

A Unified View of Entropy-Regularized Markov Decision Processes

Gergely Neu

Universitat Pompeu Fabra, Barcelona, Spain

GERGELY.NEU@GMAIL.COM

Vicenç Gómez

Universitat Pompeu Fabra, Barcelona, Spain

VICEN.GOMEZ@UPF.EDU

Anders Jonsson

Universitat Pompeu Fabra, Barcelona, Spain

ANDERS.JONSSON@UPF.EDU

Abstract

We propose a general framework for entropy-regularized average-reward reinforcement learning in Markov decision processes (MDPs). Our approach is based on extending the linear-programming formulation of policy optimization in MDPs to accommodate convex regularization functions. Our key result is showing that using the conditional entropy of the joint state-action distributions as regularization yields a dual optimization problem closely resembling the Bellman optimality equations. This result enables us to formalize a number of state-of-the-art entropy-regularized reinforcement learning algorithms as approximate variants of Mirror Descent or Dual Averaging, and thus to argue about the convergence properties of these methods. In particular, we show that the exact version of the TRPO algorithm of [Schulman et al. \(2015\)](#) actually converges to the optimal policy, while the entropy-regularized policy gradient methods of [Mnih et al. \(2016\)](#) may fail to converge to a fixed point. Finally, we illustrate empirically the effects of using various regularization techniques on learning performance in a simple reinforcement learning setup.

1. Introduction

Reinforcement learning is the discipline of model-based optimal sequential decision-making in unknown stochastic environments. In average-reward reinforcement learning, the goal is to find a behavior policy that maximizes the long-term average reward, taking into account the effect of each decision on the future evolution of the decision-making process. In known environments, this optimization problem has been studied (at least) since the influential work of [Bellman \(1957\)](#) and [Howard \(1960\)](#): the optimal behavior policy can be formulated as the solution of the Bellman optimality equations. In unknown environments with partially known or misspecified models, greedily solving these equations often results in policies that are far from optimal in the true environment. Rooted in statistical learning theory ([Vapnik, 2013](#)), the notion of

regularization offers a principled way of dealing with this issue, among many others. In particular, *entropy regularization* has proven to be one of the most successful tools of machine learning and related fields (Littlestone and Warmuth, 1994; Vovk, 1990; Freund and Schapire, 1997; Kivinen and Warmuth, 2001; Arora et al., 2012).

The idea of entropy regularization has also been used extensively in the reinforcement learning literature (Sutton and Barto, 1998; Szepesvári, 2010). Entropy-regularized variants of the classic Bellman equations and the entailing reinforcement-learning algorithms have been proposed to induce safe exploration (Fox et al., 2016) and risk-sensitive policies (Howard and Matheson, 1972; Marcus et al., 1997; Ruszczyński, 2010), or to model observed behavior of imperfect decision-makers (Ziebart et al., 2010; Ziebart, 2010; Braun et al., 2011), among others. Complementary to these approaches rooted in dynamic programming, another line of work proposes direct policy search methods attempting to optimize various entropy-regularized objectives (Williams and Peng, 1991; Peters et al., 2010; Schulman et al., 2015; Mnih et al., 2016; O’Donoghue et al., 2017), with the main goal of driving a safe online exploration procedure in an unknown Markov decision process. Notably, the state-of-the-art methods of Mnih et al. (2016) and Schulman et al. (2015) are both based on entropy-regularized policy search.

In this work, we connect these two seemingly disparate lines of work by showing a strong Lagrangian duality between the entropy-regularized Bellman equations and a certain regularized average-reward objective. Specifically, we extend the linear-programming formulation of the problem of optimization in MDPs to accommodate convex regularization functions, resulting in a convex program. We show that using the *conditional entropy* of the joint state-action distribution gives rise to a set of nonlinear equations resembling the Bellman optimality equations. Observing this duality enables us to establish a connection between regularized versions of value and policy iteration methods (Puterman and Shin, 1978) and incremental convex optimization methods like Mirror Descent (Nemirovski and Yudin, 1983; Beck and Teboulle, 2003) or Dual Averaging (Xiao, 2010; McMahan, 2014; Hazan et al., 2016; Shalev-Shwartz, 2012). For instance, the convex-optimization view we propose reveals that the TRPO algorithm of Schulman et al. (2015) and the regularized policy-gradient method of Mnih et al. (2016) are approximate versions of Mirror Descent and Dual Averaging, respectively, and that both can be interpreted as regularized policy iteration methods.

Our work provides a theoretical justification for various algorithms that were first derived heuristically. In particular, our framework reveals that the exact version of TRPO is *identical* to the MDP-E algorithm of Even-Dar et al. (2009). This establishes the fact that the policy updates of TRPO converge to the optimal policy, improving on the theoretical results claimed by Schulman et al. (2015). We also argue that our formulation is useful for pointing out possible inconsistencies of heuristic learning algorithms. In particular, we show that the approximation steps employed by Mnih et al. (2016) may break the convexity of the objective, thus possibly leading to convergence to bad local optima or even divergence. This observation is in accordance

with the very recent findings of [Asadi and Littman \(2016\)](#), who show that value iteration with poorly chosen approximate updates may lead to divergence. To complement these results, we suggest an alternative objective that can be optimized consistently, avoiding the possibility of diverging.

A similar Lagrangian duality between the Bellman equations and entropy maximization has been previously noted by [Ziebart \(2010, Sec. 5.2\)](#) and [Rawlik et al. \(2012\)](#) for a special class of *episodic* Markov decision processes where the time index within the episode is part of the state representation. In this particular setting, the convexity of the conditional entropy is more obvious. One of our key observations is pointing out the convexity of the conditional entropy of distribution functions defined over general state spaces, which enables us to develop a much broader theory of regularized Markov decision processes. We note that our theory also readily extends to discounted MDPs by replacing the stationary state-action distributions we consider by *discounted state-action occupancy measures*. For consistency, we will discuss each particular algorithm in their most natural average-reward version, noting that all conclusions remain valid in the simpler discounted and episodic settings.

The rest of the paper is organized as follows. In [Section 2](#), we provide background on average-reward Markov decision processes, briefly discussing both linear-programming and dynamic-programming derivations of the optimal control. In [Section 3](#), we provide a convex-programming formulation of regularized average-reward Markov decision processes, and show the connection to the regularized Bellman equations. [Section 4](#) provides a brief summary of the complementary dynamic-programming formulation and discusses regularized equivalents of related concepts, such as expressions of the regularized policy gradient. In [Section 5](#), we describe several existing learning algorithms in our framework. We provide an empirical evaluation of various regularization schemes in a simple reinforcement learning problem in [Section 6](#).

Notation. Given a finite set \mathcal{S} , we will often use \sum_s as shorthand for $\sum_{s \in \mathcal{S}}$, and we use $\Delta(\mathcal{S}) = \{\mu \in \mathbb{R}^{\mathcal{S}} : \sum_s \mu(s) = 1, \mu(s) \geq 0 (\forall s)\}$ to denote the set of all probability distributions on \mathcal{S} .

2. Preliminaries on Markov decision processes

We consider a finite Markov decision process (MDP) $M = (\mathcal{X}, \mathcal{A}, P, r)$, where \mathcal{X} is the finite state space, \mathcal{A} is the finite action space, $P : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \rightarrow [0, 1]$ is the transition function, with $P(y|x, a)$ denoting the probability of moving to state y from state x when taking action a , and $r : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function mapping state-action pairs to rewards.

In each round t , the learner observes state $X_t \in \mathcal{X}$, selects action $A_t \in \mathcal{A}$, moves to the next state $X_{t+1} \sim P(\cdot|X_t, A_t)$, and obtains reward $r(X_t, A_t)$. The goal is to select actions as to maximize some notion of cumulative reward. In this paper we consider the *average-reward* criterion $\liminf_{T \rightarrow \infty} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T r_t(X_t, A_t) \right]$. A *stationary state-feedback*

policy (or *policy* for short) defines a probability distribution $\pi(\cdot|x)$ over the learner’s actions in state x . MDP theory (see, e.g., [Puterman \(1994\)](#)) stipulates that under mild conditions, the average-reward criterion can be maximized by stationary policies. Throughout the paper, we make the following mild assumption about the MDP:

Assumption 1 *The MDP M is unichain: All stationary policies π induce a unique stationary distribution ν_π over the state space satisfying $\nu_\pi(y) = \sum_{x,a} P(y|x, a)\pi(a|x)\nu_\pi(x)$ for all $y \in \mathcal{X}$.*

In particular, this assumption is satisfied if all policies induce an irreducible and aperiodic Markov chain ([Puterman, 1994](#)). For ease of exposition in this section, we also make the following simplifying assumption:

Assumption 2 *The MDP M admits a single recurrent class: All stationary policies π induce stationary distributions strictly supported on the same set $\mathcal{X}' \subseteq \mathcal{X}$.*

