathrm{K}} R(\pi_t). \tag{65}$$

By Theorem B.3, a solution of the Kakade gradient flow satisfies

$$\partial_t \pi_t(a|s) = (1-\gamma)^{-1} A^{\pi_t}(s,a) \pi_t(a|s). \tag{66}$$

This explicit expression of the gradient flow allows us to provide essentially sharp convergence rates.

B.2. State-action geometry of Kakade gradient flows

Our goal is to use tools from Hessian gradient flows, but the Kakade metric is not a Hessian metric. However, it was shown that the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$ endowed with the Kakade metric is isometric to the state-action polytope endowed with the Hessian metric induced by the conditional entropy (Müller & Montúfar, 2024). This allows us to borrow from the results on Hessian gradient flows.

Definition B.6 (Conditional Fisher-Rao metric)**.** We call the metric $g^{A|S}$ on $\mathrm{int}(\mathscr{D})$ given by

$$g_\nu^{A|S}(v,w) \coloneqq v^\top \nabla^2 H_{A|S}(\nu) w \quad \text{for } v,w \in T\mathscr{D} \tag{67}$$

the *conditional Fisher-Rao metric* and the denote the corresponding gradient by $\nabla^{A|S}$.

It is elementary to check the convexity of $H_{A|S}$ on $\mathbb{R}_{>0}^{\mathcal{A}\times\mathcal{S}}$, but it also easily seen that $\nabla^2 H_{A|S}$ has zero eigenvalues. However, it can be shown that $\nabla^2 H_{A|S}$ is strictly definite on $T\mathscr{D}$ and therefore induces a Riemannian metric on $\mathrm{int}(\mathscr{D})$, see Müller & Montúfar (2024).

The following result relates the Kakade metric and Kakade divergence to the Hessian metric and Bregman divergence of the conditional entropy $H_{A|S}$.

Theorem B.7 (State-action geometry of the Kakade metric, Müller & Montúfar (2024))**.** *Consider a finite Markov decision process and let Assumption B.2 hold. Then the mapping*

$$\begin{aligned}
\Psi\colon (\mathrm{int}(\Delta_{\mathcal{A}}^{\mathcal{S}}), g^{\mathrm{K}}) &\to (\mathrm{int}(\mathscr{D}), g^{A|S}) \\
\pi &\mapsto \nu^\pi \\
\nu(\cdot|\cdot) &\leftarrow\!\shortmid \nu
\end{aligned} \tag{68}$$

between policies and state-action distributions is an isometry.

As isometries map gradient flows to gradient flows, Theorem B.7 implies that $(\pi_t)_{t\geq 0}$ solves the Kakade gradient flow $\partial_t R(\pi_t) = \nabla^{\mathrm{K}} R(\pi_t)$ if and only if $(\nu_t)_{t\geq 0} = (\nu^{\pi_t})_{t\geq 0}$ solves the conditional Fisher-Rao gradient flow $\partial_t \nu_t = \nabla^{A|S} f(\nu_t)$ if the following diagram commutes

$$\begin{tikzcd} \Delta_{\mathcal{A}}^{\mathcal{S}} \arrow[r, "\Psi"] \arrow[dr, "R"'] & \mathscr{D} \arrow[d, "f"] \\ & \mathbb{R} \end{tikzcd}, \quad \text{where} \quad \begin{tikzcd} \pi \arrow[r, mapsto] \arrow[dr, mapsto] & \nu^{\pi} \arrow[d, mapsto] \\ & R(\pi). \end{tikzcd} \tag{69}$$

The isometry property allows us to transfer the theory on Hessian gradient flows of linear programs to Kakade gradient flows despite the Kakade metric not being Hessian.

Corollary B.8 (Implications from Hessian gradient flows)**.** *Let Assumption B.2 hold, denote the set of optimal policies by* $\Pi^\star := \{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}} : R(\pi) = R^\star\}$*, and fix an initial policy* $\pi_0 \in \operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$*. Then the following statements hold:*

1. Well-posedness: *The Kakade gradient flow* (65) *admits a unique global solution* $(\pi_t)_{t\geq 0}$.
2. Central path property: *For all* $t \geq 0$ *it holds that*
$$\pi_t = \operatorname*{arg\,max}_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \Big\{ R(\pi) - t^{-1} D_{\mathrm{K}}(\pi, \pi_0) \Big\}. \tag{70}$$
3. Sublinear convergence: *For all* $t \geq 0$ *we have*
$$0 \leq R^\star - R(\pi_t) \leq \frac{\min_{\pi^\star \in \Pi^\star} D_{\mathrm{K}}(\pi^\star, \pi_0) - D_{\mathrm{K}}(\pi_t, \pi_0)}{t} \leq \frac{\min_{\pi^\star \in \Pi^\star} D_{\mathrm{K}}(\pi^\star, \pi_0)}{t}. \tag{71}$$
4. Implicit bias: *The gradient flow* $(\pi_t)_{t\geq 0}$ *converges globally and it holds that*
$$\lim_{t \to +\infty} \pi_t = \pi^\star = \operatorname*{arg\,min}_{\pi \in \Pi^\star} D_{\mathrm{K}}(\pi, \pi_0). \tag{72}$$

Proof. First, note that both 2 and 3 hold on the maximal existence interval $[0, T_\infty)$ of the gradient flow, which is a direct consequence of Appendix A.3 and Theorem B.7. Further, 4 follows from Appendix A.3 and Theorem B.7 if $T_\infty = +\infty$, hence, it remains to show well-posedness.

