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
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 function 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 $O(1/\sqrt{N})$ convergence rate to the global optimum for both Uniform and Sample-Based TRPO.

- We prove a faster rate of $O(1/N)$ for regularized MDPs. To the best of our knowledge, this 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

$$x^* \in \operatorname{arg\,min}_{x \in C} f(x),$$

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

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

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 $\omega$ 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 $O(1/\sqrt{N})$, where $N$ is the number of MD iterations (Beck and Teboulle 2003; Juditsky, Nemirovski, and others 2011), i.e., $f(x_k) - f^* \leq O(1/\sqrt{K})$, 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 $\lambda$-strongly convex functions w.r.t. the Bregman distance. We say that $f$ is $\lambda$-strongly convex w.r.t. the Bregman distance if $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \lambda B_\omega(x, y)$. For such $f$, improved convergence rate of $O(1/N)$ 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,

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

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,

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

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 $\omega$ which induce a proper Bregman distance: (a) The euclidean case, with $\omega(\cdot) := 1/2 \|\cdot\|^2_2$ and the resulting Bregman distance is the squared euclidean norm $B_\omega(x, y) = 1/2 \|x - y\|^2_2$. 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 $C$, $x_{k+1} = \operatorname{Proj}_C(x_k - t_k \nabla f(x_k))$, where $\operatorname{Proj}_C(x) = \min_{y \in C} 1/2 \|x - y\|^2_2$ 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 $\omega$ as the negative entropy, (2) has a closed form solution (Beck 2017, Example 3.71), $x_{k+1}^i = x_k^i e^{-t_k \nabla f_i(x_k)}$, where $x_k^i$ and $\nabla f_i$ are the $i$-th coordinates of $x_k$ and $\nabla f$.

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

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

(5)

This goal can be achieved using the classical operators:

$$\forall v, \pi, T^\pi v = c^\pi + \gamma P^\pi v, \text{and } \forall v, T v = \min_{\pi} T^\pi v.$$  

(6)

where $T^\pi$ is a linear operator, $T$ is the optimal Bellman operator and both $T^\pi$ and $T$ are $\gamma$-contraction mappings w.r.t. the max-norm. The fixed points of $T^\pi$ and $T$ are $v^\pi$ 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 $c^\lambda_{\pi}$ for $\lambda \geq 0$. Specifically, the cost of a policy $\pi$ on a regularized MDP translates to $c^\lambda_{\pi}(s) := c^\pi(s) + \lambda\omega(s; \pi)$ where $\omega(s; \pi) := \omega(\pi(\cdot \mid s))$ and $\omega : \Delta_A \to \mathbb{R}$ 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 $c^\pi$ is recovered. In this work we consider two choices of $\omega$: the euclidean case $\omega(s; \pi) = \frac{1}{2} \| \pi(\cdot \mid s) \|^2_2$ and non-euclidean case $\omega(s; \pi) = H(\pi(\cdot \mid s)) + \log \lambda$. By this choice we have that $0 \leq c^\lambda_{\pi}(s) \leq C_{\text{max}, \lambda}$, where $C_{\text{max}, \lambda} = C_{\text{max}} + \lambda$ and $C_{\text{max}, \lambda} = C_{\text{max}} + \lambda \log A$, for the euclidean and non-euclidean cases, respectively. With some abuse of notation we omit $\omega$ from $c^\lambda_{\pi}$.

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

$$v^\lambda_{\pi} = \min_{\pi} (I - \gamma P^\pi)^{-1} c^\lambda_{\pi} = (I - \gamma P^\pi)^{-1} c^\lambda_{\pi^*},$$  

(7)

$$\forall v, \pi, T^\pi_{\lambda} v = c^\lambda_{\pi} + \gamma P^\pi v, \text{and } \forall v, T v = \min_{\pi} T^\pi_{\lambda} v.$$  

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

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 $(5), (7)$,

$$\min_{\pi \in \Delta_A} \mathbb{E}_{s \sim v^\lambda_{\pi}} [v^\lambda_{\pi}(s)] = \min_{\pi \in \Delta_A} \mu v^\lambda_{\pi},$$  