In general, this assumption is very restrictive in that it does not allow policies to cover different parts of the state space. We stress that our results in the later sections *do not* require this assumption to hold. With the above assumptions in mind, we can define the average reward of any policy π as

$$\rho(\pi) = \lim_{T \rightarrow \infty} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T r_t(X_t, A_t) \right],$$

where $A_t \sim \pi(\cdot|X_t)$ in each round t and the existence of the limit is ensured by [Assumption 1](#). Furthermore, the average reward of any policy π can be simply written as $\rho(\pi) = \sum_{x,a} \nu_\pi(x)\pi(a|x)r(x, a)$, which is a linear function of the stationary state-action distribution $\mu_\pi = \nu_\pi\pi$. This suggests that finding the optimal policy can be equivalently written as a linear program (LP) where the decision variable is the stationary state-action distribution. Defining the set of all feasible stationary distributions as

$$\Delta = \left\{ \mu \in \Delta(\mathcal{X} \times \mathcal{A}) : \sum_b \mu(y, b) = \sum_{x,a} P(y|x, a)\mu(x, a) \quad (\forall y) \right\}, \quad (1)$$

the problem of maximizing the average reward can be written as

$$\mu^* = \arg \max_{\mu \in \Delta} \rho(\mu). \quad (2)$$

Just as a policy π induces stationary distributions ν_π and μ_π , a stationary distribution μ induces a state distribution ν_μ defined as $\nu_\mu(x) = \sum_a \mu(x, a)$ and a policy π_μ defined as $\pi_\mu(a|x) = \mu(x, a)/\nu_\mu(x)$, where the denominator is strictly positive for recurrent states by [Assumption 2](#). Since Δ is a compact polytope (non-empty by [Assumption 1](#))

the maximum in (2) is well-defined and induces an optimal policy π_{μ^*} in recurrent states. Due to Assumption 2, π_{μ^*} can be arbitrarily defined in transient states.

The linear program specified in Equation (2) is well studied in the MDP literature (see, e.g., Puterman, 1994, Section 8.8), although most commonly as the dual of the linear program

$$\begin{aligned} \min_{\rho \in \mathbb{R}} \quad & \rho & (3) \\ \text{subject to} \quad & \rho + V(x) - \sum_y P(y|x, a)V(y) \geq r(x, a), \quad \forall(x, a). & (4) \end{aligned}$$

Here, the dual variables V are commonly referred to as the *value functions*. By strong LP duality and our Assumption 1, the solution to this LP equals the optimal average reward ρ^* and the dual variables V^* at the optimum are the solution to the *average-reward Bellman optimality equations*

$$V^*(x) = \max_a \left(r(x, a) - \rho^* + \sum_y P(y|x, a)V^*(x) \right), \quad (\forall x). \quad (5)$$

Note that V^* is not unique as for any solution V , a constant shift $V - c$ for any $c \in \mathbb{R}$ is also a solution. However, we can obtain a unique solution V^* by imposing the additional constraint $\sum_{x,a} \mu^*(x, a)V^*(x) = 0$, which states that the expected value should equal 0.

3. Regularized MDPs: A convex-optimization view

Inspired by the LP formulation of the average-reward optimization problem (2), we now define a regularized optimization objective—a framework that will lead us to our main results. Our results in this section only require the mild Assumption 1. Our regularized optimization problem takes the form

$$\max_{\mu \in \Delta} \tilde{\rho}_\eta(\mu) = \max_{\mu \in \Delta} \left\{ \sum_{x,a} \mu(x, a)r(x, a) - \frac{1}{\eta}R(\mu) \right\}, \quad (6)$$

where $R(\mu) : \mathbb{R}^{\mathcal{X} \times \mathcal{A}} \rightarrow \mathbb{R}$ is a convex regularization function and $\eta > 0$ is a *learning rate* that trades off the original objective and regularization. Note that $\eta = \infty$ recovers the unregularized objective. Unlike previous work on LP formulations for MDPs, we find it useful to regard (6) as the *primal*.

We focus on two families of regularization functions: the *negative Shannon entropy* of $(X, A) \sim \mu$,

$$R_S(\mu) = \sum_{x,a} \mu(x, a) \log \mu(x, a), \quad (7)$$

and the *negative conditional entropy* of $(X, A) \sim \mu$,

$$R_C(\mu) = \sum_{x,a} \mu(x, a) \log \frac{\mu(x, a)}{\sum_b \mu(x, b)} = \sum_{x,a} \nu_\mu(x) \pi_\mu(a|x) \log \pi_\mu(a|x). \quad (8)$$

In what follows, we refer to these functions as the relative entropy and the conditional entropy. We also make use of the Bregman divergences induced by R_S and R_C which take the respective forms

$$D_S(\mu||\mu') = \sum_{x,a} \mu(x, a) \log \frac{\mu(x, a)}{\mu'(x, a)} \quad \text{and} \quad D_C(\mu||\mu') = \sum_{x,a} \mu(x, a) \log \frac{\pi_\mu(a|x)}{\pi_{\mu'}(a|x)}.$$

While the form of D_S is standard (it is the relative entropy between two state-action distributions), the fact that D_C is the Bregman divergence of R_C (or even that R_C is convex) is not immediately obvious¹. The following proposition asserts this statement, which we prove in Appendix A.1. The only work we are aware of that establishes a comparable result is the recent paper of [Neu and Gómez \(2017\)](#).

Proposition 1 *The Bregman divergence corresponding to the conditional entropy R_C is D_C . Furthermore, D_C is nonnegative on Δ , implying that R_C is convex and D_C is convex in its first argument.*

We proceed to derive the dual functions and optimal solutions to (6) for our two choices of regularization functions. Without loss of generality, we assume that the reference policy $\pi_{\mu'}$ has full support, which implies that the corresponding stationary distribution μ' is strictly positive on the recurrent set \mathcal{X}' . We only provide the derivations for the Bregman divergences; the calculations are analogous for R_S and R_C . Both of these solutions will be expressed with the help of dual variables $V : \mathbb{R}^{\mathcal{X}} \rightarrow \mathbb{R}$ which are useful to think about as *value functions*, as in the case of the LP formulation (2). We also define the corresponding *advantage functions* $A(x, a) = r(x, a) + \sum_y P(y|x, a)V(y) - V(x)$.

3.1. Relative entropy

The choice $R = D_S(\cdot||\mu')$ has been studied before by [Peters et al. \(2010\)](#) and [Zimin and Neu \(2013\)](#); we defer the proofs to Appendix A.3. The optimal state-action distribution for a given value of η is

$$\mu_\eta^*(x, a) \propto \mu'(x, a) e^{\eta A_\eta^*(x, a)}, \quad (9)$$

where A_η^* is the advantage function for the optimal dual variables V_η^* . The dual function is

$$g(V) = \frac{1}{\eta} \log \sum_{x,a} \mu'(x, a) e^{\eta A(x, a)}, \quad (10)$$

1. In the special case of loop-free episodic environments, showing the convexity of R_C is straightforward ([Lafferty et al., 2001](#); [Ziebart, 2010](#); [Rawlik et al., 2012](#)).

that now needs to be minimized on $\mathbb{R}^{\mathcal{X}}$ with no constraints in order to obtain V_η^* . By strong duality, g is convex in V and takes the value $\tilde{\rho}_\eta^* = \max_{\mu \in \Delta} \tilde{\rho}_\eta(\mu)$ at its optimum.

3.2. Conditional entropy

The choice $R = D_C(\cdot \| \mu')$ leads to our main contributions. Similar to above, the optimal policy is

$$\pi_\eta^*(a|x) \propto \pi_{\mu'}(a|x) e^{\eta A_\eta^*(x,a)}. \quad (11)$$

In this case, the dual problem closely resembles the average-reward Bellman optimality equations (5):

Proposition 2 *The dual of the optimization problem (6) when $R = D_C(\cdot \| \mu')$ is given by*

$$\begin{aligned} & \min_{\lambda \in \mathbb{R}} \quad \lambda \\ & \text{subject to} \quad V(x) = \frac{1}{\eta} \log \sum_a \pi_{\mu'}(a|x) \exp \left(\eta \left(r(x, a) - \lambda + \sum_y P(y|x, a) V(y) \right) \right), \quad (\forall x). \end{aligned}$$

We defer the proofs to Appendix A.4. Using strong duality, the optimum of the above problem is $\tilde{\rho}_\eta^*$, which implies that the optimal dual variables V_η^* are given as a solution to the system of equations

$$V_\eta^*(x) = \frac{1}{\eta} \log \sum_a \pi_{\mu'}(a|x) \exp \left(\eta \left(r(x, a) - \tilde{\rho}_\eta^* + \sum_y P(y|x, a) V_\eta^*(y) \right) \right), \quad (\forall x). \quad (12)$$

By analogy with the Bellman optimality equations (5), we call this the *regularized average-reward Bellman optimality equations*. Since $\tilde{\rho}_\eta^*$ is guaranteed to be finite (because it is the maximum of a bounded function on a compact domain), the solution to the above optimization problem is well-defined, bounded, and unique up to a constant shift (as in the case of the LP dual variables). Again, we can make the solution unique by imposing the constraint that the expected value should equal 0.