The well-posedness of the conditional Fisher-Rao gradient flow of linear programs has been established by Müller (2023), which can be carried over to the Kakade gradient flow using the isometry. Here, we offer a different proof that utilizes the central path property of the Kakade gradient flow. The local well-posedness of the gradient flow follows from the Picard-Lindelöf theorem and it remains to show that $T_\infty := \inf\{t > 0 : \inf_{s\in[0,t)} \operatorname{dist}(\pi_s, \partial\Delta_{\mathcal{A}}^{\mathcal{S}}) = 0\} = +\infty$. Assume that $T_\infty < +\infty$, then it is elementary to check that $\pi_t \to \pi_{T_\infty}$ for $t \nearrow T_\infty$, where π_{T_∞} solves the entropy-regularized problem with strength $\tau = T_\infty^{-1}$. Note that $\pi_{T_\infty} \in \operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$ contradicting $\inf_{t\in[0,T_\infty]} \operatorname{dist}(\pi_t, \partial\Delta_{\mathcal{A}}^{\mathcal{S}}) = 0$. □

Remark B.9 (Implicit and algorithmic bias). In the case of multiple optimal policies, the Kakade gradient flow will converge towards the Kakade projection of the initial policy to the set of optimal policies. The selection of gradient schemes of a particular optimizer in the case of multiple optima is commonly referred to as the *implicit* or *algorithmic bias* of the method. In the context of reinforcement learning the implicit bias was studied for discrete time natural policy gradient and policy mirror descent methods. First, for a natural actor-critic scheme in linear Markov decision processes the Kakade divergence of the optimization trajectory $(\pi_k)_{k\in\mathbb{N}}$ to the maximum entropy policy $\pi^\star$ is bounded by $D_{\mathrm{K}}(\pi^\star, \pi_k) \leq \log k + (1-\gamma)^{-2}$ (Hu et al., 2021). This control on the Bregman divergence to the maximum entropy policy ensures that the probability of selecting an optimal action $a \in A_s^\star$ decays at most like $\pi_k(a|s) \geq ck^{-1}$, but fails to identify the limiting policy.

On the other hand, a policy mirror descent scheme with decaying entropy regularization strength τ_k was studied by Li et al. (2023). Here, it is shown that for several choices of regularization strengths and stepsizes the resulting policies π_k converge to the maximum entropy optimal policy thereby characterizing the algorithmic bias of this approach. Whereas this analysis can characterize the limit of the optimization scheme it utilizes decaying explicit regularization.

In contrast, our implicit bias result works in continuous time and utilizes the correspondence between the gradient flows and regularized problems. This allows us to show convergence towards the (generalized) maximum entropy optimal policies without relying on explicit regularization.

Where Corollary B.8 provides sublinear convergence with respect to the reward function, we use this to show sublinear convergence of the advantage function.

Proposition B.10 (Sublinear convergence of advantage functions)**.** *Let Assumption B.2 hold, denote the set of optimal policies by* $\Pi^\star := \{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}} : R(\pi) = R^\star\}$*, and fix an initial policy* $\pi_0 \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ *and denote the generalized maximum entropy optimal policy by* $\pi^\star = \arg\min_{\pi\in\Pi^\star} D_{\mathrm{K}}(\pi, \pi_0)$*. Then for all* $t \geq 0$*, it holds that*

$$0 \leq V^\star(s) - V^{\pi_t}(s) \leq \frac{D_{\mathrm{K}}(\pi^\star, \pi_0)}{t} \quad \textit{for all } s \in \mathcal{S}. \tag{73}$$

In particular for any $s \in \mathcal{S}$ *and* $a \in \mathcal{A}$*, it holds that*

$$A^{\pi_t}(s,a) \leq A^\star(s,a) + \frac{D_{\mathrm{K}}(\pi^\star, \pi_0)}{t} \quad \textit{and} \tag{74}$$

$$A^{\pi_t}(s,a) \geq A^\star(s,a) - \frac{\gamma D_{\mathrm{K}}(\pi^\star, \pi_0)}{t}. \tag{75}$$

Proof. Let us fix $s \in \mathcal{S}$. By Theorem B.3 the Kakade gradient flow $(\pi_t)_{t\geq 0}$ is independent of the initial distribution μ, as long as it has full support. Let us pick some μ_n with full support and $\mu_n \to \delta_s$, then $R^{\mu_n}(\pi) \to V^\pi(s)$ for any π. By Corollary A.10 we have

$$0 \leq (R^{\mu_n})^\star - R^{\mu_n}(\pi_t) \leq \frac{D_{\mathrm{K}}(\pi^\star, \pi_0)}{t} \quad \text{for all } t \geq 0, n \in \mathbb{N},$$

which yields (73) for $n \to +\infty$. We use this to estimate

$$\begin{aligned} Q^\star(s,a) - Q^{\pi_t}(s,a) &= r(s,a) + \gamma \sum_{s'} V^\star(s') P(s'|s,a) - r(s,a) - \gamma \sum_{s'} V^{\pi_t}(s') P(s'|s,a) \\ &= \gamma \sum_{s'} (V^\star(s') - V^{\pi_t}(s')) P(s'|s,a) \\ &\leq \frac{\gamma D_{\mathrm{K}}(\pi^\star, \pi_0)}{t}. \end{aligned}$$

In particular, this implies

$$A^\star(s,a) - A^{\pi_t}(s,a) = Q^\star(s,a) - V^\star(s) - Q^{\pi_t}(s,a) + V^{\pi_t}(s) \geq V^{\pi_t}(s) - V^\star(s) \geq -\frac{D_{\mathrm{K}}(\pi^\star, \pi_0)}{t}$$

and similarly

$$A^\star(s,a) - A^{\pi_t}(s,a) \leq Q^\star(s,a) - Q^{\pi_t}(s,a) \leq \frac{\gamma D_{\mathrm{K}}(\pi^\star, \pi_0)}{t}.$$

□

C. Essentially Sharp Analysis of Entropy Regularization

We now improve the sublinear convergence guarantee from Proposition A.3 on the entropy regularization error. To this end we work with the interpretation of the solutions of the regularized problems as solutions of the Kakade gradient flow and employ the explicit expression (66) and the sublinear convergence guarantee from Proposition B.10 to establish linear convergence of Kakade gradient flows. We complement this with a lower bound that matches the upper bound up to a polynomial factor. We work in the following setting.

