Adaptive Trust Region Policy Optimization: Convergence and Faster Rates of regularized MDPs

Lior Shani[*]
Technion, Israel

Yonathan Efroni[*]
Technion, Israel

Shie Mannor
Technion, 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.

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 one another, as in Mirror Descent (MD) [Beck and Teboulle, 2003]. Despite their popularity, much less is understood in terms of their convergence guarantees and they are considered heuristics [Schulman et al., 2015, Papini et al., 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 et al., 2016], and 'softens' the Bellman equation [Chow et al., 2018, Dai et al., 2018]. Recently, Ahmed et al. [2019] empirically observed that adding entropy regularization 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 exhibits faster convergence rates for regularized 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.

[*] equal contribution

Optimization Foundations for Reinforcement Learning Workshop at NeurIPS 2019, Vancouver, Canada.

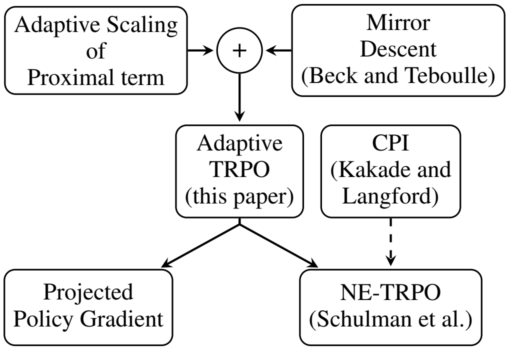
Adaptive Scaling of Proximal term
+
Mirror Descent (Beck and Teboulle)
Adaptive TRPO (this paper)
CPI (Kakade and Langford)
Projected Policy Gradient
NE-TRPO (Schulman et al.)

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

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

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.

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 et al., 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 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) (see Figure 1). In Section 3, we derive the linearized objective functions for (regularized) RL. In Section 4, using the linearized objective, we formulate Uniform TRPO, which assumes simultaneous access to the state space and that a model is given. In Section 5, we relax these assumptions and study Sample-Based TRPO. The main contributions of this paper are:

- We establish an $\tilde{O}(1/\sqrt{N})$ convergence rate to the global optimum for Sample-Based TRPO, which gives formal grounds for the NE-TRPO algorithm.
- We prove a faster rate of $\tilde{O}(1/N)$ for regularized MDPs using Sample-Based TRPO.
- The analysis of Sample-Based TRPO, unlike CPI, does not rely on improvement arguments. This allows to choose more aggressive learning rates which lead to better sample complexity.

2 Preliminaries and Notations

We consider an 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 $A = |\mathcal{A}|$. The transition kernel is $P \equiv P(s'|s,a)$, $C \equiv c(s,a)$ is a cost[†] function bounded in $[0, \mathrm{C}_{\text{max}}]$, and $\gamma \in (0,1)$ is a discount factor. Let $\pi : \mathcal{S} \rightarrow \Delta_{\mathcal{A}}^{\mathrm{S}}$ 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 $v^\pi(s) := \mathbb{E}^\pi[\sum_{t=0}^\infty \gamma^t r(s_t, \pi(s_t)) \mid s_0 = s]$. 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 π^* yielding the optimal value v^*, $v^* = \min_\pi (I - \gamma P^\pi)^{-1} c^\pi = (I - \gamma P^{\pi^*})^{-1} c^{\pi^*}$.

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_λ^π for $\lambda > 0$. Specifically, the cost of a policy π on

[†] We work with costs instead of rewards to comply with convex analysis. All results are valid for rewards.

a regularized MDP translates to $c^\pi_\lambda(s) := c^\pi(s) + \lambda\omega\left(s;\pi\right)$ where $\omega\left(s;\pi\right) := \omega(\pi(\cdot \mid s))$ 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. For $\lambda = 0$, the cost c^π is recovered. We consider two choices of ω: the **euclidean case** $\omega\left(s;\pi\right) = \frac{1}{2}\left\|\pi(\cdot \mid s)\right\|_2^2$ and **non-euclidean case** $\omega\left(s;\pi\right) = H(\pi(\cdot \mid s)) + \log A$. By this choice we have that $0 \le c^\pi_\lambda(s) \le \mathrm{C}_{\max,\lambda}$, where 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 $v^\pi_\lambda = (I - \gamma P^\pi)^{-1} c^\pi_\lambda$. The optimal value v^*_λ, optimal policy π^*_λ and Bellman operators of the regularized MDP are generalized as follows, $v^*_\lambda = \min_\pi (I - \gamma P^\pi)^{-1} c^\pi_\lambda = (I - \gamma P^{\pi^*_\lambda})^{-1} c^{\pi^*_\lambda}_\lambda$, and $T^\pi_\lambda v = c^\pi_\lambda + \gamma P^\pi v$. The q-function of a policy π for a regularized MDP is defined as $q^\pi_\lambda(s,a) = c(s,a) + \lambda\omega\left(s;\pi\right) + \gamma\sum_{s'} p^\pi(s' \mid s) v^\pi_\lambda(s')$.

A Bregman distance w.r.t. to ω is defined as $B_\omega\left(x,y\right) = \omega(x) - \omega(y) - \langle\nabla\omega(y), x - y\rangle$. Given two policies π_1, π_2, we denote their Bregman distance as $B_\omega\left(s;\pi_1,\pi_2\right) := B_\omega\left(\pi_1(\cdot \mid s), \pi_2(\cdot \mid s)\right)$ and $B_\omega\left(\pi_1,\pi_2\right) \in \mathbb{R}^S$ is the corresponding state-wise vector. The euclidean choice for ω leads to $B_\omega\left(s;\pi_1,\pi_2\right) = \frac{1}{2}\left\|\pi_1(\cdot \mid s) - \pi_2(\cdot \mid s)\right\|_2^2$, and the non-euclidean choice to $B_\omega\left(s;\pi_1,\pi_2\right) = d_{KL}(\pi_1(\cdot|s)||\pi_2(\cdot|s))$. Here, we always choose the regularization function ω to be associated with the proximity term, B_ω. This simplifies the analysis as c^π_λ is λ-strongly convex w.r.t. B_ω by definition. For more details about MD, Bregman distances and the underlying mechanism of regularization using strongly convex functions, refer to Appendix C.

The proposed algorithms always initialize π_0, the policy at the first iteration, to be a uniform distribution. 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 $\tilde{O}(\cdot)$.

3 Linear Approximation of a Policy's Value

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. In this work we consider the following objectives:

$$\min_{\pi \in \Delta^{\mathrm{S}}_{\mathcal{A}}} (I - \gamma P^\pi)^{-1} c^\pi_\lambda, \tag{1}$$

$$\min_{\pi \in \Delta^{\mathrm{S}}_{\mathcal{A}}} \mathbb{E}_{s\sim\mu}[v^\pi_\lambda(s)] = \min_{\pi \in \Delta^{\mathrm{S}}_{\mathcal{A}}} \mu v^\pi_\lambda. \tag{2}$$

Thus, we study $\langle\nabla v^\pi_\lambda, \pi' - \pi\rangle$ and $\langle\nabla\mu v^\pi_\lambda, \pi' - \pi\rangle$ in the following proposition:

Proposition 1 (Linear Approximation of a Policy's Value)**.** *Let* $\pi, \pi' \in \Delta^S_{\mathcal{A}}$. *Then,*

$$\langle\nabla_\pi v^\pi_\lambda, \pi' - \pi\rangle = (I - \gamma P^\pi)^{-1}\left(T^{\pi'}_\lambda v^\pi_\lambda - v^\pi_\lambda - \lambda B_\omega\left(\pi',\pi\right)\right), \tag{3}$$

$$\langle\nabla_\pi \mu v^\pi_\lambda, \pi' - \pi\rangle = \frac{1}{1-\gamma} d_{\mu,\pi}\left(T^{\pi'}_\lambda v^\pi_\lambda - v^\pi_\lambda - \lambda B_\omega\left(\pi',\pi\right)\right). \tag{4}$$

The proof is supplied in Appendix D. Importantly, the linear approximation is scaled by $(I - \gamma P^\pi)^{-1}$ or $\frac{1}{1-\gamma} d_{\mu,\pi}$, where $d_{\mu,\pi} = \mu(I - \gamma P^\pi)^{-1}$ is the discounted visitation frequencies. 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.

4 Uniform Trust Region Policy Optimization

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

Uniform TRPO repeats the following iterates

$$\pi_{k+1} \in \arg\min_{\pi \in \Delta^{\mathrm{S}}_{\mathcal{A}}} \langle\nabla v^{\pi_k}_\lambda, \pi - \pi_k\rangle + \frac{1}{t_k}(I - \gamma P^{\pi_k})^{-1} B_\omega\left(\pi,\pi_k\right). \tag{5}$$

This update rule resembles MD's updating-rule (see Appendix C). 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,

$$\pi_{k+1} \in \arg\min_{\pi\in\Delta_{\mathcal{A}}^{\mathrm{S}}} (I-\gamma P^{\pi_k})^{-1} \underbrace{\left(T^\pi_\lambda v^{\pi_k}_\lambda - v^{\pi_k}_\lambda + \big(1/t_k - \lambda\big) B_\omega(\pi,\pi_k)\right)}_{(*)}. \tag{6}$$

Because $(I-\gamma P^{\pi_k})^{-1} \geq 0$ component-wise, minimizing (6) 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 E.1). Meaning, the updated policy for all $s \in \mathcal{S}$ is

$$\pi_{k+1}(\cdot \mid s) \in \arg\min_{\pi\in\Delta_{\mathcal{A}}} t_k T^\pi_\lambda v^{\pi_k}_\lambda(s) + (1-\lambda t_k) B_\omega(s; \pi, \pi_k), \tag{7}$$

which is the update rule of Uniform TRPO (see Algorithm 1 in Appendix A). Importantly, this update rule is a direct consequence of choosing the adaptive scaling for the Bregman distance in (5), and without it, the trust region problem would involve optimizing over $\Delta_{\mathcal{A}}^{\mathrm{S}}$. The following theorem formalizes the convergence of the unregularized and regularized cases of Uniform TRPO, to the same rates of MD for the convex and strongly convex cases, respectively:

Theorem 2 (Convergence of Uniform TRPO in simplified form)**.** *Let $\{\pi_N\}_{N\geq 0}$ be the sequence generated by Uniform TRPO. Then, for a proper choice of step size, for all $N \geq 1$, The unregularized and regularized versions of Uniform TRPO converge with rates of $\|v^{\pi_N} - v^*\|_\infty \leq \tilde{O}(1/\sqrt{N})$ and $\|v^{\pi_N}_\lambda - v^*_\lambda\|_\infty \leq \tilde{O}(1/N)$, respectively. The full version is given in Appendix F, Theorem 11.*

Theorem 2 establishes that regularization allows a faster rate of $\tilde{O}(1/N)$. Note that using such regularization leads to a 'biased' solution: Generally $\left\|v^{\pi^*_\lambda} - v^*\right\|_\infty > 0$, i.e., the optimal policy of the regularized MDP evaluated on the original MDP is not necessarily optimal. Yet, regularizing the problem can make it easier to solve, in the sense Uniform TRPO converges faster.