(8)

where $\mu(\cdot)$ is a probability measure over the state space. Using this objective, one wishes to find a policy $\pi$ which minimizes the expectation of $v^\lambda_{\pi}(s)$ under a measure $\mu$. 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 $\omega$ to be associated with the Bregman distance used, $B_\omega$. This simplifies the analysis as $c^\lambda_{\pi}$ is $\lambda$-strongly convex w.r.t. $B_\omega$ by definition. Given two policies $\pi_1, \pi_2$, 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 $\omega$ 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 $\omega$-dependent constant, $C_{\omega} = \sqrt{A}$ in the euclidean case, and $C_{\omega} = 1$ 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 $(x, y) = \sum_{s,a} x(s, a) y(s, a)$, and the fixed-state inner product is $(x(s, \cdot), y(s, \cdot)) = \sum_a x(s, a) y(s, a)$. Lastly, when $x \in \mathbb{R}^{S \times S \times A}$ (e.g., first claim of Proposition 1) the inner product $(x, y)$ is a vector in $\mathbb{R}^S$ where $(x, y)(s) := (x(s, \cdot, \cdot), y)$, 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, $(\nabla f(x), x' - x)$, 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 $\Delta^\lambda_A$. Thus, we study $\langle \nabla v^\lambda_{\pi}, \pi' - \pi \rangle$ and $\langle \nabla \mu v^\lambda_{\pi}, \pi' - \pi \rangle$, for which the following proposition gives a closed form:

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

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

(9)

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

(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 $(1/1 - \gamma) d_{\mu, \pi, \pi}$, 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.
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 \((\lambda)\) can be solved, i.e., an optimal policy which jointly minimizes the vector \(v_{\lambda}\) 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 \((\lambda)\) 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 \(O(1/\sqrt{N})\) and \(O(1/N)\) for the unregularized and regularized case, respectively.

**Algorithm 1 Uniform TRPO**

initialize: \(t_k, \gamma, \lambda, \pi_0\) is the uniform policy.

for \(k = 0, 1, \ldots\) do

\[
v_{\pi_k} = (I - \gamma P_{\pi_k})^{-1}c_{\lambda}
\]

for \(\forall s \in S\) do

for \(\forall a \in \mathcal{A}\) do

\[
q_{\pi_k}^{s,a}(s, a) \leftarrow c_{\lambda}^s(s, a) + \gamma \sum_{s'} p(s'|s, a) v_{\pi_k}^{s'}(s')
\]

end for

\(\pi_{k+1}(\cdot | s) \leftarrow \text{PolicyUpdate}(\cdot | s), q_{\pi_k}^{s,a}(s, a), t_k, \lambda)\)

end for

end for

Uniform TRPO repeats the following iterates

\[
\pi_{k+1} \in \arg\min_{\pi \in \Delta_{\mathcal{A}}} \left\{ \langle \nabla v_{\pi_k}^s, \pi - \pi_k \rangle + \frac{1}{t_k} (I - \gamma P_{\pi_k})^{-1} B_{\omega} (\pi, \pi_k) \right\}. \quad (11)
\]

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,

\[
\pi_{k+1} \in \arg\min_{\pi \in \Delta_{\mathcal{A}}} (I - \gamma P_{\pi_k})^{-1} \left( T_{\lambda} v_{\pi_k}^s - v_{\pi_k}^s + \frac{1}{t_k} (I - \gamma P_{\pi_k})^{-1} B_{\omega} (\pi, \pi_k) \right). \quad (12)
\]

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

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

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 \(\Delta_{\mathcal{A}}\).

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 \(\omega\) and the induced Bregman distance \(B_{\omega}\). In the euclidean case, \(\omega(\cdot) = \frac{1}{\gamma} \parallel \cdot \parallel_{\ell_2}^2\) (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 \(\omega\) (Section 2).

With the analogy to MD (2) in mind, one would expect Uniform TRPO, to converge with rates \(O(1/\sqrt{N})\) and \(O(1/N)\) 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 \(\lambda = 0\), \(t_k = \frac{(1 - \gamma t)}{C_{max} \sqrt{k+1}}\), then

\[
\parallel v^{\pi_k} - v^* \parallel_\infty \leq O \left( \frac{C_{\omega,1} C_{max}}{(1 - \gamma)^2 \sqrt{N}} \right).
\]

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

\[
\parallel v^{\pi_k} - v^* \parallel_\infty \leq O \left( \frac{C_{\omega,2}^2 C_{max} \lambda^2}{\lambda (1 - \gamma)^3 N} \right).
\]

Theorem 2 establishes that regularization allows faster convergence of \(O(1/N)\). It is important to note such regularization leads to a ‘biased’ solution: Generally \(\parallel v^{\pi_k} - v^* \parallel_\infty > 0\), where we denote \(\pi^*_k\) 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 $v^\pi$. 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 $v^\pi(s)$ under a measure $\mu$, $\min_{\pi \in \Delta^v} \mathbb{E}_{\nu \sim \mu}[v^\pi(s)] = \min_{\pi \in \Delta^v} \mu v^\pi$.