Setting C.1. *Consider a finite discounted Markov decision process* $(\mathcal{S}, \mathcal{A}, P, \gamma, r)$*, an initial distribution* $\mu \in \Delta_{\mathcal{S}}$ *and fix a policy* $\pi_0 \in \mathrm{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$*, assume that Assumption B.2 holds and denote the optimal reward by* $R^\star := \max\{R(\pi) : \pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}\}$*. Further, let* $(\pi_t)_{t\geq 0}$ *be the unique global solution of the Kakade policy gradient flow (65) or equivalently the solutions of the entropy-regularized problems*

$$\pi_t = \arg\max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \Big\{ R(\pi) - t^{-1} D_{\mathrm{K}}(\pi, \pi_0) \Big\}. \tag{76}$$

We denote the generalized maximum entropy optimal policy by $\pi^\star = \arg\min_{\pi\in\Pi^\star} D_{\mathrm{K}}(\pi,\pi_0)$*, where* $\Pi^\star \coloneqq \{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}} : R(\pi) = R^\star\}$ *denotes the set of optimal policies. Finally, we set*

$$\Delta \coloneqq -(1-\gamma)^{-1}\max\left\{A^\star(s,a) : A^\star(s,a) \neq 0, s \in \mathcal{S}, a \in \mathcal{A}\right\}. \tag{77}$$

Note that $\Delta > 0$ unless every action is optimal in every state. We can interpret Δ as the minimal suboptimality of a suboptimal action under $A^\star$.

C.1. Convergence in value

First, we study the entropy regularization error in the objective function, meaning that we study the reward achieved by the optimal regularized policies.

Theorem C.2 (Convergence in value)**.** *Consider Setting C.1 and set* $c \coloneqq (1-\gamma)^{-1}D_{\mathrm{K}}(\pi^\star,\pi_0)$*. For any* $t \geq 1$ *it holds that*

$$R^\star - R(\pi_t) \leq \frac{2\|r\|_\infty}{1-\gamma}\cdot e^{-\Delta(t-1)+c\log t} \quad \textit{as well as} \tag{78}$$

$$R^\star - R(\pi_t) \geq \Delta\cdot\left(\min_{s\in\mathcal{S}} d^{\pi_t}(s)\sum_{a\notin A_s^\star}\pi_0(a|s)\right)\cdot e^{-\Delta(t-1)-\gamma c\log t-2\|r\|_\infty}. \tag{79}$$

Note that the coefficient of the lower bound depends on t. However, the coefficient does not become arbitrarily small as $\min_s d^{\pi_t}(s) \to \min_s d^{\pi^\star}(s) > 0$ for $t \to +\infty$. Further, the policies $(\pi_t)_{t\geq 0}$ do not depend on μ, where d^{π_t} does. If we choose μ to be the uniform distribution, then $d^\pi(s) \geq (1-\gamma)|\mathcal{S}|^{-1}$, which yields a lower bound of

$$R^\star - R(\pi_t) \geq (1-\gamma)\Delta|\mathcal{S}|^{-1}\sum_{a\notin A_s^\star}\pi_0(a|s)\cdot e^{-\Delta(t-1)-\gamma c\log t-2\|r\|_\infty},$$

where the coefficient is independent of t.

The above result ensures that the probability $\pi_t(a|s)$ of selecting a suboptimal action a decays exponentially which implies exponential convergence of the reward $R(\pi_t)$ achieved by the policies.

We combine the explicit expression (66) as well as the sublinear convergence of the advantage function towards the optimal advantage function $A^\star$ to bound the individual entries of the policies along the gradient flow trajectory.

Lemma C.3. *Consider Setting C.1. Then for all* $t \geq t_0 > 0$*,* $s \in \mathcal{S}$*, and* $a \in \mathcal{A}$ *it holds that*

$$\pi_t(a|s) \leq \pi_{t_0}(a|s)\exp\left(\frac{A^\star(s,a)(t-t_0) + D_{\mathrm{K}}(\pi^\star,\pi_0)\log(\frac{t}{t_0})}{1-\gamma}\right) \tag{80}$$

$$\pi_t(a|s) \geq \pi_{t_0}(a|s)\exp\left(\frac{A^\star(s,a)(t-t_0) - \gamma D_{\mathrm{K}}(\pi^\star,\pi_0)\log(\frac{t}{t_0})}{1-\gamma}\right) \tag{81}$$

and for $t \geq 1$ *it holds that*

$$\pi_t(a|s) \leq \pi_0(a|s)\exp\left(\frac{A^\star(s,a)(t-1) + D_{\mathrm{K}}(\pi^\star,\pi_0)\log t + 2\|r\|_\infty}{1-\gamma}\right) \tag{82}$$

$$\pi_t(a|s) \geq \pi_0(a|s)\exp\left(\frac{A^\star(s,a)(t-1) - \gamma D_{\mathrm{K}}(\pi^\star,\pi_0)\log t - 2\|r\|_\infty}{1-\gamma}\right). \tag{83}$$

Proof. As the policies $(\pi_t)_{t\geq 0}$ solve the Kakade policy gradient flow in $\Delta_{\mathcal{A}}^{\mathcal{S}}$ we have

$$\partial_t\pi_t(a|s) = (1-\gamma)^{-1}A^{\pi_t}(s,a)\pi_t(a|s). \tag{84}$$

By Proposition B.10 it holds that

$$(1-\gamma)^{-1}\left(A^\star(s,a) - \frac{\gamma D_{\mathrm{K}}(\pi^\star,\pi_0)}{t}\right)\pi_t(a|s) \leq \partial_t\pi_t(a|s) \leq (1-\gamma)^{-1}\left(A^\star(s,a) + \frac{D_{\mathrm{K}}(\pi^\star,\pi_0)}{t}\right)\pi_t(a|s).$$