4. Dynamic programming in regularized MDPs

We now present a dynamic-programming view of the regularized optimization problem (Bertsekas, 2007) for the choice $R = D_C(\cdot \| \mu')$. Similar derivations have been done several times for *discounted and episodic MDPs* (Littman and Szepesvári, 1996; Ruzsչyński, 2010; Azar et al., 2011; Rawlik et al., 2012; Asadi and Littman, 2016; Fox et al., 2016), but we are not aware of any work that considers the average-reward case. That said, the generalization is straightforward, and the existence and unicity of

the optimal solution to the Bellman optimality equation (12) follows from our results in the previous section.

We first define the *regularized Bellman equations* for an arbitrary policy π and a reference policy π' :

$$V_\eta^\pi(x) = \sum_a \pi(a|x) \left(r(x, a) - \frac{1}{\eta} \log \frac{\pi(a|x)}{\pi'(a|x)} - \tilde{\rho}_\eta(\pi) + \sum_y P(y|x, a) V_\eta^\pi(y) \right) \quad (\forall x), \quad (13)$$

where $\tilde{\rho}_\eta(\pi)$ is the regularized average reward of policy π defined as in Equation (6). By our Assumption 1 and Proposition 4.2.4 of Bertsekas (2007), it is easy to show that this system of equations has a unique solution satisfying the additional constraint $\sum_{x,a} \mu_\pi(x, a) V_\eta^\pi(x) = 0$. We also define the *Bellman optimality operator* $T_\eta^{*\pi'}$ and the *Bellman operator* $T_\eta^{\pi|\pi'}$ that correspond to the Bellman equations (12) and (13), respectively, as well as the greedy policy operator $G_\eta^{\pi'}$ that corresponds to Equation (11) (for completeness, the formal definitions appear in Appendix B).

We include two results that are useful for deriving approximate dynamic programming algorithms. We first provide a counterpart to the *performance-difference lemma* (Burnetas and Katehakis, 1997, Prop. 1, Kakade and Langford, 2002, Lemma 6.1, Cao, 2007). This statement will rely on the regularized advantage function A_η^π defined for each policy π as

$$A_\eta^\pi(x, a) = r(x, a) - \frac{1}{\eta} \log \frac{\pi(a|x)}{\pi'(a|x)} - \tilde{\rho}_\eta(\pi) + \sum_y P(y|x, a) V_\eta^\pi(y) - V_\eta^\pi(x),$$

where V_η^π is the regularized value function corresponding to π with baseline π' .

Lemma 3 *For any pair of policies π, π^* , we have*

$$\tilde{\rho}_\eta(\pi^*) - \tilde{\rho}_\eta(\pi) = \sum_{x,a} \mu_{\pi^*}(x, a) A_\eta^\pi(x, a) - \frac{1}{\eta} D_C(\mu_{\pi^*} \| \mu_\pi).$$

For completeness, we provide the simple proof in Appendix B.1.

Second, we provide an expression for the gradient of $\tilde{\rho}_\eta$, thus providing a regularized counterpart of the *policy gradient theorem* of Sutton et al. (1999). To formalize this statement, let us consider a policy π_θ parametrized by a vector $\theta \in \mathbb{R}^d$ and assume that the gradient $\nabla \pi_\theta(a|x)$ exists for all x, a and all θ . The form of the policy gradient is given by the following lemma, which we prove in Appendix B.2:

Lemma 4 *Assume that $\frac{\partial \pi_\theta(a|x)}{\partial \theta_i} / \pi_\theta(a|x) > 0$ for all θ_i, x, a . The gradient of $\tilde{\rho}_\eta$ exists and satisfies*

$$\nabla \tilde{\rho}_\eta(\theta) = \sum_{x,a} \mu_{\pi_\theta}(x, a) \nabla \log \pi_\theta(a|x) A_\eta^{\pi_\theta}(x, a).$$

5. Algorithms

In this section we derive several reinforcement learning algorithms based on our results. For clarity of presentation, we assume that the MDP M is fully known, an assumption that we later relax in the experimental evaluation. We will study a generic sequential optimization framework where a sequence of policies π_k are computed iteratively. Inspired by the online convex optimization literature (see, e.g., [Shalev-Shwartz, 2012](#); [Hazan et al., 2016](#)) and by our convex-optimization formulation, we study two families of algorithms: Mirror Descent and Dual Averaging (also known as Follow-the-Regularized-Leader).

5.1. Iterative policy optimization by Mirror Descent

A direct application of the Mirror Descent algorithm ([Nemirovski and Yudin, 1983](#); [Beck and Teboulle, 2003](#); [Martinet, 1978](#); [Rockafellar, 1976](#)) to our case is defined as

$$\mu_{k+1} = \arg \max_{\mu \in \Delta} \left\{ \rho(\mu) - \frac{1}{\eta} D_R(\mu \| \mu_k) \right\}, \quad (14)$$

where D_R is the Bregman divergence associated with the convex regularization function R . We now proceed to show how various learning algorithms can be recovered from this formulation.

5.1.1. MIRROR DESCENT WITH THE RELATIVE ENTROPY

We first remark that the Relative Entropy Policy Search (REPS) algorithm of [Peters et al. \(2010\)](#) can be formulated as an instance of Mirror Descent with the Bregman divergence D_S . This is easily seen by comparing the form of the update rule (14) with the problem formulation of [Peters et al. \(2010, pp. 2\)](#), with the slight difference that our regularization is additive and theirs is enforced as a constraint. It is easy to see that this only amounts to a change in learning rate. This connection is not new: it has been first shown by [Zimin and Neu \(2013\)](#)², and has been recently rediscovered by [Montgomery and Levine \(2016\)](#). Independently of each other, [Zimin and Neu \(2013\)](#) and [Dick et al. \(2014\)](#) both show that Mirror Descent achieves near-optimal regret guarantees in an online learning setup where the transition function is known, but the reward function is allowed to change arbitrarily between decision rounds. This implies that REPS duly converges to the optimal policy in our setup.

5.1.2. MIRROR DESCENT WITH THE CONDITIONAL ENTROPY

We next show that the Dynamic Policy Programming (DPP) algorithm of [Azar et al. \(2012\)](#) and the Trust-Region Policy Optimization (TRPO) algorithm of [Schulman et al. \(2015\)](#) are both approximate variants of Mirror Descent with the Bregman

2. Although they primarily referred to Mirror Descent as the “Proximal Point Algorithm” following ([Rockafellar, 1976](#); [Martinet, 1978](#)).

divergence D_C . To see this, note that a full Mirror Descent update requires computing the optimal value function V_η^* for the baseline μ_k , e.g. by regularized value iteration or regularized policy iteration (see Appendix B). Since a full update for V_η^* is expensive, DPP and TRPO provide two ways to approximate it. We remark that the algorithm of Rawlik et al. (2012) can also be viewed as an instance of Mirror Descent for the finite-horizon episodic setting, in which the exact update can be computed efficiently by dynamic programming.

Dynamic Policy Programming. We first claim that each iteration of DPP is a *single regularized value iteration step*: Starting from the previous value function V_k , it extracts the greedy policy $\pi_{k+1} = G_\eta^{\pi_k}[V_k]$ and applies the Bellman optimality operator $T_\eta^{*|\pi_k}$ to obtain $V_{k+1} = T_\eta^{*|\pi_k}[V_k]$. This follows from comparing the form of DPP presented in Appendix A of Azar et al. (2012): their update rules (19) and (20) precisely match the discounted analogue of our expressions (26) in Appendix B with $\pi' = \pi_k$. The convergence guarantees proved by Azar et al. (2012) demonstrate the soundness of this approximate update.

Trust-Region Policy Optimization. Second, we claim that each iteration of TRPO is a *single policy iteration step*: TRPO first fully evaluates the policy π_k to compute its *unregularized* value function $V_k = V_\infty^{\pi_k}$ and then extracts the regularized greedy policy $\pi_{k+1} = G_\eta^{\pi_k}[V_k]$ with π_k as a baseline. This can be seen by inspecting the TRPO update³ that takes the form

$$\pi_{k+1} = \arg \max_{\pi} \left\{ \sum_x \nu_{\pi_k}(x) \sum_a \pi(a|x) \left(A_\infty^{\pi_k}(x, a) - \frac{1}{\eta} \log \frac{\pi(a|x)}{\pi_k(a|x)} \right) \right\}.$$

This objective approximates Mirror Descent by ignoring the effect of changing the policy on the state distribution. Surprisingly, using our formalism, this update can be expressed in closed form as

$$\pi_{k+1}(a|x) \propto \pi_k(a|x) e^{\eta A_\infty^{\pi_k}(x, a)}.$$

We present the detailed derivations in Appendix B.3. A particularly interesting consequence of this result is that TRPO is *completely equivalent* to the MDP-E algorithm of Even-Dar et al. (2009) (see also (Neu et al., 2010, 2014)), which is known to minimize regret in an online setting, thus implying that TRPO also converges to the optimal policy in the stationary setting. This guarantee is much stronger than the ones provided by Schulman et al. (2015), who only claim that TRPO produces a monotonically improving sequence of policies (which may still converge to a suboptimal policy).

