Lagrangian Duality in Reinforcement Learning

Pranay Pasula
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
pasula@berkeley.edu

Abstract

Although duality is used extensively in certain fields, such as supervised learning in machine learning, it has been much less explored in others, such as reinforcement learning (RL). In this paper, we show show how duality is involved in a variety of RL work, from that which spearheaded the field, such as Richard Bellman's value iteration, to that which was done within just the past few years yet has already had significant impact, such as TRPO, A3C, and GAIL. We show that duality is *not* uncommon in reinforcement learning, especially when *value iteration*, or *dynamic programming*, is used or when first or second order approximations are made to transform initially intractable problems into tractable convex programs.

1 Introduction

The study of optimization problems that are dual to certain initial problems has led to key insights, including efficient ways to bound or exactly solve these original problems [5]. Though duality is used extensively in certain fields, such as supervised learning in machine learning, it has been much less explored in others, such as reinforcement learning (RL). In this paper, we will show how duality is involved in a variety of RL work, from that which spearheaded the field [3, 11, 4] to that which has been done within the past few years but has already had significant impact [19, 10, 14].

We show that duality is *not* uncommon in reinforcement learning, especially when *value iteration*, or *dynamic programming*, is used or when first or second order approximations are made to transform initially intractable problems into tractable convex programs [19, 12, 1]. We touch on a number of works that span a wide range of RL paradigms but focus on some works that have been particularly influential, such as Trust Region Policy Optimization (TRPO) [19], Asynchronous Advantage Actor-Critic (A3C) [14], and Generative Adversarial Imitation Learning (GAIL) [10]. In some cases duality is used as a theoretical tool to prove certain results or to gain insight into the meaning of the problem involved. In other cases duality is leveraged to employ gradient-based methods over some *dual* space, as is done in alternating direction method of multipliers (ADMM) [6], mirror descent [2], and dual averaging [22, 8].

We hope this work will encourage others to more strongly consider duality in reinforcement learning, which we believe to be an often underexplored yet promising line of reasoning.

2 Preliminaries

We consider a standard reinforcement learning framework in which we have a Markov decision process (MDP) $M = (\mathcal{S}, \mathcal{A}, P, R, T, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the distribution representing the transition probabilities of the agent arriving in state s' after taking action a while in state s, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function representing the reward the agent obtains by taking action a while in state s, $T > 0$ is the time horizon and can be either finite or infinite, and $\gamma \in [0, 1]$ is the multiplicative rate at which future rewards are discounted per time step. For simplicity, we will usually set $\gamma = 1$ and drop this term. Throughout this paper we

will use the terms *trajectory*, *behavior*, and *demonstration* interchangeably, all of which refer to a finite or infinite sequence of states and actions $\{s_1, a_1, s_2, a_2, \ldots, s_T, a_T\}$.

We assume the reader has at least an introductory understanding of duality theory and defer its exposition to [5].

3 Duality in Dynamic Programming

3.1 Unregularized Markov Decision Processes

Duality has well-established history in the theory underlying fundamental classes of RL algorithms, such as value iteration and policy optimization. In particular, the Bellman backup operator $\mathcal{T}_\pi$ [3] induced by a policy π and applied to Q functions Q_π induced by π represents the dynamic programming update

$$Q_\pi(s_t, a_t) = R(s_t, a_t) + \sum_{s_{t+1} \in \mathcal{S}} \left[P(s_{t+1} \mid s_t, a_t) \sum_{a_{t+1} \in \mathcal{A}} Q_\pi(s_{t+1}, a_{t+1}) \right],$$

which under mild conditions [18] converges to the fixed point Q_π.

Since the goal in reinforcement learning is to find the optimal policy π^* that maximizes the expected sum of rewards obtained by the agent, an equivalent problem is to find π^* that maximizes the average reward $\bar{R}$ obtained by the agent over timesteps $t = 1, 2, 3, \ldots$

MDP theory allows us to write the average reward $\bar{R}(\pi)$, obtained by following policy π, in terms of R, π, and the stationary state distribution $\nu_\pi(x)$ under π. Namely, we have

$$\bar{R}(\pi) = \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \nu_\pi(s) \pi(a \mid s) R(s, a).$$