Now, Grönwall's inequality yields

$$\pi_T(a|s) \leq \pi_{t_0}(a|s) \exp\left(\frac{A^\star(s,a)(T-t_0) + D_{\mathrm{K}}(\pi^\star,\pi_0)\int_{t_0}^T t^{-1}\mathrm{d}t}{1-\gamma}\right)$$
$$= \pi_{t_0}(a|s) \exp\left(\frac{A^\star(s,a)(T-t_0) + D_{\mathrm{K}}(\pi^\star,\pi_0)\log T - D_{\mathrm{K}}(\pi^\star,\pi_0)\log t_0}{1-\gamma}\right)$$

as well as

$$\pi_T(a|s) \geq \pi_{t_0}(a|s) \exp\left(\frac{A^\star(s,a)(T-t_0) - \gamma D_{\mathrm{K}}(\pi^\star,\pi_0)\int_{t_0}^T t^{-1}\mathrm{d}t}{1-\gamma}\right)$$
$$= \pi_{t_0}(a|s) \exp\left(\frac{A^\star(s,a)(T-t_0) - \gamma D_{\mathrm{K}}(\pi^\star,\pi_0)\log T + \gamma D_{\mathrm{K}}(\pi^\star,\pi_0)\log t_0}{1-\gamma}\right).$$

Further, note that $\|A^\pi\|_\infty \leq \|Q^\pi\|_\infty + \|V^\pi\|_\infty \leq 2\|r\|_\infty$ and hence Grönwall yields

$$\pi_0(a|s)e^{-\frac{2\|r\|_\infty t_0}{1-\gamma}} \leq \pi_{t_0}(a|s) \leq \pi_0(a|s)e^{\frac{2\|r\|_\infty t_0}{1-\gamma}}$$

and choosing $t_0 = 1$ finishes the proof. □

Lemma C.4 (Sub-optimality gap)**.** *Consider a discrete discounted Markov decision process. Then for any policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ it holds that*

$$\Delta \cdot \min_{s\in\mathcal{S}} \sum_{s\in\mathcal{S}, a\notin A_s^\star} d^\pi(s)\pi(a|s) \leq R^\star - R(\pi) \leq \frac{2\|r\|_\infty}{1-\gamma} \sum_{s\in\mathcal{S}, a\notin A_s^\star} d^\pi(s)\pi(a|s), \tag{85}$$

where $A_s := \{a \in \mathcal{A} : A^\star(s,a) = 0\}$ denotes the set of optimal actions in $s \in \mathcal{S}$.

Proof. By the performance difference Lemma E.1, we have

$$(1-\gamma)(R^\star - R(\pi)) = -\sum_{s\in\mathcal{S}, a\notin A_s^\star} d^\pi(s)\pi(a|s)A^\star(s,a) \leq \|A^\star\|_\infty \sum_{s\in\mathcal{S}, a\notin A_s^\star} d^\pi(s)\pi(a|s)$$

Further, note that $|A^\star(s,a)| \leq |Q^\star(s,a)| + |V^\star(s)| \leq 2\|r\|_\infty$. The lower bound follows with an analog argument as

$$(1-\gamma)(R^\star - R(\pi)) = -\sum_{s\in\mathcal{S}, a\notin A_s^\star} d^\pi(s)\pi(a|s)A^\star(s,a) \geq (1-\gamma)\Delta \sum_{s\in\mathcal{S}, a\notin A_s^\star} d^\pi(s)\pi(a|s).$$

□

Proof of Theorem C.2. We use Lemma C.4 together with Lemma C.3 and estimate

$$\sum_{s\in\mathcal{S}, a\notin A_s^\star} d^{\pi_t}(s)\pi_t(a|s) \leq \sum_{s\in\mathcal{S}, a\notin A_s^\star} d^{\pi_t}(s)\pi_1(a|s)e^{-\Delta(t-1)+\gamma c\log t} \leq e^{-\Delta(t-1)+\gamma c\log t},$$

which yields (78). For the lower bound, we fix $s_0 \in \mathcal{S}$ and $a_0 \in \mathcal{A}$ with $(1-\gamma)^{-1}A^\star(s_0,a_0) = -\Delta$ and estimate

$$\sum_{s\in\mathcal{S}, a\notin A_s^\star} d^{\pi_t}(s)\pi_t(a|s) \geq \sum_{s\in\mathcal{S}, a\notin A_s^\star} d^{\pi_t}(s)\pi_0(a|s)e^{(1-\gamma)^{-1}A^\star(s,a)(t-1)-c\log t-2\|r\|_\infty}$$
$$\geq d^{\pi_t}(s_0) \sum_{a\notin A_{s_0}^\star} \pi_0(a|s_0)e^{-\Delta(t-1)-c\log t-2\|r\|_\infty}$$
$$\geq \min_{s\in\mathcal{S}} \left\{ d^{\pi_t}(s) \sum_{a\notin A_s^\star} \pi_0(a|s) \right\} \cdot e^{-\Delta(t-1)-c\log t-2\|r\|_\infty}$$

□

C.2. Convergence of policies

Having studied the decay of the suboptimality gap $R^\star - R(\pi_t)$, we now give sharp bounds for convergence of the policies measured in the Kakade divergence.