3. As in the case of REPS, we discuss here the additive-regularization version of the algorithm. The entropy-constrained update actually implemented by Schulman et al. (2015) only differs in the learning rate.

5.2. Iterative policy optimization by Dual Averaging

We next study algorithms arising from the Dual Averaging scheme (Xiao, 2010; McMahan, 2014), commonly known as Follow-the-Regularized-Leader in online learning (Shalev-Shwartz, 2012; Hazan et al., 2016). This algorithm is defined by the iteration

$$\mu_{k+1} = \arg \max_{\mu \in \Delta} \left\{ \rho(\mu) - \frac{1}{\eta_k} R(\mu) \right\}, \quad (15)$$

where η_k is usually an increasing sequence to ensure convergence in the limit. We are unaware of any pure instance of dual averaging using relative entropy, and only discuss conditional entropy below.

5.2.1. DUAL AVERAGING WITH THE CONDITIONAL ENTROPY

Just as for Mirror Descent, a full update (15) requires computing the optimal value function V_η^* . Various approximations of this update have been long studied in the RL literature—see, e.g., (Littman and Szepesvári, 1996) (with additional discussion by (Asadi and Littman, 2016)), (Perkins and Precup, 2002; Ruszczyński, 2010; Petrik and Subramanian, 2012; Fox et al., 2016). In this section, we focus on the state-of-the-art algorithms of Mnih et al. (2016) and O’Donoghue et al. (2017) that were originally derived from an optimization formulation resembling our Equation (6). Our main insight is that this algorithm can be adjusted to have a dynamic-programming interpretation and a convergence guarantee.

Entropy-regularized policy gradients. The A3C algorithm of Mnih et al. (2016) aims to maximize

$$\rho(\pi) - \frac{1}{\eta_k} \sum_x \nu_{\pi_k}(x) \sum_a \pi(a|x) \log \pi(a|x)$$

by taking policy gradient steps. Interestingly, our formalism implies a connection between TRPO and A3C. Due to Lemma 4, the gradient of $\rho(\pi_\theta)$ with respect to θ coincides with the gradient of $\sum_x \nu_{\pi_k}(x) \sum_a \pi_\theta(a|x) A_\infty^{\pi_k}(x, a)$, so A3C actually attempts to optimize the objective

$$\sum_x \nu_{\pi_k}(x) \sum_a \pi_\theta(a|x) \left(A_\infty^{\pi_k}(x, a) - \frac{1}{\eta_k} \log \pi_\theta(a|x) \right). \quad (16)$$

This objective can be seen as the dual-averaging counterpart of the TRPO objective. As in the case of TRPO, the maximizer of this objective can be computed in closed form as

$$\pi_{k+1}(a|x) \propto e^{\eta_k A_\infty^{\pi_k}}.$$

Unlike for TRPO, we are not aware of any convergence results for A3C, and we believe the algorithm does not converge. Indeed, the objective (16) is non-convex in either

of the natural parameters μ or π , which can cause premature convergence to a bad local optimum. An even more serious concern is that the objective function changes between iterations, so gradient descent may fail to converge to *any* stationary point. This problem is avoided by TRPO since the sum of the TRPO objectives is a sensible optimization objective (Even-Dar et al., 2009, Theorem 4.1). However, there is no such clear interpretation for the objective (16).

O’Donoghue et al. (2017, Section 3.1) study the stationary points of the objective (16) and, similarly to us, show a connection between a certain type of value function and a policy achieving the stationary point. However, they do not show that this stationary point is unique nor that gradient descent converges to a stationary point⁴. As we argue above, this may very well not be the case. These observations are consistent with those of Asadi and Littman (2016), who show that softmax policy updates may lead to inconsistent behavior when used in tandem with unregularized advantage functions.

To overcome these issues, we advocate for directly optimizing the objective (15) instead of (16) via gradient descent. Due to the fact that (15) is convex in μ and to standard results regarding dual averaging (McMahan, 2014), this scheme is guaranteed to converge to the optimal policy. Estimating the gradients can be done analogously as for the unregularized objective, by our Lemma 4.

6. Experiments

In this section we analyze empirically several of the algorithms described in the previous section, with the objective of illustrating the interplay of regularization and model-estimation error in a simple reinforcement learning setting. We consider an iterative setup where in each episode $k = 1, 2, \dots, N$, we execute a policy π_k , observe the sample transitions and update the estimated model via maximum likelihood. We focus on the regularization aspect, with no other approximation error than that introduced by model estimation. It is important to emphasize that the comparison may not extend to other variants of the algorithms or in the presence of other sources of approximation.

We consider a simple MDP, defined on a grid (Fig. 1, left), where an agent has four possible actions (up, down, left and right) that succeed with probability 0.9 but fail with probability 0.1. In case of failure, the agent does not move or goes to any random adjacent location. Negative (or positive) rewards are given after hitting a wall (or reaching one of the white diamond locations, respectively). In both cases, the agent is sent back to one of the starting locations (marked with ‘X’ in the figure).

The reward of the diamonds is proportional to the distance from the starting locations. Therefore, the challenge of this experiment is to discover the path towards the top-right reward while learning the dynamics incrementally, and then exploit it.

4. Strictly speaking, O’Donoghue et al. (2017) do not even show that a stationary point (16) exists.

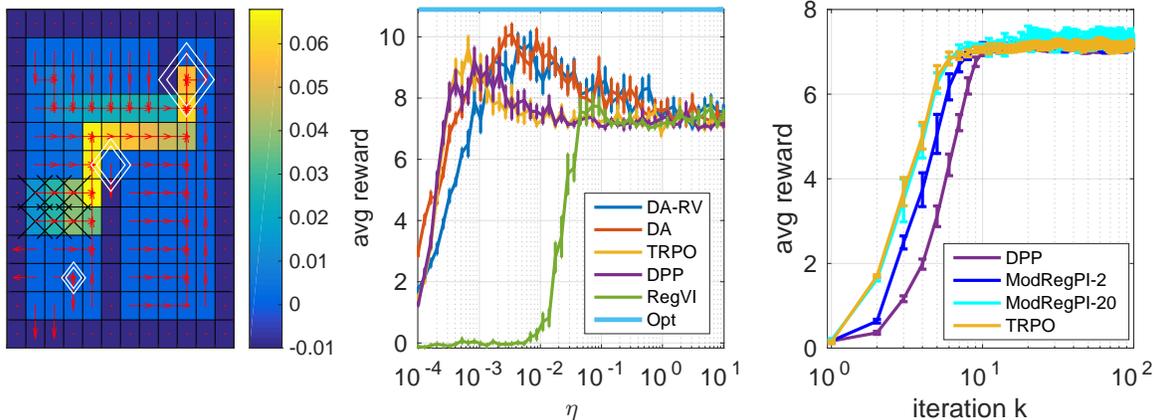


Figure 1: **Left:** the MDP used for evaluation. Reward is -0.1 at the walls and $5, 50, 200$ at the diamonds. The optimal policy is indicated by red arrows. The cell colors correspond to the stationary state distribution for open locations. **Middle:** Average reward as a function of the learning rate η for all algorithms (see text for details). Number of iterations N and samples per iteration S are $N = S = 500$. Results are taken over 20 random runs per value of η . **Right:** Performance of DPP, TRPO and two version of modified regularized Policy Iteration for a fixed $\eta \approx 0.1$.

Note that the optimal agent ignores the intermediate reward at the center and even prefers to hit a wall in locations too far away from the largest reward (bottom-left).

We fix the number of iterations N and samples per iteration S and analyze the average reward of the final policy as a function of η . We compare the following algorithms: regularized Value Iteration with a fixed reference uniform policy and fixed η (RegVI); several variants of approximate Mirror Descent, including DPP and TRPO (Section 5.1); and two Dual Averaging methods (DA and DA-RV). DA corresponds to optimizing the objective (16), which is not guaranteed to lead to an optimal policy, and DA-RV corresponds to the iteration (15), which has convergence guarantees (Section 5.2). For both variants, we use a linear annealing schedule $\eta_k = \eta \cdot k$.

Fig. 1 (middle) shows results as a function of η . The maximum reward is depicted in blue at the top. For very small η (strong regularization), all algorithms perform poorly and do not even reach the intermediate reward. In contrast, for very large η , they converge prematurely to the greedy policy that exploits the intermediate reward. Typically, for an intermediate value of η , the algorithms occasionally discover the optimal path and exploit it. Note that this is not the case for RegVI, which never obtains the optimal policy. This shows that using both a fixed value of η and a fixed reference policy is a bad choice. In this MDP, we observe that the performance of both Dual Averaging methods (DA and DA-RV) is very similar, and in general slightly better than the approximate Mirror Descent variants.

