Adaptive Trust Region Policy Optimization: Global Convergence and Faster Rates for Regularized MDPs

Lior Shani,[†] Yonathan Efroni,[†] Shie Mannor
[†]equal contribution
Technion - Israel Institute of Technology
Haifa, Israel

Abstract

Trust region policy optimization (TRPO) is a popular and empirically successful policy search algorithm in Reinforcement Learning (RL) in which a surrogate problem, that restricts consecutive policies to be 'close' to one another, is iteratively solved. Nevertheless, TRPO has been considered a heuristic algorithm inspired by Conservative Policy Iteration (CPI). We show that the adaptive scaling mechanism used in TRPO is in fact the natural "RL version" of traditional trust-region methods from convex analysis. We first analyze TRPO in the planning setting, in which we have access to the model and the entire state space. Then, we consider sample-based TRPO and establish $\tilde{O}(1/\sqrt{N})$ convergence rate to the global optimum. Importantly, the adaptive scaling mechanism allows us to analyze TRPO in *regularized MDPs* for which we prove fast rates of $\tilde{O}(1/N)$, much like results in convex optimization. This is the first result in RL of better rates when regularizing the instantaneous cost or reward.

1 Introduction

The field of Reinforcement learning (RL) (Sutton and Barto 2018) tackles the problem of learning how to act optimally in an unknown dynamic environment. The agent is allowed to apply actions on the environment, and by doing so, to manipulate its state. Then, based on the rewards or costs it accumulates, the agent learns how to act optimally. The foundations of RL lie in the theory of Markov Decision Processes (MDPs), where an agent has an access to the model of the environment and can plan to act optimally.

Trust Region Policy Optimization (TRPO): Trust region methods are a popular class of techniques to solve an RL problem and span a wide variety of algorithms including Non-Euclidean TRPO (NE-TRPO) (Schulman et al. 2015) and Proximal Policy Optimization (Schulman et al. 2017). In these methods a sum of two terms is iteratively being minimized: a linearization of the objective function and a proximity term which restricts two consecutive updates to be 'close' to each other, as in Mirror Descent (MD) (Beck and Teboulle 2003). In spite of their popularity, much less is understood in terms of their convergence guarantees and they are considered heuristics (Schulman et al. 2015; Papini, Pirotta, and Restelli 2019) (see Figure 1).

TRPO and Regularized MDPs: Trust region methods are often used in conjunction with regularization. This is commonly done by adding the negative entropy to the instantaneous cost (Nachum et al. 2017; Schulman et al. 2017). The intuitive justification for using entropy regularization is that it induces inherent exploration (Fox, Pakman, and Tishby 2016), and the advantage of 'softening' the Bellman equation (Chow, Nachum, and Ghavamzadeh 2018; Dai et al. 2018). Recently, Ahmed et al. (2019) empirically observed that adding entropy regularization results in a smoother objective which in turn leads to faster convergence when the learning rate is chosen more aggressively. Yet, to the best of our knowledge, there is no finite-sample analysis that demonstrates faster convergence rates for regularization in MDPs. This comes in stark contrast to well established faster rates for strongly convex objectives w.r.t. convex ones (Nesterov 1998). In this work we refer to regularized MDPs as describing a more general case in which a strongly convex function is added to the immediate cost.

The goal of this work is to bridge the gap between the practicality of trust region methods in RL and the scarce theoretical guarantees for standard (unregularized) and regularized MDPs. To this end, we revise a fundamental question in this context:

What is the proper form of the proximity term in trust region methods for RL?

In Schulman et al. (2015), two proximity terms are suggested which result in two possible versions of trust region methods for RL. The first (Schulman et al. 2015, Algorithm 1) is motivated by Conservative Policy Iteration (CPI) (Kakade and others 2003) and results in an improving and thus converging algorithm in its exact error-free version. Yet, it seems computationally infeasible to produce a sample-based version of this algorithm. The second algorithm, with an adaptive proximity term which depends on the current policy (Schulman et al. 2015, Equation 12), is described as a heuristic approximation of the first, with no convergence guarantees, but leads to NE-TRPO, currently among the most popular algorithms in RL (see Figure 1).

In this work, we focus on tabular discounted MDPs and

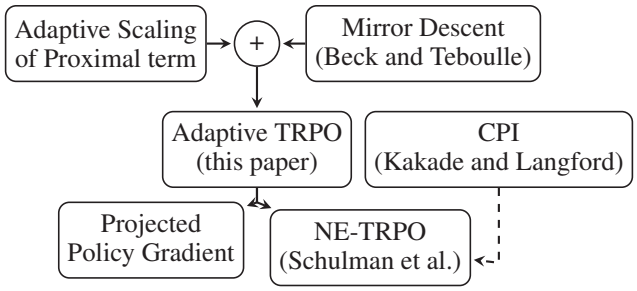

Figure 1: The adaptive TRPO: a solid line implies a formal relation; a dashed line implies a heuristic relation.

study a general TRPO method which uses the latter adaptive proximity term. Unlike the common belief, we show this adaptive scaling mechanism is 'natural' and imposes the structure of RL onto traditional trust region methods from convex analysis. We refer to this method as adaptive TRPO, and analyze two of its instances: NE-TRPO (Schulman et al. 2015, Equation 12) and Projected Policy Gradient (PPG), as illustrated in Figure 1. In Section 2, we review results from convex analysis that will be used in our analysis. Then, we start by deriving in Section 4 a closed form solution of the linearized objective functions for RL. In Section 5, using the closed form of the linearized objective, we formulate and analyze Uniform TRPO. This method assumes simultaneous access to the state space and that a model is given. In Section 6, we relax these assumptions and study Sample-Based TRPO, a sample-based version of Uniform TRPO, while building on the analysis of Section 5. The main contributions of this paper are:

- We establish $\tilde{O}(1/\sqrt{N})$ convergence rate to the global optimum for both Uniform and Sample-Based TRPO.
- We prove a faster rate of $\tilde{O}(1/N)$ for regularized MDPs. To the best of our knowledge, it is the first evidence for faster convergence rates using regularization in RL.
- The analysis of Sample-Based TRPO, unlike CPI, does not rely on improvement arguments. This allows us to choose a more aggressive learning rate relatively to CPI which leads to an improved sample complexity even for the unregularized case.

