The Geometry of Nonlinear Reinforcement Learning

Nikola Milosevic NMILOSEVIC@CBS.MPG.DE
Nico Scherf NSCHERF@CBS.MPG.DE
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig; Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig

Editors: Michael Bleher, Freya Jensen, Levin Maier, Diaaeldin Taha, and Anna Wienhard

Abstract

Reward maximization, safe exploration, and intrinsic motivation are often studied as separate objectives in reinforcement learning (RL). We present a unified geometric framework, that views these goals as instances of a single optimization problem on the space of achievable long-term behavior in an environment. Within this framework, classical methods such as policy mirror descent, natural policy gradient, and trust-region algorithms naturally generalize to nonlinear utilities and convex constraints. We illustrate how this perspective captures robustness, safety, exploration, and diversity objectives, and outline open challenges at the interface of geometry and deep RL.

1. Introduction

Classical Reinforcement Learning (RL) is formalized as maximizing the expected cumulative reward in a Markov Decision Process (MDP). While this linear formulation has driven major advances, many real-world problems demand richer objectives: respecting safety constraints (Dai et al., 2023), encouraging exploration (Hazan et al., 2019), or balancing multiple goals (Sun et al., 2024; Kolev et al., 2025; Grillotti et al., 2024). Such requirements often correspond to *nonlinear* utility functionals or convex constraints on the agent's long-term behavior (Zhang et al., 2020; Zahavy et al., 2021b).

A natural space to formulate these problems is the manifold of *discounted state-action occupancy measures*, Ω in Figure 1. In this space, standard reward maximization is a linear program (Altman, 1999; Puterman, 2014). Nonlinear utilities and constraints deform this picture into a general nonlinear program, where dynamic programming no longer applies. However, when the utility is concave and constraints are convex, the problem remains a so-called *convex MDP*, which can be solved using a combination of online learning with certain reinforcement learning algorithms (Zhang et al., 2020; Zahavy et al., 2021b).

The challenge, especially in deep RL, is that the mapping from policy parameters θ to the occupancy measure of a policy $\omega_{\pi_\theta} \in \Omega$ is highly non-convex. Thus the overall optimization landscape is non-convex, regardless of the utility's form, and insights from the online convex optimization literature do not necessarily apply. Our key insight is that the form of the utility functional does not change the problem's solvability by deep on-policy actor-critic methods. These algorithms already resort to local, iterative updates due to the non-convexity of the policy parameterization. Therefore, extending the utility from linear to nonlinear does

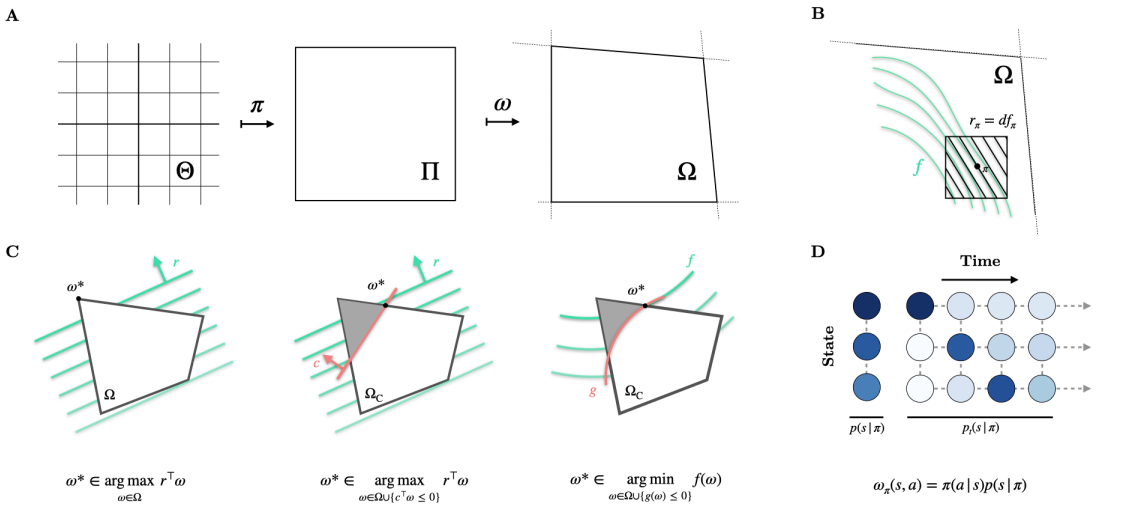

Figure 1: Nonlinear MDPs are nonlinear programs on the occupancy space Ω. A: The mapping from policy parameters θ to occupancies ω is a diffeomorphism. B: Nonlinear MDPs with differentiable utilities can be approximated by a linear MDP locally around the current policy π. The respective reward r_π is obtained as the differential of f at π. C: Standard MDPs (left) and constrained MDPs (middle) are linear programs, while convex MDPs (right) are convex programs on Ω. D: Intuitively, the occupancy measure can be understood as the probability measure obtained by "marginalizing out" the time variable using a geometric distribution.

not fundamentally change the nature of the optimization problem they solve, as long as the functional is differentiable w.r.t the occupancy measure.

In this work, we demonstrate how this "utility-agnostic" view could be formalized geometrically and how it leads to more principled algorithm design for general-utility RL. Specifically, we show that standard actor-critic methods are instances of *mirror descent* on the occupancy manifold. We then leverage the underlying Hessian geometry of this framework to generalize these methods to the nonlinear constrained case, yielding a practical and scalable algorithm for solving general nonlinear MDPs.

Related Work. Nonlinear formulations of reinforcement learning (RL) have attracted significant attention as they generalize classical reward maximization and allow for richer objectives and constraints. In the unconstrained case with a *concave* utility, the problem can be reformulated as a two-player game around the standard MDP linear program (Zahavy et al., 2021b). Here, a *reward player* iteratively selects reward functions to minimize a payoff, while a *policy player* seeks to maximize it (Miryoosefi et al., 2019; Hazan et al., 2019; Nachum and Dai, 2020; Zhang et al., 2020; Zahavy et al., 2021b; Geist et al., 2022). From the policy

player's perspective, this corresponds to an MDP with non-stationary rewards, enabling the use of policy optimization algorithms tailored to such settings, especially methods inspired by mirror descent and proximal methods from online convex optimization (Shani et al., 2020; Tomar et al., 2022; Lan, 2023). For completeness, Zahavy et al. (2021b) also briefly discuss convex constraints. However, to date, theory that treats the general constrained case, as well as solution methods for deep reinforcement learning for nonlinear utilities remain largely unexplored. On-policy policy gradient-based approaches for unconstrained general utility maximization have been previously suggested (Zhang et al., 2020; Kumar et al., 2022), although not from the geometric perspective presented here. Beyond policy gradients, Mutti et al. (2023); Moreno et al. (2025); Santos et al. (2025) recently examined sample complexity and practical challenges for concave utilities, Gemp et al. (2025) study convex Markov games, and mirror descent has been generalized to regularized and concave-utility RL settings (Geist et al., 2022; Zhan et al., 2023). Finally, our approach is inspired by the Hessian geometry of optimization algorithms (Duistermaat, 1999; Alvarez et al., 2004; Raskutti and Mukherjee, 2015), which was recently applied to policy optimization (Müller and Montúfar, 2023; Milosevic et al., 2024, 2025).

2. Actor-Critic Methods for Nonlinear Decision Making

We consider a controlled Markov process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \mu, \gamma)$, see Appendix A. For any stationary policy $\pi : s \mapsto p(a)$, the discounted state-action occupancy measure is

$$\omega_\pi(s,a) := \pi(a|s) \left[(1-\gamma) \sum_{t=0}^{\infty} \gamma^t p_t(s|\pi) \right], \tag{1}$$

where $p_t(s|\pi)$ is the probability of state s at time t given policy π. The set of all achievable occupancies forms the *occupancy manifold* Ω, a convex polytope defined by a system of linear equations, see (Kallenberg, 1994) and Appendix B. As a subset of the probability simplex, Ω can be endowed with a (curved) Riemannian structure (Amari, 1982; Ay et al., 2017), which we will discuss below and in Appendix E.

We define a *Nonlinear MDP* as the following optimization problem on this manifold:

$$\boldsymbol{\omega}^* \in \arg\max_{\boldsymbol{\omega} \in \Omega} \{ f(\boldsymbol{\omega}) \mid g_i(\boldsymbol{\omega}) \leq 0, \forall i \}, \tag{N-MDP}$$

where f and g_i are continuously differentiable functions, and each g_i is convex. This framework subsumes standard (constrained) MDPs, where f, g_i are linear (Altman, 1999), and convex MDPs (Zahavy et al., 2021b) when f is concave and all g_i are convex. Otherwise, local methods are required. This formulation is rich enough to express objectives from imitation learning (Ho and Ermon, 2016), exploration (Hazan et al., 2019), safety (Achiam et al., 2017), and control-as-inference (Toussaint and Storkey, 2006; Ziebart et al., 2008; Rawlik et al., 2012).