We also show an interesting relationship between the Mirror Descent approximations. Our analysis in Section 5.1.2 suggests an entire array of algorithms lying between DPP and TRPO, just as Modified Policy Iteration lies between Value Iteration and Policy Iteration (Puterman and Shin, 1978; Scherrer et al., 2012). Fig. 1 (right) illustrates this idea, showing the convergence of DPP and TRPO for a fixed value of η . TRPO tends to converge faster than DPP to a locally optimal policy, since DPP uses a single value update per iteration. Using more value updates leads to a modified regularized Policy Iteration algorithm (we call it ModRegPI-2 and ModRegPI-20, for 2 and 20 updates, respectively) that interpolates between DPP and TRPO.

7. Conclusion

We have presented a unifying view of entropy-regularized MDPs from a convex-optimization perspective. We believe that such unifying theories can be very useful in moving a field forward: We recall that in the field of online learning theory, the convex-optimization view has enabled a unified treatment of many existing algorithms and acts today as the primary framework for deriving new algorithms (see the progress from Cesa-Bianchi and Lugosi (2006) through Shalev-Shwartz (2012) to Hazan et al. (2016)). In this paper, we argued that the convex-optimization view may also be very useful in analyzing algorithms for reinforcement learning: In particular, we demonstrated how this framework can be used to provide theoretical justification for state-of-the-art reinforcement learning algorithms, and how it can highlight potential problems with them. We expect that this newly-found connection will also open the door for constructing more advanced reinforcement learning algorithms by borrowing further ideas from the convex optimization literature, such as Composite Objective Mirror Descent Duchi et al. (2010) and Regularized Dual Averaging (Xiao, 2010).

Finally, we point out that our work *does not* provide a statistical justification for using entropy regularization in reinforcement learning. In the case of online learning in known Markov decision processes with changing reward functions, entropy-regularization has been known to yield near-optimal learning algorithms (Even-Dar et al., 2009; Neu et al., 2010, 2014; Zimin and Neu, 2013; Dick et al., 2014). It remains to be seen if this technique also provably helps in driving the exploration process in unknown Markov decision processes.

References

- S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: A meta-algorithm and applications. *Theory of Computing*, 8:121–164, 2012.
- Kavosh Asadi and Michael L. Littman. A new softmax operator for reinforcement learning. *CoRR*, abs/1612.05628, 2016.

- Mohammad Gheshlaghi Azar, Vicenç Gómez, and Hilbert J Kappen. Dynamic policy programming with function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 119–127, 2011.
- Mohammad Gheshlaghi Azar, Vicenç Gómez, and Hilbert J Kappen. Dynamic policy programming. *Journal of Machine Learning Research*, 13(Nov):3207–3245, 2012.
- A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.
- R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, New Jersey, 1957.
- D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, Belmont, MA, 3 edition, 2007.
- Daniel A Braun, Pedro A Ortega, Evangelos Theodorou, and Stefan Schaal. Path integral control and bounded rationality. In *Adaptive Dynamic Programming And Reinforcement Learning (ADPRL), 2011 IEEE Symposium on*, pages 202–209. IEEE, 2011.
- A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for Markov Decision Processes. 22(1):222–255, 1997.
- Xi-Ren Cao. *Stochastic Learning and Optimization: A Sensitivity-Based Approach*. Springer, New York, 2007.
- N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.
- T. Dick, A. György, and Cs. Szepesvári. Online learning in markov decision processes with changing cost sequences. In *ICML 2014*, 2014.
- John C Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *COLT*, pages 14–26, 2010.
- E. Even-Dar, S. M. Kakade, and Y. Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.
- Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 2016.
- Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55: 119–139, 1997.

- Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.
- R. A. Howard. *Dynamic Programming and Markov Processes*. The MIT Press, Cambridge, MA, 1960.
- Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Management science*, 18(7):356–369, 1972.
- Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002)*, pages 267–274, 2002.
- J. Kivinen and M. Warmuth. Relative loss bounds for multidimensional regression problems. *Machine Learning*, 45:301–329, 2001.
- John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML 2001*, pages 282–289, 2001. URL citeseer.ist.psu.edu/lafferty01conditional.html.
- N. Littlestone and M. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.
- M.L. Littman and Cs. Szepesvári. A Generalized Reinforcement Learning Model: Convergence and applications. In *Int. Conf. on Machine Learning*, pages 310–318, 1996.
- Steven I Marcus, Emmanuel Fernández-Gaucherand, Daniel Hernández-Hernandez, Stefano Coraluppi, and Pedram Fard. Risk sensitive markov decision processes. In *Systems and control in the twenty-first century*, pages 263–279. Springer, 1997.
- B. Martinet. Perturbation des méthodes d’optimisation. applications. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 12(2):153–171, 1978. URL <http://eudml.org/doc/193317>.
- H Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *arXiv preprint arXiv:1403.3465*, 2014.
- Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937, 2016.
- William H Montgomery and Sergey Levine. Guided policy search via approximate mirror descent. In *NIPS-29*, pages 4008–4016, 2016.

- A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.
- G. Neu, A. György, and Cs. Szepesvári. The online loop-free stochastic shortest-path problem. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 231–243, 2010.
- Gergely Neu and Vicenç Gómez. Fast rates for online learning in Linearly Solvable Markov Decision Processes. *To appear in COLT*, 2017.
- Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online Markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59:676–691, 2014.
- Brendan O’Donoghue, Remi Munos, Koray Kavukcuoglu, and Volodymyr Mnih. PGQ: Combining policy gradient and Q-learning. In *5th International Conference on Learning Representations*, 2017.
- Theodore J. Perkins and Doina Precup. A convergent form of approximate policy iteration. In *Advances in Neural Information Processing Systems*, pages 1595–1602, 2002.
- Jan Peters, Katharina Mülling, and Yasemin Altun. Relative entropy policy search. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 1607–1612, 2010.
- Marek Petrik and Dharmashankar Subramanian. An approximate solution method for large risk-averse markov decision processes. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 805–814, 2012.
- Martin L. Puterman. *Markov Decision Processes*. Wiley, 1994. ISBN 978-0471727828. URL http://books.google.com/books/about/Markov_decision_processes.html?id=Y-gmAQAIAAJ.
- Martin L. Puterman and Moon Chirl Shin. Modified policy iteration algorithms for discounted Markov decision processes. *Management Science*, 24:1127–1137, 1978.
- Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. In *Proceedings of Robotics: Science and Systems VIII*, 2012.
- R. Tyrrell Rockafellar. Monotone Operators and the Proximal Point Algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.
- Andrzej Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Mathematical programming*, 125(2):235–261, 2010.

- B. Scherrer, V. Gabillon, M. Ghavamzadeh, and M. Geist. Approximate modified policy iteration. In *ICML 2012*, pages 1207–1214, 2012.
- John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1889–1897, 2015.
- Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.
- R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS-12*, pages 1057–1063, 1999.
- R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.
- Cs. Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.
- Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.
- V. Vovk. Aggregating strategies. In *Proceedings of the third annual workshop on Computational learning theory (COLT)*, pages 371–386, 1990.
- Ronald J Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.
- Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.
- Brian D Ziebart. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, Carnegie Mellon University, 2010.
- Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. Modeling interaction via the principle of maximum causal entropy. In *International Conference on Machine Learning (ICML)*, pages 1247–1254, 2010.
- A. Zimin and G. Neu. Online learning in episodic Markovian decision processes by relative entropy policy search. In *NIPS-26*, pages 1583–1591, 2013.

Appendix A. Complementary Technical Results

A.1. Convexity of the negative conditional entropy

Let us consider the joint state-action distribution μ on the finite set $\mathcal{X} \times \mathcal{A}$. We denote $\nu_\mu(x) = \sum_a \mu(x, a)$ and $\pi_\mu(a|x) = \mu(x, a)/\nu_\mu(x)$ for all x, a . We study the negative conditional entropy of $(X, A) \sim \mu$ as a function of μ :

$$R_C(\mu) = \sum_{x,a} \mu(x, a) \log \frac{\mu(x, a)}{\sum_b \mu(x, b)} = \sum_{x,a} \mu(x, a) \log \frac{\mu(x, a)}{\nu_\mu(x)}.$$

We will study the Bregman divergence D_{R_C} corresponding to R_C :

$$D_{R_C}(\mu \parallel \mu') = R_C(\mu) - R_C(\mu') - \nabla R_C(\mu')^\top (\mu - \mu'),$$

where the inner product between two vectors $v, w \in \mathbb{R}^{\mathcal{X} \times \mathcal{A}}$ is $w^\top v = \sum_{x,a} v(x, a)w(x, a)$. Our aim is to show that D_{R_C} is nonnegative, which will imply the convexity of R_C .