5 Sample-Based TRPO

In the previous section we analyzed Uniform TRPO, which uniformly minimizes the vector v^π. However, in case of a large state space it can be computationally infeasible to apply Uniform TRPO as it requires access to the entire state space and to a model of the environment, which is usually absent. In this work, we construct a sample-based algorithm which minimizes a *scalar objective* of the form $\min_{\pi\in\Delta_{\mathcal{A}}^{\mathrm{S}}} \mu v^\pi_\lambda$ (2), which is widely used in the RL literature [Sutton et al., 2000, Kakade and Langford, 2002, Schulman et al., 2015], instead of a vector in (1). Using this objective, one wishes to find a policy π which minimizes the expectation of $v^\pi_\lambda(s)$ under a probability measure μ.

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 et al. [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, we make the following assumption, similarly to CPI:

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]. 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.

In this section, we derive *Sample-Based TRPO*, and establish high-probability convergence guarantees in a batch setting. Sample-Based TRPO assumes access to a ν-restart model. Meaning, it can only access sampled trajectories and restarts according to the distribution ν. Importantly, on expectation, its updating rule is **exactly the 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.

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 $U(\mathcal{A})$ 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). 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 F.2. Based on this data, Sample-Based TRPO updates the policy at the end of the k-th episode, by the following proximal problem,

$$\pi_{k+1} \in \arg\min_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} \frac{1}{M} \sum\nolimits_{m=1}^{M} \langle \hat{\nabla} \nu v_\lambda^{\pi_k}[m], \pi(\cdot \,|\, s_m) - \pi_k(\cdot \,|\, s_m) \rangle + \frac{1}{t_k(1-\gamma)} B_\omega\left(s_m; \pi, \pi_k\right), \quad (8)$$

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

The following proposition shows the expectation of the proximal objective of Sample-Based TRPO (8) is the proximal objective of Uniform TRPO (5), scaled by the measure ν. This fact motivates us to study this algorithm, anticipating it inherits the convergence guarantees of Uniform TRPO,

Proposition 3 (Uniform 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,*

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

Like Uniform 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 TRPO which requires access to all states in $\mathcal{S}$, and is possible due to the *sample-based adaptive scaling* of the Bregman distance: due to the sampling process, the estimated gradient and the Bregman distance at a state s are scaled by the empirical frequency of visits to s, which is on expectation d_{ν,π_k}, the scaling term observed in (4). 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). For any non-visited state-action pair, $\hat{q}_\lambda^{\pi_k}(s,a) = 0$. Thus, Sample-Based TRPO updates the policy for all $s \in \mathcal{S}_M^k$ by a simplified update rule:

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

The full algorithm is found in Appendix A (Alg. 2). Instantiating the *PolicyUpdate* procedure is equivalent to choosing ω and the induced Bregman distance B_ω: In the euclidean case, $\omega(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, we get the PPG variant (Alg. 3) and in the non-euclidean case, $\omega(\cdot) = H(\cdot)$, we get NE-TRPO (Alg. 4). This comes in complete analogy to the fact that Projected Gradient Descent and Exponentiated Gradient Descent are instances of MD with similar ω choices. Generalizing the proof technique of Uniform TRPO and using standard concentration inequalities, we derive a high-probability convergence guarantee for Sample-Based TRPO (for the proof see Appendix F):

Theorem 4 (Convergence Rate: Sample-Based TRPO). *Let $\{\pi_k\}_{k\geq 0}$ be the sequence generated by Sample-Based TRPO, using $M_k \geq O\Big(\frac{A^2 \, \mathrm{C}_{\mathit{max},\lambda}^2 (S \log 2A + \log 1/\delta)}{(1-\gamma)^2 \epsilon^2}\Big)$ samples in each iteration, and $\{\mu v_{\textbf{\textit{best}},\lambda}^k\}_{k\geq 0}$ be the sequence of best achieved values, $\mu v_{\textbf{\textit{best}},\lambda}^N :=$ $\arg\min_{k=0,\ldots,N} \mu v_\lambda^{\pi_k} - \mu v_\lambda^*$. Then, with probability greater than $1-\delta$ for every $\epsilon > 0$ the following holds for all $N \geq 1$:*

1. (Unregularized) $\mu v_{\textbf{\textit{best}}}^N - \mu v^* \leq O\Big(\frac{C_{\omega,1} \, \mathrm{C}_{\mathit{max}}}{(1-\gamma)^2 \sqrt{N}} + \frac{C^{\pi^*} \epsilon}{(1-\gamma)^2}\Big)$*, for stepsize* $t_k = \frac{(1-\gamma)}{C_{\omega,1} \, \mathrm{C}_{\mathit{max}} \sqrt{k+1}}$.
2. (Regularized) $\mu v_{\textbf{\textit{best}},\lambda}^N - \mu v_\lambda^* \leq O\Big(\frac{C_{\omega,1}^2 C_{\omega,2} \, {\mathrm{C}_{\mathit{max},\lambda}}^2}{\lambda(1-\gamma)^3 N} + \frac{C^{\pi^*} \epsilon}{(1-\gamma)^2}\Big)$*, for stepsize* $t_k = \frac{1}{\lambda(k+2)}$.

In the euclidean case $C_{\omega,1} = \sqrt{A}, C_{\omega,2} = 1$ and in the non-euclidean case $C_{\omega,1} = 1, C_{\omega,2} = A^2$.

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* required for a $\frac{C^{\pi^*}\epsilon}{(1-\gamma)^2}$ error, the same as the error of CPI, is given in Table 1. 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)**.** *CPI can be interpreted as a sample-based Conditional Gradient Descent 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 either 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 Needed?)**.** *Unlike CPI, Sample-Based TRPO does not rely on improvement arguments. 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.*

6 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 [Sutton et al., 2000, Bhatnagar et al., 2009, Pirotta et al., 2013, Dai et al., 2018, Papini et al., 2019]. There, convergence to a local optimum was established under different conditions and several aspects 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 et al. [2017] the authors analyzed entropy regularized MDPs from a linear programming perspective for average-reward MDPs. Yet, convergence rates were not provided.

Geist et al. [2019] addressed regularized MDPs using MD-like updates in an approximate PI scheme. The authors focused 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 Uniform TRPO (7) with a fixed learning rate, $t_k = 1$. In this paper, we argued in favor of viewing this update rule as an instance of the more general update rule (5), i.e., MD with an adaptive proximity term. This view allowed us to analyze adaptive Sample-Based TRPO, which does not require uniform access to the state space. Also, we proved Sample-Based TRPO inherits the same asymptotic performance guarantees of CPI: the quality of the policy Sample-Based TRPO outputs depends on C^{π^*}. The results of Geist et al. [2019] in the approximate setting led to worse concentrability coefficients, C_q^i, which can be infinite even when C^{π^*} is finite [Scherrer, 2014] and depends on the worst case of all policies.

Recently, Agarwal et al. [2019] analyzed a variant of PPG under the assumption of exact gradients and uniform access to the state space. Their proven convergence rate depends on both S and C^{π^*} whereas Sample-Based TRPO with exact gradients ($\epsilon = 0$) does not depend on S nor on C^{π^*}. Also, the authors did not establish faster rates for regularized MDPs. Importantly, their PPG algorithm is *different* than the one we study, which can explain the discrepancy between our results: It uses the update $\pi_{k+1} \in P_{\Delta_{\mathcal{A}}^{\mathcal{S}}}(\pi_k - \eta\nabla\mu v^{\pi_k})$, whereas, the PPG studied in this work applies a different update rule using *adaptive scaling* of the Bregman distance (see (7), (8), in the exact euclidean case).

7 Conclusions

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 does, 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.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, pages 151–160, 2019.

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

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

Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Yinlam Chow, Ofir Nachum, and Mohammad Ghavamzadeh. Path consistency learning in tsallis entropy regularized mdps. In *International Conference on Machine Learning*, pages 978–987, 2018.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1125–1134, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

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*, pages 202–211. AUAI Press, 2016.

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

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

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pages 267–274, 2002.

Sham Machandranath Kakade et al. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

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.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Trust-pcl: An off-policy trust region method for continuous control. *arXiv preprint arXiv:1707.01891*, 2017.

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

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

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

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

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

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Adaptive step-size for policy gradient methods. In *Advances in Neural Information Processing Systems*, pages 1394–1402, 2013.