Actor-Critic as Mirror Descent Our central claim is that modern actor-critic algorithms are in principle suited for nonlinear utility maximization. They can be viewed as an

approximation of mirror descent (MD) on the occupancy manifold Ω, implemented entirely in policy space. The general mirror descent update step to maximize a utility f on Ω is:

$$\omega_{\pi_{k+1}} \in \arg\max_{\omega_\pi \in \Omega} \langle \nabla f(\omega_{\pi_k}),\ \omega_\pi - \omega_{\pi_k} \rangle - \eta_k^{-1} D_\phi(\omega_\pi \| \omega_{\pi_k}), \tag{2}$$

where the first term is a linear approximation of the change in utility on Ω, and the second term is a Bregman divergence D_ϕ that regularizes the size of the update step according to a geometry induced by a potential function ϕ on Ω. Commonly used proximal actor-critic methods like PPO (Schulman et al., 2017b), TRPO (Schulman et al., 2017a) and more recent improvements (Peng et al., 2019; Abdolmaleki et al., 2018; Haarnoja et al., 2024) solve the surrogate update:

$$\pi_{k+1} \in \arg\max_{\pi \in \Pi} \mathbb{E}_{s\sim\omega_k, a\sim\pi}[A_{\pi_k}(s,a)] - \eta_k^{-1} \mathbb{E}_{s\sim\omega_k} \mathrm{KL}(\pi_k(\cdot|s) \,\|\, \pi(\cdot|s)), \tag{3}$$

where the surrogate advantage term, $\mathbb{E}[A_{\pi_k}(s,a)]$, is a first-order approximation of the utility change. Further, the KL regularization term corresponds to the Bregman divergence in (2) generated by the *negative conditional entropy* potential, $\phi(\omega) = \sum_{s,a} \omega(s,a) \log \pi(a|s)$, see (Neu et al., 2017). The approximation of MD arises because $\mathbb{E}[A_{\pi_k}(s,a)]$ is the utility change, $\langle \nabla f(\omega_{\pi_k}), \omega_\pi - \omega_{\pi_k} \rangle$ up to first order in π if we use a *pseudo-reward* of the form $r_k(s,a) = \partial f(\omega_{\pi_k})/\partial_\omega$, which are the components of $\mathrm{d}f_{\pi_k}$, the differential of f at π_k. We refer the reader to Appendix D.3 for a more precise statement.

3. The Geometry of Nonlinear Actor-Critic Methods

This equivalence reveals the utility-agnostic and inherently geometric nature of actor-critic methods. The proximal policy update Eq. 3 only depends on the differential of f at the current π, not on its global properties. This can be used to turn nonlinear MDPs into a collection of locally linear MDPs using the differential $\mathrm{d}f_\pi$ as the linear functional, i.e. using the components of the differential as the reward function entries: $r_\pi(s,a) = (\partial f/\partial_\omega)(s,a)$. Whether f is linear or nonlinear, proximal actor-critic algorithms perform the same local linearization and regularized update steps in policy parameter space. Only Q-Learning does not work out-of-the box, because the Bellman optimality equation does not refer to the optimal policy of the N-MDP. Since the mapping $\theta \mapsto \pi \mapsto \omega_\pi \mapsto f(\omega_\pi)$ is highly non-convex, we are always seeking a local optimum. Thus, proximal methods are naturally suited to solve general N-MDPs, or at least as well as they are for solving linear MDPs.

Geometry Matters. While the nonlinearity of the utility function does not change the nature of the policy optimization problem, considering its underlying geometry can indeed improve convergence properties and stability of the updates. To achieve stable updates, we must ensure proximity to the previous iterate, so the MDP remains locally linear. Because changes in utilities and constraints in the N-MDP are directly affected only by the occupancy, the geometry of occupancy space and its relation to the utility function ultimately set our requirements for proximity. We argue that the natural geometry to consider is the *Hessian* geometry induced by a particular type of convex potential $\phi : \Omega \to \mathbb{R} \cap \{+\infty\}$, see Appendix E. We demonstrate in Figure 2 the effectiveness of the resulting *Hessian Policy Gradient* (HPG) for a *constrained diversity* problem, see Appendix D.2 for details.

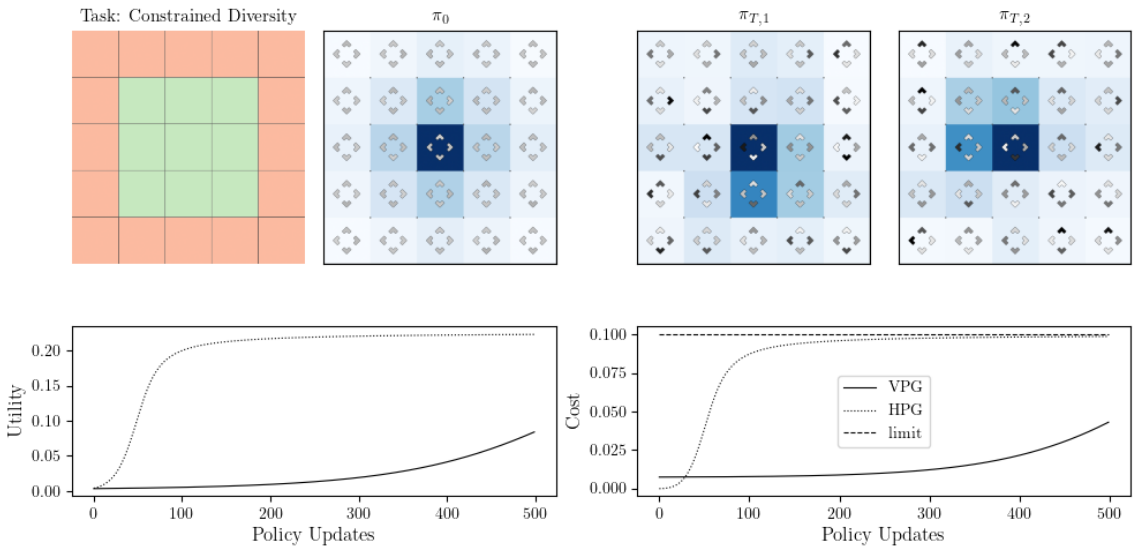

Figure 2: A nonlinear MDP in a 5x5 open gridworld environent. The task is to approximately imitate an initial policy (π_0) that visits the green squares often and avoids the red squares (top left), implemented as a constraint (0.1 bits in Jensen-Shannon divergence). The imitation should be performed using a diverse mixture of policies (2 in this case). The diversity metric is a commonly used convex utility, measuring the mutual information between a binary policy label and the resulting state distribution. A close-to-optimal solution is shown on the top right, obtained with a Hessian optimization approach (HPG, see main text), and compared with the vanilla Lagrangian policy gradient (VPG) in the bottom figures.

4. Conclusion and Outlook

We have outlined a geometric framework for solving nonlinear and constrained MDPs via implicit optimization on occupancy space. This unifies reward maximization, safety, and intrinsic motivation objectives under a common optimization view, enabling principled algorithm design. A possible future direction is to extend the geometric theory, especially to continuous spaces as in e.g. (Ay et al., 2017; Aubin-Frankowski et al., 2022), and to use geometric arguments to derive convergence (Müller and Montúfar, 2023) and policy improvement guarantees (Schulman et al., 2017a) for parametric policies. Another crucial aspect is to understand the practical implications of the nonlinear problem domain for deep function approximation. Scalable off-policy, model-based, and offline variants are required to make N-MDP solution tractable and reliable in practice, and respective theories must accompany them. Off-policy and offline methods are particularly interesting, as they leverage Q-Learning and experience replay to improve sample efficiency, which requires either understanding how Q-Learning can be properly localized, or by addressing the general existence of Bellman operators, see e.g. (Neu et al., 2017; Geist et al., 2019).

Acknowledgements

N.S. is supported by BMFTR (Federal Ministry of Research, Technology and Space) through ACONITE (16IS22065) and the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI.) Leipzig and by the European Union and the Free State of Saxony through BIOWIN.

References

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*, 2018.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization, 2017.

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

Eitan Altman. *Constrained Markov Decision Processes*. CRC Press, Taylor & Francis Group, 1999. URL https://api.semanticscholar.org/CorpusID:14906227.

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.

Shun-Ichi Amari. Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics*, pages 357–385, 1982.

Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror descent with relative smoothness in measure spaces, with application to sinkhorn and em. *Advances in Neural Information Processing Systems*, 35:17263–17275, 2022.