We begin by computing the partial derivative of $R_C(\mu)$ with respect to $\mu(x, a)$:

$$\frac{\partial R_C(\mu)}{\partial \mu(x, a)} = \log \frac{\mu(x, a)}{\nu_\mu(x)} + 1 - \sum_b \frac{\mu(x, b)}{\nu_\mu(x)} = \log \frac{\mu(x, a)}{\nu_\mu(x)},$$

where we used the fact that $\partial \nu_\mu(x) / \partial \mu(x, a) = 1$ for all a . With this expression, we have

$$\begin{aligned} R_C(\mu') + \nabla R_C(\mu')^\top (\mu - \mu') &= \sum_{x,a} \mu'(x, a) \log \frac{\mu'(x, a)}{\nu_{\mu'}(x)} + \sum_{x,a} (\mu(x, a) - \mu'(x, a)) \log \frac{\mu'(x, a)}{\nu_{\mu'}(x)} \\ &= \sum_{x,a} \mu(x, a) \log \frac{\mu'(x, a)}{\nu_{\mu'}(x)}. \end{aligned}$$

Thus, the Bregman divergence takes the form

$$\begin{aligned} D_{R_C}(\mu \parallel \mu') &= \sum_{x,a} \mu(x, a) \left(\log \frac{\mu(x, a)}{\nu_\mu(x)} - \log \frac{\mu'(x, a)}{\nu_{\mu'}(x)} \right) \\ &= \sum_{x,a} \mu(x, a) \log \frac{\pi_\mu(a|x)}{\pi_{\mu'}(a|x)} = \sum_x \nu(x) \sum_a \pi_\mu(a|x) \log \frac{\pi_\mu(a|x)}{\pi_{\mu'}(a|x)}. \end{aligned}$$

This proves that the Bregman divergence corresponding to R_C coincides with D_C , as claimed. To conclude the proof, note that D_C is the average relative entropy between the distributions π_μ and $\pi_{\mu'}$ —that is, a sum of positive terms. Indeed, this shows that D_C is nonnegative on the set of state-action distributions $\Delta(\mathcal{X} \times \mathcal{A})$, proving that $R_C(\mu)$ is convex. \blacksquare

A.2. Derivation of optimal policies

Here we prove the results stated in Equations (9)-(11) and Proposition 2, which give the expressions for the dual optimization problems and the optimal solutions corresponding to the primal optimization problem (6), for the two choices of regularization function $D_S(\cdot\|\mu')$ and $D_C(\cdot\|\mu')$. We start with generic derivations that will be helpful for analyzing both cases and then turn to studying the individual regularizers.

Recall that the primal optimization objective in (6) is given by

$$\max_{\mu \in \Delta} \tilde{\rho}_\eta(\mu) = \max_{\mu \in \Delta} \left\{ \sum_{x,a} \mu(x,a)r(x,a) - \frac{1}{\eta}R(\mu) \right\},$$

where Δ , the feasible set of stationary distributions, is defined by the following constraints:

$$\sum_b \mu(y,b) = \sum_{x,a} \mu(x,a)P(y|x,a), \quad \forall y \in \mathcal{X}, \quad (17)$$

$$\sum_{x,a} \mu(x,a) = 1, \quad (18)$$

$$\mu(x,a) \geq 0, \quad \forall (x,a) \in \mathcal{X} \times \mathcal{A}. \quad (19)$$

We begin by noting that for all state-action pairs where $\mu'(x,a) = 0$, the optimal solution $\mu_\eta^*(x,a)$ will also be zero, thanks to the form of our regularized objective. Thus, without loss of generality, we will assume that all states are recurrent under μ' : $\mu'(x,a) > 0$ holds for all state-action pairs.

For any choice of regularizer R , the Lagrangian of the primal (6) is given by

$$\begin{aligned} \mathcal{L}(\mu; V, \lambda, \varphi) &= \sum_{x,a} \mu(x,a)r(x,a) - \frac{1}{\eta}R(\mu) + \sum_y V(y) \left(\sum_{x,a} \mu(x,a)P(y|x,a) - \sum_b \mu(y,b) \right) \\ &\quad + \lambda \left(1 - \sum_{x,a} \mu(x,a) \right) + \sum_{x,a} \varphi(x,a)\mu(x,a) \\ &= \sum_{x,a} \mu(x,a) \left(r(x,a) + \sum_y P(y|x,a)V(y) - V(x) - \lambda + \varphi(x,a) \right) - \frac{1}{\eta}R(\mu) + \lambda \\ &= \sum_{x,a} \mu(x,a) (A(x,a) - \lambda + \varphi(x,a)) - \frac{1}{\eta}R(\mu) + \lambda, \end{aligned}$$

where V , λ and φ are the Lagrange multipliers⁵, and A is the advantage function for V . Setting the gradient of the Lagrangian with respect to μ to 0 yields the system of

5. Technically, these are KKT multipliers as we also have inequality constraints. However, these will be eliminated by means of complementary slackness in the next sections.

equations

$$\begin{aligned}
0 &= \frac{\partial \mathcal{L}}{\partial \mu(x, a)} = (A(x, a) - \lambda + \varphi(x, a)) - \frac{1}{\eta} \frac{\partial R(\mu)}{\partial \mu(x, a)}, \\
&\Leftrightarrow \frac{\partial R(\mu)}{\partial \mu(x, a)} = \eta (A(x, a) - \lambda + \varphi(x, a)), \tag{20}
\end{aligned}$$

for all x, a . By the first-order stationary condition, the unique optimal solution μ_η^* satisfies this system of equations. To obtain the final solution we have to compute the optimal values V_η^* , λ_η^* and φ_η^* of the Lagrange multipliers by optimizing the dual optimization objective $g(V, \lambda, \varphi) = \mathcal{L}(\mu_\eta^*; V, \lambda, \varphi)$, and insert into the expression for μ_η^* . V and λ are unconstrained in the dual, while φ satisfies $\varphi(x, a) \geq 0$ for each $(x, a) \in \mathcal{X} \times \mathcal{A}$. We give the derivations for each regularizer below.

A.3. The relative entropy

Here we prove the results for $R(\mu) = D_S(\mu \| \mu') = \sum_{x,a} \mu(x, a) \log \frac{\mu(x, a)}{\mu'(x, a)}$. The gradient of R is

$$\frac{\partial R(\mu)}{\partial \mu(x, a)} = \log \frac{\mu(x, a)}{\mu'(x, a)} + 1.$$

The optimal state-action distribution μ_η^* is now directly given by Equation (20):

$$\mu_\eta^*(x, a) = \mu'(x, a) \exp(\eta (A(x, a) - \lambda + \varphi(x, a)) - 1). \tag{21}$$

For μ_η^* to belong to Δ , it has to satisfy Constraints (18) and (19). Because of the exponent in (21), $\mu_\eta^*(x, a) \geq 0$ trivially holds for any choice of $\varphi(x, a)$, and complementary slackness implies $\varphi_\eta^*(x, a) = 0$ for each (x, a) . Eliminating φ and inserting μ_η^* into Constraint (18) gives us

$$\begin{aligned}
1 &= \sum_{x,a} \mu'(x, a) \exp(\eta A(x, a)) e^{-\eta \lambda - 1}, \\
&\Leftrightarrow \lambda = \frac{1}{\eta} \left(\log \sum_{x,a} \mu'(x, a) \exp(\eta A(x, a)) - 1 \right). \tag{22}
\end{aligned}$$

Since the value of λ is uniquely determined by (22), we can optimize the dual over V only. The dual function is given by

$$\begin{aligned}
g(V) &= \mathcal{L}(\mu_\eta^*; V, \lambda) = \sum_{x,a} \mu_\eta^*(x, a) \left(A(x, a) - \lambda - \frac{1}{\eta} \log \frac{\mu_\eta^*(x, a)}{\mu'(x, a)} \right) + \lambda = \frac{1}{\eta} + \lambda \\
&= \frac{1}{\eta} \log \sum_{x,a} \mu'(x, a) \exp(\eta A(x, a)).
\end{aligned}$$

This is precisely the dual given in Equation (10). Note that this dual function has no associated constraints. The expression for the optimal state-action distribution in Equation (9) is obtained by inserting the advantage function A_η^* corresponding to the optimal value function V_η^* into (21).