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

Bruno Scherrer and Matthieu Geist. 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*, pages 35–50. Springer, 2014.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

List of Appendices

A Algorithms

By instantiating the *PolicyUpdate* procedure with Algorithms 3 and 4 we get the PPG and NE-TRPO instances of Uniform and Sample-Based TRPO. The algorithms are discussed and analyzed in the following sections: Algorithm 1 is described Section 4 and analyzed Appendix E; Algorithm 2 is described in Section 5 and analyzed in Appendix F; Algorithms 3 and 4 are analyzed in Appendix E.2.

Algorithm 1 Uniform TRPO

```
Initialize: $t_k$, $\gamma$, $\lambda$, $\pi_0$ is the uniform policy.
  for $k = 0, 1, ...$ do
    $v^{\pi_k} = (I - \gamma P^{\pi})^{-1} c^{\pi}_{\lambda}$
    for $\forall s \in \mathcal{S}$ do
      for $\forall a \in \mathcal{A}$ do
        $q^{\pi_k}_{\lambda}(s,a) \leftarrow c^{\pi}_{\lambda}(s,a) + \gamma \sum_{s'} p(s'|s,a) v^{\pi_k}_{\lambda}(s')$
      end for
      $\pi_{k+1}(\cdot|s) \leftarrow$ PolicyUpdate($\pi_k(\cdot|s), q^{\pi_k}_{\lambda}(s,\cdot), t_k, \lambda$)
    end for
  end for
```

Algorithm 2 Sample-Based TRPO

```
Initialize: $t_k, \gamma, \lambda, \pi_0, \epsilon, \delta > 0$
  for $k = 0, 1, ...$ do
    $\mathcal{S}^k_M = \{\}, \forall s, a, \hat{q}^{\pi_k}_{\lambda}(s,a) = 0, n_k(s,a) = 0$
    $M_k \geq \tilde{O}(\frac{A^2 \, \mathrm{C}^2_{\text{max},\lambda}(S \log 2A + \log 1/\delta)}{(1-\gamma)^2 \epsilon^2})$ # See Appendix F.5
    # Sample Trajectories
    for $m = 1, .., M_k$ do
      Sample $s_m \sim d_{\nu,\pi_k}(\cdot)$
      Sample $a_m \sim U(\mathcal{A})$
      $\hat{q}^{\pi_k}_{\lambda}(s_m, a_m, m)$=Truncated Rollout of $q^{\pi_k}_{\lambda}(s_m, a_m)$
      $\hat{q}^{\pi_k}_{\lambda}(s_m, a_m) \leftarrow \hat{q}^{\pi_k}_{\lambda}(s_m, a_m) + \hat{q}^{\pi_k}_{\lambda}(s_m, a_m, m)$
      $n_k(s_m, a_m) \leftarrow n_k(s_m, a_m) + 1$
      $\mathcal{S}^k_M = \mathcal{S}^k_M \cup \{s_m\}$
    end for
    # Update Next Policy
    for $\forall s \in \mathcal{S}^k_M$ do
      for $\forall a \in \mathcal{A}$ do
        $\hat{q}^{\pi_k}_{\lambda}(s,a) \leftarrow \frac{A \hat{q}^{\pi_k}_{\lambda}(s,a)}{\sum_a n_k(s,a)}$
      end for
      $\pi_{k+1}(\cdot|s) \leftarrow$ PolicyUpdate($\pi_k(\cdot|s), \hat{q}^{\pi_k}_{\lambda}(s,\cdot), t_k, \lambda$)
    end for
  end for
```

Algorithm 3 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 4 PolicyUpdate: NE-TRPO

```
  input: $\pi(\cdot \mid s), q(s,\cdot), t_k, \lambda$
  for $a \in \mathcal{A}$ do
    $\pi(a|s) \leftarrow \frac{\pi(a|s) e^{-t_k (q(s,a) + \lambda \log \pi(a|s))}}{\sum\limits_{a' \in \mathcal{A}} \pi(a'|s) e^{-t_k \left(q(s,a') + \lambda \log \pi(a'|s)\right)}}$
  end for
return $\pi(\cdot \mid s)$
```

B Detailed Notations

In this section we elaborate more than the notation section (Section 2) in main body of the paper.

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 $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}^{\mathrm{S}}$ 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

$$v^* = \min_\pi (I - \gamma P^\pi)^{-1} c^\pi = (I - \gamma P^{\pi^*})^{-1} c^{\pi^*}. \tag{9}$$

This goal can be achieved using the classical operators:

$$\forall v, \pi,\ T^\pi v = c^\pi + \gamma P^\pi v, \text{and } \forall v,\ Tv = \min_\pi T^\pi v, \tag{10}$$

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,

$$\begin{aligned} v^*_\lambda &= \min_\pi (I - \gamma P^\pi)^{-1} c^\pi_\lambda = (I - \gamma P^{\pi^*_\lambda})^{-1} c^{\pi^*_\lambda}_\lambda, \\ \forall v, \pi,\ T^\pi_\lambda v &= c^\pi_\lambda + \gamma P^\pi v, \text{and } \forall v,\ T_\lambda v = \min_\pi T^\pi_\lambda v. \end{aligned} \tag{11}$$

As Bellman operators for MDPs, both T^π_λ, T are γ-contractions with fixed points $v^\pi_\lambda, v^*_\lambda$ [Geist et al., 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, both criteria (9), (11), 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 and to a model of the environment, which is usually absent. In this work, we relax these assumptions, and construct a sample-based algorithm which minimizes a *scalar objective* instead of a vector in (9), (11), of the following form,

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

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 $B_\omega(s; \pi_1, \pi_2) = d_{KL}(\pi_1(\cdot|s)||\pi_2(\cdot|s))$.

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

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.

The proposed algorithms always initialize π_0, the policy at the first iteration, to be a uniform distribution. 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 $\langle x, y\rangle(s) := \langle x(s,\cdot,\cdot), y\rangle = \sum_{s',a} x(s,s',a)y(s,a)$, with some abuse of notation.

C Mirror Descent in Convex Optimization

This section serves as a background section about the Mirror Descent algorithm. We recommend reading this short brief in order to understand the basis of the regularization mechanism used in this paper for regularized MDPs, and its relation to the Convex Optimization literature.

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

$$x^* \in \arg\min_{x\in C} f(x), \tag{12}$$

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

$$x_{k+1} \in \arg\min_{x\in C} \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{t_k} B_\omega\left(x, x_k\right), \tag{13}$$

where $B_\omega\left(x, x_k\right) := \omega(x) - \omega(x_k) - \langle \nabla\omega(x_k), x - x_k\rangle$ is the Bregman distance associated with a strongly convex ω and t_k is a stepsize (see Appendix C.1). In the general convex case, MD converges to the optimal solution of (12) with a rate of $\tilde{O}(1/\sqrt{N})$, where N is the number of MD iterations [Beck and Teboulle, 2003, Juditsky et al., 2011], i.e., $f(x_k) - f^* \le \tilde{O}(1/\sqrt{k})$, 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 $f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \lambda B_\omega\left(y, x\right)$. For such f, improved convergence rate of $\tilde{O}(1/N)$ can be obtained [Juditsky et al., 2011, Nedic and Lee, 2014]. Thus, instead of using MD to optimize a convex f, one can consider the following regularized problem,

$$x^* = \arg\min_{x\in C} f(x) + \lambda g(x), \tag{14}$$

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,

$$x_{k+1} = \arg\min_{x\in C} \langle \nabla F_\lambda(x_k), x - x_k\rangle + \frac{1}{t_k} B_\omega\left(x, x_k\right). \tag{15}$$

Solving (15) allows faster convergence, at the expense of adding a bias to the solution of (12). 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 $\omega(\cdot) = \frac{1}{2}\left\|\cdot\right\|_2^2$ and the resulting Bregman distance is the squared euclidean norm $B_\omega\left(x, y\right) = \frac{1}{2}\left\|x - y\right\|_2^2$. In this case, (13) 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, $x_{k+1} = P_c\left(x_k - t_k\nabla f(x_k)\right)$, where $P_c(x) = \min_{y\in C} \frac{1}{2}\left\|x - y\right\|_2^2$ 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, $B_\omega\left(x, y\right) = d_{KL}(x||y)$. 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, (13) 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.

C.1 Assumptions of Mirror Descent

Assumption 2 (properties of Bregman distance)**.**

(A) ω *is proper closed and convex.*

(B) ω *is differentiable over dom*$(\partial\omega)$*.*

(C) $C \subseteq$ *dom*(ω)

(D) $\omega + \delta_C$ *is* σ*-strongly convex (*$\sigma > 0$*)*

Assumption 2 is the main assumption regarding the underlying Bregman distance used in Mirror Descent. In our analysis, we have two common choice of ω: a) the negative entropy function, denoted as $H(\cdot)$, for which the corresponding Bregman distance is $B_\omega(\cdot,\cdot) = d_{KL}(\cdot||\cdot)$. b) the euclidean norm $\omega(\cdot) = \frac{1}{2}\|\cdot\|^2$, for which the resulting Bregman distance is the euclidean distance. The convex optimization domain C is in our case $\Delta_{\mathcal{A}}^{\mathrm{S}}$, the state-wise unit simplex over the space of actions. For both choices, the assumption holds. Finally, $\delta_C(x)$ is an extended real valued function which describes the optimization domain C. It is defined as follows: For $x \in C$, $\delta_C(x) = 0$. For $x \notin C$, $\delta_C(x) = \infty$. For more details, see [Beck, 2017].

We go on to define the second assumption regarding the optimization problem:

Assumption 3.

(A) $f : \mathbb{E} \to (-\infty, \infty]$ *is proper closed.*

(B) $C \subseteq \mathbb{E}$ *is nonempty closed and convex.*

(C) $C \subseteq$ *int*$($*dom*$(f))$*.*

(D) The optimal set of (P) is nonempty.