Theorem C.5 (Convergence of policies)**.** *Consider Setting C.1 and set $c \coloneqq (1-\gamma)^{-1} D_{\mathrm{K}}(\pi^\star, \pi_0)$. For any $t \geq 1$ it holds that*

$$D_{\mathrm{K}}(\pi^\star, \pi_t) \leq \frac{e^{-\Delta(t-1)+c\log t}}{1-e^{-\Delta(t-1)+c\log t}} \quad \textit{as well as} \tag{86}$$

$$D_{\mathrm{K}}(\pi^\star, \pi_t) \geq \min_{s\in\mathcal{S}, a\notin A_s^\star} d^\star(s)\pi_0(a|s) \cdot e^{-\Delta(t-1)-\gamma c\log t - 2\|r\|_\infty}, \tag{87}$$

where the upper bound holds if $e^{-\Delta(t-1)+c\log t} < 1$ which is satisfied for $t > 0$ large enough.

Note that the denominator of the upper bound converges to 1 for $t \to +\infty$. We will use the following auxiliary result in the proof of Theorem C.5.

Lemma C.6 (Information projection onto faces)**.** *Consider a finite set $\mathcal{X}$ and the face $F \coloneqq \{\mu \in \Delta_{\mathcal{X}} : \mu_x = 0 \textit{ for all } x \notin X\}$ of the simplex $\Delta_{\mathcal{X}}$ for some $X \subseteq \mathcal{X}$. Then*

$$\min_{\mu\in F} D_{\mathrm{KL}}(\mu,\nu) = -\log\left(\sum_{x\in X}\nu_x\right). \tag{88}$$

In particular, for any $\mu, \nu \in \Delta_{\mathcal{X}}$ we have

$$D_{\mathrm{KL}}(\mu,\nu) \geq -\log\left(\sum_{x\in\mathrm{supp}(\mu)}\nu_x\right). \tag{89}$$

Proof. We set $g(\mu) \coloneqq D_{\mathrm{KL}}(\mu,\nu)$ and consider the information projection $\hat{\nu} = \arg\min_{\mu\in F} D_{\mathrm{KL}}(\mu,\nu)$ of $\nu \in \Delta_{\mathcal{X}}$ onto F, which is characterized by

$$v^\top \nabla g(\hat{\nu}) = 0 \quad \text{for all } v \in TF = \mathrm{span}\{\nu_{x_1} - \nu_{x_2} : x_1, x_2 \in X\}.$$

As $\partial_x g(\mu) = \log(\frac{\mu_x}{\nu_x}) + 1$ this is equivalent to

$$\log\left(\frac{\hat{\nu}_{x_1}}{\nu_{x_1}}\right) = \log\left(\frac{\hat{\nu}_{x_2}}{\nu_{x_2}}\right) \quad \text{for all } x_1, x_2 \in X.$$

This implies $\frac{\hat{\nu}_x}{\nu_x}$ is constant for $x \in X$ and hence we obtain

$$\hat{\nu}_x = \begin{cases} \frac{\nu_x}{\sum_{x'\in X}\nu_{x'}} & \text{for } x \in X \\ 0 & \text{for } x \notin X. \end{cases}$$

Setting $c^{-1} \coloneqq \sum_{x\in X}\nu_x$ we obtain

$$\min_{\mu\in F} D_{\mathrm{KL}}(\mu,\nu) = D_{\mathrm{KL}}(\hat{\nu},\nu) = \sum_{x\in X} c\nu_x \log\left(\frac{c\nu_x}{\nu_x}\right) = \log c.$$

To show (89), we choose $X \coloneqq \mathrm{supp}(\mu)$ and use $D_{\mathrm{KL}}(\mu,\nu) \geq \min_{\xi\in F} D_{\mathrm{KL}}(\xi,\nu)$. □

Lemma C.7. *Consider a finite MDP and let $\pi^\star \in \Pi^\star$ be the Kakade projection of $\pi \in \mathrm{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$ onto the set of optimal policies $\Pi^\star$. Then it holds that*

$$\sum_{s\in\mathcal{S}} d^{\pi^\star}(s) \sum_{a\notin A_s^\star} \pi(a|s) \leq D_{\mathrm{K}}(\pi^\star,\pi) \leq \frac{\max_{s\in\mathcal{S}}\sum_{a\notin A_s^\star}\pi(a|s)}{1-\max_{s\in\mathcal{S}}\sum_{a\notin A_s^\star}\pi(a|s)}, \tag{90}$$

where $A_s \coloneqq \{a \in \mathcal{A} : A^\star(s,a) = 0\}$ denotes the set of optimal actions in $s \in \mathcal{S}$.

Proof. We first prove the upper bound. To this end, we consider the state-wise information projection

$$\hat{\pi}(\cdot|s) = \arg\min \left\{ D_{\mathrm{KL}}(\mu, \pi(\cdot|s)) : \mu \in \Delta_{\mathcal{A}}, \mathrm{supp}(\mu) \subseteq A_s^\star \right\}$$

of the policy π to the set of optimal policies. By concavity we have $-\log(1-h) \leq \frac{h}{1-h}$ for $h < 1$ and using Lemma C.6 we estimate

$$\begin{aligned}
D_{\mathrm{K}}(\pi^\star, \pi) &\leq D_{\mathrm{K}}(\hat{\pi}, \pi) \\
&= -\sum_{s\in\mathcal{S}} d^{\hat{\pi}}(s) D_{\mathrm{KL}}(\hat{\pi}(\cdot|s), \pi_t(\cdot|s)) \\
&= -\sum_{s\in\mathcal{S}} d^{\hat{\pi}}(s) \log\left(\sum_{a\in A_s^\star} \pi(a|s)\right) \\
&= -\sum_{s\in\mathcal{S}} d^{\hat{\pi}}(s) \log\left(1 - \sum_{a\notin A_s^\star} \pi(a|s)\right) \\
&\leq \sum_{s\in\mathcal{S}} d^{\hat{\pi}}(s) \cdot \frac{\sum_{a\notin A_s^\star} \pi(a|s)}{1 - \sum_{a\notin A_s^\star} \pi(a|s)}. \\
&\leq \sum_{s\in\mathcal{S}} d^{\hat{\pi}}(s) \cdot \frac{\max_{s'\in\mathcal{S}} \sum_{a\notin A_{s'}^\star} \pi(a|s')}{1 - \max_{s'\in\mathcal{S}} \sum_{a\notin A_{s'}^\star} \pi(a|s')}. \\
&= \frac{\max_{s\in\mathcal{S}} \sum_{a\notin A_s^\star} \pi(a|s)}{1 - \max_{s\in\mathcal{S}} \sum_{a\notin A_s^\star} \pi(a|s)}.
\end{aligned}$$