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

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Jacob Burbea and C Rao. On the convexity of some divergence measures based on entropy functions. *IEEE Transactions on Information Theory*, 28(3):489–495, 2003.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback, 2023. URL https://arxiv.org/abs/2310.12773.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural computation*, 5(4):613–624, 1993.

JJ Duistermaat. On hessian riemannian structures. *Asian Journal of Mathematics*, 5(1): 79–91, 1999.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Eugene A Feinberg and Adam Shwartz. *Handbook of Markov decision processes: methods and applications*, volume 40. Springer Science & Business Media, 2012.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR, 2019.

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: the mean-field game viewpoint, 2022. URL https://arxiv.org/abs/2106.03787.

Ian Gemp, Andreas Alexander Haupt, Luke Marris, Siqi Liu, and Georgios Piliouras. Convex markov games: A new frontier for multi-agent reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=yIfCq03hsM.

Samuel J Gershman. The successor representation: its computational logic and neural substrates. *Journal of Neuroscience*, 38(33):7193–7200, 2018.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

Luca Grillotti, Maxence Faldor, Borja G León, and Antoine Cully. Quality-diversity actor-critic: learning high-performing and diverse behaviors via value and successor features critics. *arXiv preprint arXiv:2403.09930*, 2024.

Tuomas Haarnoja, Ben Moran, Guy Lever, Sandy H Huang, Dhruva Tirumala, Jan Humplik, Markus Wulfmeier, Saran Tunyasuvunakool, Noah Y Siegel, Roland Hafner, et al. Learning agile soccer skills for a bipedal robot with deep reinforcement learning. *Science Robotics*, 9 (89):eadi8022, 2024.

Steven Hansen, Will Dabney, Andre Barreto, Tom Van de Wiele, David Warde-Farley, and Volodymyr Mnih. Fast task inference with variational intrinsic successor features. *arXiv preprint arXiv:1906.05030*, 2019.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Audrey Huang and Nan Jiang. Occupancy-based policy gradient: Estimation, convergence, and optimality. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Nq8enbbaP2.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, page 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

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

Lodewijk CM Kallenberg. Survey of linear programming for standard and nonstandard markovian control problems. part i: Theory. *Zeitschrift für Operations Research*, 40:1–42, 1994.

Pavel Kolev, Marin Vlastelica, and Georg Martius. Dual-force: Enhanced offline diversity maximization under imitation constraints. *arXiv preprint arXiv:2501.04426*, 2025.

Navdeep Kumar, Kaixin Wang, Kfir Levy, and Shie Mannor. Policy gradient for reinforcement learning with general utilities, 2022.

Saurabh Kumar, Aviral Kumar, Sergey Levine, and Chelsea Finn. One solution is not all you need: Few-shot extrapolation via structured maxent rl. *Advances in Neural Information Processing Systems*, 33:8198–8210, 2020.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1): 1059–1106, 2023.

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

Nikola Milosevic, Johannes Müller, and Nico Scherf. Embedding safety into rl: A new take on trust region methods. *arXiv preprint arXiv:2411.02957*, 2024.

Nikola Milosevic, Johannes Müller, and Nico Scherf. Central path proximal policy optimization. *arXiv preprint arXiv:2506.00700*, 2025.

Sobhan Miryoosefi, Kianté Brantley, Hal Daume III, Miro Dudik, and Robert E Schapire. Reinforcement learning with convex constraints. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/873be0705c80679f2c71fbf4d872df59-Paper.pdf.

Bianca Marin Moreno, Khaled Eldowa, Pierre Gaillard, Margaux Brégère, and Nadia Oudjane. Online episodic convex reinforcement learning, 2025. URL https://arxiv.org/abs/2505.07303.

Johannes Müller. Geometry of optimization in markov decision processes and neural network-based pde solvers. 2024.

Johannes Müller and Guido Montúfar. Geometry and convergence of natural policy gradient methods. *Information Geometry*, pages 1–39, 2023.

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Challenging common assumptions in convex reinforcement learning, 2023. URL https://arxiv.org/abs/2202.01511.

Ofir Nachum and Bo Dai. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Frank Nielsen and Sylvain Boltz. The burbea-rao and bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8):5455–5466, 2011.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2019. URL https://arxiv.org/abs/1910.00177.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015. doi: 10.1109/TIT.2015.2388583.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. *Proceedings of Robotics: Science and Systems VIII*, 2012.

Pedro Pinto Santos, Alberto Sardinha, and Francisco S. Melo. The number of trials matters in infinite-horizon general-utility markov decision processes. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=I4jNAbqHnM.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization, 2017a.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017b. URL https://arxiv.org/abs/1707.06347.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8604–8613. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/shani20a.html.

Hirohiko Shima. *The geometry of Hessian structures*. World Scientific, 2007.

Jingkai Sun, Qiang Zhang, Yiqun Duan, Xiaoyang Jiang, Chong Cheng, and Renjing Xu. Prompt, plan, perform: Llm-based humanoid control via quantized imitation learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16236–16242. IEEE, 2024.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBO5SvgSt1.

Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. *Advances in Neural Information Processing Systems*, 34:13–23, 2021.

Marc Toussaint and Amos Storkey. Probabilistic inference for solving discrete and continuous state markov decision processes. In *Proceedings of the 23rd international conference on Machine learning*, pages 945–952, 2006.

Tom Zahavy, Brendan O'Donoghue, Andre Barreto, Volodymyr Mnih, Sebastian Flennerhag, and Satinder Singh. Discovering diverse nearly optimal policies with successor features. *arXiv preprint arXiv:2106.00669*, 2021a.

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, 2021b.

Tom Zahavy, Yannick Schroecker, Feryal Behbahani, Kate Baumli, Sebastian Flennerhag, Shaobo Hou, and Satinder Singh. Discovering policies with domino: Diversity optimization maintaining near optimality. *arXiv preprint arXiv:2205.13521*, 2022.

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, 2023. URL https://arxiv.org/abs/2105.11066.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

Appendix A. Reinforcement Learning as Linear Programming

Every stationary policy π in a CMP induces a discounted state-action occupancy measure $\omega_\pi \in \Omega \subset \Delta_{\mathcal{S}\times\mathcal{A}}$, which indicates the relative frequencies of visiting a state-action pair, discounted by how far the event lies in the future. We will refer to this measure as the *state-action occupancy* for short, and its marginal in the state variable will be called the *state occupancy*.

Definition A.1 *The* state-action occupancy measure *is defined as*

$$\omega_\pi(s) \coloneqq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}_\pi(s_t = s)\pi(a|s), \tag{4}$$

where $\mathbb{P}_\pi(s_t = s)$ is the probability of observing the environment in state s at time t given the agent follows policy π.

Note that similar measures can be introduced for the average-reward setting (Zahavy et al., 2021b), and for arbitrary state-action spaces (Laroche and Des Combes, 2023). The following property of the occupancy underscores its utility in policy optimization.

Lemma A.2 *Given trajectories $\tau = (s_0, a_0, s_1, a_1, ..., s_T)$ generated in the CMP $(\mathcal{S}, \mathcal{A}, P, \mu, \gamma)$ with policy π, for a bounded function $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ it holds that*

$$(1-\gamma)\mathbb{E}_{\tau\sim\pi,\mu}\left[\sum_{t=0}^{\infty}\gamma^t f(s_t, a_t)\right] = \mathbb{E}_{s,a\sim d_\pi^\mu}[f(s,a)] \tag{5}$$

Proof

$$\begin{aligned}
(1-\gamma)\mathbb{E}_{\tau\sim\pi,\mu}\left[\sum_{t=0}^{\infty}\gamma^t f(s_t, a_t)\right] &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{s_t,a_t\sim\pi,\mu}[f(s_t,a_t)] && (6)\\
&= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \sum_{s,a}\mathbb{P}(s_t = s, a_t = a)f(s,a) && (7)\\
&= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \sum_{s,a}\mathbb{P}(s_t = s)\pi(a|s)f(s,a) && (8)\\
&= \sum_{s,a}(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s)\pi(a|s)f(s,a) && (9)\\
&= \mathbb{E}_{s,a\sim\omega_\pi^\mu}[f(s,a)], && (10)
\end{aligned}$$

where we were able to swap the order of the infinite sum and expectation in the first line, since the sum converges uniformly, and $\mathbb{P}(s_t = s, a_t = a)$ is the probability of observing the state s and the action a at time t given the initial distribution μ and the policy π. ■

Lemma A.3 *The occupancy can also be written as*

$$\omega_\pi(s,a) = (1-\gamma)\mathbb{E}_{s',a'\sim\omega_\pi}\left[\delta_{s,a}(s',a')\right] \tag{11}$$

and

$$\omega_\pi(s,a) = (1-\gamma)\mathbb{E}_{\tau\sim\pi,\mu}\left[\sum_{t=0}^{\infty}\gamma^t\delta_{s,a}(s_t,a_t)\right], \tag{12}$$

where $\delta_{s,a}$ is the indicator or Dirac distribution at (s,a).