D Policy Gradient, and Directional Derivatives for Regularized MDPs

In this section we re-derive the Policy Gradient Theorem [Sutton et al., 2000] for regularized MDPs when tabular representation is used. Meaning, we explicitly calculate the derivative $\nabla_\pi v_\lambda^\pi(s)$. Based on this result, we derive the directional derivative, or the linear approximation of the objective functions, $\langle \nabla_\pi v_\lambda^\pi(s), \pi - \pi' \rangle$, $\langle \nabla_\pi \mu v_\lambda^\pi(s), \pi - \pi' \rangle$.

D.1 Extended Value Functions

To formally study $\nabla_\pi v_\lambda^\pi(s)$ we need to define value functions v^π when π is outside of the simplex $\Delta_{\mathcal{A}}^{\mathrm{S}}$, since when $\pi(a \mid s)$ changes infinitesimally, $\pi(\cdot \mid s)$ does not remain a valid probability distribution. To this end, we study *extended value functions* denoted by $v(y) \in \mathbb{R}^S$ for $y \in \mathbb{R}^{S\times A}$, and denote $v_s(y)$ as the component of $v(y)$ which corresponds to the state s. Furthermore, we define the following cost and dynamics,

$$c_{\lambda,s}^y := \sum_{a'} y(a' \mid s)(c(s,a) + \lambda\omega_s(y)),$$
$$p_{s,s'}^y := \sum_{a'} y(a' \mid s)p(s' \mid s, a'),$$

where $\omega_s(y) := \omega(y(\cdot \mid s))$ for $\omega : \mathbb{R}^A \to \mathbb{R}$, $p^y \in \mathbb{R}^{S\times S}$ and $c_\lambda^y \in \mathbb{R}^S$.

Definition 1 (Extended value and q functions.)**.** *An extended value function is a mapping* $v : \mathbb{R}^{S\times A} \to \mathbb{R}^S$*, such that for* $y \in \mathbb{R}^{S\times A}$

$$v(y) := \sum_{t=0}^{\infty} \gamma^t (p^y)^t c_\lambda^y, \tag{16}$$

Similarly, an extended q-function is a mapping $q : \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$, such that its s, a element is given by

$$q_{s,a}(y) := c(s,a) + \lambda\omega_s(y) + \gamma \sum_{s'} p(s' \mid s,a) v^y_{s'}, \tag{17}$$

When $y \in \Delta^{\mathrm{S}}_{\mathcal{A}}$ is a policy, π, we denote $v(\pi) := v^\pi_\lambda \in \mathbb{R}^S, q(\pi) = q^\pi_\lambda \in \mathbb{R}^{S\times A}$.

Note that in this section we use different notations than the rest of the paper, in order to generalize the discussion and keep it out of the regular RL conventions.

The following proposition establishes that $v(y)$ the fixed point of a corresponding Bellman operator when y is close to the simplex component-wise.

Lemma 5. *Let $y \in \{y' \in \mathbb{R}^{S\times A} : \forall s, \sum_a |y'(a \mid s)| < \frac{1}{\gamma}\}$. Define the operator $T^y : \mathbb{R}^S \to \mathbb{R}^S$, such that for any $v \in \mathbb{R}^S$,*

$$(T^y v)_s := c^y_{\lambda,s} + \gamma \sum_{s'} p^y_{s,s'} v_{s'}.$$

Then,

1. *T^y is a contraction operator in the max norm.*
2. *Its fixed-point is $v(y)$ and satisfies $v_s(y) = (T^y v(y))_s$.*

Proof. We start by proving the first claim. Unlike in classical results on MDPs, y is not a policy. However, since it is not 'too far' from being a policy we get the usual contraction property by standard proof techniques.

Let $v', v \in \mathbb{R}^S$, and assume $(T^y v')_s \geq (T^y v)_s$.

$$\begin{aligned}
(T^y v')_s - (T^y v)_s &= \gamma \sum_a y(a \mid s) \sum_{s'} p(s' \mid s,a)(v'_{s'} - v_{s'}) \\
&\leq \gamma \sum_a y(a \mid s) \sum_{s'} p(s' \mid s,a) \left\| v'_{s'} - v_{s'} \right\|_\infty \\
&= \gamma \left\| v'_{s'} - v_{s'} \right\|_\infty \sum_a y(a \mid s) \\
&\leq \gamma \left\| v'_{s'} - v_{s'} \right\|_\infty \sum_a |y(a \mid s)| \\
&< \left\| v'_{s'} - v_{s'} \right\|_\infty .
\end{aligned}$$

In the fourth relation we used the assumption that $\gamma \sum_a |y(a \mid s)| < 1$. Repeating the same proof for the other case where $(T^y v')_s < (T^y v)_s$, concludes the proof of the first claim.

To prove the second claim, we use the definition of $v(y)$.

$$\begin{aligned}
v(y) &:= \sum_{t=0}^\infty \gamma^t (p^y)^t c^y_\lambda \\
&= c^y_\lambda + \sum_{t=1}^\infty \gamma^t (p^y)^t c^y_\lambda \\
&= c^y_\lambda + \gamma p^y \left(\sum_{t=0}^\infty \gamma^t (p^y)^t c^y_\lambda \right) \\
&= c^y_\lambda + \gamma p^y v(y).
\end{aligned}$$

In the third relation we used the distributive property of matrix multiplication and in the forth relation we used the definition of $v(y)$. Thus, $v(y) = T^y v(y)$, i.e., $v(y)$ is the fixed point of the operator T^y. □

D.2 Policy Gradient Theorem for Regularized MDPs

We now derive the Policy Gradient Theorem for regularized MDPs for tabular policy representation. Specifically, we use the notion of an extended value function and an extended q-functions defined in the previous section.

Lemma 6. *Let $y \in \{y' \in \mathbb{R}^{S\times A} : \forall s, \sum_a |y'(a \mid s)| < \frac{1}{\gamma}\}$. Then,*

$$v_s(y) = \sum_{a'} y(a' \mid s) q_{s,a'}(y)$$

Proof. Using (17), we get

$$\begin{aligned}\sum_{a'} y(a' \mid s) q_{s,a'}(y) &= \sum_{a'} y(a' \mid s)(c(s,a) + \lambda\omega_s(y)) + \gamma \sum_{s'} p(s' \mid s, a') v_{s'}(y) \\ &= c^y_{\lambda,s} + \gamma \sum_{s'} p(s' \mid s, a') v_{s'}(y) = v_s(y),\end{aligned}$$

where the last equality is by the fixed-point property of Lemma 5. □

We now derive the Policy Gradient Theorem for extended (regularized) value functions.

Theorem 7 (Policy Gradient for Extended Regularized Value Functions)**.** *Let $y \in \{y : \forall s, \sum_a |y(a \mid s)| < \frac{1}{\gamma}\}$. Furthermore, consider a fixed s, a and $\bar{s}$. Then,*

$$\partial_{y_{\bar{s},\bar{a}}} v_s(y) = \sum_{t=0}^{\infty} \gamma^t p^y_t(s_t \mid s) \delta_{\bar{s},s_t} \left(q_{s,\bar{a}}(y) + \lambda \partial_{y_{s,\bar{a}}} \omega_s(y) \left(\sum_{a'} y(a' \mid s) \right) \right),$$

where $p^y(s_t \mid s) = \sum_{s_1,..,s_t} p^y(s_t \mid s_{t-1}) \cdots p^y(s_1 \mid s)$, and $p^y_t(s_0 \mid s) = 1$.

Proof. Following similar derivation to the original Policy Gradient Theorem [Sutton et al., 2000], for every s,

$$\begin{aligned}&\partial_{y_{\bar{s},\bar{a}}} v_s(y) \\ &= \sum_{a'} (\partial_{y_{\bar{s},\bar{a}}} y(a' \mid s)) q_{s,a'}(y) + y(a' \mid s) \partial_{y_{\bar{s},\bar{a}}} q_{s,a'}(y) \\ &= \sum_{a'} \delta_{s,\bar{s}} \delta_{a',\bar{a}} q_{s,a'}(y) + y(a' \mid s) \partial_{y_{\bar{s},\bar{a}}} q_{s,a'}(y).\end{aligned}$$

We now explicitly write the last term,

$$\begin{aligned}&\partial_{y_{\bar{s},\bar{a}}} q_{s,a'}(y) \\ &= \partial_{y_{\bar{s},\bar{a}}} \left(c(s,a') + \lambda\omega_s(y) + \gamma \sum_{s'} p(s' \mid s, a') v_{s'}(y) \right) \\ &= \lambda \delta_{s,\bar{s}} \partial_{y_{s,\bar{a}}} \omega_s(y) + \gamma \sum_{s'} p(s' \mid s, a') \partial_{y_{\bar{s},\bar{a}}} v_{s'}(y).\end{aligned}$$

Plugging this back yields,

$$\begin{aligned}&\partial_{y_{\bar{s},\bar{a}}} v_s(y) \\ &= \sum_{a'} \delta_{s,\bar{s}} \delta_{a',\bar{a}} q_{s,a'}(y) + \lambda y(a' \mid s) \delta_{s,\bar{s}} \partial_{y_{s,\bar{a}}} \omega_s(y) \\ &\quad + \gamma \sum_{s'} \sum_{a'} y(a' \mid s) p(s' \mid s, a') \partial_{y_{\bar{s},\bar{a}}} v_{s'}(y) \\ &= \sum_{a'} \delta_{s,\bar{s}} \delta_{a',\bar{a}} q_{s,a'}(y) + \lambda y(a' \mid s) \delta_{s,\bar{s}} \partial_{y_{s,\bar{a}}} \omega_s(y) + \gamma \sum_{s'} p^y(s' \mid s) \partial_{y_{\bar{s},\bar{a}}} v_{s'}(y).\end{aligned}$$

Iteratively applying this relation yields

$$\partial_{y_{\bar{s},\bar{a}}} v_s(y) = \sum_{t=0}^{\infty} \gamma^t p_t^y(s_t \mid s) \delta_{\bar{s},s_t} \left(q_{s,\bar{a}}(y) + \lambda \partial_{y_{s,\bar{a}}} \omega_s(y) \left(\sum_{a'} y(a' \mid s) \right) \right),$$

where,

$$p^y(s_t \mid s) = \sum_{s_1,..,s_t} p^y(s_t \mid s_{t-1}) \cdots p^y(s_1 \mid s),$$

and $p_t^y(s_0 \mid s) = 1$. □

Returning to the specific notation for RL, defined in Section B, by setting $y = \pi$, i.e., when y is a policy, we get the Policy Gradient Theorem for regularized MDPs, since for all s, $\sum_{a'} \pi(a' \mid s) = 1$.