The stationary state distribution ν_π is linear in the stationary state-action distribution $\mu_\pi = \nu_\pi \pi$, which suggests that the problem of finding the optimal policy can be formulated as a linear program (LP) with decision variable μ over the probability simplex $\mathcal{C}$:

$$\mu^* = \arg\max_{\mu \in \mathcal{C}} \bar{R}(\mu). \tag{1}$$

[18] shows that by assuming $\mu^* \in \mathcal{C}$, which holds when the MDP has only one recurrent class, (1) is the dual of the LP

$$\begin{aligned} \max_{\bar{R} \in \mathbb{R}} \quad & \bar{R} \\ \text{subj. to} \quad & \bar{R} - R(s, a) + V(s) - \sum_{s'} P(s' \mid s, a) V(s'), \qquad \text{for all } s, a \in \mathcal{S} \times \mathcal{A}, \end{aligned} \tag{2}$$

where the dual variable $V(\cdot) = \sum_a P(a \mid \cdot) Q(\cdot\,, a)$ is known as the *value function*. Since this is a linear program, if the optimal average reward $\bar{R}^*$ is attained, strong duality holds, and it can be shown that the optimal value function V^* is the solution to the *average-reward Bellman equations*

$$V^*(s) = \max_a \left(R(s, a) - \bar{R}^* + \sum_{s'} P(s' \mid s, a) V^*(s) \right), \quad \text{for all } s \in \mathcal{S}, \tag{3}$$

and that V^* is a fixed point of the Bellman operator $\mathcal{T}_{\pi^*}$ under π^*.

3.2 Regularized Markov Decision Processes

[16] defines a regularized objective $\bar{R}_\eta(\mu)$ inspired by the linear program (2) and proposes the concave optimization problem

$$\max_{\mu \in \mathcal{C}} \bar{R}_\eta(\mu) = \max_{\mu \in \mathcal{C}} \left\{ \sum_{s,a} \left[\mu(s,a) R(s,a) \right] - \frac{1}{\eta} W(\mu) \right\}, \tag{4}$$

where $\eta > 0$ is the scale parameter that trades-off between the original objective and regularization and $W : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \to \mathbb{R}$ is a convex regularization function.

[16] shows that by using Bregman divergences $D_S(\mu \parallel \mu')$, induced by the *negative Shannon entropy*, and $D_C(\mu \parallel \mu')$, induced by the *conditional entropy*, for some reference distribution μ' as choices of $W(\mu)$ in (4), the dual to (4) is similar to the average-reward Bellman equations (3). To note, these choices of W allow us to perform *mirror descent* to find both the primal and dual optimal values.

Furthermore, [16] shows that the popular reinforcement learning algorithms Trust Region Policy Optimization (TRPO) [19] and Asynchronous Advantage Actor-Critic (A3C) [14] can be formulated as optimization problems with the form of (4) for particular choices of W. We will revisit each of these when we discuss TRPO in Section 5.1 and A3C in Section 5.2.

3.2.1 Relative entropy Bregman divergence

Namely, [17, 25, 16] show that for $W := D_S(\cdot \parallel \mu')$, we have

$$\mu_\eta(s,a)^* \propto \mu'(s,a) e^{\eta \left(R(s,a) + \sum_{s'} P(s'|s,a) V^*(s') - V^*(s) \right)}, \tag{5}$$

and the dual function

$$g(V) = \frac{1}{\eta} \log \sum_{s,a} \mu'(s,a) e^{\eta \left(R(s,a) + \sum_{s'} P(s'|s,a) V^*(s') - V^*(s) \right)}. \tag{6}$$

Assuming that strong duality holds, we have

$$g(V_\eta^*) = \bar{R}_\eta^* = \max_{\mu \in \mathcal{C}} \bar{R}_\eta(\mu) \tag{7}$$

3.2.2 Conditional entropy Bregman divergence

[16] shows that for $W := D_C(\cdot \parallel \mu')$, we have

$$\pi_\eta^*(a \mid s) \propto \pi_{\mu'}(a \mid s) e^{\eta \left(R(s,a) + \sum_{s'} P(s'|s,a) V^*(s') - V^*(s) \right)} \tag{8}$$

and the dual problem of (4) is

$$\begin{aligned} \max_{\lambda \in \mathbb{R}} \quad & \lambda \\ \text{subj. to} \quad & V(s) = \frac{1}{\eta} \log \sum_a \pi_{\mu'}(a \mid s) e^{\eta \left(R(s,a) - \lambda + \sum_{s'} P(s'|s,a) V(s') \right)}, \quad \text{for all } s \in \mathcal{S}, \end{aligned} \tag{9}$$

which has dual optimal solution

$$V_\eta^*(s) = \frac{1}{\eta} \log \sum_a \pi_{\mu'}(a \mid s) e^{\eta \left(R(s,a) - \bar{R}_\eta^* + \sum_{s'} P(s'|s,a) V_\eta^*(s') \right)}, \quad \text{for all } s' \in \mathcal{S}. \tag{10}$$

4 Duality in Learning from Demonstration

Learning from demonstration (LfD) is a paradigm in which an agent is given demonstrations from some presumably expert agent and aims to imitate those demonstrations as closely as possible or learn the reward function that the expert was trying to maximize. LfD can be broken into three subfields, behavioral cloning, adversarial imitation learning, and inverse reinforcement learning.