The state-action occupancy is useful, since it lets us abstract away the recursive structure of the expected return, and simplify expressions that involve expectations of discounted infinite sums in time. Instead, we can focus on the optimization problems at hand. For finite MDPs, it is well-known that an optimal policy can be found, by first identifying an occupancy that maximizes the expected discounted return

$$\omega^* \in \max_{\omega} \mathbb{E}_{s,a\sim\omega}[r(s,a)] \quad \text{subject to } \omega\in\Omega, \tag{13}$$

and then conditioning on the state to obtain an optimal policy $\pi^*(s,a) = \omega^*(s,a)/\sum_a \omega^*(s,a)$. Here, Ω is the set of feasible state-action measures given the CMP parameters (Feinberg and Shwartz, 2012). Next, we want to characterize the state-action space Ω in more detail, and show that it is a convex polytope, which can be described by a set of linear constraints.

Appendix B. Characterizing state-action space

The set of state-action occupancies which are feasible under the CMP dynamics forms a convex polytope in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ (Kallenberg, 1994), which is characterized as the solution set of the *Bellman flow equations*.

Proposition B.1 (Bellman flow constraints) *The occupancy of the stationary policy π in the CMP with parameters μ,γ,P must be a probability measure over the sample space $\mathcal{S}\times\mathcal{A}$, and it must satisfy the Bellman flow equations*

$$\sum_a \omega(s,a) = (1-\gamma)\mu(s) + \gamma\sum_{s',a'}\omega(s',a')p(s|s',a'),$$

which is a linear system of equations in the variables $\omega(s,a)$, for all $s\in\mathcal{S}$ and $a\in\mathcal{A}$. We can write

$$\Omega = \{\boldsymbol{\omega} : \boldsymbol{\omega}\in\mathbb{R}_{\geq 0}^{|\mathcal{S}||\mathcal{A}|},\ \boldsymbol{B\omega} = \boldsymbol{\mu}\}, \tag{14}$$

where $\boldsymbol{B}$ is a matrix that depends only on the CMP parameters.

Proof Recall that

$$\omega_\pi(s,a) = \pi(a|s)\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}_\pi(s_t=s)\right],$$

which is a probability measure, since $\pi(a|s)$ is a probability measure over actions given the state, $\mathbb{P}_\pi(s_t = s)$ is the probability of observing the state s at time t under the policy π, and $(1-\gamma)\gamma^t$ is the geometric distribution over time-steps. This means that we can write $\omega_\pi(s,a) = p(a|s,\pi)\sum_{t=0}^\infty p(s|t,\pi)p(t|\gamma)$, which is clearly a probability distribution obtained by marginalization.

Note that, due to the markov property of the CMP, we can write the probability of observing the state s at time t as

$$\begin{aligned}\mathbb{P}_\pi(s_t = s) &= \sum_{s',a'} \mathbb{P}_\pi(s_t = s|s_{t-1} = s', a_{t-1} = a')\mathbb{P}_\pi(s_{t-1} = s', a_{t-1} = a') \\ &= \sum_{s',a'} P(s|s',a')\pi(a'|s')\mathbb{P}_\pi(s_{t-1} = s'),\end{aligned}$$

where $P(s|s',a')$ is the transition kernel of the CMP. Hence, splitting into the first time-step and the rest of the trajectory, we can write

$$\begin{aligned}\omega_\pi(s,a) &= \pi(a|s)\left[(1-\gamma)\mu(s) + (1-\gamma)\sum_{t=1}^\infty \gamma^t \mathbb{P}_\pi(s_t = s),\right] \\ &= \pi(a|s)\left[(1-\gamma)\mu(s) + \gamma\sum_{s',a'} P(s|s',a')\left[(1-\gamma)\sum_{t'=0}^\infty \gamma^{t'}\mathbb{P}_\pi(s_{t'} = s')\pi(a'|s')\right]\right] \\ &= \pi(a|s)\left[(1-\gamma)\mu(s) + \gamma\sum_{s',a'} P(s|s',a')\omega(s',a')\right],\end{aligned}$$

where we re-set the time index to $t' = t-1$ in the second line, and swapped again expectation and uniformly convergent sum. The Bellman flow equations are obtained by a sum over actions

$$\sum_a \omega(s,a) = (1-\gamma)\mu(s) + \gamma\sum_{s',a'} p(s|s',a')\omega(s',a').$$

For the final claim, we rewrite the equations in matrix form:

$$\boldsymbol{B}\boldsymbol{\omega} = \boldsymbol{\mu}. \tag{15}$$

Here, $\boldsymbol{B} = (1-\gamma)^{-1}(\mathbf{1}_a - \gamma\boldsymbol{P})$ is a $|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|$-matrix with $\mathbf{1}_a$ being a $|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|$-matrix with ones along state-action combinations with the same state, and $\boldsymbol{P}$ is the $|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|$-matrix of the Markov kernel. ■

The Bellman flow equations are thus a linear system of equations in the variables $\omega(s,a)$, for all $s\in\mathcal{S}$ and $a\in\mathcal{A}$, which can be solved by standard linear programming techniques, e.g. the simplex method or interior point methods. The occupancy measure ω is a probability measure, i.e. $\omega(s,a)\geq 0$ and $\sum_{s,a}\omega(s,a) = 1$, which is equivalent to the constraints $\boldsymbol{\omega}\in\Delta(\mathcal{S}\times\mathcal{A})$,

where $\Delta(\mathcal{S} \times \mathcal{A})$ is the probability simplex over the state-action pairs. The Bellman flow equations can be interpreted as a set of constraints on the occupancy measure, which ensure that the probability mass remains consistent with the transition probabilities.

Note that we can express the expected discounted return as

$$\mathbb{E}_{s,a\sim d_\pi^\mu}[r(s,a)] = \boldsymbol{r}^\top \boldsymbol{\omega} \tag{16}$$

where $\boldsymbol{d}, \boldsymbol{f} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, up to multiplication of $(1-\gamma)^{-1}$. Since the Bellman flow equations and the occupancy matching problem from above are linear in $\boldsymbol{d}$, we can express it as a linear program

$$\boldsymbol{\omega}^* \in \arg\max_{\boldsymbol{\omega}} \{\boldsymbol{r}^\top \boldsymbol{\omega} \mid \boldsymbol{\omega} \in \Delta(\mathcal{S} \times \mathcal{A}),\ \boldsymbol{B}\boldsymbol{\omega} = \boldsymbol{\mu}\}. \tag{17}$$

A common geometric interpretation, see e.g. (Ay et al., 2017), is of $\boldsymbol{r}$ as a covector and $\boldsymbol{\omega}$ as a vector in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, the expectation being the standard inner product between the function (covector) and the occupancy probability measure (vector). Here, $\boldsymbol{r}$ can also be thought of as a constant one-form acting on the space of occupancies. In the nonlinear utility case, df_π is a nonlinear one-form that we will use to define a *local* or *intrinsic* reward $r_\pi := df_\pi$ which depends on the policy and define a collection of locally linear MDPs on the tangent space $T\Omega$.

Appendix C. Successor Representation

The occupancy is related to the well-known *successor representation* (Dayan, 1993) of the policy in an MDP, and it will play a role in the occupancy gradient lemma below.

Definition C.1 (Successor Representation) *The successor representation is defined as*

$$M^\pi(s,a|s',a') \coloneqq \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \delta_{s',a'}(s_t,a_t) | s_0 = s, a_0 = a\right], \tag{18}$$

where $\delta_{s',a'}(s_t,a_t)$ is the indicator function that is 1 if $(s_t,a_t) = (s',a')$ and 0 otherwise.

The successor representation and the occupancy measure are equivalent up to the initial state distribution μ, meaning

$$\omega_{\pi,\mu}(s,a) = (1-\gamma)\sum_{s',a'} M_\pi(s,a|s',a')\pi(a'|s')\mu(s') \tag{19}$$

and euqivalently, the successor representation can be expressed as a conditional occupancy measure

$$M^\pi(s,a|s',a') = (1-\gamma)^{-1}\omega_{\pi,\delta_{s',a'}}(s,a). \tag{20}$$

Intuitively, the successor representation tells us the *discounted* probability of witnessing the event (s,a) conditional on starting at (s',a') and following π thereafter. It follows, that the successor representation must also satisfy a Bellman flow equation

$$\omega(s,a|s',a') = (1-\gamma)\delta_{s',a'}(s,a) + \gamma \sum_{s',a'} \pi(a|s)P(s|s',a')\omega(s',a'|s,a), \tag{21}$$

or

$$M(s,a|s',a') = \delta_{s',a'}(s,a) + \gamma \sum_{s',a'} \pi(a|s)P(s|s',a')M(s',a'|s,a), \tag{22}$$

which is the backward Bellman equation for M, see also (Touati and Ollivier, 2021).