Starting from the seminal work on CPI, it is common to assume access to the environment in the form of a $\nu$-restart model. Using a $\nu$-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 $(s_0, r_0, s_1, r_1, \ldots)$ by following a policy $\pi_k$. As mentioned in Kakade and others (2003), a $\nu$-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).

$$ C^\pi := \left\| \frac{d_{\nu,\pi}}{\nu} \right\|_{\infty} = \max_{s \in S} \left\| \frac{d_{\nu,\pi}(s)}{\nu(s)} \right\| < \infty. $$

The term $C^\pi$ 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, $C^\pi$ 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 $\mu v^\pi$ (8) instead of minimizing the vector $v^\pi$ (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.

The goal of minimizing the objective $\mu v^\pi$, Exact TRPO repeats the following iterates

$$ \pi_{k+1} \in \arg \min_{\pi \in \Delta^v} \left\{ \langle \nabla v^\pi, \pi - \pi_k \rangle + \frac{1}{t_k(1-\gamma)} d_{\nu,\pi_k} B_{\omega} (\pi, \pi_k) \right\}, \quad (14) $$

Its update rule resembles MD’s update rule (11), but uses the $\nu$-restart distribution for the linearized term. Unlike in MD (2), the Bregman distance is scaled by an adaptive scaling factor $d_{\nu,\pi_k}$, using $\nu$ and the policy $\pi_k$ 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 $\pi, \pi_k \in \Delta^v$

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

Meaning, the proximal objective solved in each iteration of Exact TRPO (14) is the expectation w.r.t. the measure $\nu$ 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 $S_{d_{\nu,\pi_k}} = \{ s : d_{\nu,\pi_k}(s) > 0 \}$ as the set of these states. Then, Exact TRPO is equivalent to the following update rule (see Appendix D.2).

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

i.e., it has the same updates as Uniform TRPO, but updates only states in $S_{d_{\nu,\pi_k}}$. 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 $\mu v^\pi$ (8). Differently from Exact TRPO which requires an access to a model and to simultaneous updates in all states in $S_{d_{\nu,\pi_k}}$, Sample-Based TRPO assumes access to a $\nu$-restart model. Meaning, it can only access sampled trajectories and restarts according to the distribution $\nu$.

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 $a_m \sim U(A)$ where
Algorithm 4 Sample-Based TRPO

initialize: \( t_k, \gamma, \lambda, \pi_0 \) is the uniform policy, \( \epsilon, \delta > 0 \)
for \( k = 0, 1, \ldots \) do
\[
S_k^m = \{ \varphi(s, a, q_{\pi_k}^m(s, a), 0, n_k(s, a) = 0 \}
\]
\[
M_k \geq \frac{A^2 C_{c_{\max}}^2 (\log A + \log 1/\delta)}{1 - \gamma^2} \] # Appendix E.5

Sample Trajectories
for \( m = 1, \ldots, M_k \) do
\[
\hat{q}_{\pi_k}^m(s, a) = \text{Truncated rollout of } q_{\pi_k}^m(s, a)
\]
\[
\hat{q}_{\lambda}^m(s, a, m) = \hat{q}_{\pi_k}^m(s, a, m) + \hat{q}_{\lambda}^m(s, a, m)
\]
\[
n_k(s, a, m) = n_k(s, a) + 1
\]
\[
S_k^m = S_k^m \cup \{ s \}
\]
end for

# Update Next Policy
for \( \forall s \in S_k^m \) do
\[
\hat{q}_{\pi_k}^m(s, a) \leftarrow A_q \hat{q}_{\pi_k}^m(s, a, m)/\sum_a n_k(s, a)
\]
end for

\[
\pi_{k+1}(s) = \text{PolicyUpdate}(\pi_k, s, t_k, \lambda)
\]
end for
end for

\( U(A) \) is the uniform distribution on the set \( A \). Then, by following the current policy \( \pi_k \), it estimates \( q_{\pi_k}^m(s, a, m) \) using a roll-out (possibly truncated in the infinite horizon case). We denote this estimate as \( \hat{q}_{\pi_k}^m(s, a, m) \) and observe it is (nearly) an unbiased estimator of \( q_{\pi_k}^m(s, a, m) \). 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,

\[
\pi_{k+1} \in \arg\min_{\pi \in \Delta_A} \left\{ \frac{1}{M} \sum_{m=1}^M \frac{1}{t_k(1-\gamma)} B_\omega(s_m, \pi, \pi_k) + \langle \hat{q}_{\pi_k}^m(s, \cdot) - \pi_k(\cdot | s_m) \rangle \right\}
\]

where the estimation of the gradient is \( \hat{q}_{\pi_k}^m(s, \cdot) \) for \( \forall s \in A \)