2 Mirror Descent in Convex Optimization

Mirror descent (MD) (Beck and Teboulle 2003) is a well known first-order trust region optimization method for solving constrained convex problems, i.e, for finding

$$x^* \in \arg\min_{x \in C} f(x), \tag{1}$$

where f is a convex function and C is a convex compact set. In each iteration, MD minimizes a linear approximation of the objective function, using the gradient $\nabla f(x_k)$, together with a proximity term by which the updated x_{k+1} is 'close' to x_k. Thus, it is considered a trust region method, as the iterates are 'close' to one another. The iterates of MD are

$$x_{k+1} \in \arg\min_{x \in C} \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{t_k} B_\omega(x, x_k), \tag{2}$$

where $B_\omega(x, x_k) := \omega(x) - \omega(x_k) - \langle \nabla \omega(x_k), x - x_k \rangle$ is the Bregman distance associated with a strongly convex ω and t_k is a stepsize (see Appendix A). In the general convex case, MD converges to the optimal solution of (1) with a rate of $\tilde{O}(1/\sqrt{N})$, where N is the number of MD iterations (Beck and Teboulle 2003; Juditsky, Nemirovski, and others 2011), i.e., $f(x_k) - f^* \leq \tilde{O}(1/\sqrt{k})$, where $f^* = f(x^*)$.

The convergence rate can be further improved when f is a part of special classes of functions. One such class is the set of λ-strongly convex functions w.r.t. the Bregman distance. We say that f is λ-strongly convex w.r.t. the Bregman distance if $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \lambda B_\omega(y, x)$. For such f, improved convergence rate of $\tilde{O}(1/N)$ can be obtained (Juditsky, Nemirovski, and others 2011; Nedic and Lee 2014). Thus, instead of using MD to optimize a convex f, one can consider the following regularized problem,

$$x^* = \arg\min_{x \in C} f(x) + \lambda g(x), \tag{3}$$

where g is a strongly convex regularizer with coefficient $\lambda > 0$. Define $F_\lambda(x) := f(x) + \lambda g(x)$, then, each iteration of MD becomes,

$$x_{k+1} = \arg\min_{x \in C} \langle \nabla F_\lambda(x_k), x - x_k \rangle + \frac{1}{t_k} B_\omega(x, x_k). \tag{4}$$

Solving (4) allows faster convergence, at the expense of adding a bias to the solution of (1). Trivially, by setting $\lambda = 0$, we go back to the unregularized convex case.

In the following, we consider two common choices of ω which induce a proper Bregman distance: (a) **The euclidean case**, with $\omega(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ and the resulting Bregman distance is the squared euclidean norm $B_\omega(x, y) = \frac{1}{2}\|x - y\|_2^2$. In this case, (2) becomes the *Projected Gradient Descent* algorithm (Beck 2017, Section 9.1), where in each iteration, the update step goes along the direction of the gradient at x_k, $\nabla f(x_k)$, and then projected back to the convex set C, $x_{k+1} = P_c(x_k - t_k \nabla f(x_k))$, where $P_c(x) = \min_{y \in C} \frac{1}{2}\|x - y\|_2^2$ is the orthogonal projection operator w.r.t. the euclidean norm.

(b) **The non-euclidean case**, where $\omega(\cdot) = H(\cdot)$ is the negative entropy, and the Bregman distance then becomes the Kullback-Leibler divergence, $B_\omega(x, y) = d_{KL}(x||y)$. In this case, MD becomes the *Exponentiated Gradient Descent* algorithm. Unlike the euclidean case, where we need to project back into the set, when choosing ω as the negative entropy, (2) has a closed form solution (Beck 2017, Example 3.71), $x_{k+1}^i = \frac{x_k^i e^{-t_k \nabla_i f(x_k)}}{\sum_j x_k^j e^{-t_k \nabla_j f(x_k)}}$, where x_k^i and $\nabla_i f$ are the i-th coordinates of x_k and ∇f.

3 Preliminaries and Notations

We consider the infinite-horizon discounted MDP which is defined as the 5-tuple $(\mathcal{S}, \mathcal{A}, P, C, \gamma)$ (Sutton and Barto 2018), where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action sets with

cardinality of $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$, respectively. The transition kernel is $P \equiv P(s'|s,a)$, $C \equiv c(s,a)$ is a cost function bounded in $[0, \mathrm{C}_{\max}]$*, and $\gamma \in (0,1)$ is a discount factor. Let $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ be a stationary policy, where $\Delta_{\mathcal{A}}$ is the set probability distributions on $\mathcal{A}$. Let $v^\pi \in \mathbb{R}^S$ be the value of a policy π, with its $s \in \mathcal{S}$ entry given by $v^\pi(s) := \mathbb{E}^\pi[\sum_{t=0}^\infty \gamma^t r(s_t, \pi(s_t)) \mid s_0 = s]$, and $\mathbb{E}^\pi[\cdot \mid s_0 = s]$ denotes expectation w.r.t. the distribution induced by π and conditioned on the event $\{s_0 = s\}$. It is known that $v^\pi = \sum_{t=0}^\infty \gamma^t (P^\pi)^t c^\pi = (I - \gamma P^\pi)^{-1} c^\pi$, with the component-wise values $[P^\pi]_{s,s'} := P(s' \mid s, \pi(s))$ and $[c^\pi]_s := c(s, \pi(s))$. Our goal is to find a policy π^* yielding the optimal value v^* such that

$$v^* = \min_\pi (I - \gamma P^\pi)^{-1} c^\pi = (I - \gamma P^{\pi^*})^{-1} c^{\pi^*}. \quad (5)$$

This goal can be achieved using the classical operators:

$$\forall v, \pi,\ T^\pi v = c^\pi + \gamma P^\pi v, \text{and } \forall v,\ Tv = \min_\pi T^\pi v, \quad (6)$$

where T^π is a linear operator, T is the optimal Bellman operator and both T^π and T are γ-contraction mappings w.r.t. the max-norm. The fixed points of T^π and T are v^π and v^*.

A large portion of this paper is devoted to analysis of regularized MDPs: A regularized MDP is an MDP with a shaped cost denoted by c^π_λ for $\lambda \geq 0$. Specifically, the cost of a policy π on a regularized MDP translates to $c^\pi_\lambda(s) := c^\pi(s) + \lambda \omega(s;\pi)$ where $\omega(s;\pi) := \omega(\pi(\cdot \mid s))$ and $\omega : \Delta_{\mathcal{A}} \to \mathbb{R}$ is a 1-strongly convex function. We denote $\omega(\pi) \in \mathbb{R}^S$ as the corresponding state-wise vector. See that for $\lambda = 0$, the cost c^π is recovered. In this work we consider two choices of ω: the **euclidean case** $\omega(s;\pi) = \frac{1}{2}\|\pi(\cdot \mid s)\|_2^2$ and **non-euclidean case** $\omega(s;\pi) = H(\pi(\cdot \mid s)) + \log A$. By this choice we have that $0 \leq c^\pi_\lambda(s) \leq \mathrm{C}_{\max,\lambda}$ where $\mathrm{C}_{\max,\lambda} = \mathrm{C}_{\max} + \lambda$ and $\mathrm{C}_{\max,\lambda} = \mathrm{C}_{\max} + \lambda \log A$, for the euclidean and non-euclidean cases, respectively. With some abuse of notation we omit ω from $\mathrm{C}_{\max,\lambda}$.