One can demonstrate several interesting properties of the SR including that it has the following closed-form matrix-valued expression in finite MDPs

$$\mathbf{M}^\pi = [\mathbf{I} - \gamma \mathbf{P}^\pi]^{-1} \tag{23}$$

where $\mathbf{M}^\pi$, and $\mathbf{P}^\pi$ are $|S||A| \times |S||A|$ matrices, and $\mathbf{I}$ is the indicator matrix of the proper dimensions. The inverse exists for $0 \leq \gamma < 1$ and equivalently the Neumann series for $\gamma \mathbf{P}^\pi$ converges to $\mathbf{M}^\pi$

$$\mathbf{M}^\pi = [\mathbf{I} + \gamma \mathbf{P}^\pi + (\gamma \mathbf{P}^\pi)^2 + ...]. \tag{24}$$

As the return of a policy can be obtained as an inner product between occupancy and reward, the equivalent relation holds for M and the on-policy Q-Function:

$$Q_\pi(s,a) = \sum_{s',a'} r(s',a') M_\pi(s',a'|s,a). \tag{25}$$

The successor representation features prominently in the study of neural representations (Gershman, 2018), as well as in transfer learning (Barreto et al., 2017) and intrinsic control algorithms (Hansen et al., 2019) in deep reinforcement learning. Interestingly, it also appears in the differential of the map between policy and state-action spaces. The following result will be relevant in the general version of the policy gradient theorem below, and has been derived as an expression for the jacobian of the map $\pi \mapsto \omega$.

Theorem C.2 (Occupancy Gradient Theorem, Lemma 3.10 in (Müller, 2024)) *Let π be a stationary policy in the finite CMP $(\mathcal{S}, \mathcal{A}, P, \mu, \gamma)$, then the gradient of the occupancy measure w.r.t. the policy parameters θ is given by*

$$\nabla_\theta \omega_\pi(s,a) = (1-\gamma)\mathbb{E}_{s',a' \sim \omega}\left[\nabla_\theta \log \pi_\theta(a'|s') M_\pi(s,a|s',a')\right], \tag{26}$$

where ∇_θ is the gradient w.r.t. the policy parameters θ, and M_π is the successor representation given π.

Proof For finite MDPs, a proof based on differential geometry can be found in Müller and Montúfar (2023). However, we also present another proof based on the classical policy gradient proof by Agarwal et al. (2019).

Note that

$$\nabla_\theta \omega_\pi(s,a) = (1-\gamma)\mathbb{E}_{s_0 \sim \mu} \sum_{a_0} \nabla_\theta \pi_\theta(a_0|s_0) M_{\pi_\theta}(s,a|s_0,a_0). \tag{27}$$

Hence, we can focus on the gradient term and proceed by telescoping the product rule:

$$
\begin{aligned}
&\nabla_\theta \sum_{a_0} \pi_\theta(a_0|s_0) M_{\pi_\theta}(s,a|s_0,a_0) \\
&\quad= \sum_{a_0} \pi_\theta(a_0|s_0) \nabla_\theta \log \pi_\theta(a_0|s_0)\, M_{\pi_\theta}(s,a|s_0,a_0) \\
&\qquad + \sum_{a_0} \pi_\theta(a_0|s_0) \nabla_\theta \Big(\delta_{s_0,a_0}(s,a) + \gamma \sum_{s_1} P(s_1|s_0,a_0) \sum_{a_1} \pi_\theta(a_1|s_1) M_{\pi_\theta}(s,a|s_1,a_1) \Big) \\
&\quad= \sum_{a_0} \pi_\theta(a_0|s_0) \nabla_\theta \log \pi_\theta(a_0|s_0)\, M_{\pi_\theta}(s,a|s_0,a_0) \\
&\qquad + \gamma \sum_{a_0} \pi_\theta(a_0|s_0) P(s_1|s_0,a_0) \nabla_\theta \sum_{a_1} \pi_\theta(a_1|s_1) M_{\pi_\theta}(s,a|s_1,a_1) \\
&\quad= \sum_{a_0} \pi_\theta(a_0|s_0) \nabla_\theta \log \pi_\theta(a_0|s_0)\, M_{\pi_\theta}(s,a|s_0,a_0) \\
&\qquad + \gamma \sum_{a_0} \pi_\theta(a_0|s_0) P(s_1|s_0,a_0) \Big[\sum_{a_1} \pi_\theta(a_1|s_1) \nabla_\theta \log \pi_\theta(a_1|s_1)\, M_{\pi_\theta}(s,a|s_1,a_1) \\
&\qquad\quad + \gamma \sum_{a_1} \pi_\theta(a_1|s_1) P(s_2|s_1,a_1) \nabla_\theta \sum_{a_2} [\dots] \Big] \\
&\quad= \mathbb{E}_{a_0\sim\pi_\theta} \left[\nabla \log \pi_\theta(a_0|s_0) M_{\pi_\theta}(s,a|s_0,a_0) + \gamma\, \mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)} \left[\nabla_\theta \log \pi_\theta(a_1|s_1) M_{\pi_\theta}(s,a|s_1,a_1) + \gamma \mathbb{E}[\dots] \right] \middle| s_0 \right] \\
&\quad= \mathbb{E}_{\tau\sim\pi_\theta} \left[\nabla \log \pi_\theta(a_0|s_0) M_{\pi_\theta}(s,a|s_0,a_0) + \gamma\, \nabla_\theta \log \pi_\theta(a_1|s_1) M_{\pi_\theta}(s,a|s_1,a_1) + \gamma^2 \dots \middle| s_0 \right].
\end{aligned}
$$

The final result is obtained by taking the expectation over the initial state distribution $\mu(s_0)$ and the policy $\pi_\theta(a_0|s_0)$ on both sides, applying Lemma A.2 to obtain the expression as an expectation over ω, and multiplying both sides by $(1-\gamma)^{-1}$.

$$
\begin{aligned}
\nabla_\theta \omega(s,a) &= (1-\gamma)\, \mathbb{E}_{s_0\sim d_0, a_0\sim\pi} \left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t) M_{\pi_\theta}(s,a|s_t,a_t) \right] \\
&= (1-\gamma)^{-1}\, \mathbb{E}_{s,a'\sim\omega} \Big[\nabla_\theta \log \pi_\theta(a'|s')\, M_{\pi_\theta}(s,a|s',a') \Big].
\end{aligned}
$$

■

Another occupancy-based policy gradient theorem has been shown by Huang and Jiang (2024), but it is a distinct result which does not involve the successor representation.

Appendix D. Nonlinear MDPs

We now extend the above framework to nonlinear MDPs (N-MDPs) with nonlinear utilities and constraints:

$$
\boldsymbol{\omega}^* \in \arg\max_{f\boldsymbol{\omega}\in\Omega} \{ f(\boldsymbol{\omega}) \mid g(\boldsymbol{\omega}) \leq 0 \}, \tag{28}
$$

where $\Omega = \{\boldsymbol{\omega} \in \Delta(\mathcal{S}\times\mathcal{A}),\ \boldsymbol{B}\boldsymbol{\omega} = \boldsymbol{\mu}\}$ from Eq. 17. Further, we assume that $f : \Omega \to \mathbb{R}$ is at least once continuously differentiable and $g : \Omega \to \mathbb{R}$ at least twice. The following

result is the basis for the first-order equivalence between occupancy-space mirror descent and proximal actor-critic methods.

Lemma D.1 (Utility Gradient) *Given an N-MDP, the gradient of the nonlinear differentiable utility f w.r.t the policy parameters θ is given by*

$$\nabla_\theta f(\omega_\theta) = \mathbb{E}_{s,a\sim\omega_\pi}[\log \pi_\theta(a|s) A_{\pi_\theta}(s,a)] \tag{29}$$

where $A_{\pi_\theta}(s,a)$ is the advantage function of policy π_θ for the linear MDP with reward $r_{\pi_\theta} = \mathrm{d}f_{\pi_\theta}$.