A.4. The conditional entropy

We next prove the results for $R(\mu) = D_C(\mu||\mu') = \sum_{x,a} \mu(x,a) \log \frac{\pi_\mu(a|x)}{\pi_{\mu'}(a|x)}$. The gradient of R is

$$\frac{\partial R(\mu)}{\partial \mu(x,a)} = \log \frac{\pi_\mu(a|x)}{\pi_{\mu'}(a|x)} + \sum_b \frac{\mu(x,b)}{\pi_\mu(b|x)} \cdot \frac{\partial \pi_\mu(b|x)}{\partial \mu(x,a)}.$$

Since the policy is defined as $\pi_\mu(a|x) = \frac{\mu(x,a)}{\nu_\mu(x)}$, its gradient with respect to μ is

$$\frac{\partial \pi_\mu(b|x)}{\partial \mu(x,a)} = \frac{\mathbb{I}_{\{a=b\}}}{\nu_\mu(x)} - \frac{\mu(x,b)}{\nu_\mu(x)^2} = \frac{1}{\nu_\mu(x)} (\mathbb{I}_{\{a=b\}} - \pi_\mu(b|x)), \quad \forall x \in \mathcal{X}, a, b \in \mathcal{A}.$$

Inserting into the expression for the gradient of R yields

$$\begin{aligned} \frac{\partial R(\mu)}{\partial \mu(x,a)} &= \log \frac{\pi_\mu(a|x)}{\pi_{\mu'}(a|x)} + \sum_b \frac{\pi_\mu(b|x)}{\pi_\mu(b|x)} (\mathbb{I}_{\{a=b\}} - \pi_\mu(b|x)) \\ &= \log \frac{\pi_\mu(a|x)}{\pi_{\mu'}(a|x)} + 1 - \sum_b \pi_\mu(b|x) = \log \frac{\pi_\mu(a|x)}{\pi_{\mu'}(a|x)}. \end{aligned}$$

The optimal policy π_μ^* is now directly given by Equation (20):

$$\pi_\eta^*(a|x) = \pi_{\mu'}(a|x) \exp(\eta(A(x,a) - \lambda + \varphi(x,a))). \quad (23)$$

For μ_η^* to belong to Δ , it has to satisfy Constraint (19). Because of the exponent in (23) and the fact that $\mu_\eta^*(x,a) \propto \pi_\eta^*(a|x)$, $\mu_\eta^*(x,a) \geq 0$ trivially holds for any choice of $\varphi(x,a)$, implying that $\varphi_\eta^*(x,a) = 0$ for each (x,a) by complementary slackness. Since $\pi_\eta^*(a|x) = \frac{\mu_\eta^*(x,a)}{\nu_\eta^*(x)}$, we also obtain the following set of constraints:

$$\sum_a \pi_\eta^*(a|x) = \sum_a \frac{\mu_\eta^*(x,a)}{\nu_\eta^*(x)} = \frac{\nu_\eta^*(x)}{\nu_\eta^*(x)} = 1, \quad \forall x \in \mathcal{X}.$$

Inserting the expression for π_η^* yields

$$1 = \sum_a \pi_{\mu'}(a|x) \exp(\eta A(x,a)) e^{-\eta\lambda}, \quad \forall x \in \mathcal{X}.$$

If we expand the expression for $A(x,a)$ and rearrange the terms we obtain

$$V(x) = \frac{1}{\eta} \log \sum_a \pi_{\mu'}(a|x) \exp \left(\eta \left(r(x,a) - \lambda + \sum_y P(y|x,a) V(y) \right) \right), \quad \forall x \in \mathcal{X}. \quad (24)$$

The dual function is obtained by inserting the expression for μ^* into the Lagrangian:

$$g(V, \lambda) = \mathcal{L}(\mu_\eta^*, V, \lambda) = \sum_{x,a} \mu_\eta^*(a|x) \left(A(x, a) - \lambda - \frac{1}{\eta} \log \frac{\pi_\eta^*(a|x)}{\pi_{\mu'}(a|x)} \right) + \lambda = \lambda. \quad (25)$$

Together, Equations (24) and (25) define the dual optimization problem in Proposition 2. The expression for the optimal policy in Equation (11) is obtained by inserting the optimal advantage function A_η^* into (23).

We remark that to recover the optimal stationary state-action distribution μ_η^* , we would have to insert the expression for the optimal policy π_η^* into Constraints (17) and (18), and solve for the stationary state distribution ν_η^* . However, this is not necessary since μ_η^* and ν_η^* are not required to solve the dual function or to compute the optimal policy.

Appendix B. The regularized Bellman operators

In this section, we define the *regularized Bellman operator* $T_{\pi|\pi'}$ corresponding to the policy π and regularized with respect to baseline π' as

$$T_\eta^{\pi|\pi'}[V](x) = \sum_a \pi(a|x) \left(r(x, a) - \log \frac{\pi(a|x)}{\pi'(a|x)} + \sum_{x'} P(x'|x, a)V(x') \right) \quad (\forall x).$$

Similarly, we define the *regularized Bellman optimality operator* $T_{*\pi'}$ with respect to baseline π' as

$$T_\eta^{*|\pi'}[V](x) = \frac{1}{\eta} \log \sum_a \pi'(a|x) \exp \left(\eta \left(r(x, a) + \sum_y P(y|x, a)V(y) \right) \right) \quad (\forall x),$$

and the *regularized greedy policy* with respect to the baseline π' as

$$G_\eta^{\pi'}[V](a|x) \propto \pi'(a|x) \exp \left(\eta \left(r(x, a) + \sum_y P(y|x, a)V(y) - V(x) \right) \right).$$

With these notations, we can define the *regularized relative value iteration* algorithm with respect to π' by the iteration

$$\pi_{k+1} = G_\eta^{\pi'}[V_k] \quad V_{k+1}(x) = T_\eta^{*|\pi'}[V_k](x) - \delta_{k+1} \quad (26)$$

for some δ_{k+1} lying between the minimal and maximal values of $T_\eta^{*|\pi'}[V_k]$. A common technique is to fix a reference state x' and choose $\delta_{k+1} = T_\eta^{*|\pi'}[V_k](x')$.

Similarly, we can define the *regularized policy iteration* algorithm by the iteration

$$\pi_{k+1} = G_\eta^{\pi'}[V_k] \quad V_{k+1}(x) = \left(T_\eta^{\pi_{k+1}|\pi'} \right)^\infty [V_k](x) - \delta_{k+1}, \quad (27)$$

with δ_{k+1} defined analogously.

For establishing the convergence of the above procedures, it is crucial to ensure that the operator $T_\eta^{*|\pi'}$ is a *non-expansion*: For any value functions V_1 and V_2 , we need to ensure

$$\left\| T_\eta^{*|\pi'}[V_1] - T_\eta^{*|\pi'}[V_2] \right\| \leq \|V_1 - V_2\|$$

for some norm. We state the following result claiming that the above requirement indeed holds and present the simple proof below. We note that analogous results have been proven several times in the literature, see, e.g., (Fox et al., 2016; Asadi and Littman, 2016).

Proposition 5 $T_\eta^{*|\pi'}$ is a non-expansion for the supremum norm $\|f\|_\infty = \max_x |f(x)|$.

Proof For simplicity, let us introduce the notation $Q_1(x, a) = r(x, a) + \sum_y P(y|x, a)V_1(y)$, with Q_2 defined analogously, and $\Delta = Q_1 - Q_2$. We have

$$\begin{aligned} T_\eta^{*|\pi'}[V_1](x) - T_\eta^{*|\pi'}[V_2](x) &= \frac{1}{\eta} \left(\log \sum_a \pi'(a|x) \exp(\eta Q_1(x, y)) - \log \sum_a \pi'(a|x) \exp(\eta Q_2(x, y)) \right) \\ &= \frac{1}{\eta} \left(\log \frac{\sum_a \pi'(a|x) \exp(\eta Q_1(x, y))}{\sum_a \pi'(a|x) \exp(\eta Q_2(x, y))} \right) \\ &= \frac{1}{\eta} \left(\log \frac{\sum_a \pi'(a|x) \exp(\eta (Q_2(x, y) + \Delta(x, y)))}{\sum_a \pi'(a|x) \exp(\eta Q_2(x, y))} \right) \\ &= \frac{1}{\eta} \log \sum_a p(x, a) \exp(\eta \Delta(x, a)) \quad (\text{with an appropriately defined } p) \\ &\leq \frac{1}{\eta} \log \max_a \exp(\eta \Delta(x, a)) \\ &= \max_a \Delta(x, a) = \max_a \sum_y P(y|x, a) (V_1(y) - V_2(y)) \\ &\leq \max_y |V_1(y) - V_2(y)|. \end{aligned}$$

With an analogous technique, we can also show the complementary inequality

$$T_\eta^{*|\pi'}[V_2](x) - T_\eta^{*|\pi'}[V_1](x) \leq \max_y |V_2(y) - V_1(y)|,$$

which concludes the proof. ■

Together with the easily-seen fact that $T_\eta^{*|\pi'}$ is continuous, this result immediately implies that $T_\eta^{*|\pi'}$ has a fixed point by Brouwer's fixed-point theorem. Furthermore, this insight allows us to treat the value iteration method (26) as an instance of *generalized value iteration*, as defined by Littman and Szepesvári (1996).