The value of a stationary policy π on the regularized MDP is $v^\pi_\lambda = (I - \gamma P^\pi)^{-1} c^\pi_\lambda$. Furthermore, the optimal value v^*_λ, optimal policy π^*_λ and Bellman operators of the regularized MDP are generalized as follows,

$$v^*_\lambda = \min_\pi (I - \gamma P^\pi)^{-1} c^\pi_\lambda = (I - \gamma P^{\pi^*_\lambda})^{-1} c^{\pi^*_\lambda}_\lambda, \quad (7)$$
$$\forall v, \pi,\ T^\pi_\lambda v = c^\pi_\lambda + \gamma P^\pi v, \text{and } \forall v,\ T_\lambda v = \min_\pi T^\pi_\lambda v.$$

As Bellman operators for MDPs, both T^π_λ, T are γ-contractions with fixed points $v^\pi_\lambda, v^*_\lambda$ (Geist, Scherrer, and Pietquin 2019). Denoting $c^\pi_\lambda(s,a) = c(s,a) + \lambda\omega(s;\pi)$, the q-function of a policy π for a regularized MDP is defined as $q^\pi_\lambda(s,a) = c^\pi_\lambda(s,a) + \gamma \sum_{s'} p^\pi(s' \mid s) v^\pi_\lambda(s')$.

When the state space is small and the dynamics of environment is known (5), (7) can be solved using DP approaches. However, in case of a large state space it is expected to be computationally infeasible to apply such algorithms as they require access to the entire state space. In this work, we construct a sample-based algorithm which minimizes the following *scalar objective* instead of (5), (7),

$$\min_{\pi \in \Delta^{\mathrm{S}}_{\mathcal{A}}} \mathbb{E}_{s\sim\mu}[v^\pi_\lambda(s)] = \min_{\pi \in \Delta^{\mathrm{S}}_{\mathcal{A}}} \mu v^\pi_\lambda, \quad (8)$$

where $\mu(\cdot)$ is a probability measure over the state space. Using this objective, one wishes to find a policy π which minimizes the expectation of $v^\pi_\lambda(s)$ under a measure μ. This objective is widely used in the RL literature (Sutton et al. 2000; Kakade and Langford 2002; Schulman et al. 2015).

Here, we always choose the regularization function ω to be associated with the Bregman distance used, B_ω. This simplifies the analysis as c^π_λ is λ-strongly convex w.r.t. B_ω by definition. Given two policies π_1, π_2, we denote their Bregman distance as $B_\omega(s;\pi_1,\pi_2) := B_\omega(\pi_1(\cdot \mid s), \pi_2(\cdot \mid s))$ and $B_\omega(\pi_1,\pi_2) \in \mathbb{R}^S$ is the corresponding state-wise vector. The euclidean choice for ω leads to $B_\omega(s;\pi_1,\pi_2) = \frac{1}{2}\|\pi_1(\cdot \mid s) - \pi_2(\cdot \mid s)\|_2^2$, and the non-euclidean choice to $B_\omega(s;\pi_1,\pi_2) = d_{KL}(\pi_1(\cdot \mid s)||\pi_2(\cdot \mid s))$. In the results we use the following ω-dependent constant, $C_{\omega,1} = \sqrt{A}$ in the euclidean case, and $C_{\omega,1} = 1$ in the non-euclidean case.

For brevity, we omit constant and logarithmic factors when using $O(\cdot)$, and omit any factors other than non-logarithmic factors in N, when using $\tilde{O}(\cdot)$. For $x, y \in \mathbb{R}^{S\times A}$, the state-action inner product is $\langle x, y\rangle = \sum_{s,a} x(s,a)y(s,a)$, and the fixed-state inner product is $\langle x(s,\cdot), y(s,\cdot)\rangle = \sum_a x(s,a)y(s,a)$. Lastly, when $x \in \mathbb{R}^{S\times S\times A}$ (e.g., first claim of Proposition 1) the inner product $\langle x, y\rangle$ is a vector in $\mathbb{R}^S$ where $\langle x, y\rangle(s) := \langle x(s,\cdot,\cdot), y\rangle$, with some abuse of notation.

4 Linear Approximation of a Policy's Value

As evident from the updating rule of MD (2), a crucial step in adapting MD to solve MDPs is studying the linear approximation of the objective, $\langle \nabla f(x), x' - x\rangle$, i.e., the directional derivative in the direction of an element from the convex set. The objectives considered in this work are (7), (8), and the optimization set is the convex set of policies $\Delta^{\mathrm{S}}_{\mathcal{A}}$. Thus, we study $\langle \nabla v^\pi_\lambda, \pi' - \pi\rangle$ and $\langle \nabla \mu v^\pi_\lambda, \pi' - \pi\rangle$, for which the following proposition gives a closed form:

Proposition 1 (Linear Approximation of a Policy's Value)**.** *Let $\pi, \pi' \in \Delta^S_A$, and $d_{\mu,\pi} = (1-\gamma)\mu(I - \gamma P^\pi)^{-1}$. Then,*

$$\langle \nabla_\pi v^\pi_\lambda, \pi' - \pi\rangle = (I - \gamma P^\pi)^{-1}\left(T^{\pi'}_\lambda v^\pi_\lambda - v^\pi_\lambda - \lambda B_\omega(\pi', \pi)\right), \quad (9)$$

$$\langle \nabla_\pi \mu v^\pi_\lambda, \pi' - \pi\rangle = \frac{1}{1-\gamma} d_{\mu,\pi}\left(T^{\pi'}_\lambda v^\pi_\lambda - v^\pi_\lambda - \lambda B_\omega(\pi', \pi)\right). \quad (10)$$

The proof, supplied in Appendix B, is a direct application of a Policy Gradient Theorem (Sutton et al. 2000) derived for regularized MDPs. Importantly, the linear approximation is scaled by $(I - \gamma P^\pi)^{-1}$ or $\frac{1}{1-\gamma} d_{\mu,\pi}$, the discounted visitation frequency induced by the current policy. In what follows, we use this understanding to properly choose an *adaptive scaling* for the proximity term of TRPO, which allows us to use methods from convex optimization.

*We work with costs instead of rewards to comply with convex analysis. All results are valid to the case where a reward is used.