Proof We apply the chain rule to the gradient of $\theta \mapsto \omega_{\pi_\theta} \mapsto f$ as by Zhang et al. (2020). Note that the partial derivatives $\partial_\omega f$ are the components of the reward functional from the preceding lemma. Substituting $\nabla_\theta \omega$ using the occupancy gradient theorem from Eq. 26, we obtain

$$\nabla_\theta f(\omega_{\pi_\theta}) = \sum_{s,a} \frac{\partial f}{\partial \omega(s,a)} \nabla_\theta \omega(s,a) \tag{30}$$

$$= (1-\gamma)\mathbb{E}_{s',a'\sim\omega}\left[\nabla_\theta \log \pi_\theta(a'|s') \sum_{s,a} M_{\pi_\theta}(s,a|s',a') r_{\pi_\theta}(s,a)\right] \tag{31}$$

$$= (1-\gamma)\mathbb{E}_{s',a'\sim\omega}\left[\nabla_\theta \log \pi_\theta(a'|s') \sum_{s,a} Q_{\pi_\theta}(s',a')\right], \tag{32}$$

where Q_{π_θ} is the on-policy Q-Function for policy π_θ and reward $r_{\pi_\theta}(s,a) = \mathrm{d}f_{\pi_\theta}$. The final result follows from adding the on-policy value function as a state-dependent baseline:

$$\nabla_\theta f(\omega_{\pi_\theta}) = (1-\gamma)\mathbb{E}_{s',a'\sim\omega}\left[\nabla_\theta \log \pi_\theta(a'|s') Q_{\pi_\theta}(s',a')\right] \tag{33}$$

$$= (1-\gamma)\mathbb{E}_{s',a'\sim\omega}\left[\nabla_\theta \log \pi_\theta(a'|s') Q_{\pi_\theta}(s',a') - V_{\pi_\theta}(s')\right] \tag{34}$$

$$= (1-\gamma)\mathbb{E}_{s',a'\sim\omega}\left[\nabla_\theta \log \pi_\theta(a'|s') A_{\pi_\theta}(s',a')\right] \tag{35}$$

■

D.1. Examples of Nonlinear MDPs

Nonlinear MDPs cover some practically relevant control learning objectives, such as behavioral mutual information, state entropy, divergence-based imitation learning, and control-as-inference objectives, see (Zahavy et al., 2021b). Some policy learning objectives consider mixtures of two or more policies, and we are going to take a geometric perspective on characterizing them. Therefore, we consider the Burbeo-Rao divergence, a common symmetrization of the Bregman divergence, and its extension to mixtures of probability distributions (Nielsen and Boltz, 2011; Burbea and Rao, 2003).

Definition D.2 (Dispersion of a policy mixture) *Let us consider a set of policies π_i with corresponding state-action occupancies $\omega_i(s,a)$, and let*

$$\omega(s,a) = \sum_i z_i \omega_i(s,a), \text{ where } z_i \geq 0, \sum_i z_i = 1, \tag{36}$$

be a finite mixture of occupancies, see e.g. (Peng et al., 2019; Laroche and Des Combes, 2023), and let ϕ be a Legendre-type function on Ω. The ϕ-dispersion *of a policy mixture $\bar{\boldsymbol{\pi}} = \{\pi_1, \dots, \pi_N\}$ is defined as*

$$d_\phi(\bar{\boldsymbol{\pi}}) = \sum_i z_i \phi(\omega_i) - \phi(\omega), \tag{37}$$

i.e. the Burbea-Rao divergence of ω generated by ϕ.

The dispersion can be used to compactly characterize popular reward-free objectives like DIAYN (Gregor et al., 2016; Eysenbach et al., 2018), and GAIL (Ho and Ermon, 2016).

It can be shown that decision problems with objectives of this form correspond to convex MDPs (Zahavy et al., 2021b), which are known to be solvable using policy optimization methods. To see this, consider the following properties of the polic dispersion.

Proposition D.3 (Jensen-Bregman) *The* ϕ-dispersion *of a policy mixture $\bar{\boldsymbol{\pi}} = \{\pi_1, \dots, \pi_N\}$ with respective occupancy mixture $\bar{\omega} = \sum_i z_i \omega_i$ can be expressed as*

$$d_\phi(\bar{\boldsymbol{\pi}}) = \mathbb{E}_z[\mathrm{D}_h(\omega_z, \bar{\omega})], \tag{38}$$

which is an average of Bregman divergences of the same generator, and reduces to the mutual information between z and (s, a) for the negative conditional entropy. Further, it is jointly strictly convex in ω_z for strictly convex h.

Proof

$$\sum_i z_i D_h(\omega_i || \bar{\omega}) = \sum_i z_i h(\omega_i) - \sum_i z_i h(\bar{\omega}) - \sum_i z_i \nabla h(\bar{\omega})(\omega_i - \bar{\omega}) \tag{39}$$

$$= \sum_i z_i h(\omega_i) - \sum_i z_i h(\bar{\omega}) - \nabla h(\bar{\omega})(\bar{\omega} - \bar{\omega}) \tag{40}$$

$$= \sum_i z_i h(\omega_i) - h\left(\sum_i z_i \omega_i\right). \tag{41}$$

The second claim follows from the definition of mutual information between (s, a)-realizations following $\bar{\omega}$, and the skill z:

$$I(z; s, a) = H(\bar{\omega}) - H(\bar{\omega}|z) \tag{42}$$

$$= H(\bar{\omega}) - \sum_i z_i H(\omega_i) \tag{43}$$

$$= \sum_i z_i \sum \omega_i \log \omega_i - \sum \bar{\omega} \log \bar{\omega} \tag{44}$$

$$= \sum_i z_i \left(\sum \omega_i \log \omega_i - \sum \omega_i \log \bar{\omega}\right) \tag{45}$$

$$= \sum_i z_i D_{KL}(\omega_i || \bar{\omega}), \tag{46}$$

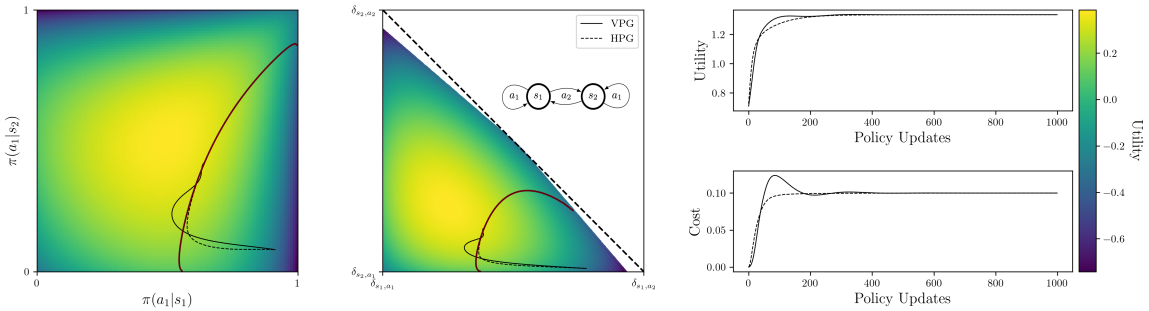

Figure 3: A constrained diversity problem in an MDP with two states and two actions (middle figure, top right), in policy space (left), in linearly projected occupancy space (middle), and a comparison between Lagrangian Vanilla Policy Gradient (VPG) and Hessian Policy Gradient (HPG) in terms of sample efficiency curves. The utility functional is discounted state-action entropy in bits.

which by the above identity is the MPD of the negative conditional entropy. Joint convexity follows from the convexity of $D_{KL}(\cdot||\bar{\omega})$ and the last line being a convex combination of convex functions. ■

In the context of information geometry, $\bar{\omega}$ is also known as the Bregman centroid (Nielsen and Boltz, 2011).

In the following, we demonstrate some examples in theory and in a small scale simulation, see Figure 3.

Example D.4 (GAIL (Ho and Ermon, 2016)) *Set* $N = 2$ *and* $\bar{\omega}(s,a) = \frac{1}{2}\omega^{\pi}(s,a) + \frac{1}{2}\omega^{E}(s,a)$*, then*

$$\max_{\pi} d_{-H}(\pi, \pi^E), \tag{47}$$

where π^E *is an expert policy and* $-H$ *is the negative entropy, is a nonlinear MDP and a* convex *MDP in particular. The reward is* $r_\pi(s,a) = \sum_i z_i[\log z_i - \log p(i|s,a)]$*, which is approximated using an adversarial objective in practice.*

Example D.5 (DIAYN (Gregor et al., 2016; Eysenbach et al., 2018)) *Set* $z = \mathrm{Unif}(\{1,...,N\})$*, then the DIAYN (Eysenbach et al., 2018) objective*

$$\max_{\bar{\boldsymbol{\pi}}} d_{-H_s}(\bar{\boldsymbol{\pi}}), \tag{48}$$

forms a nonlinear MDP, where $-H_s$ *is the negative entropy of the state occupancy* $-H_s = -H[\sum_a \omega(\cdot,a)]$*. The reward is* $r_\pi(s,a) = \sum_i z_i[\log p_\pi(i|s) - \log z_i]$*. In this case,* $p(i|s,\pi)$ *is the Bayesian posterior for* $\sum_i z_i \omega_{\pi_i}(s,a)$*, which is approximated with a variational objective in practice.*

Example D.6 (Maximum Entropy Exploration (Hazan et al., 2019))
Set $z = \text{Unif}(\{1, ..., N\})$*, then the Maximum Entropy Exploration objective*

$$\max_{\pi} H[\omega^{\pi}], \tag{49}$$

is a nonlinear MDP, in particular a convex one. Here, H *is the entropy, and the reward is* $r_{\pi}(s, a) = -(\log \omega_{\pi}(s, a) + 1)$.