Similar to the upper bound, we use $-\log(1-h) \geq h$ for $h < 1$ and Lemma C.6 to estimate

$$\begin{aligned}
D_{\mathrm{K}}(\pi^\star, \pi) &= \sum_{s\in\mathcal{S}} d^\star(s) D_{\mathrm{KL}}(\pi^\star(\cdot|s), \pi(\cdot|s)) \\
&\geq -\sum_{s\in\mathcal{S}} d^\star(s) \log\left(\sum_{a\in A_s^\star} \pi(a|s)\right) \\
&= -\sum_{s\in\mathcal{S}} d^\star(s) \log\left(1 - \sum_{a\notin A_s^\star} \pi(a|s)\right) \\
&\geq \sum_{s\in\mathcal{S}} d^\star(s) \sum_{a\notin A_s^\star} \pi(a|s).
\end{aligned}$$

□

Proof of Theorem C.5. As expected, we work with Lemma C.7. By Lemma C.3 we have

$$\sum_{a\notin A_s^\star} \pi_t(a|s) \leq \sum_{a\notin A_s^\star} \pi_1(a|s) e^{-\Delta(t-1)+c\log t} \leq e^{-\Delta(t-1)+c\log t}$$

and thus

$$\frac{\max_{s\in\mathcal{S}} \sum_{a\notin A_s^\star} \pi_t(a|s)}{1 - \max_{s\in\mathcal{S}} \sum_{a\notin A_s^\star} \pi_t(a|s)} \leq \frac{e^{-\Delta(t-1)+c\log t}}{1 - e^{-\Delta(t-1)+c\log t}}$$

if $e^{-\Delta(t-1)+c\log t} < 1$ and Lemma C.7 yields the upper bound. For the lower bound, we fix $s_0 \in \mathcal{S}$ and $a_0 \in \mathcal{A}$ with $A^\star(s,a) = -\Delta$ and estimate

$$\begin{aligned}\sum_{s\in\mathcal{S}} d^\star(s) \sum_{a\notin A_s^\star} \pi_t(a|s) &\geq \sum_{s\in\mathcal{S}} d^\star(s) \sum_{a\notin A_s^\star} \pi_0(a|a) e^{A^\star(s,a)(t-1)-\gamma c\log t - 2\|r\|_\infty} \\ &\geq d^\star(s_0)\pi_0(a_0|s_0) e^{-\Delta(t-1)-\gamma c\log t - 2\|r\|_\infty} \\ &\geq \min_{s\in\mathcal{S}, a\notin A_s^\star} d^\star(s)\pi_0(a|s) \cdot e^{-\Delta(t-1)-\gamma c\log t - 2\|r\|_\infty}.\end{aligned}$$

□

D. Overall Error Analysis for Regularized Natural Policy Gradients

Entropy regularization is commonly added to encourage exploration and accelerate the optimization process, however, the unregularized objective is still the objective criterion one wishes to optimize. Hence, it is a natural question to ask what accuracy one can achieve with a budget of k iterations with a method that aims to optimize the regularized reward. One way to approach this is to use the error decomposition

$$0 \leq R^\star - R(\pi_k) = R^\star - R(\pi_\tau^\star) + R(\pi_\tau^\star) - R_\tau(\pi_\tau^\star) + R_\tau(\pi_\tau^\star) - R_\tau(\pi_k) + R_\tau(\pi_k) - R(\pi_k), \tag{91}$$

Applying the $R(\pi) - R_\tau(\pi) = O(\tau)$ bound on the entropy regularization error gives

$$0 \leq R^\star - R(\pi_k) = O(\tau) + R_\tau(\pi_\tau^\star) - R_\tau(\pi_k) = O(\tau + e^{-\tau\eta k}) = O\left(\frac{\log k}{\eta k}\right) \tag{92}$$

for entropy-regularized natural policy gradients with stepsize $\eta = \frac{\log k}{k} > 0$, see Cen et al. (2021). See also Sethi et al. (2024) for an $O(\frac{1}{t})$ guarantee of entropy-regularized natural policy gradient flows with $\tau = \frac{1}{t}$. In contrast, unregularized natural policy gradient achieves an exponential convergence rate of $\tilde{O}(e^{-\Delta\eta k})$, see Khodadadian et al. (2022); Liu et al. (2024). We use our tight estimate on the regularization error and obtain the following improved guarantee for entropy-regularized natural policy gradients.

Theorem D.1 (Overall error analysis)**.** *Consider a regularization strength $\tau \in (0,1]$ and consider the entropy-regularized reward $R_\tau(\pi) = R(\pi) - \tau D_{\mathrm{K}}(\pi, \pi_{\text{unif}})$, π_{unif} denotes the uniform policy, meaning $\pi_{\text{unif}}(a|s) = |A|^{-1}$ for all $a \in \mathcal{A}$, $s \in \mathcal{S}$. We denote the optimal entropy regularized policy by $\pi_\tau^\star = \arg\max_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}} R_\tau(\pi)$ and the maximum entropy optimal policy by $\pi^\star$, meaning $\pi^\star(a|s) = |A_s^\star|^{-1}$ if $a \in A_s^\star$ is an optimal action. Assume that $(\pi_k)_{k\in\mathbb{N}}$ be the iterates produced by natural policy ascent with a log-linear tabular policy parametrization with stepsize $\eta > 0$. Then it holds that*