5 Uniform Trust Region Policy Optimization

In this section we formulate *Uniform TRPO*, a trust region *planning* algorithm with an adaptive proximity term by which (7) can be solved, i.e., an optimal policy which jointly minimizes the vector v_λ^π is acquired. We show that the presence of the adaptive term simplifies the update rule of Uniform TRPO and then analyze its performance for both the regularized ($\lambda > 0$) and unregularized ($\lambda = 0$) cases. Despite the fact (7) is not a convex optimization problem, the presence of the adaptive term allows us to use techniques applied for MD in convex analysis and establish convergence to the global optimum with rates of $\tilde{O}(1/\sqrt{N})$ and $\tilde{O}(1/N)$ for the unregularized and regularized case, respectively.

Algorithm 1 Uniform TRPO

initialize: t_k, γ, λ, π_0 is the uniform policy.
  for $k = 0, 1, ...$ **do**
    $v^{\pi_k} = (I - \gamma P^{\pi_k})^{-1} c_\lambda^{\pi_k}$
    for $\forall s \in \mathcal{S}$ **do**
      for $\forall a \in \mathcal{A}$ **do**
        $q_\lambda^{\pi_k}(s,a) \leftarrow c_\lambda^{\pi_k}(s,a) + \gamma \sum_{s'} p(s'|s,a) v_\lambda^{\pi_k}(s')$
      end for
      $\pi_{k+1}(\cdot|s) \leftarrow$ PolicyUpdate$(\pi(\cdot|s), q_\lambda^{\pi_k}(s,\cdot), t_k, \lambda)$
    end for
  end for

Uniform TRPO repeats the following iterates

$$\pi_{k+1} \in \arg\min_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} \Big\{ \langle \nabla v_\lambda^{\pi_k}, \pi - \pi_k \rangle + \frac{1}{t_k} (I - \gamma P^{\pi_k})^{-1} B_\omega(\pi, \pi_k) \Big\}. \quad (11)$$

The update rule resembles MD's updating-rule (2). The updated policy minimizes the linear approximation while being not 'too-far' from the current policy due to the presence of $B_\omega(\pi, \pi_k)$. However, and unlike MD's update rule, the Bregman distance is scaled by the adaptive term $(I - \gamma P^{\pi_k})^{-1}$. Applying Proposition 1, we see why this adaptive term is so natural for RL,

$$\pi_{k+1} \in \arg\min_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} (I - \gamma P^{\pi_k})^{-1} \overbrace{\Big(T_\lambda^\pi v_\lambda^{\pi_k} - v_\lambda^{\pi_k} + \Big(\frac{1}{t_k} - \lambda \Big) B_\omega(\pi, \pi_k) \Big)}^{(*)}. \quad (12)$$

Since $(I - \gamma P^{\pi_k})^{-1} \geq 0$ component-wise, minimizing (12) is equivalent to minimizing the vector $(*)$. This results in a simplified update rule: instead of minimizing over $\Delta_{\mathcal{A}}^{\mathrm{S}}$ we minimize over $\Delta_{\mathcal{A}}$ for each $s \in \mathcal{S}$ independently (see Appendix C.1). For each $s \in \mathcal{S}$ the policy is updated by

$$\pi_{k+1}(\cdot \,|\, s) \in \arg\min_{\pi \in \Delta_{\mathcal{A}}} t_k T_\lambda^\pi v_\lambda^{\pi_k}(s) + (1 - \lambda t_k) B_\omega(s; \pi, \pi_k). \quad (13)$$

This is the update rule of Algorithm 1. Importantly, the update rule is a direct consequence of choosing the adaptive scaling for the Bregman distance in (11), and without it, the trust region problem would involve optimizing over $\Delta_{\mathcal{A}}^{\mathrm{S}}$.

Algorithm 2 PolicyUpdate: PPG

input: $\pi(\cdot \mid s), q(s,\cdot), t_k, \lambda$
for $a \in \mathcal{A}$ **do**
  $\pi(a|s) \leftarrow \pi(a \mid s) - \frac{t_k}{1 - \lambda t_k} q(s,a)$
end for
$\pi(\cdot|s) \leftarrow P_{\Delta_{\mathcal{A}}}(\pi(\cdot \mid s))$
return $\pi(\cdot \mid s)$

Algorithm 3 PolicyUpdate: NE-TRPO

input: $\pi(\cdot \mid s), q(s,\cdot), t_k, \lambda$
for $a \in \mathcal{A}$ **do**
  $\pi(a \mid s) \leftarrow \frac{\pi(a|s) \exp(-t_k(q(s,a) + \lambda \log \pi_k(a|s)))}{\sum_{a' \in \mathcal{A}} \pi(a'|s) \exp(-t_k(q(s,a') + \lambda \log \pi_k(a'|s)))}$
end for
return $\pi(\cdot \mid s)$

By instantiating the *PolicyUpdate* procedure with Algorithms 2 and 3 we get the PPG and NE-TRPO, respectively, which are instances of Uniform TRPO. Instantiating PolicyUpdate is equivalent to choosing ω and the induced Bregman distance B_ω. In the euclidean case, $\omega(\cdot) = \frac{1}{2} \|\cdot\|_2^2$ (Alg. 2), and in the non-euclidean case, $\omega(\cdot) = H(\cdot)$ (Alg. 3). This comes in complete analogy to the fact Projected Gradient Descent and Exponentiated Gradient Descent are instances of MD with similar choices of ω (Section 2).

With the analogy to MD (2) in mind, one would expect Uniform TRPO, to converge with rates $\tilde{O}(1/\sqrt{N})$ and $\tilde{O}(1/N)$ for the unregularized and regularized cases, respectively, similarly to MD. Indeed, the following theorem formalizes this intuition for a proper choice of learning rate. The proof of Theorem 2 extends the techniques of Beck (2017) from convex analysis to the non-convex optimization problem (5), by relying on the adaptive scaling of the Bregman distance in (11) (see Appendix C).

Theorem 2 (Convergence Rate: Uniform TRPO)**.** *Let $\{\pi_k\}_{k \geq 0}$ be the sequence generated by Uniform TRPO. Then, the following holds for all $N \geq 1$:*

1. (Unregularized) *Let $\lambda = 0$, $t_k = \frac{(1-\gamma)}{C_{\omega,1} C_{max} \sqrt{k+1}}$, then*

$$\|v^{\pi_N} - v^*\|_\infty \leq O\left(\frac{C_{\omega,1} \mathrm{C}_{max}}{(1-\gamma)^2 \sqrt{N}} \right).$$

2. (Regularized) *Let $\lambda > 0$, $t_k = \frac{1}{\lambda(k+2)}$, then*

$$\|v_\lambda^{\pi_N} - v_\lambda^*\|_\infty \leq O\left(\frac{C_{\omega,1}^2 \, \mathrm{C}_{max,\lambda}{}^2}{\lambda (1-\gamma)^3 N} \right).$$

Theorem 2 establishes that regularization allows faster convergence of $\tilde{O}(1/N)$. It is important to note using such regularization leads to a 'biased' solution: Generally $\left\| v^{\pi_\lambda^*} - v^* \right\|_\infty > 0$, where we denote π_λ^* as the optimal policy of the regularized MDP. In other words, the optimal

policy of the regularized MDP evaluated on the unregularized MDP is not necessarily the optimal one. However, when adding such regularization to the problem, it becomes easier to solve, in the sense Uniform TRPO converges faster (for a proper choice of learning rate).