D.2. Implementation Details

In the computational experiments and implementations we use what we call the *Hessian Policy Gradient* (HPG) updates, a generalization of NPG (Kakade, 2001), and its constrained counterparts (Milosevic et al., 2024) to constrained nonlinear MDPs. The HPG update is defined as

$$\theta_{k+1} = \theta_k + \eta_k \mathbf{H}_{\phi}^{\dagger}(\theta_k) \nabla_{\theta} f(\omega_{\pi_{\theta_k}}), \tag{50}$$

where $\mathbf{H}_{\phi}^{\dagger}$ is a pseudo-inverse of the Hessian of the potential ϕ with respect to the policy parameters θ, and where $\nabla_{\theta} f(\omega_{\pi_{\theta_k}})$ is estimated as the policy gradient for the intrinsic reward at π. In our small scale experiments, it was sufficient to use out-of-the-box auto-differentiation (pytorch) for computing the Hessian and gradient, and a least-squares solver for the pseudo-inverse.

Example: Constrained Diversity In the toy experiments, we demonstrate that policy gradient convergence greatly benefits from Hessian updates, especially in nonlinear constrained settings. Many practical objectives combine diversity with quality constraints (Kumar et al., 2020; Zahavy et al., 2021a, 2022; Sun et al., 2024; Kolev et al., 2025; Grillotti et al., 2024) in a constrained policy optimization problem of the form

$$\max_{\boldsymbol{\pi}} \text{ Diversity}(\boldsymbol{\pi}) \text{ s.t. Quality}(\boldsymbol{\pi}) \geq \alpha Q^{*}. \tag{51}$$

These metrics are often expressible as f and g in problem N-MDP, allowing direct application of the Hessian approach. In the gridworld experiment in Figure 2, we chose the DIAYN objective with two policies as the diversity metric and the GAIL objective with a threshold of 0.1 as the quality metric. In the two dimensional experiments in Figure 3, we chose the maximum entropy objective as the diversity metric instead.

D.3. Mirror Descent Equivalence

First, we give a short introduction to Bregman divergences, which are part of the definition of mirror descent. For this, we consider a Lagendre potential ϕ over a convex subset of Euclidean space $C \subseteq \mathbb{R}^d$ with a non-empty interior $\text{int}(C)$. Then, the *Bregman divergence* induced by ϕ is

$$D_{\phi}(x||y) \coloneqq \phi(x) - \phi(y) - \nabla\phi(y)^{\top}(x - y), \tag{52}$$

which is well defined for $x \in C, y \in \text{int}(C)$. Intuitively, the Bregman divergence measures the difference between ϕ and its linearization at y. The strict convexity of ϕ ensures that $D_{\phi}(x||y) \geq 0$ and $D_{\phi}(x||y) = 0$ if and only if $x = y$.

An important Bregman divergence is the Kullback-Leibler (KL) divergence between finite probability measures $\sum_i p_i = \sum_i q_i = 1$

$$\mathrm{KL}(p||q) \coloneqq \sum_{i=1}^{d} p_i \log \frac{p_i}{q_i}. \tag{53}$$

Proposition D.7 *Let the Policy Mirror Descent (PMD) objective be*

$$J_{\mathrm{PMD}}(\pi) = \langle \nabla f(\omega_k), \omega_\pi - \omega_k \rangle - \frac{1}{\eta} D_\phi(\omega_\pi \| \omega_k) \tag{54}$$

and the standard surrogate objective be

$$J_{\mathrm{SURR}}(\pi) = \mathbb{E}_{s\sim\omega_k, a\sim\pi}[A_k(s,a)] - \frac{1}{\eta}\mathbb{E}_{s\sim\omega_k}[\mathrm{KL}(\pi_k(\cdot|s)\|\pi(\cdot|s))] \tag{55}$$

where the advantage function A_k is computed using the intrinsic reward $r_k := \partial f(\omega_k)/\partial\omega$, and the potential function ϕ is the negative conditional entropy, $\phi(\omega) = \sum_{s,a}\omega(s,a)\log\omega(s,a)/\sum_a \omega(s,a)$. Then, the Taylor expansions of $J_{\mathrm{PMD}}(\pi_\theta)$ and $J_{\mathrm{SURR}}(\pi_\theta)$ around the current policy parameters θ_k are identical up to first order. Furthermore, the regularizers are identical up to second order.

We will show this by explicitly computing the gradients and Hessians of both objectives, evaluated at $\theta = \theta_k$. Let $\pi_k = \pi_{\theta_k}$ and $\omega_k = d_{\pi_k}$. We rely on the following properties of the KL-divegence:

$$\begin{aligned}\nabla_\theta \mathrm{KL}(\pi_k\|\pi_\theta)|_{\theta_k} &= 0\\ \nabla^2_\theta \mathrm{KL}(\pi_k\|\pi_\theta)|_{\theta_k} &= F_s(\theta_k)\end{aligned}$$

where $F_s(\theta_k)$ is the Fisher Information Matrix (FIM) of the policy $\pi(\cdot|s)$ at state s. In general, the Bregman divergence $D_\phi(\omega_\pi\|\omega_k)$ is minimized (with value 0) at $\pi = \pi_k$. Its gradient and Hessian with respect to θ at θ_k are:

$$\begin{aligned}\nabla_\theta D_\phi(\omega_\pi\|\omega_k)|_{\theta_k} &= 0\\ \nabla^2_\theta D_\phi(\omega_\pi\|\omega_k)|_{\theta_k} &= J_\omega(\theta_k)^T H_\phi(\omega_k) J_\omega(\theta_k)\end{aligned}$$

where $J_\omega(\theta_k) = \nabla_\theta \omega_\pi|_{\theta_k}$ is the Jacobian of $\pi \mapsto \theta$ and $H_\phi(\omega_k) = \nabla^2_d\phi|_{\omega_k}$ is the Hessian of the potential function. For our chosen potential ϕ, this product is the FIM weighted by the state-visitation frequencies: $J_\omega^T H_\phi J_\omega = \mathbb{E}_{s\sim\omega_k}[F_s(\theta_k)] = F_k$.

Surrogate Objective We compute the gradient of J_{SURR} and evaluate it at θ_k:

$$\begin{aligned}\nabla_\theta J_{\mathrm{SURR}}(\pi_\theta)|_{\theta_k} &= \nabla_\theta \mathbb{E}_{s\sim\omega_k, a\sim\pi_\theta}[A_k(s,a)]|_{\theta_k} - \frac{1}{\eta}\nabla_\theta \mathbb{E}_{s\sim\omega_k}[\mathrm{KL}(\pi_k\|\pi_\theta)]|_{\theta_k}\\ &= \mathbb{E}_{s\sim\omega_k, a\sim\pi_k}[\nabla_\theta \log\pi_\theta(a|s)|_{\theta_k} A_k(s,a)] - \frac{1}{\eta}\cdot 0\\ &= \mathbb{E}_{s\sim\omega_k, a\sim\pi_k}[\nabla_\theta \log\pi_\theta(a|s)|_{\theta_k} A_k(s,a)]\end{aligned}$$

The expectation $\mathbb{E}_{s\sim\omega_k}$ in the KL term is constant with respect to θ as it uses the fixed distribution from the previous policy. The Hessian of J_{SURR} at θ_k is

$$\begin{aligned}\nabla^2_\theta J_{\text{SURR}}(\pi_\theta)|_{\theta_k} &= \nabla^2_\theta \mathbb{E}_{s\sim\omega_k, a\sim\pi_\theta}[A_k(s,a)]|_{\theta_k} - \frac{1}{\eta}\nabla^2_\theta \mathbb{E}_{s\sim\omega_k}[\text{KL}(\pi_k\|\pi_\theta)]|_{\theta_k} \\ &= \nabla^2_\theta \mathbb{E}_{s\sim\omega_k, a\sim\pi_\theta}[A_k(s,a)]|_{\theta_k} - \frac{1}{\eta}\mathbb{E}_{s\sim\omega_k}[F_s(\theta_k)] \\ &= \mathbb{E}_{s\sim\omega_k, a\sim\pi_k}[\nabla^2_\theta \log\pi(a|s)|_{\theta_k} A_k(s,a)] - \frac{1}{\eta}F_k\end{aligned}$$

Mirror Descent Objective We compute the gradient of J_{PMD} and evaluate it at θ_k. Note that $\nabla f(\omega_k)$ and ω_k are constants.