Corollary 8 (Policy Gradient for Regularized MDPs)**.** *Let* $\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}$. *Then,* $\nabla_\pi v^\pi \in \mathbb{R}^{S \times S \times A}$ *and*

$$\nabla_\pi v^\pi(s, \bar{s}, \bar{a}) := \nabla_{\pi(\bar{a}|\bar{s})} v_\lambda^\pi(s) = \sum_{t=0}^{\infty} \gamma^t p^\pi(s_t = \bar{s} \mid s) \left(\lambda \partial_{\pi(\bar{a}|\bar{s})} \omega^\pi(\bar{s}) + q_\lambda^\pi(\bar{s}, \bar{a}) \right).$$

D.3 The Linear Approximation of the Policy's Value and The Directional Derivative for Regularized MDPs

In this section, we derive the directional derivative in policy space for regularized MDPs with tabular policy representation.

The linear approximation of the value function of the policy π', around the policy π, is given by

$$v_\lambda^{\pi'} \approx v_\lambda^\pi + \langle \nabla_\pi v_\lambda^\pi, \pi' - \pi \rangle$$

In the MD framework, we take the $\arg\min$ w.r.t. to this linear approximation. Note that the minimizer is independent on the zeroth term, v_λ^π, and thus the optimization problem depends only on the directional derivative, $\langle \nabla_\pi v_\lambda^\pi, \pi' - \pi \rangle$. To keep track with the MD formulation, we chose to refer to Proposition 1 as the 'linear approximation of a policy's value', even though it is actually the directional derivative.

Proposition 1 (Linear Approximation of a Policy's Value)**.** *Let* $\pi, \pi' \in \Delta_A^S$. *Then,*

$$\langle \nabla_\pi v_\lambda^\pi, \pi' - \pi \rangle = (I - \gamma P^\pi)^{-1} \left(T_\lambda^{\pi'} v_\lambda^\pi - v_\lambda^\pi - \lambda B_\omega(\pi', \pi) \right), \tag{3}$$

$$\langle \nabla_\pi \mu v_\lambda^\pi, \pi' - \pi \rangle = \frac{1}{1-\gamma} d_{\mu,\pi} \left(T_\lambda^{\pi'} v_\lambda^\pi - v_\lambda^\pi - \lambda B_\omega(\pi', \pi) \right). \tag{4}$$

See that (4) is a vector in $\mathbb{R}^S$, whereas (3) is a scalar.

Proof. We start by proving the first claim. Consider the inner product, $\left\langle \nabla_{\pi(\cdot|\bar{s})} v^\pi(s), \pi'(\cdot \mid \bar{s}) - \pi(\cdot \mid \bar{s}) \right\rangle$. By the linearity of the inner product and using Corollary 8 we get,

$$\begin{aligned} &\left\langle \nabla_{\pi(\cdot|\bar{s})} v^\pi(s), \pi'(\cdot \mid \bar{s}) - \pi(\cdot \mid \bar{s}) \right\rangle \\ &= \sum_{t=0}^{\infty} \gamma^t p^\pi(s_t = \bar{s} \mid s) \left\langle \lambda \nabla_{\pi(\cdot|\bar{s})} \omega(\bar{s}; \pi) + q_\lambda^\pi(\bar{s}, \cdot), \pi'(\cdot \mid \bar{s}) - \pi(\cdot \mid \bar{s}) \right\rangle \\ &= \sum_{t=0}^{\infty} \gamma^t p^\pi(s_t = \bar{s} \mid s) \left(\lambda \left\langle \nabla_{\pi(\cdot|\bar{s})} \omega(\bar{s}; \pi), \pi'(\cdot \mid \bar{s}) - \pi(\cdot \mid \bar{s}) \right\rangle + \langle q_\lambda^\pi(\bar{s}, \cdot), \pi'(\cdot \mid \bar{s}) - \pi(\cdot \mid \bar{s}) \rangle \right), \end{aligned} \tag{18}$$

The following relations hold.

$$
\begin{aligned}
&\langle q_\lambda^\pi(\bar{s},\cdot), \pi'(\cdot \mid \bar{s}) - \pi(\cdot \mid \bar{s})\rangle \\
&= \langle q_\lambda^\pi(\bar{s},\cdot), \pi'(\cdot \mid \bar{s})\rangle - \langle q_\lambda^\pi(\bar{s},\cdot), \pi(\cdot \mid \bar{s})\rangle \\
&= \sum_{a'} \pi'(a' \mid \bar{s}) \left(c(\bar{s},a) + \lambda\omega\left(\bar{s};\pi\right) + \gamma \sum_{s'} P(s' \mid \bar{s},a) v_\lambda^\pi(s') \right) \\
&\quad - \sum_{a'} \pi(a' \mid \bar{s}) \left(c(\bar{s},a) + \lambda\omega\left(\bar{s};\pi\right) + \gamma \sum_{s'} P(s' \mid \bar{s},a) v_\lambda^\pi(s') \right) \\
&= \sum_{a'} \pi'(a' \mid \bar{s}) \left(c(\bar{s},a) + \lambda\omega\left(\bar{s};\pi\right) + \gamma \sum_{s'} P(s' \mid \bar{s},a) v_\lambda^\pi(s') \right) - v_\lambda^\pi(\bar{s}) \\
&= \sum_{a'} \pi'(a' \mid \bar{s}) \left(c(\bar{s},a) + \lambda\omega\left(\bar{s};\pi'\right) - \lambda\omega\left(\bar{s};\pi'\right) + \lambda\omega\left(\bar{s};\pi\right) + \gamma \sum_{s'} P(s' \mid \bar{s},a) v_\lambda^\pi(s') \right) - v_\lambda^\pi(\bar{s}) \\
&= \sum_{a'} \pi'(a' \mid \bar{s}) \left(c(\bar{s},a) + \lambda\omega\left(\bar{s};\pi'\right) + \gamma \sum_{s'} P(s' \mid \bar{s},a) v_\lambda^\pi(s') \right) - v_\lambda^\pi(\bar{s}) + \lambda(\omega\left(\bar{s};\pi\right) - \omega\left(\bar{s};\pi'\right)) \\
&= c_\lambda^{\pi'}(\bar{s}) + \gamma \sum_{s'} P^{\pi'}(s' \mid \bar{s}) v_\lambda^\pi(s') - v_\lambda^\pi(\bar{s}) + \lambda(\omega\left(\bar{s};\pi\right) - \omega\left(\bar{s};\pi'\right)) \\
&= (T_\lambda^{\pi'} v_\lambda^\pi)(\bar{s}) - v_\lambda^\pi(\bar{s}) + \lambda(\omega\left(\bar{s};\pi\right) - \omega\left(\bar{s};\pi'\right))
\end{aligned} \tag{19}
$$

The third relation holds by the fixed-point property of v_λ^π, and the last relation is by the definition of the regularized Bellman operator.

Plugging this back into (18), we get,

$$
\begin{aligned}
&\left\langle \nabla_{\pi(\cdot|\bar{s})} v^\pi(s), \pi'(\cdot \mid \bar{s}) - \pi(\cdot \mid \bar{s}) \right\rangle \\
&= \sum_{t=0}^{\infty} \gamma^t p^\pi(s_t = \bar{s} \mid s) \times \\
&\Big(-\lambda \left(\omega\left(s;\pi'\right) - \omega\left(s;\pi\right) - \left\langle \nabla_{\pi(\cdot|\bar{s})} \omega\left(\bar{s};\pi\right), \pi'(\cdot \mid \bar{s}) - \pi(\cdot \mid \bar{s}) \right\rangle \right) + (T_\lambda^{\pi'} v_\lambda^\pi)(\bar{s}) - v_\lambda^\pi(\bar{s}) \Big) \\
&= \sum_{t=0}^{\infty} \gamma^t p^\pi(s_t = \bar{s} \mid s) \left((T_\lambda^{\pi'} v_\lambda^\pi)(\bar{s}) - v_\lambda^\pi(\bar{s}) - \lambda B_\omega\left(\bar{s};\pi',\pi\right) \right)
\end{aligned} \tag{20}
$$

Thus, we have that

$$
\begin{aligned}
\langle \nabla_\pi v^\pi(s), \pi' - \pi \rangle &:= \sum_{\bar{s}} \sum_{a} \left(\nabla_{\pi(a|\bar{s})} v^\pi(s), \pi'(a \mid \bar{s}) - \pi(a \mid \bar{s}) \right) \\
&= \sum_{\bar{s}} \left\langle \nabla_{\pi(\cdot|\bar{s})} v^\pi(s), \pi'(\cdot \mid \bar{s}) - \pi(\cdot \mid \bar{s}) \right\rangle \\
&= \sum_{\bar{s}} \sum_{t=0}^{\infty} \gamma^t p^\pi(s_t = \bar{s} \mid s) \left((T_\lambda^{\pi'} v_\lambda^\pi)(\bar{s}) - v_\lambda^\pi(\bar{s}) - \lambda B_\omega\left(\bar{s};\pi',\pi\right) \right) \\
&= \sum_{\bar{s}} (I - \gamma P^\pi)^{-1}_{s,\bar{s}} \left((T_\lambda^{\pi'} v_\lambda^\pi)(\bar{s}) - v_\lambda^\pi(\bar{s}) - \lambda B_\omega\left(\bar{s};\pi',\pi\right) \right) \\
&= \left[(I - \gamma P^\pi)^{-1} \left(T_\lambda^{\pi'} v_\lambda^\pi - v_\lambda^\pi - \lambda B_\omega\left(\pi',\pi\right) \right) \right](s).
\end{aligned}
$$

Where the third relation is by (20), the forth by defining the matrix $\sum_{t=0}^{\infty} \gamma^t P^\pi = (I - \gamma P^\pi)^{-1}$, and the fifth by the definition of matrix-vector product.