In the next section, we extend the analysis of Uniform TRPO to Sample-Based TRPO, and relax the assumption of having access to the entire state space in each iteration, while still securing similar convergence rates in N.

6 Exact and Sample-Based TRPO

In the previous section we analyzed Uniform TRPO, which uniformly minimizes the vector v^π. Practically, in large-scale problems, such an objective is infeasible as one cannot access the entire state space, and less ambitious goal is usually defined (Sutton et al. 2000; Kakade and Langford 2002; Schulman et al. 2015). The objective usually minimized is the *scalar objective* (8), the expectation of $v^\pi_\lambda(s)$ under a measure μ, $\min_{\pi\in\Delta^{\mathrm{S}}_{\mathcal{A}}} \mathbb{E}_{s\sim\mu}[v^\pi_\lambda(s)] = \min_{\pi\in\Delta^{\mathrm{S}}_{\mathcal{A}}} \mu v^\pi_\lambda$.

Starting from the seminal work on CPI, it is common to assume access to the environment in the form of a *ν-restart model*. Using a ν-restart model, the algorithm interacts with an MDP in an episodic manner. In each episode k, the starting state is sampled from the initial distribution $s_0 \sim \nu$, and the algorithm samples a trajectory $(s_0, r_0, s_1, r_1, ...)$ by following a policy π_k. As mentioned in Kakade and others (2003), a ν-restart model is a weaker assumption than an access to the true model or a generative model, and a stronger assumption than the case where no restarts are allowed.

To establish global convergence guarantees for CPI, Kakade and Langford (2002) have made the following assumption, which we also assume through the rest of this section:

Assumption 1 (Finite Concentrability Coefficient)**.** $C^{\pi^*} := \left\| \frac{d_{\mu,\pi^*}}{\nu} \right\|_\infty = \max_{s\in\mathcal{S}} \left| \frac{d_{\mu,\pi^*}(s)}{\nu(s)} \right| < \infty.$

The term C^{π^*} is known as a concentrability coefficient and appears often in the analysis of policy search algorithms (Kakade and Langford 2002; Scherrer and Geist 2014; Bhandari and Russo 2019). Interestingly, C^{π^*} is considered the 'best' one among all other existing concentrability coefficients in approximate Policy Iteration schemes (Scherrer 2014), in the sense it can be finite when the rest of them are infinite.

6.1 Warm Up: Exact TRPO

We split the discussion on the sample-based version of TRPO: we first discuss *Exact TRPO* which minimizes the scalar μv^π_λ (8) instead of minimizing the vector v^π_λ (7) as Uniform TRPO, while having an exact access to the gradients. Importantly, its updating rule is **the same update rule used in NE-TRPO** (Schulman et al. 2015, Equation 12), which uses the adaptive proximity term, and is described there as a heuristic. Specifically, there are two minor discrepancies between NE-TRPO and Exact TRPO: 1) We use a penalty formulation instead of a constrained optimization problem. 2) The policies in the Kullback-Leibler divergence are reversed. Exact TRPO is a straightforward adaptation of Uniform TRPO to solve (8) instead of (7) as we establish in Proposition 3. Then, in the next section, we extend Exact TRPO to a sample-based version with provable guarantees.

With the goal of minimizing the objective μv^π_λ, Exact TRPO repeats the following iterates

$$\pi_{k+1} \in \arg\min_{\pi\in\Delta^{\mathrm{S}}_{\mathcal{A}}} \Big\{ \langle \nabla \nu v^{\pi_k}_\lambda, \pi - \pi_k \rangle + \frac{1}{t_k(1-\gamma)} d_{\nu,\pi_k} B_\omega(\pi, \pi_k) \Big\}, \quad (14)$$

Its update rule resembles MD's update rule (11), but uses the ν-restart distribution for the linearized term. Unlike in MD (2), the Bregman distance is scaled by an adaptive scaling factor d_{ν,π_k}, using ν and the policy π_k by which the algorithm interacts with the MDP. This update rule is motivated by the one of Uniform TRPO analyzed in previous section (11) as the following straightforward proposition suggests (Appendix D.2):

Proposition 3 (Uniform to Exact Updates)**.** *For any* $\pi, \pi_k \in \Delta^{\mathrm{S}}_{\mathcal{A}}$

$$\nu\Big(\langle \nabla v^{\pi_k}_\lambda, \pi - \pi_k \rangle + \frac{1}{t_k} (I - \gamma P^{\pi_k})^{-1} B_\omega(\pi, \pi_k) \Big) = \langle \nabla \nu v^{\pi_k}_\lambda, \pi - \pi_k \rangle + \frac{1}{t_k(1-\gamma)} d_{\nu,\pi_k} B_\omega(\pi, \pi_k).$$

Meaning, the proximal objective solved in each iteration of Exact TRPO (14) is the expectation w.r.t. the measure ν of the objective solved in Uniform TRPO (11).

Similarly to the simplified update rule for Uniform TRPO (12), by using the linear approximation in Proposition 1, it can be easily shown that using the adaptive proximity term allows to obtain a simpler update rule for Exact TRPO. Unlike Uniform TRPO which updates all states, Exact TRPO updates only states for which $d_{\nu,\pi_k}(s) > 0$. Denote $\mathcal{S}_{d_{\nu,\pi_k}} = \{s : d_{\nu,\pi_k}(s) > 0\}$ as the set of these states. Then, Exact TRPO is equivalent to the following update rule (see Appendix D.2), $\forall s \in \mathcal{S}_{d_{\nu,\pi_k}}$:

$$\pi_{k+1}(\cdot \mid s) \in \arg\min_\pi t_k T^\pi_\lambda v^{\pi_k}_\lambda(s) + (1 - \lambda t_k) B_\omega(s; \pi, \pi_k),$$

i.e., it has the same updates as Uniform TRPO, but updates only states in $\mathcal{S}_{d_{\nu,\pi_k}}$. Exact TRPO converges with similar rates for both the regularized and unregularized cases, as Uniform TRPO. These are formally stated in Appendix D.