In behavioral cloning (BC), the agent aims to mimic the expert demonstrations without reward signal nor interaction with the environment. Though attractively simple, BC is susceptible to compounding errors due to covariate shift and lack of feedback. Since an agent that performs BC learns only from the demonstrations provided to it, BC usually requires a large amount of data that captures both the behaviors under optimal conditions as well as corrective actions that rectify suboptimal behaviors.

In adversarial imitation learning, a discriminator is given unlabeled demonstrations from the agent and expert. For each demonstration the discriminator is trained to classify whether the demonstration was generated by the agent or by the expert. The agent's goal is to fool the discriminator as often as possible, and by training to maximize this objective, the agent learns to generate demonstrations that are similar to the expert demonstrations. This is akin to the generative adversarial network (GAN) framework [9], and indeed, we will soon see that adversarial imitation learning can be see as GAN training [10]. Unlike in behavioral cloning, in adversarial imitation learning the agent generates trajectories by interacting with the environment.

In inverse reinforcement learning (IRL), the primary objective is to recover the reward function that the expert agent was optimizing for. Often simultaneously, an agent is trained to maximize an iteration of this reward function and generates demonstrations comparable to expert demonstrations at convergence. A major downside to IRL is that it is typically posed as a two-loop problem, in which the inner loop requires performing an expensive reinforcement learning procedure.

4.1 Generative Adversarial Imitation Learning (GAIL)

Generative Adversarial Imitation Learning [10] is a framework for learning an expert policy from expert demonstrations while avoiding the need to recover the optimal reward function. However, it assumes that the expert agent has optimized for some optimal reward function $R^* : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$ from which a unique optimal policy $\pi^* : \mathcal{S} \longrightarrow \mathcal{A}$ can be learned. In fact, since we assume that $R^* \in \text{IRL}(\pi^*)$ and $\pi^* \in \text{RL}(R^*)$, we have

$$R^* \in \arg\max_{R \in \mathcal{R}} \mathbb{E}_{\pi^*} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right].$$

$$\pi^* \in \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R^*(s_t, a_t) \right].$$

IRL optimizes for R^* and π^* simultaneously by solving the saddle-point optimization problem

$$\max_{R \in \mathcal{R}} \left(\min_{\pi \in \Pi} \mathbb{E}_{(s,a) \sim \pi} \left[R(s,a) \right] \right) - \mathbb{E}_{(s,a) \sim \pi^*} \left[R(s,a) \right].$$

Maximum causal entropy IRL [24, 23] aims to recover a policy that not only performs well but also captures diverse behaviors. It does so by modifying the objective directly above to also maximize the causal entropy H of the recovered policy.

[10] shows that under mild conditions, the dual to this modified problem is actually a problem that aims to match the stationary state-action distributions of the expert policy and the policy being trained. Through this result, they propose an adversarial imitation learning framework with guaranteed convergence results as well as a robust, practical imitation learning algorithm.

5 Duality in Policy Search

5.1 Trust Region Policy Optimization (TRPO)

Policy gradient (PG) algorithms [21, 20] are a family of algorithms that aim to recover the optimal policy π^* with respect to some reward function R, usually by interacting with the environment, obtaining Monte Carlo estimates of the discounted rewards, and computing the gradient of the policy with respect to these rewards so that gradient-based optimization over the policy space converges to π^*.

In vanilla PG, the policy being trained undergoes gradient updates that are based on a loss function with respect to the policy parameters. Though the policy is kept close in parameter space, in practice vanilla PG can be highly unstable, and policy performance can vary significantly from iteration to iteration. [19] proposes Trust Region Policy Optimization (TRPO), which overcomes this issue by constraining the expected Kullback-Leibler divergence from the old policy θ^{k+1} to the new policy θ^k over the states encounted while running the old policy. Namely, [19] proposes a heuristic approximation to the solution of the theoretical TRPO problem,

$$\begin{aligned} \theta_{k+1} = \quad & \underset{\theta \in \Theta}{\arg\max} \quad \mathcal{L}(\theta, \theta^k) \\ & \text{subj. to} \quad \bar{D}_{KL}(\theta \parallel \theta^k) \leq \delta, \end{aligned} \tag{11}$$