We now argue that regularized value iteration converges to the fixed point of $T_\eta^{*|\pi'}$. If the initial value function V_0 is bounded, then so is V_k for each k since the operator

$T_\eta^{*|\pi'}$ is a non-expansion. Similar to Section 3, we assume without loss of generality that the initial reference policy π_0 has full support, i.e. $\pi_0(a|x) > 0$ for each recurrent state x and each action a . Inspecting the greedy policy operator $G_\eta^{\pi'}$, it is easy to show by induction that π_k has full support for each k . In particular, $\pi_{k+1}(a|x)$ only equals 0 if either $\pi_k(a|x)$ equals 0 or if the exponent $A_k(x, a)$ equals $-\infty$, which is only possible if V_k is unbounded.

Now, since π_k has full support for each k , any trajectory always has a small probability of reaching a given recurrent state. We can now use a similar argument as Bertsekas (2007, Prop. 4.3.2) to show that regularized value iteration converges to the fixed point for $T_\eta^{*|\pi'}$.

B.1. The proof of Lemma 3

Let μ and μ^* be the respective stationary distributions of π and π^* . The statement follows easily from using the definition of A_η^π :

$$\begin{aligned} \sum_{x,a} \mu^*(x, a) A_\eta^\pi(x, a) &= \sum_{x,a} \mu^*(x, a) \left(r(x, a) - \frac{1}{\eta} \log \frac{\pi(a|x)}{\pi'(a|x)} - \tilde{\rho}_\eta(\mu) + \sum_y P(y|x, a) V_\eta^\pi(y) - V_\eta^\pi(x) \right) \\ &= \tilde{\rho}_\eta(\mu^*) + \frac{1}{\eta} \sum_{x,a} \mu^*(x, a) \log \frac{\pi^*(a|x)}{\pi(a|x)} - \tilde{\rho}_\eta(\mu) \\ &\quad + \sum_{x,a} \mu^*(x, a) \left(\sum_y P(y|x, a) V_\eta^\pi(y) - V_\eta^\pi(x) \right) \\ &= \tilde{\rho}_\eta(\mu^*) - \tilde{\rho}_\eta(\mu) + \frac{1}{\eta} D_C(\mu^* \|\mu), \end{aligned}$$

where the last step follows from the stationarity of μ' and the definition of D_C . \blacksquare

B.2. Regularized policy gradient

Here we prove Lemma 4 which gives the gradient of the regularized average reward $\tilde{\rho}_\eta(\theta)$ when the policy π_θ is parameterized on θ . Following Sutton et al. (1999), we first compute the gradient of $V_\eta^{\pi_\theta}$:

$$\begin{aligned} \frac{\partial V_\eta^{\pi_\theta}(x)}{\partial \theta_i} &= \sum_a \frac{\partial \pi_\theta(a|x)}{\partial \theta_i} \left(r(x, a) - \frac{1}{\eta} \log \frac{\pi_\theta(a|x)}{\pi'(a|x)} - \tilde{\rho}_\eta(\theta) + \sum_y P(y|x, a) V_\eta^{\pi_\theta}(y) \right) \\ &\quad + \sum_a \pi_\theta(a|x) \left(-\frac{1}{\eta \pi_\theta(a|x)} \frac{\partial \pi_\theta(a|x)}{\partial \theta_i} - \frac{\partial \tilde{\rho}_\eta}{\partial \theta_i} + \sum_y P(y|x, a) \frac{\partial V_\eta^{\pi_\theta}(y)}{\partial \theta_i} \right). \end{aligned}$$

Rearranging the terms gives us

$$\begin{aligned}
 \frac{\partial \tilde{\rho}_\eta}{\partial \theta_i} &= \sum_a \frac{\partial \pi_\theta(a|x)}{\partial \theta_i} \left(r(x, a) - \frac{1}{\eta} \log \frac{\pi_\theta(a|x)}{\pi'(a|x)} - \tilde{\rho}_\eta(\theta) + \sum_y P(y|x, a) V_\eta^{\pi_\theta}(y) - \frac{1}{\eta} \right) \\
 &+ \sum_a \pi_\theta(a|x) \sum_y P(y|x, a) \frac{\partial V_\eta^{\pi_\theta}(y)}{\partial \theta_i} - \frac{\partial V_\eta^{\pi_\theta}(x)}{\partial \theta_i} \\
 &= \sum_a \frac{\partial \pi_\theta(a|x)}{\partial \theta_i} A_\eta^{\pi_\theta}(x, a) + \sum_a \pi_\theta(a|x) \sum_y P(y|x, a) \frac{\partial V_\eta^{\pi_\theta}(y)}{\partial \theta_i} - \frac{\partial V_\eta^{\pi_\theta}(x)}{\partial \theta_i}.
 \end{aligned}$$

The last equality follows from the fact that since $\sum_a \partial \pi_\theta(a|x)/\partial \theta_i = 0$ for each x , we can add any state-dependent constant to the multiplier of $\partial \pi_\theta(a|x)/\partial \theta_i$; adding the term $\frac{1}{\eta} - V_\eta^{\pi_\theta}(x)$ results in the given expression. Summing both sides over the stationary state distribution ν_{π_θ} yields

$$\begin{aligned}
 \frac{\partial \tilde{\rho}_\eta}{\partial \theta_i} &= \sum_x \nu_{\pi_\theta}(x) \frac{\partial \tilde{\rho}_\eta}{\partial \theta_i} = \sum_{x,a} \nu_{\pi_\theta}(x) \frac{\partial \pi_\theta(a|x)}{\partial \theta_i} A_\eta^{\pi_\theta}(x, a) \\
 &+ \sum_y \sum_{x,a} \nu_{\pi_\theta}(x) \pi_\theta(a|x) P(y|x, a) \frac{\partial V_\eta^{\pi_\theta}(y)}{\partial \theta_i} - \sum_x \nu_{\pi_\theta}(x) \frac{\partial V_\eta^{\pi_\theta}(x)}{\partial \theta_i} \\
 &= \sum_{x,a} \nu_{\pi_\theta}(x) \frac{\partial \pi_\theta(a|x)}{\partial \theta_i} A_\eta^{\pi_\theta}(x, a) + \sum_y \nu_{\pi_\theta}(y) \frac{\partial V_\eta^{\pi_\theta}(y)}{\partial \theta_i} - \sum_x \nu_{\pi_\theta}(x) \frac{\partial V_\eta^{\pi_\theta}(x)}{\partial \theta_i} \\
 &= \sum_{x,a} \nu_{\pi_\theta}(x) \frac{\partial \pi_\theta(a|x)}{\partial \theta_i} A_\eta^{\pi_\theta}(x, a).
 \end{aligned}$$

To conclude the proof it is sufficient to note that

$$\mu_{\pi_\theta}(x, a) \frac{\partial \log \pi_\theta(a|x)}{\partial \theta_i} = \frac{\mu_{\pi_\theta}(x, a)}{\pi_\theta(a|x)} \frac{\partial \pi_\theta(a|x)}{\partial \theta_i} = \nu_{\pi_\theta}(x) \frac{\partial \pi_\theta(a|x)}{\partial \theta_i}.$$

■

B.3. The closed form of the TRPO update

Here we derive the closed-form solution of the TRPO update. To do so, we first briefly summarize the mechanism of the algorithm. The main idea of [Schulman et al. \(2015\)](#) is replacing $\rho(\mu')$ by the surrogate

$$L^\pi(\pi') = \rho(\pi) + \sum_x \nu_\pi(x) \sum_a \pi'(a|x) A_\infty^\pi(x, a),$$

where A_∞^π is the unregularized advantage function corresponding to policy π .⁶ Furthermore, TRPO uses the regularization term

$$D_{\text{TRPO}}(\mu \parallel \mu') = \sum_x \nu_{\mu'}(x) \sum_a \pi_\mu(a|x) \log \frac{\pi_\mu(a|x)}{\pi_{\mu'}(a|x)}.$$

The difference between $L + D_{\text{TRPO}}$ and $\rho + D_C$ is that the approximate version ignores the impact of changing the policy π on the stationary distribution. Given this surrogate objective, TRPO approximately computes the distribution

$$\mu_{k+1} = \arg \max_{\mu \in \Delta} \left\{ L_{\mu_k}(\mu) - \frac{1}{\eta} D_{\text{TRPO}}(\mu \parallel \mu_k) \right\}. \quad (28)$$

Observing that the TRPO policy update can be expressed equivalently as

$$\pi_{k+1} = \arg \max_{\pi} \left\{ \sum_x \nu_{\mu_k}(x) \sum_a \pi(a|x) \left(A_\infty^{\pi_k}(x, a) - \frac{1}{\eta} \log \frac{\pi(a|x)}{\pi_k(a|x)} \right) \right\},$$

we can see that the policy update can be expressed in closed form as

$$\pi_{k+1}(a|x) \propto \pi_k(a|x) e^{\eta A_\infty^{\pi_k}(x, a)}.$$

This update then can be seen as a regularized greedy step with respect to the value function of the previous policy π_k .

6. This form is inspired by the well-known identity we state as Lemma 3.