6.2 Sample-Based TRPO

In this section we derive and analyze the *sample-based* version of Exact TRPO, and establish high-probability convergence guarantees in a batch setting. Similarly to the previous section, we are interested in minimizing the scalar objective μv^π_λ (8). Differently from Exact TRPO which requires an access to a model and to simultaneous updates in all states in $\mathcal{S}_{d_{\nu,\pi_k}}$, *Sample-Based TRPO* assumes access to a ν-restart model. Meaning, it can only access sampled trajectories and restarts according to the distribution ν.

Sample-Based TRPO samples M_k trajectories per episode. In every trajectory of the k-th episode, it first samples $s_m \sim d_{\nu,\pi_k}$ and takes an action $a_m \sim U(\mathcal{A})$ where

Algorithm 4 Sample-Based TRPO

initialize: t_k, γ, λ, π_0 is the uniform policy, $\epsilon, \delta > 0$
  for $k = 0, 1, ...$ **do**
    $\mathcal{S}_M^k = \{\}, \forall s, a,\ \hat{q}_\lambda^{\pi_k}(s,a) = 0, n_k(s,a) = 0$
    $M_k \geq \tilde{O}(\frac{A^2\, \mathrm{C}_{\max,\lambda}^2 (S \log 2A + \log 1/\delta)}{(1-\gamma)^2 \epsilon^2})$ # Appendix E.5
    # Sample Trajectories
    for $m = 1, .., M_k$ **do**
      Sample $s_m \sim d_{\nu,\pi_k}(\cdot)$, $a_m \sim U(\mathcal{A})$
      $\hat{q}_\lambda^{\pi_k}(s_m, a_m, m)$= Truncated rollout of $q_\lambda^{\pi_k}(s_m, a_m)$
      $\hat{q}_\lambda^{\pi_k}(s_m, a_m) \leftarrow \hat{q}_\lambda^{\pi_k}(s_m, a_m) + \hat{q}_\lambda^{\pi_k}(s_m, a_m, m)$
      $n_k(s_m, a_m) \leftarrow n_k(s_m, a_m) + 1$
      $\mathcal{S}_M^k = \mathcal{S}_M^k \cup \{s_m\}$
    end for
    # Update Next Policy
    for $\forall s \in \mathcal{S}_M^k$ **do**
      for $\forall a \in \mathcal{A}$ **do**
        $\hat{q}_\lambda^{\pi_k}(s,a) \leftarrow A \hat{q}_\lambda^{\pi_k}(s,a) / (\sum_a n_k(s,a))$
      end for
      $\pi_{k+1}(\cdot|s)$ =PolicyUpdate($\pi_k(\cdot|s), \hat{q}_\lambda^{\pi_k}(s,\cdot), t_k, \lambda$)
    end for
  end for

$U(\mathcal{A})$ is the uniform distribution on the set $\mathcal{A}$. Then, by following the current policy π_k, it estimates $q_\lambda^{\pi_k}(s_m, a_m)$ using a rollout (possibly truncated in the infinite horizon case). We denote this estimate as $\hat{q}_\lambda^{\pi_k}(s_m, a_m, m)$ and observe it is (nearly) an unbiased estimator of $q_\lambda^{\pi_k}(s_m, a_m)$. We assume that each rollout runs sufficiently long so that the bias is small enough (the sampling process is fully described in Appendix E.2). Based on this data, Sample-Based TRPO updates the policy at the end of the k-th episode, by the following proximal problem,

$$\pi_{k+1} \in \arg\min_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} \Big\{ \frac{1}{M} \sum_{m=1}^{M} \frac{1}{t_k(1-\gamma)} B_\omega\left(s_m; \pi, \pi_k\right) + \langle \hat{\nabla} \nu v_\lambda^{\pi_k}[m], \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m) \rangle \Big\}, \quad (15)$$

where the estimation of the gradient is $\hat{\nabla} \nu v_\lambda^{\pi_k}[m] := \frac{1}{1-\gamma}(A \hat{q}_\lambda^{\pi_k}(s_m, \cdot, m) \mathbb{1}\{\cdot = a_m\} + \lambda \nabla \omega\left(s_m; \pi_k\right))$.

The following proposition motivates the study of this update rule and formalizes its relation to Exact TRPO:

Proposition 4 (Exact to Sample-Based Updates)**.** *Let $\mathcal{F}_k$ be the σ-field containing all events until the end of the $k-1$ episode. Then, for any $\pi, \pi_k \in \Delta_{\mathcal{A}}^{\mathrm{S}}$ and every sample m,*

$$\begin{aligned} &\langle \nabla \nu v_\lambda^{\pi_k}, \pi - \pi_k \rangle + \frac{1}{t_k(1-\gamma)} d_{\nu,\pi_k} B_\omega\left(\pi, \pi_k\right) \\ &= \mathbb{E}\Big[\langle \hat{\nabla} \nu v_\lambda^{\pi_k}[m], \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m) \rangle \\ &\qquad + \frac{1}{t_k(1-\gamma)} B_\omega\left(s_m; \pi, \pi_k\right) \mid \mathcal{F}_k \Big]. \end{aligned}$$

Meaning, the expectation of the proximal objective of Sample-Based TRPO (15) is the proximal objective of Exact TRPO (14). This fact motivates us to study this algorithm, anticipating it inherits the convergence guarantees of its exacted counterpart.

Like Uniform and Exact TRPO, Sample-Based TRPO has a simpler update rule, in which, the optimization takes place on every *visited state* at the k-th episode. This comes in contrast to Uniform and Exact TRPO which require access to all states in $\mathcal{S}$ or $\mathcal{S}_{d_{\nu,\pi_k}}$, and is possible due to the *sample-based adaptive scaling* of the Bregman distance. Let $\mathcal{S}_M^k$ be the set of visited states at the k-th episode, $n(s,a)$ the number of times $(s_m, a_m) = (s,a)$ at the k-th episode, and

$$\hat{q}_\lambda^{\pi_k}(s,a) = \frac{A}{\sum_a n(s,a)} \sum_{i=1}^{n(s,a)} \hat{q}_\lambda^{\pi_k}(s,a,m_i),$$

is the empirical average of all rollout estimators for $q_\lambda^{\pi_k}(s,a)$ gathered in the k-th episode (m_i is the episode in which $(s_m, a_m) = (s,a)$ for the i-th time). If the state action pair (s,a) was not visited at the k-th episode then $\hat{q}_\lambda^{\pi_k}(s,a) = 0$. Given these definitions, Sample-Based TRPO updates the policy for all $s \in \mathcal{S}_M^k$ by a simplified update rule:

$$\pi_{k+1}(\cdot \mid s) \in \arg\min_{\pi} t_k \langle \hat{q}_\lambda^{\pi_k}(s,\cdot) + \lambda \nabla \omega\left(s; \pi_k\right), \pi \rangle + B_\omega\left(s; \pi, \pi_k\right),$$

As in previous sections, the euclidean and non-euclidean choices of ω correspond to a PPG and NE-TRPO instances of Sample-Based TRPO. The different choices correspond to instantiating PolicyUpdate with the subroutines 2 or 3. Generalizing the proof technique of Exact TRPO and using standard concentration inequalities, we derive a high-probability convergence guarantee for Sample-Based TRPO (see Appendix E). An additional important lemma for the proof is Lemma 27 provided in the appendix. This lemma bounds the change $\nabla \omega(\pi_k) - \nabla \omega(\pi_{k+1})$ between consecutive episodes by a term proportional to t_k. Had this bound been t_k-independent, the final results would deteriorate significantly.