where the *surrogate advantage* function $\mathcal{L}(\theta, \theta^k)$ quantifies how well the policy π_θ performs compared with the old policy π_{θ^k} using trajectories obtained by running the old policy,

$$\mathcal{L}(\theta, \theta^k) = \mathop{\mathbb{E}}_{(s,a)\sim\pi_{\theta^k}} \left[\frac{\pi_\theta(a \mid s)}{\pi_{\theta^k}(a \mid s)} \left(R(s,a) - \mathop{\mathbb{E}}_{a\sim\pi_{\theta^k}(\cdot|s)} [R(\cdot \mid s)] \right) \right],$$

and

$$\bar{D}_{KL}(\theta \parallel \theta^k) = \mathop{\mathbb{E}}_{s\sim\pi_{\theta^k}} \left[D_{KL}(\pi_\theta(\cdot \mid s) \parallel \pi_{\theta^k}(\cdot \mid s)) \right].$$

Interestingly, by using a linear approximation to the objective function and a quadratic approximation to the constraint in (11), we get exactly the natural policy gradient [12] update,

$$\begin{aligned} \theta^{k+1} = \quad & \underset{\theta \in \Theta}{\arg\max} \quad \nabla_\theta \mathcal{L}(\theta, \theta^k)\Big|_{\theta=\theta^k} \cdot (\theta - \theta^k) \\ & \text{subj. to} \quad (\theta - \theta^k)^T H(\theta^k)(\theta - \theta^k) \leq \delta, \end{aligned} \tag{12}$$

where H is the Hessian matrix of the constraint in (11) with respect to θ. Using duality theory, this problem has the analytical solution

$$\theta^{k+1} = \theta^k + \alpha \sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1} g, \tag{13}$$

where α is determined through backtracking line search and $g = \nabla_\theta \mathcal{L}(\theta, \theta^k)\Big|_{\theta=\theta^k}$.

Computing or storing H^{-1} may be prohibitively expensive, but according to the form of (13), finding $H^{-1}g$ is sufficient, and computing or storing this matrix-vector product is relatively cheap. TRPO uses the conjugate gradient method to obtain $H^{-1}g$ while circumventing computation or storage of H^{-1}, enabling its use in practical settings.

5.1.1 Trust Region Policy Optimization as approximated Mirror Descent

[16] shows that the TRPO update with the form

$$\pi_{k+1} = \arg\max_{\pi} \left\{ \sum_x \nu_{\pi_k}(s) \sum_a \pi(a \mid s) \left(R(s,a) + \sum_{s'} P(s' \mid s,a) V^{\pi_k}_{\eta=\infty}(s') - V^{\pi_k}_{\eta=\infty}(s) - \frac{1}{\eta} \log \frac{\pi(a \mid s)}{\pi_k(a \mid s)} \right) \right\} \quad (14)$$

approximates mirror descent and that this update can be expressed in closed form as

$$\pi_{k+1}(a \mid s) \propto \pi_k(a \mid s) e^{\eta \left(R(s,a) + \sum_{s'} P(s' \mid s,a) V^{\pi_k}_{\eta=\infty}(s') - V^{\pi_k}_{\eta=\infty}(s) \right)}.$$

[19] claims that the theoretical TRPO updates are monotonic improvements, but do not claim whether TRPO converges to an optimal policy. However, through duality theory, [16] shows that the theoretical TRPO updates indeed converge to the optimal policy π^*.

5.2 A3C as Dual Averaging

[16] shows that the A3C algorithm [14] aims to optimize the objective

$$\sum_x \nu_{\pi_k}(s) \sum_a \pi(a \mid s) \left(R(s,a) + \sum_{s'} P(s' \mid s,a) V^{\pi_k}_{\eta=\infty}(s') - V^{\pi_k}_{\eta=\infty}(s) - \frac{1}{\eta^k} \log \pi_\theta(a \mid s) \right),$$

which can be interpreted as a dual-averaging [22, 8] variant of the TRPO objective in (14).

Again through the use of duality theory, [16] conjectures that A3C does *not* converge and proposes a modified method with convergence guarantees.

5.3 Constrained Policy Optimization (CPO)

[1] uses an approximated update similar to (12) but with an additional constraint that is affine in θ to make safety guarantees that hold in expectation. The dual problem in CPO is similar to that of TRPO, and the optimal update to θ can be derived analytically through duality theory here as well.

5.4 A Lyapunov-based approach to Safe Reinforcement Learning

[7] puts forth a primal-dual subgradient method that obtains a policy that satisfies certain safety constraints in expectation.