$$\begin{aligned}\nabla_\theta J_{\text{PMD}}(\pi_\theta)|_{\theta_k} &= \nabla_\theta \langle \nabla f(\omega_k), \omega_\pi\rangle|_{\theta_k} - \frac{1}{\eta}\nabla_\theta D_\phi(\omega_\pi\|\omega_k)|_{\theta_k} \\ &= J_\omega(\theta_k)^T \nabla f(\omega_k) - \frac{1}{\eta}\cdot 0 \\ &= \nabla_\theta f(\omega_\pi)|_{\theta_k}\end{aligned}$$

By the Utility Gradient Lemma 29, this is exactly

$$\nabla_\theta J_{\text{PMD}}(\pi_\theta)|_{\theta_k} = \mathbb{E}_{s\sim\omega_k, a\sim\pi_k}[\nabla_\theta \log\pi_\theta(a|s)|_{\theta_k} A_k(s,a)]$$

The gradients match. The Hessian of J_{PMD} at θ_k is

$$\nabla^2_\theta J_{\text{PMD}}(\pi_\theta)|_{\theta_k} = \nabla^2_\theta \langle \nabla f(\omega_k), \omega_\pi\rangle|_{\theta_k} - \frac{1}{\eta}\nabla^2_\theta D_\phi(\omega_\pi\|\omega_k)|_{\theta_k}$$

For the second term, we use the Hessian property of the Bregman divergence:

$$\nabla^2_\theta D_\phi(\omega_\pi\|\omega_k)|_{\theta_k} = F_k$$

The Hessians of the regularizers also match. The Hessian of the first term is related to the Hessian of the surrogate, though not identical. However, in the policy optimization literature, one makes the approximation that second order changes in the occupancy are negligible, focusing on the second-order term from the regularizer (Kakade and Langford, 2002; Schulman et al., 2017a).

D.3.1. Constrained Policy Geometry

This equivalence suggests that one can approximate arbitrary mirror descent schemes up to second order, by considering modified versions of TRPO (Schulman et al., 2017a). In C-TRPO (Milosevic et al., 2024), the authors consider mirror functions of the form

$$\Phi_{\text{C}}(\omega) = \Phi_{\text{K}}(\omega) + \Phi_{\text{B}}(\omega) \tag{56}$$

$$\coloneqq \sum_{s,a}\omega(s,a)\log\pi_\omega(a|s) + \sum_{i=1}^{m}\beta_i \ell\left(b_i - \sum_{s,a}\omega(s,a)c(s,a)\right), \tag{57}$$

where $\omega \in \mathcal{K}_{\text{safe}}$ is a feasible state-action occupancy, Φ_{K} is the conditional entropy, and Φ_{B} and $\ell\colon \mathbb{R}_{>0} \to \mathbb{R}$ are convex functions with $\ell'(x) \to +\infty$ for $x \searrow 0$. Possible candidates for ℓ are $-\log(x)$ and $x\log(x)$, corresponding to a logarithmic barrier and entropy, respectively. This results in an *intractable* policy divergence with Hessian

$$H_{\text{C}}(\theta) = \mathbb{E}_{s\sim\omega_\pi} F(\theta) + \sum_i \beta_i \phi''(b_i - V_{c_i}(\theta)) \nabla^2_\theta V_{c_i}(\theta)\Big|_{\theta=\theta_k}.$$

However, the authors show that the following divergence with the same Hessian can be derived using the common first-order advantage approximation, but for the cost instead of the rewardx:

$$D_{\text{C}}(\pi||\pi_k) = \bar{D}_{KL}(\pi||\pi_k) + \beta(\ell(B_k - \mathbb{A}) - \ell(B_k) - \ell'(B_k)\cdot \mathbb{A}) \tag{58}$$

where

$$\mathbb{A} = \sum_s \omega_{\pi_k}(s) \sum_a \pi(a|s) A_c^{\pi_k}(s,a)$$

is the surrogate cost advantage, and $\bar{D}_{KL}(\pi||\pi_k)$ is the expected KL-Divergence between the policies w.r.t π_k's state occupancy measure. We introduce $B_k = b - V_c(\pi_k)$, i.e. the budget until constraint violation, and focus on a single constraint to reduce notational clutter.

Convex Constraints. The mirror descent perspective provides a principled way to handle the convex constraints in (N-MDP). Any Legendre-type potential ϕ not only defines a Bregman divergence but also endows the manifold Ω with a Hessian geometry via its metric tensor. This gives rise to a generalized natural policy gradient (Müller, 2024). Following (Milosevic et al., 2024, 2025), we can incorporate constraints directly into the geometry by using a *barrier potential*:

$$b(\omega) = \phi(\omega) + \beta \sum_i \ell(g_i(\omega)),$$

where ϕ is the original potential (e.g., negative conditional entropy) and ℓ is a Legendre-type barrier, such as $\ell(x) = -\log(-x)$. Performing natural policy gradient descent with respect to the Hessian metric induced by this new potential b ensures that iterates remain strictly feasible while preserving convergence guarantees (Alvarez et al., 2004).

Appendix E. Hessian and Information Geometries of the Occupancy Space

The set of all valid state-action occupancy measures for a given CMP forms the natural domain for formulating nonlinear RL problems. It can be characterized as

$$\Omega = \{\omega \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} : \sum_a \omega(s,a) - \gamma \sum_{s',a'} P(s|s',a')\omega(s',a') = (1-\gamma)\mu(s) \quad \forall s \in S\}.$$

Geometrically, Ω is the intersection of an affine subspace of $\mathcal{R}^{|\mathcal{S}||\mathcal{A}|}$ and the probability simplex $\Delta_{\mathcal{S}\times\mathcal{A}}$. This subspace of $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is defined by the set of probability measures ω, that satisfy the linear Bellman flow equations.

Fisher-Rao Geometry The ambient space, the simplex $\Delta_{\mathcal{S}\times\mathcal{A}}$, possesses a canonical and well-studied Riemannian structure known as the Fisher-Rao geometry. This geometry is induced by the Hessian of the negative joint entropy potential function:

$$\phi_{FR}(\omega) = \sum_{s,a} \omega(s,a) \log \omega(s,a)$$

The Hessian, $H_{\phi_{FR}}(\omega)$, defines the Fisher metric on Ω. The natural parameters corresponding to this potential are the log-probabilities of the occupancy measure itself:

$$\eta_{s,a} = \frac{\partial \phi_{FR}}{\partial \omega(s,a)} = \log \omega(s,a) + 1.$$

The Bregman divergence generated by ϕ_{FR} is the Kullback-Leibler (KL) divergence, $D_{KL}(\omega' \| \omega)$.

When this standard geometry is restricted to the manifold Ω, i.e.

$$\eta_{s,a} = \log \sum_{s',a'} M_\pi(s,a|s',a')\pi(a'|s')\mu(s'),$$

it provides a principled way to measure distances between valid occupancy measures, however the resulting exponential family is curved (Amari, 1982). A significant practical challenge is that this geometry requires access to evaluations of the log-occupancy measure, $\log \omega(s,a)$, which are not directly available to a learning agent that only controls its policy, and it is in general difficult to estimate from data.

Kakade Geometry In practice, we do not optimize over ω directly. Instead, we parameterize the occupancy manifold indirectly through a policy π, as the mapping from a policy to its occupancy measure, ω_π, is a diffeomorphism under regularity conditions (Müller, 2024). This provides an alternative and more practical set of coordinates. The corresponding potential function is the **negative conditional entropy**:

$$\phi_K(\omega) = \sum_{s,a} \omega(s,a) \log \pi(a|s)$$

where the policy $\pi(a|s)$ is recovered from the occupancy measure via $\pi(a|s) = \omega(s,a) / \sum_{a'} \omega(s,a')$. The natural parameters for this geometry are the, still curved, log-probabilities of the policy:

$$\eta_{s,a} = \frac{\partial \phi_K}{\partial \omega(s,a)} = \log \pi(a|s)$$

This is the geometry implicitly used by the Natural Policy Gradient (Kakade, 2001) and its variants.

Motivation for a General Hessian Framework The Fisher-Rao geometry is theoretically elegant but practically difficult, while the policy-induced view is practical and directly aligned with the learning problem, but does not come with some of the convenient results of the Fisher-Rao geometry. Rather than committing to a single, fixed geometry, we can consider the choice of geometry as a degree of freedom in algorithm design. This motivates framing policy optimization within the more general theory of *Hessian structures* (Shima, 2007). In this framework, any Legendre-type potential function ϕ can be used to define a valid Riemannian structure on Ω. By carefully selecting ϕ, e.g. to incorporate constraints via barrier functions as we did in Milosevic et al. (2024), or to counteract the curvature of the utility function as in mirror descent, we can design novel and more powerful policy update steps that are tailored to the specific structure of the nonlinear MDP we aim to solve.