$$R^\star - R(\pi_{k+1}) \leq \frac{2\|r\|_\infty e^\Delta}{1-\gamma} \cdot \tau^{-c} e^{-\Delta\tau^{-1}} + \frac{2\cdot|\mathcal{S}|\cdot\|r\|_\infty C^{\frac{1}{2}}}{1-\gamma} \cdot \tau^{-\frac{1}{2}} e^{-\frac{\eta\tau(k-1)}{2}}, \tag{93}$$

where $c = D_{\mathrm{K}}(\pi^\star, \pi_{\text{unif}}) = \sum_{s\in\mathcal{S}} d^{\pi^\star}(s) \log\frac{|\mathcal{A}|}{|A_s^\star|} \leq \log|\mathcal{A}|$ and

$$C = \|Q_\tau^{\pi_\tau^\star} - Q_\tau^{\pi^{(0)}}\|_\infty + 2\tau\left(1 - \frac{\eta\tau}{1-\gamma}\right)\|\log\pi_\tau^\star - \log\pi_0\|_\infty. \tag{94}$$

In particular, choosing $\tau = \sqrt{\frac{2\Delta}{\eta k}}$ yields

$$R^\star - R(\pi_k) = O\left(\left((\eta k)^{\frac{c}{2}} + (\eta k)^{\frac{1}{4}}\right) \cdot e^{-\sqrt{\frac{\Delta\eta k}{2}}}\right). \tag{95}$$

In practice, one does not have access to the problem-dependent constant $\Delta > 0$ without solving the reward optimization problem. However, setting $\tau = (\alpha\eta k)^{-\frac{1}{2}}$ for some $\alpha > 0$ yields an overall error estimate of

$$R^\star - R(\pi_k) = O\left(\left(\frac{\eta k}{\alpha}\right)^{\frac{c}{2}} e^{-\Delta\sqrt{\alpha\eta k}} + \left(\frac{\eta k}{\alpha}\right)^{\frac{1}{4}} e^{-\sqrt{\alpha\eta k}}\right). \tag{96}$$

In our proof, we use the following stability result on the reward function.

Proposition D.2 (Lipschitz-continuity of the reward Müller (2023))**.** *It holds that*

$$|R(\pi_1) - R(\pi_2)| \leq \frac{\|r\|_\infty}{1-\gamma} \cdot \|\pi_1 - \pi_2\|_1 \quad \textit{for all } \pi_1, \pi_2 \in \Delta_{\mathcal{A}}^{\mathcal{S}}. \tag{97}$$

Now we can upper bound the suboptimality gap $R^\star - R(\pi)$ of a policy in terms of $R^\star - R(\pi_\tau^\star)$ and the distance between π and $\pi_\tau^\star$.

Lemma D.3. *For any policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ we have*

$$0 \leq R^\star - R(\pi) \leq R^\star - R(\pi_\tau^\star) + \frac{\sqrt{2} \cdot |\mathcal{S}| \cdot \|r\|_\infty}{1-\gamma} \cdot \|\log \pi_\tau^\star - \log \pi\|_\infty^{\frac{1}{2}}. \tag{98}$$

Proof. First, note that $R^\star - R(\pi) = R^\star - R(\pi_\tau^\star) + R(\pi_\tau^\star) - R(\pi)$. Using the Lipschitz continuity of the reward from Proposition D.2 as well as Pinsker's and Jensen's inequality we estimate

$$\begin{aligned}
R(\pi_\tau^\star) - R(\pi) &\leq \frac{\|r\|_\infty}{1-\gamma} \cdot \|\pi_\tau^\star - \pi\|_1 \\
&= \frac{\|r\|_\infty}{1-\gamma} \cdot \sum_{s \in \mathcal{S}} \|\pi_\tau^\star(\cdot|s) - \pi(\cdot|s)\|_1 \\
&\leq \frac{\sqrt{2} \cdot \|r\|_\infty}{1-\gamma} \cdot \sum_{s \in \mathcal{S}} D_{\mathrm{KL}}(\pi_\tau^\star(\cdot|s), \pi(\cdot|s))^{\frac{1}{2}} \\
&= \frac{\sqrt{2} \cdot \|r\|_\infty}{1-\gamma} \cdot \sum_{s \in \mathcal{S}} \left(\sum_{a \in \mathcal{A}} \pi_\tau^\star(a|s) \log \frac{\pi_\tau^\star(a|s)}{\pi(a|s)} \right)^{\frac{1}{2}} \\
&\leq \frac{\sqrt{2} \cdot \|r\|_\infty}{1-\gamma} \cdot \sum_{s \in \mathcal{S}} \left(\sum_{a \in \mathcal{A}} \pi_\tau^\star(a|s) \|\log \pi_\tau^\star - \log \pi\|_\infty \right)^{\frac{1}{2}} \\
&= \frac{\sqrt{2} \cdot |\mathcal{S}| \cdot \|r\|_\infty}{1-\gamma} \cdot \|\log \pi_\tau^\star - \log \pi\|_\infty^{\frac{1}{2}}.
\end{aligned}$$

□

Lemma D.3 can by used in combination with any result bounding $D_{\mathrm{K}}(\pi_\tau^\star, \pi_k)$ for a policy optimization technique. If only bounds on $R_\tau(\pi_\tau^\star) - R_\tau(\pi_k)$ are available, one can also use the local bound $D_{\mathrm{K}}(\pi_\tau^\star, \pi) \leq \omega\tau^{-1}(R_\tau(\pi_\tau^\star) - R_\tau(\pi))$, which holds in a neighborhood of $\pi_\tau^\star$ that depends on $\omega \in (0,1)$, see (Müller & Montúfar, 2024, Lemma 29). Here, we limit our discussion to entropy-regularized natural policy gradient methods, which are known to converge linearly.