Theorem 5 (Convergence Rate: Sample-Based TRPO)**.** *Let $\{\pi_k\}_{k \geq 0}$ be the sequence generated by Sample-Based TRPO, using $M_k \geq O\Big(\frac{A^2\, \mathrm{C}_{max,\lambda}^2 (S \log A + \log 1/\delta)}{(1-\gamma)^2 \epsilon^2}\Big)$ samples in each iteration, and $\{\mu v_{\textbf{\textit{best}}}^k\}_{k \geq 0}$ be the sequence of best achieved values, $\mu v_{\textbf{\textit{best}}}^N := \arg\min_{k=0,...,N} \mu v_\lambda^{\pi_k} - \mu v_\lambda^*$. Then, with probability greater than $1-\delta$ for every $\epsilon > 0$ the following holds for all $N \geq 1$:*

1. (Unregularized) *Let $\lambda = 0$, $t_k = \frac{(1-\gamma)}{C_{\omega,1}\, \mathrm{C}_{max}\, \sqrt{k+1}}$, then*

$$\mu v_{\textbf{\textit{best}}}^N - \mu v^* \leq O\Big(\frac{C_{\omega,1}\, \mathrm{C}_{max}}{(1-\gamma)^2 \sqrt{N}} + \frac{C^{\pi^*} \epsilon}{(1-\gamma)^2} \Big).$$

2. (Regularized) *Let $\lambda > 0$, $t_k = \frac{1}{\lambda(k+2)}$, then*

$$\mu v_{\textbf{\textit{best}}}^N - \mu v_\lambda^* \leq O\Big(\frac{C_{\omega,1}^2 C_{\omega,2}\, \mathrm{C}_{max,\lambda}{}^2}{\lambda (1-\gamma)^3 N} + \frac{C^{\pi^*} \epsilon}{(1-\gamma)^2} \Big).$$

Where $C_{\omega,2} = 1$ for the euclidean case, and $C_{\omega,2} = A^2$ for the non-euclidean case.

Method	Sample Complexity
TRPO (this work)	$\frac{C_{\omega,1}^2 A^2 C_{\max}^4 \left(S+\log\frac{1}{\delta}\right)}{(1-\gamma)^3\epsilon^4}$
Regularized TRPO (this work)	$\frac{C_{\omega,1}^2 C_{\omega,2} A^2 C_{\max,\lambda}^4 \left(S+\log\frac{1}{\delta}\right)}{\lambda(1-\gamma)^4\epsilon^3}$
CPI (Kakade and Langford)	$\frac{A^2 C_{\max}^4 \left(S+\log\frac{1}{\delta}\right)}{(1-\gamma)^5\epsilon^4}$

Table 1: The sample complexity of Sample-Based TRPO (TRPO) and CPI. For TRPO, the best policy so far is returned, where for CPI, the last policy π_N is returned.

Similarly to Uniform TRPO, the convergence rates are $\tilde{O}(1/\sqrt{N})$ and $\tilde{O}(1/N)$ for the unregularized and regularized cases, respectively. However, the Sample-Based TRPO converges to an approximate solution, similarly to CPI. The *sample complexity* for a $\frac{C^{\pi^*}\epsilon}{(1-\gamma)^2}$ error, the same as the error of CPI, is given in Table 6.2. Interestingly, Sample-Based TRPO has better polynomial sample complexity in $(1-\gamma)^{-1}$ relatively to CPI. Importantly, **the regularized versions have a superior sample-complexity in** ϵ, which can explain the empirical success of using regularization.

Remark 1 (Optimization Perspective)**.** *From an optimization perspective, CPI can be interpreted as a sample-based Conditional Gradient Descent (Frank-Wolfe) for solving MDPs (Scherrer and Geist 2014). With this in mind, the two analyzed instances of Sample-Based TRPO establish the convergence of sample-based projected and exponentiated gradient descent methods for solving MDPs: PPG and NE-TRPO. It is well known that a convex problem can be solved with any one of the three aforementioned methods. The convergence guarantees of CPI together with the ones of Sample-Based TRPO establish the same holds for RL.*

Remark 2 (Is Improvement and Early Stopping Needed?)**.** *Unlike CPI, Sample-Based TRPO does not rely on improvement arguments or early stopping. Even so, its asymptotic performance is equivalent to CPI, and its sample complexity has better polynomial dependence in* $(1-\gamma)^{-1}$*. This questions the necessity of ensuring improvement for policy search methods, heavily used in the analysis of these methods, yet less used in practice, and motivated by the analysis of CPI.*

7 Related Works

The empirical success of policy search and regularization techniques in RL (Peters and Schaal 2008; Mnih et al. 2016; Schulman et al. 2015; 2017) led to non-negligible theoretical analysis of these methods. Gradient based policy search methods were mostly analyzed in the function approximation setting, e.g., (Sutton et al. 2000; Bhatnagar et al. 2009; Pirotta, Restelli, and Bascetta 2013; Dai et al. 2018; Papini, Pirotta, and Restelli 2019; Bhandari and Russo 2019). There, convergence to a local optimum was established under different conditions and several aspects of policy search methods were investigated. In this work, we study a trust-region based, as opposed to gradient based, policy search method in tabular RL and establish global convergence guarantees. Regarding regularization in TRPO, in Neu, Jonsson, and Gómez (2017) the authors analyzed entropy regularized MDPs from a linear programming perspective for average-reward MDPs. Yet, convergence rates were not supplied, as opposed to this paper.