5.5 Guided Policy Search (GPS)

[13] can be used to train complex high-dimensional policies without optimizing the policies directly in high-dimensional parameter spaces. GPS trains a student policy to imitate a teacher policy that is also being trained to produce behaviors that the the student can demonstrate. This approach formulates the Lagrangian of a specified cost function and takes gradient steps on primal and dual variables in an alternating fashion.

Furthermore, [15] shows that GPS can be seen as an approximate variant of mirror descent. More generally, [17, 25] show that the form of the update in GPS can formulated as mirror descent with the Bregman divergence D_S induced by the relative entropy of the stationary state-action distribution μ_π corresponding to the policy π.

6 Conclusion

The study and application of duality theory has led to key advances in the understanding and efficiency of solving optimization problems in a number of fields. However, duality hasn't received comparable

focus in reinforcement learning even though much of the influential work in RL has relied on it. We've shown that duality arises more often in RL than one may think, with benefits including increased interpretability, convergence guarantees, and state-of-the-art advances in algorithmic performance and efficiency. Going forward we hope others will more strongly consider duality while addressing reinforcement learning problems.

References

[1] Joshua Achiam et al. "Constrained policy optimization". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 22–31.

[2] Amir Beck and Marc Teboulle. "Mirror descent and nonlinear projected subgradient methods for convex optimization". In: *Operations Research Letters* 31.3 (2003), pp. 167–175.

[3] Richard Bellman. "Dynamic programming". In: *Science* 153.3731 (1966), pp. 34–37.

[4] Dimitri P Bertsekas et al. *Dynamic programming and optimal control*. Vol. 1. 2. Athena scientific Belmont, MA, 1995.

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

[6] Stephen Boyd et al. "Distributed optimization and statistical learning via the alternating direction method of multipliers". In: *Foundations and Trends® in Machine learning* 3.1 (2011), pp. 1–122.

[7] Yinlam Chow et al. "A lyapunov-based approach to safe reinforcement learning". In: *Advances in neural information processing systems*. 2018, pp. 8092–8101.

[8] John C Duchi, Alekh Agarwal, and Martin J Wainwright. "Dual averaging for distributed optimization: Convergence analysis and network scaling". In: *IEEE Transactions on Automatic control* 57.3 (2011), pp. 592–606.

[9] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[10] Jonathan Ho and Stefano Ermon. "Generative adversarial imitation learning". In: *Advances in neural information processing systems*. 2016, pp. 4565–4573.

[11] Ronald A Howard. "Dynamic programming and markov processes." In: (1960).

[12] Sham M Kakade. "A natural policy gradient". In: *Advances in neural information processing systems*. 2002, pp. 1531–1538.

[13] Sergey Levine and Vladlen Koltun. "Guided policy search". In: *International Conference on Machine Learning*. 2013, pp. 1–9.

[14] Volodymyr Mnih et al. "Asynchronous methods for deep reinforcement learning". In: *International conference on machine learning*. 2016, pp. 1928–1937.

[15] William H Montgomery and Sergey Levine. "Guided policy search via approximate mirror descent". In: *Advances in Neural Information Processing Systems*. 2016, pp. 4008–4016.

[16] Gergely Neu, Anders Jonsson, and Vicenç Gómez. "A unified view of entropy-regularized markov decision processes". In: *arXiv preprint arXiv:1705.07798* (2017).

[17] Jan Peters, Katharina Mulling, and Yasemin Altun. "Relative entropy policy search". In: *Twenty-Fourth AAAI Conference on Artificial Intelligence*. 2010.

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

[19] John Schulman et al. "Trust region policy optimization". In: *International conference on machine learning*. 2015, pp. 1889–1897.

[20] David Silver et al. "Deterministic policy gradient algorithms". In: 2014.

[21] Richard S Sutton et al. "Policy gradient methods for reinforcement learning with function approximation". In: *Advances in neural information processing systems*. 2000, pp. 1057–1063.

[22] Lin Xiao. "Dual averaging methods for regularized stochastic learning and online optimization". In: *Journal of Machine Learning Research* 11.Oct (2010), pp. 2543–2596.

[23] Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. “Modeling interaction via the principle of maximum causal entropy”. In: (2010).

[24] Brian D Ziebart et al. “Maximum entropy inverse reinforcement learning.” In: 2008.

[25] Alexander Zimin and Gergely Neu. “Online learning in episodic Markovian decision processes by relative entropy policy search”. In: *Advances in neural information processing systems*. 2013, pp. 1583–1591.