To prove the second claim, multiply both sides of the first relation (3) by μ. For the LHS we get,

$$\begin{aligned}\sum_s \mu(s)\left\langle \nabla_{\pi(\cdot|\bar{s})} v^\pi(s), \pi'(\cdot \mid \bar{s}) - \pi(\cdot \mid \bar{s})\right\rangle &= \left\langle \sum_s \mu(s) \nabla_{\pi(\cdot|\bar{s})} v^\pi(s), \pi'(\cdot \mid \bar{s}) - \pi(\cdot \mid \bar{s})\right\rangle \\ &= \left\langle \nabla_{\pi(\cdot|\bar{s})} \sum_s \mu(s) v^\pi(s), \pi'(\cdot \mid \bar{s}) - \pi(\cdot \mid \bar{s})\right\rangle \\ &= \left\langle \nabla_{\pi(\cdot|\bar{s})} \mu v^\pi, \pi'(\cdot \mid \bar{s}) - \pi(\cdot \mid \bar{s})\right\rangle.\end{aligned}$$

In the first and second relation we used the linearity of the inner product and the derivative, and in the third relation the definition of μv^π. Lastly, observe that multiplying the RHS by μ yields $\mu(I - \gamma P^\pi)^{-1} = \frac{1}{1-\gamma} d_{\mu,\pi}$. □

E Uniform Trust Region Policy Optimization

In this Appendix, we derive the Uniform TRPO algorithm (Algorithm 1) and prove its convergence for both the unregularized and regularized versions. Both Uniform Projected Policy Gradient and Uniform NE-TRPO are instances of Uniform TRPO, by a proper choice of the Bregman distance. In Appendix E.1, we explicitly show that the iterates

$$\pi_{k+1} \in \operatorname*{arg\,min}_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} \left\{ \langle \nabla v_\lambda^{\pi_k}, \pi - \pi_k \rangle + \frac{1}{t_k}(I - \gamma P^{\pi_k})^{-1} B_\omega(\pi, \pi_k) \right\}, \tag{21}$$

result in algorithm 1. In Appendix E.2, we derive the updates of the *PolicyUpdate* procedure, Algorithms 3 and 4. Then, we turn to analyze Uniform TRPO and its instances in Appendix E.3. Specifically, we derive the fundamental inequality for Unifom TRPO, similarly to the fundamental inequality for Mirror Descent (Beck, 2017, Lemma-9.13). Although the objective is not convex, we show that due to the adaptive scaling, by applying the linear approximation of the value of regularized MDPs (Proposition 1), we can repeat similar derivation to that of MD, with some modifications. Finally, in Appendix E.4, we go on to prove convergence rates for both the unregularized ($\lambda = 0$) and regularized ($\lambda > 0$) versions of Uniform TRPO, using a right choice of stepsizes.

E.1 Uniform TRPO Update Rule

In each TRPO step, we solve the following optimization problem:

$$\begin{aligned}\pi_{k+1} &\in \operatorname{arg\,min}_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} \left\{ \langle \nabla v_\lambda^{\pi_k}, \pi - \pi_k \rangle + \frac{1}{t_k}(I - \gamma P^{\pi_k})^{-1} B_\omega(\pi, \pi_k) \right\} \\ &\in \operatorname{arg\,min}_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} \left\{ (I - \gamma P^{\pi_k})^{-1}(T_\lambda^\pi v_\lambda^{\pi_k} - v_\lambda^{\pi_k} - \lambda B_\omega(\pi, \pi_k)) + \frac{1}{t_k}(I - \gamma P^{\pi_k})^{-1} B_\omega(\pi, \pi_k) \right\} \\ &\in \operatorname{arg\,min}_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} \left\{ (I - \gamma P^{\pi_k})^{-1}(T_\lambda^\pi v_\lambda^{\pi_k} - v_\lambda^{\pi_k} + \left(\frac{1}{t_k} - \lambda\right) B_\omega(\pi, \pi_k)) \right\} \\ &\in \operatorname{arg\,min}_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} \left\{ T_\lambda^\pi v_\lambda^{\pi_k} - v_\lambda^{\pi_k} + \left(\frac{1}{t_k} - \lambda\right) B_\omega(\pi, \pi_k) \right\},\end{aligned}$$

where the second transition holds by plugging in the linear approximation (Proposition 1), and the last transition holds since $(I - \gamma P^{\pi_k})^{-1} > 0$ and does not depend on π. Thus, we have,

$$\pi_{k+1} \in \operatorname{arg\,min}_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} \{ t_k(T_\lambda^\pi v_\lambda^{\pi_k} - v_\lambda^{\pi_k}) + (1 - \lambda t_k) B_\omega(\pi, \pi_k) \} \tag{22}$$

By discarding terms which do not depend on π, we get

$$\pi_{k+1} \in \operatorname*{arg\,min}_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} \{ t_k T_\lambda^\pi v_\lambda^{\pi_k} + (1 - \lambda t_k) B_\omega(\pi, \pi_k) \} \tag{23}$$

We are now ready to write (7), using the fact that (23), can be written as the following state-wise optimization problem: For every $s \in \mathcal{S}$,

$$\pi_{k+1}(\cdot \mid s) \in \operatorname*{arg\,min}_{\pi \in \Delta_{\mathcal{A}}} \{ t_k T_\lambda^\pi v_\lambda^{\pi_k}(s) + (1 - \lambda t_k) B_\omega(s; \pi, \pi_k) \}$$

E.2 The *PolicyUpdate* procedure: Algorithms 3 and 4

Next, we write the solution for the optimization problem for each of the cases:

By plugging Lemma 20 into (22)

$$\pi_{k+1} \in \underset{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}}{\arg\min}\{t_k\langle q_\lambda^{\pi_k} + \lambda\nabla\omega(\pi_k), \pi - \pi_k\rangle + B_\omega\left(\pi, \pi_k\right)\}$$

Or again in a state-wise form,

$$\pi_{k+1}(\cdot \mid s) \in \underset{\pi \in \Delta_{\mathcal{A}}}{\arg\min}\{t_k\langle q_\lambda^{\pi_k}(s,\cdot) + \lambda\nabla\omega\left(s;\pi_k\right), \pi - \pi_k(\cdot \mid s)\rangle + B_\omega\left(s;\pi,\pi_k\right)\} \tag{24}$$

Using (24), we can plug in the solution of the MD iteration for each of the different cases.

Euclidean Case:

For ω chosen to be the L_2 norm, the solution to (24) is the orthogonal projection. For all $s \in \mathcal{S}$ the policy is updated according to

$$\begin{aligned}\pi_{k+1}(\cdot|s) &= P_{\Delta_{\mathcal{A}}}(\pi_k(\cdot|s) - t_k q_\lambda^{\pi_k}(s,\cdot) - \lambda t_k \pi_k(\cdot \mid s)) \\ &= P_{\Delta_{\mathcal{A}}}((1-\lambda t_k)\pi_k(\cdot|s) - t_k q_\lambda^{\pi_k}(s,\cdot)),\end{aligned}$$

where $P_{\Delta_{\mathcal{A}}}$ is the orthogonal projection operator over the simplex. Refer to [Beck, 2017] for details.

Finally, dividing by the constant $1 - \lambda t_k$ does not change the optimizer. Thus,

$$\pi_{k+1}(\cdot|s) = P_{\Delta_{\mathcal{A}}}\left(\pi_k(\cdot|s) - \frac{t_k}{1-\lambda t_k} q_\lambda^{\pi_k}(s,\cdot)\right), \tag{25}$$

Non-Euclidean Case:

For ω chosen to be the negative entropy, (24) has the following analytic solution for all $s \in \mathcal{S}$,

$$\begin{aligned}\pi_{k+1}(\cdot \mid s) &\in \underset{\pi \in \Delta_{\mathcal{A}}}{\arg\min}\{t_k\langle q_\lambda^{\pi_k}(s,\cdot) + \lambda\nabla H(\pi_k(\cdot \mid s), \pi - \pi_k(\cdot \mid s)\rangle + d_{KL}(\pi(\cdot \mid s)||\pi_k(\cdot \mid s))\} \\ &\in \underset{\pi \in \Delta_{\mathcal{A}}}{\arg\min}\{\langle t_k q_\lambda^{\pi_k}(s,\cdot) - (1-\lambda t_k)\nabla H(\pi_k(\cdot \mid s), \pi - \pi_k(\cdot \mid s)\rangle + H(\pi(\cdot \mid s)) - H_k(\pi(\cdot \mid s))\} \\ &\in \underset{\pi \in \Delta_{\mathcal{A}}}{\arg\min}\{\langle t_k q_\lambda^{\pi_k}(s,\cdot) - (1-\lambda t_k)\nabla H(\pi_k(\cdot \mid s), \pi\rangle + H(\pi(\cdot \mid s))\}\end{aligned}$$

where the first transition is by substituting ω and the Bregman distance, the second is by the definition of the Bregman distance, and the last transition is by omitting constant factors.