In Geist, Scherrer, and Pietquin (2019) different aspects of regularized MDPs were studied, especially, when combined with MD-like updates in an approximate PI scheme (with partial value updates). The authors focus on update rules which require uniform access to the state space of the form $\pi_{k+1} = \arg\min_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}} \langle q_k, \pi - \pi_k\rangle + B_\omega(\pi, \pi_k)$, similarly to the simplified update rule of Uniform TRPO (13) with a fixed learning rate, $t_k = 1$. In this paper, we argued it is instrumental to view this update rule as an instance of the more general update rule (11), i.e., MD with an adaptive proximity term. This view allowed us to formulate and analyze the adaptive Sample-Based TRPO, which does not require uniform access to the state space. Moreover, we proved Sample-Based TRPO inherits the same asymptotic performance guarantees of CPI. Specifically, the quality of the policy Sample-Based TRPO outputs depends on the concentrability coefficient C^{π^*}. The results of Geist, Scherrer, and Pietquin (2019) in the approximate setting led to a worse concentrability coefficient, C_q^i, which can be infinite even when C^{π^*} is finite (Scherrer 2014) as it depends on the worst case of all policies.

In a recent work of Agarwal et al. (2019), Section 4.2, the authors study a variant of Projected Policy Gradient Descent and analyze it under the assumption of exact gradients and uniform access to the state space. The proven convergence rate depends on both S and C^{π^*} whereas the convergence rate of Exact TRPO (Section 6.1) does not depend on S nor on C^{π^*} (see Appendix D.4), and is similar to the guarantees of Uniform TRPO (Theorem 2). Furthermore, the authors do not establish faster rates for regularized MDPs. It is important to note their projected policy gradient algorithm is *different* than the one we study, which can explain the discrepancy between our results. Their projected policy gradient updates by $\pi_{k+1} \in P_{\Delta_{\mathcal{A}}^{\mathcal{S}}}(\pi_k - \eta\nabla\mu v^{\pi_k})$, whereas, the Projected Policy Gradient studied in this work applies a different update rule based on the adaptive scaling of the Bregman distance.

Lastly, in another recent work of Liu et al. (2019) the authors established global convergence guarantees for a sampled-based version of TRPO when neural networks are used as the q-function and policy approximators. The sample complexity of their algorithm is $O(\epsilon^{-8})$ (as opposed to $O(\epsilon^{-4})$ we obtained) neglecting other factors. It is an interesting question whether their result can be improved.

8 Conclusions and Future Work

We analyzed the Uniform and Sample-Based TRPO methods. The first is a planning, trust region method with an adaptive proximity term, and the latter is an RL sample-based version of the first. Different choices of the proximity term led to two instances of the TRPO method: PPG and NE-TRPO. For both, we proved $\tilde{O}(1/\sqrt{N})$ convergence rate

to the global optimum, and a faster $\tilde{O}(1/N)$ rate for regularized MDPs. Although Sample-Based TRPO does not necessarily output an improving sequence of policies, as CPI, its best policy in hindsight does improve. Furthermore, the asymptotic performance of Sample-Based TRPO is equivalent to that of CPI, and its sample complexity exhibits better dependence in $(1-\gamma)^{-1}$. These results establish the popular NE-TRPO (Schulman et al. 2015) should not be interpreted as an approximate heuristic to CPI but as a viable alternative.

In terms of future work, an important extension of this study is deriving algorithms with linear convergence, or, alternatively, establish impossibility results for such rates in RL problems. Moreover, while we proved positive results on regularization in RL, we solely focused on the question of optimization. We believe that establishing more positive as well as negative results on regularization in RL is of value. Lastly, studying further the implication of the adaptive proximity term in RL is of importance due to the empirical success of NE-TRPO and its now established convergence guarantees.

9 Acknowledgments

We would like to thank Amir Beck for illuminating discussions regarding Convex Optimization and Nadav Merlis for helpful comments. This work was partially funded by the Israel Science Foundation under ISF grant number 1380/16.

References

Agarwal, A.; Kakade, S. M.; Lee, J. D.; and Mahajan, G. 2019. Optimality and approximation with policy gradient methods in markov decision processes. *arXiv preprint arXiv:1908.00261*.

Ahmed, Z.; Le Roux, N.; Norouzi, M.; and Schuurmans, D. 2019. Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, 151–160.

Beck, A., and Teboulle, M. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters* 31(3):167–175.

Beck, A. 2017. *First-order methods in optimization*, volume 25. SIAM.

Bhandari, J., and Russo, D. 2019. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*.

Bhatnagar, S.; Sutton, R. S.; Ghavamzadeh, M.; and Lee, M. 2009. Natural actor–critic algorithms. *Automatica* 45(11):2471–2482.

Chow, Y.; Nachum, O.; and Ghavamzadeh, M. 2018. Path consistency learning in tsallis entropy regularized mdps. In *International Conference on Machine Learning*, 978–987.

Dai, B.; Shaw, A.; Li, L.; Xiao, L.; He, N.; Liu, Z.; Chen, J.; and Song, L. 2018. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In Dy, J., and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1125–1134. Stockholmsmässan, Stockholm Sweden: PMLR.

Fox, R.; Pakman, A.; and Tishby, N. 2016. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 202–211. AUAI Press.

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

Juditsky, A.; Nemirovski, A.; et al. 2011. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. *Optimization for Machine Learning* 121–148.

Kakade, S., and Langford, J. 2002. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, 267–274.

Kakade, S. M., et al. 2003. *On the sample complexity of reinforcement learning*. Ph.D. Dissertation, University of London London, England.

Liu, B.; Cai, Q.; Yang, Z.; and Wang, Z. 2019. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937.

Nachum, O.; Norouzi, M.; Xu, K.; and Schuurmans, D. 2017. Trust-pcl: An off-policy trust region method for continuous control. *arXiv preprint arXiv:1707.01891*.

Nedic, A., and Lee, S. 2014. On stochastic subgradient mirror-descent algorithm with weighted averaging. *SIAM Journal on Optimization* 24(1):84–107.

Nesterov, Y. 1998. *Introductory lectures on convex programming volume i: Basic course*. Springer, New York, NY.

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

Papini, M.; Pirotta, M.; and Restelli, M. 2019. Smoothing policies and safe policy gradients. *arXiv preprint arXiv:1905.03231*.

Peters, J., and Schaal, S. 2008. Natural actor-critic. *Neurocomputing* 71(7-9):1180–1190.

Pirotta, M.; Restelli, M.; and Bascetta, L. 2013. Adaptive step-size for policy gradient methods. In *Advances in Neural Information Processing Systems*, 1394–1402.

Scherrer, B., and Geist, M. 2014. Local policy search in a convex space and conservative policy iteration as boosted policy search. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 35–50. Springer.

Scherrer, B. 2014. Approximate policy iteration schemes: a comparison. In *International Conference on Machine Learning*, 1314–1322.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *International Conference on Machine Learning*, 1889–1897.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sutton, R. S., and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S.; McAllester, D. A.; Singh, S. P.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, 1057–1063.