Theorem D.4 (Convergence of entropy-regularized NPG,Cen et al. (2021))**.** *Consider natural policy gradient with a tabular softmax policy parametrization, a fixed regularization strength $\tau > 0$ and stepsize $\eta > 0$ and denote the iterates of the natural policy gradient updates by $(\pi_k)_{k \in \mathbb{N}}$. Then for any $k \in \mathbb{N}$ it holds that*

$$\|Q_\tau^{\pi_\tau^\star} - Q_\tau^{\pi_{k+1}}\|_\infty \leq C(1-\eta\tau)^k \quad \textit{and} \tag{99}$$

$$\|\log \pi_\tau^\star - \log \pi_{k+1}\|_\infty \leq 2C\tau^{-1}(1-\eta\tau)^k, \tag{100}$$

where C is defined in (94)

Now we can provide the following estimate on the performance of entropy-regularized natural policy gradients measured in the unregularized reward.

Proof of Theorem D.1. Through a direct combination of Lemma D.3, Theorem C.2, and Theorem D.4 we obtain (93). □

Recently, a sharp asymptotic analysis of entropy-regularized natural policy gradient methods has been conducted showing $R_\tau^\star - R(\pi_k) = O((1+\eta\tau)^{-2k})$ compared to the $O((1-\eta\tau)^k)$ convergence of Theorem D.4. However, an analogous

estimate on the convergence of the policies as well as a control on the entry times after which the rate holds is missing, which prevents us from using it in our analysis. Unregularized natural policy gradients essentially achieve $O(e^{-\Delta\eta k})$ convergence rate (Khodadadian et al., 2022). Hence, although Theorem D.1 improves existing $O(\frac{\log k}{k})$ guarantees, the provided rate is still asymptotically slower compared to unregularized natural policy gradients. Where we consider a fixed regularization and step size, $O(\gamma^k)$-convergence was established when exponentially decreasing the regularization and increasing the step size (Li et al., 2023), which can also be achieved by unregularized policy mirror descent, where the rate $O(\gamma^k)$ is known to be optimal (Johnson et al., 2024). For the small stepsize limit $\eta \to 0$ the updates of the regularized natural policy gradient scheme follow the regularized Kakade gradient flow recently studied by Kerimkulov et al. (2023). These results complement the discrete-time analysis and ensure that $D_{\mathrm{K}}(\pi_\tau^\star, \pi_t) \leq e^{-\tau t} D_{\mathrm{K}}(\pi_\tau^\star, \pi_0)$, see (Kerimkulov et al., 2023, Equation (61)). An analog treatment to the discrete-time case yields an overall estimate of $O(e^{-\sqrt{2^{-1}\Delta t}})$ for the regularized flow with strength $\tau = \sqrt{2\Delta t^{-1}}$, compared to the existing $O(t^{-1})$ guarantee by Sethi et al. (2024).

E. Auxiliary Results

E.1. Performance difference lemma

We recall an expression of the difference between the rewards of two policies in terms of the advantage function and the state-action distribution and for a proof we refer to (Agarwal et al., 2021, Lemma 2).

Lemma E.1 (Performance difference)**.** *For any two policies $\pi_1, \pi_2 \in \Delta_{\mathcal{S}}^{\mathcal{A}}$ it holds that*

$$R(\pi_1) - R(\pi_2) = \frac{\langle \nu^{\pi_1}, A^{\pi_2} \rangle_{\mathcal{S}\times\mathcal{A}}}{1-\gamma}. \tag{101}$$

E.2. Pythagorean theorem in Bregman divergences

For the sake of completeness, we provide the proof of a generalized Pythagorean theorem.

Proposition E.2 (Pythagoras for Bregman divergences)**.** *Consider a convex differentiable function $\phi\colon \Omega \to \mathbb{R}$ defined on a convex set $\Omega \subseteq \mathbb{R}^d$. Further, consider a convex and closed subset $X \subseteq \Omega$, fix $y \in \Omega$ as well as a Bregman projection*

$$\hat{y} \in \operatorname*{arg\,min}_{x\in X} D_\phi(x, y). \tag{102}$$

Then for any $x \in X$ we have

$$D_\phi(x, y) \geq D_\phi(x, \hat{y}) + D_\phi(\hat{y}, y) \tag{103}$$

If further $X = \Omega \cap \mathcal{L}$ for some affine space $\mathcal{L} \subseteq \mathbb{R}^d$ and if $\hat{y} \in \operatorname{int}(X)$ we have

$$D_\phi(x, y) = D_\phi(x, \hat{y}) + D_\phi(\hat{y}, y) \tag{104}$$

Proof. Set $g(x) \coloneqq D_\phi(x, y)$, then by the first order stationarity condition for constrained convex optimization, it holds that

$$0 \geq \nabla g(\hat{y})^\top (\hat{y} - x) = (\nabla\phi(\hat{y}) - \nabla\phi(y))^\top (\hat{y} - x) \quad \text{for all } x \in X, \tag{105}$$

where we used the definition of the Bregman divergence. We use this to estimate

$$\begin{aligned} D_\phi(x, \hat{y}) + D_\phi(\hat{y}, y) &= \phi(x) - \phi(\hat{y}) - \nabla\phi(\hat{y})^\top (x - \hat{y}) + \phi(\hat{y}) - \phi(y) - \nabla\phi(y)^\top (\hat{y} - y) \\ &\leq \phi(x) - \nabla\phi(y)^\top (x - \hat{y}) - \phi(y) - \nabla\phi(y)^\top (\hat{y} - y) \\ &= \phi(x) - \phi(y) - \nabla\phi(y)^\top (x - y) \\ &= D_\phi(x, y). \end{aligned} \tag{106}$$

If $X = \Omega \cap \mathcal{L}$ for an affine $\mathcal{L} \subseteq \mathbb{R}^d$ and $\hat{y} \in \operatorname{int}(X)$, then equality holds in (105) and thus in (106). □