By using (Beck, 2017, Example 3.71), we get

$$\pi_{k+1}\left(a|s\right) = \frac{\pi_k\left(a \mid s\right)e^{-t_k q_\lambda^{\pi_k}(s,a) - \lambda t_k \nabla_{\pi_k(a|s)} H(\pi_k(\cdot|s))}}{\sum_{a'} \pi_k(a' \mid s)e^{-t_k q_\lambda^{\pi_k}(s,a') - \lambda t_k \nabla_{\pi_k(a'|s)} H(\pi_k(\cdot|s))}}.$$

Now, using the derivative of the negative entropy function $H(\cdot)$, we have that for every s, a,

$$\pi_{k+1}\left(a|s\right) = \frac{\pi_k\left(a \mid s\right)e^{-t_k\left(q_\lambda^{\pi_k}(s,a) - \lambda\log\pi_k(a|s)\right)}}{\sum_{a'} \pi_k(a' \mid s)e^{-t_k\left(q_\lambda^{\pi_k}(s,a') - \lambda\log\pi_k(a'|s)\right)}}, \tag{26}$$

which concludes the result.

E.3 Fundamental Inequality for Uniform TRPO

Central to the following analysis is Lemma 9, which we prove in this section. This lemma replaces Lemma [Beck, 2017][9.13] from which it inherits its name, for the RL non-convex case. It has two main differences relatively to Lemma [Beck, 2017][9.13]: (a) The inequality is in vector form (statewise). (b) The non-convexity of f demands replacing the gradient inequality with different proof mechanism, i.e., the directional derivative in RL (see Proposition 1).

Lemma 9 (fundamental inequality for Uniform TRPO)**.** *Let $\{\pi_k\}_{k\geq 0}$ be the sequence generated by the uniform TRPO method with stepsizes $\{t_k\}_{k\geq 0}$. Then, for every π and $k \geq 0$,*

$$\begin{aligned}
&t_k(I-\gamma P^{\pi})\left(v_\lambda^{\pi_k} - v_\lambda^{\pi}\right) \\
&\leq (1-\lambda t_k)B_\omega\left(\pi,\pi_k\right) - B_\omega\left(\pi,\pi_{k+1}\right) + \lambda t_k(\omega(\pi_k)-\omega(\pi_{k+1})) + \frac{t_k^2 h_\omega^2}{2}e,
\end{aligned}$$

where h_ω is defined in the second claim of Lemma 21, and e *is a vector of ones.*

Proof. First, notice that assumptions 2 and 3 hold. Assumption 2 is a regular assumption on the Bregman distance, which holds trivially both in the euclidean and non-euclidean case, where the optimization domain is the $\Delta_{\mathcal{A}}^{\mathrm{S}}$. Assumption 3 deals with the optimization problem itself and is similar to (Beck, 2017, Assumption 9.1) over $\Delta_{\mathcal{A}}$. The only difference is that in our case, the optimization objective v^π is non-convex.

Define $\psi(\pi) \equiv t_k(I-\gamma P^{\pi_k})\langle \nabla v_\lambda^{\pi_k}, \pi\rangle + \delta_{\Delta_{\mathcal{A}}^{\mathrm{S}}}(\pi)$ where $\delta_{\Delta_{\mathcal{A}}^{\mathrm{S}}}(\pi) = 0$ when $\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}$ and infinite otherwise. Observe it is a convex function in π, as a sum of two convex functions: The first term is linear in π for any $\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}$, and thus convex, and $\delta_{\Delta_{\mathcal{A}}^{\mathrm{S}}}(\pi)$ is convex since $\Delta_{\mathcal{A}}^{\mathrm{S}}$ is a convex set. Applying the non-euclidean second prox theorem (Theorem 27), with $a = \pi_k$, $b = \pi_{k+1}$, we get that for any $\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}$,

$$\langle \nabla\omega(\pi_k) - \nabla\omega(\pi_{k+1}), \pi - \pi_{k+1}\rangle \leq t_k(I-\gamma P^{\pi_k})\langle \nabla v_\lambda^{\pi_k}, \pi - \pi_{k+1}\rangle \tag{27}$$

By the three-points lemma (26),

$$\langle \nabla\omega(\pi_k) - \nabla\omega(\pi_{k+1}), \pi - \pi_{k+1}\rangle = B_\omega\left(\pi,\pi_{k+1}\right) + B_\omega\left(\pi_{k+1},\pi_k\right) - B_\omega\left(\pi,\pi_k\right),$$

which, combined with (27), gives,

$$B_\omega\left(\pi,\pi_{k+1}\right) + B_\omega\left(\pi_{k+1},\pi_k\right) - B_\omega\left(\pi,\pi_k\right) \leq t_k(I-\gamma P^{\pi_k})\langle \nabla v_\lambda^{\pi_k}, \pi - \pi_{k+1}\rangle.$$

Therefore, by simple algebraic mainpulation, we get

$$\begin{aligned}
&t_k(I-\gamma P^{\pi_k})\langle \nabla v_\lambda^{\pi_k}, \pi_k - \pi\rangle \\
&\leq B_\omega\left(\pi,\pi_k\right) - B_\omega\left(\pi,\pi_{k+1}\right) - B_\omega\left(\pi_{k+1},\pi_k\right) + t_k(I-\gamma P^{\pi_k})\langle \nabla v_\lambda^{\pi_k}, \pi_k - \pi_{k+1}\rangle \\
&= B_\omega\left(\pi,\pi_k\right) - B_\omega\left(\pi,\pi_{k+1}\right) - B_\omega\left(\pi_{k+1},\pi_k\right) + t_k\left(T_\lambda^{\pi_k} v_\lambda^{\pi_k} - T_\lambda^{\pi_{k+1}} v_\lambda^{\pi_k}\right) + \lambda t_k B_\omega\left(\pi_{k+1},\pi_k\right),
\end{aligned} \tag{28}$$

where the last equality is due to Proposition 1, and using $(I-\gamma P^{\pi_k})(I-\gamma P^{\pi_k})^{-1} = I$.

Rearranging we get

$$\begin{aligned}
&t_k(I-\gamma P^{\pi_k})\langle \nabla v_\lambda^{\pi_k}, \pi_k - \pi\rangle \\
&\leq B_\omega\left(\pi,\pi_k\right) - B_\omega\left(\pi,\pi_{k+1}\right) - (1-\lambda t_k)B_\omega\left(\pi_{k+1},\pi_k\right) + t_k\left(T_\lambda^{\pi_k} v_\lambda^{\pi_k} - T_\lambda^{\pi_{k+1}} v_\lambda^{\pi_k}\right) \\
&\leq B_\omega\left(\pi,\pi_k\right) - B_\omega\left(\pi,\pi_{k+1}\right) - \frac{1-\lambda t_k}{2}\left\|\pi_{k+1} - \pi_k\right\|^2 + t_k\left(T_\lambda^{\pi_k} v_\lambda^{\pi_k} - T_\lambda^{\pi_{k+1}} v_\lambda^{\pi_k}\right),
\end{aligned} \tag{29}$$

where the last inequality follows since the Bregman distance is 1-strongly-convex for our choices of B_ω (e.g., Beck, 2017, Lemma 9.4(a)).

Furthermore, for every state $s \in \mathcal{S}$,

$$
\begin{aligned}
& t_k \left(T_\lambda^{\pi_k} v_\lambda^{\pi_k} - T_\lambda^{\pi_{k+1}} v_\lambda^{\pi_k}\right)(s) \\
& = t_k \lambda (\omega(s;\pi_k) - \omega(s;\pi_{k+1})) \\
& + \sum_a t_k \left(\pi_k(a|s) - \pi_{k+1}(a|s)\right)\left(c(s,a) + \gamma \sum_{s'} p(s'|s,a) v_\lambda^{\pi_k}(s')\right) \\
& = t_k \lambda (\omega(s;\pi_k) - \omega(s;\pi_{k+1})) \\
& + \left\langle \frac{t_k}{\sqrt{1-\lambda t_k}} (c(s,\cdot) + \gamma \sum_{s'} p(s'|s,\cdot) v_\lambda^{\pi_k}(s')), \sqrt{1-\lambda t_k}(\pi_k(\cdot|s) - \pi_{k+1}(\cdot|s)) \right\rangle \\
& \le \lambda t_k (\omega(s;\pi_k) - \omega(s;\pi_{k+1}))) \\
& + \frac{1-\lambda t_k}{2} \left\|\pi_{k+1} - \pi_k\right\|^2 + \frac{t_k^2}{2(1-\lambda t_k)} \left\| c(s,\cdot) + \gamma \sum_{s'} p(s'|s,\cdot) v_\lambda^{\pi_k}(s') \right\|_*^2 \\
& \le \lambda t_k (\omega(s;\pi_k) - \omega(s;\pi_{k+1})) + \frac{1-\lambda t_k}{2} \left\|\pi_{k+1} - \pi_k\right\|^2 + \frac{t_k^2 h_\omega^2}{2(1-\lambda t_k)},
\end{aligned}
$$

where the first inequality is due to the Fenchel's inequality on the convex $\|\cdot\|^2$ and its convex conjugate $\|\cdot\|_*^2$, and the last equality uses the fact that $\|c(s,\cdot) + \gamma \sum_{s'} p(s'|s,\cdot) v_\lambda^{\pi_k}(s')\|_* \le \|c_\lambda(s,\cdot) + \gamma \sum_{s'} p(s'|s,\cdot) v_\lambda^{\pi_k}(s')\|_* = \|q_\lambda^{\pi_k}(s,\cdot)\|_*$, and using the repsective bound in Lemma 21.