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 \( F_k \) be the \( \sigma \)-field containing all events until the end of the \( k \)-1 episode. Then, for any \( \pi, \pi_k \in \Delta_A \) and every sample \( m \),

\[
\langle \hat{q}_{\pi_k}^m(s, \cdot) - \pi_k(\cdot | s_m) \rangle = \mathbb{E} \left[ \langle \hat{q}_{\pi_k}^m(s, \cdot) - \pi_k(\cdot | s_m) \rangle | F_k \right]
\]

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 \( S \) or \( S_d \), and is possible due to the sample-based adaptive scaling of the Bregman distance. Let \( S_k^m \) be the set of visited states at the \( k \)-th episode, \( \pi(s, a) \) the number of times \( (s, a, m) = (s, a) \) at the \( k \)-th episode, and

\[
\hat{q}_{\pi_k}^m(s, a) = \frac{A}{\sum_{a} n(s, a)} \sum_{i=1}^{n(s, a)} \hat{q}_{\pi_k}^m(s, a, m_i)
\]

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

\[
\pi_{k+1}(s) = \arg\min_{\pi \in \Delta_A} \langle q_{\pi_k}^m(s, \cdot) + \lambda \nabla_\omega (s, \pi_k), \pi \rangle + B_\omega(s, \pi, \pi_k)
\]

As in previous sections, the euclidean and non-euclidean choices of \( \omega \) 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 \( M_k \geq O(\frac{A^2 C_{c_{\max}}^2 (\log A + \log 1/\delta)}{(1-\gamma^2)^2}) \) samples in each iteration, and \( \{ \mu_{k \text{ best}} \}_{k \geq 0} \) be the set of best achieved values, \( \mu_{k \text{ best}} := \arg\min_{\pi_k = \pi_0, \ldots, \pi_{k \text{ best}} \in \Pi} \mu_{\pi_k} - \mu^* \). Then, with probability greater than \( 1 - \delta \) for every \( \epsilon > 0 \) the following holds for all \( N \geq 1 \):

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

\[
\mu_{k \text{ best}} - \mu^* \leq O \left( \frac{C_{\omega,1} C_{\max} \epsilon}{(1-\gamma)^2 \sqrt{N}} + \frac{C_{\pi}^* \epsilon}{(1-\gamma)^2} \right).
\]

2. (Regularized) Let \( \lambda > 0 \), \( t_k = \frac{1}{\lambda (1-\gamma^2)} \), then

\[
\mu_{k \text{ best}} - \mu_{\pi_k}^* \leq O \left( \frac{C_{\omega,1}^2 C_{\omega,2} C_{\max,\lambda}^2}{\lambda (1-\gamma)^2 N} + \frac{C_{\pi}^* \epsilon}{(1-\gamma)^2} \right).
\]

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Sample Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRPO (this work)</td>
<td>$C_{\pi}^2\epsilon^2A^T_{\pi}C_{\pi}A_{\pi}(S+\log \mu)^2 \frac{1}{(1-\gamma)^{2c}}$</td>
</tr>
<tr>
<td>Regularized TRPO (this work)</td>
<td>$C_{\pi}^2\epsilon^2\lambda C^T_{\pi\pi}A_{\pi}(S+\log \mu)^2 \frac{1}{(1-\gamma)^{2c}}$</td>
</tr>
<tr>
<td>CPI (Kakade and Langford)</td>
<td>$A^T_{\pi}C_{\pi}(S+\log \mu)^2 \frac{1}{(1-\gamma)^{2c}}$</td>
</tr>
</tbody>
</table>

Similarly to Uniform TRPO, the convergence rates are $O(1/\sqrt{N})$ and $O(1/N)$ 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 $\epsilon$ 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 $\epsilon$, which can explain the empirical success of using regularization.

Remark 1 (Optimization Perspective). From an optimization perspective, CPI can be interpreted as an instance of the Policy 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 $\pi_{k+1} = \arg\min_{\pi} \Delta(\pi_k, \pi - \pi_k) + B_\omega(\pi_k, \pi_k)$, 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 $C_{\pi\pi}^\pi$. The results of Geist, Scherrer, and Pietquin (2019) in the approximate setting led to a worse concentrability coefficient, $C_{\mu\mu}^\pi$, which can be infinite even when $C_{\pi\pi}^\pi$ 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 $C_{\pi\pi}^\pi$ whereas the convergence rate of Exact TRPO (Section 6.1) does not depend on $S$ nor on $C_{\pi\pi}^\pi$ (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 $\pi_{k+1} \in P_{\Delta(\pi_k, \eta Nv_{\pi_k})}$, 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 $O(e^{-\delta})$ (as opposed to $O(1/\epsilon^2)$ 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 $O(1/\sqrt{N})$ 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


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


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.