Plugging the last inequality into (29),

$$
t_k (I - \gamma P^{\pi_k}) \langle \nabla v_\lambda^{\pi_k}, \pi_k - \pi \rangle \le \lambda t_k (\omega(\pi_k) - \omega(\pi_{k+1})) + B_\omega(\pi, \pi_k) - B_\omega(\pi, \pi_{k+1}) + \frac{t_k^2 h_\omega^2}{2(1-\lambda t_k)} e,
$$

where e is a vector of all ones.

By using Proposition 1 on the LHS, we get,

$$
\begin{aligned}
& - t_k (T^\pi v^{\pi_k} - v^{\pi_k} - \lambda B_\omega(\pi, \pi_k)) \le \lambda t_k (\omega(\pi_k) - \omega(\pi_{k+1})) + B_\omega(\pi, \pi_k) - B_\omega(\pi, \pi_{k+1}) + \frac{t_k^2 h_\omega^2}{2(1-\lambda t_k)} e \\
\iff & - t_k (T^\pi v^{\pi_k} - v^{\pi_k}) \le \lambda t_k (\omega(\pi_k) - \omega(\pi_{k+1})) + (1 - \lambda t_k) B_\omega(\pi, \pi_k) - B_\omega(\pi, \pi_{k+1}) + \frac{t_k^2 h_\omega^2}{2(1-\lambda t_k)} e.
\end{aligned}
$$

Lastly,

$$
\begin{aligned}
& t_k (I - \gamma P^\pi)(v_\lambda^{\pi_k} - v_\lambda^\pi) = -t_k (T^\pi v^{\pi_k} - v^{\pi_k}) \\
& \le (1 - \lambda t_k) B_\omega(\pi, \pi_k) - B_\omega(\pi, \pi_{k+1}) + \lambda t_k (\omega(\pi_k) - \omega(\pi_{k+1})) + \frac{t_k^2 h_\omega^2}{2(1-\lambda t_k)} e,
\end{aligned}
$$

where the first relation holds by the second claim in Lemma 25. □

E.4 Proof of Theorem 2

Before proving the theorem, we establish that the policy improves in k for the chosen learning rates.

Lemma 10 (Uniform TRPO Policy Improvement). *Let $\{\pi_k\}_{k \ge 0}$ be the sequence generated by Uniform TRPO. Then, for both the euclidean and non-euclidean versions of the algorithm, for any $\lambda \ge 0$, the value improves for all k,*

$$
v_\lambda^{\pi_k} \ge v_\lambda^{\pi_{k+1}}.
$$

Proof. Restating (28), we have that for any π,

$$
\begin{aligned}
& t_k (I - \gamma P^{\pi_k}) \langle \nabla v_\lambda^{\pi_k}, \pi_k - \pi \rangle \\
& \le B_\omega(\pi, \pi_k) - B_\omega(\pi, \pi_{k+1}) - B_\omega(\pi_{k+1}, \pi_k) + t_k \left(T_\lambda^{\pi_k} v_\lambda^{\pi_k} - T_\lambda^{\pi_{k+1}} v_\lambda^{\pi_k}\right) + \lambda t_k B_\omega(\pi_{k+1}, \pi_k).
\end{aligned}
$$

Plugging the closed form of the directional derivative (Proposition (1)), setting $\pi = \pi_k$, using $B_\omega\left(\pi_k, \pi_k\right) = 0$, we get,

$$t_k\left(T_\lambda^{\pi_k} v_\lambda^{\pi_k} - T_\lambda^{\pi_{k+1}} v_\lambda^{\pi_k}\right) \geq B_\omega\left(\pi_k, \pi_{k+1}\right) + B_\omega\left(\pi_{k+1}, \pi_k\right)(1 - \lambda t_k). \tag{30}$$

The choice of the learning rate and the fact that the Bregman distance is non negative ($\lambda > 0$, $\lambda t_k = \frac{1}{k+2} \leq 1$ and for $\lambda = 0$ the RHS of (30) is positive) implies that

$$\begin{aligned} & v_\lambda^{\pi_k} - T_\lambda^{\pi_{k+1}} v_\lambda^{\pi_k} = \left(T_\lambda^{\pi_k} v_\lambda^{\pi_k} - T_\lambda^{\pi_{k+1}} v_\lambda^{\pi_k}\right) \geq 0 \\ & \rightarrow v_\lambda^{\pi_k} \geq T_\lambda^{\pi_{k+1}} v_\lambda^{\pi_k}. \end{aligned} \tag{31}$$

Applying iteratively $T_\lambda^{\pi_{k+1}}$ and using its monotonicty we obtain,

$$v_\lambda^{\pi_k} \geq T_\lambda^{\pi_{k+1}} v_\lambda^{\pi_k} \geq (T_\lambda^{\pi_{k+1}})^2 v_\lambda^{\pi_k} \geq \cdots \geq \lim_{n\to\infty} (T_\lambda^{\pi_{k+1}})^n v_\lambda^{\pi_k} = v_\lambda^{\pi_{k+1}},$$

where in the last relation we used the fact $T_\lambda^{\pi_{k+1}}$ is a contraction operator and its fixed point is $v_\lambda^{\pi_{k+1}}$ which proves the claim. □

For the sake of completeness and readability, we restate here Theorem 2, this time including the full theorem with all logarithmic factors:

Theorem 11 (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 $\lambda = 0$, $t_k = \frac{(1-\gamma)}{C_{\omega,1} \mathrm{C}_{max} \sqrt{k+1}}$ then*

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

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

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

Where $C_{\omega,1} = \sqrt{A}, C_{\omega,3} = 1$ for the euclidean case, and $C_{\omega,1} = 1, C_{\omega,3} = \log A$ for the non-euclidean case.

We are now ready to prove Theorem 11, while following arguments from (Beck, 2017, Theorem 9.18).

The Unregularized case

Proof. Applying Lemma 9 with $\pi = \pi^*$ and $\lambda = 0$ (the unregularized case) and let $e \in \mathbb{R}^{\mathcal{S}}$, a vector ones, the following relations hold.

$$t_k (I - \gamma P^{\pi^*}) \left(v^{\pi_k} - v^*\right) \leq B_\omega\left(\pi^*, \pi_k\right) - B_\omega\left(\pi^*, \pi_{k+1}\right) + \frac{t_k^2 h_\omega^2}{2} e \tag{32}$$

Summing the above inequality over $k = 0, 1, ..., N$, and noticing we get a telescopic sum gives

$$\begin{aligned} \sum_{k=0}^N t_k (I - \gamma P^{\pi^*}) \left(v^{\pi_k} - v^*\right) &\leq B_\omega\left(\pi^*, \pi_0\right) - B_\omega\left(\pi^*, \pi_{N+1}\right) + \sum_{k=0}^N \frac{t_k^2 h_\omega^2}{2} e \\ &\leq B_\omega\left(\pi^*, \pi_0\right) + \sum_{k=0}^N \frac{t_k^2 h_\omega^2}{2} e \\ &\leq \left\|B_\omega\left(\pi^*, \pi_0\right)\right\|_\infty e + \sum_{k=0}^N \frac{t_k^2 h_\omega^2}{2} e \end{aligned}$$

where the second relation holds since $B_\omega(\pi^*, \pi_{N+1}) \geq 0$ component-wise. From which we get the following relations,

$$\begin{aligned}
&(I - \gamma P^{\pi^*}) \sum_{k=0}^{N} t_k \left(v^{\pi_k} - v^*\right) \leq \left\|B_\omega(\pi^*, \pi_0)\right\|_\infty e + \sum_{k=0}^{N} \frac{t_k^2 h_\omega^2}{2} e \\
\iff &\sum_{k=0}^{N} t_k \left(v^{\pi_k} - v^*\right) \leq (I - \gamma P^{\pi^*})^{-1} \left(\left\|B_\omega(\pi^*, \pi_0)\right\|_\infty e + \sum_{k=0}^{N} \frac{t_k^2 h_\omega^2}{2} e \right) \\
\iff &\sum_{k=0}^{N} t_k \left(v^{\pi_k} - v^*\right) \leq \frac{\left\|B_\omega(\pi^*, \pi_0)\right\|_\infty}{1-\gamma} e + \sum_{k=0}^{N} \frac{t_k^2 h_\omega^2}{2(1-\gamma)} e. \qquad (33)
\end{aligned}$$

In the second relation we multiplied both sides of inequality by $(I - \gamma P^{\pi^*})^{-1} \geq 0$ component-wise. In the third relation we used $(I - \gamma P^\pi)^{-1} e = \frac{1}{1-\gamma} e$ for any π. By Lemma (10) the policies are improving, from which, we get

$$\left(v_\lambda^{\pi_N} - v^*\right) \sum_{k=0}^{N} t_k \leq \sum_{k=0}^{N} t_k \left(v^{\pi_k} - v^*\right). \qquad (34)$$

Combining (33), (34) , and dividing by $\sum\limits_{k=0}^{N} t_k$ we get the following component-wise inequality,

$$v_\lambda^{\pi_N} - v^* \leq \frac{\left\|B_\omega(\pi^*, \pi_0)\right\|_\infty + \frac{h_\omega^2}{2} \sum\limits_{k=0}^{N} t_k^2}{(1-\gamma) \sum\limits_{k=0}^{N} t_k} e$$

By plugging in the stepsizes, $t_k = \frac{1}{h_\omega \sqrt{k+1}}$ we get,

$$v_\lambda^{\pi_N} - v^* \leq O\left(\frac{h_\omega}{1-\gamma} \frac{\left\|B_\omega(\pi^*, \pi_0)\right\|_\infty + \sum\limits_{k=0}^{N} \frac{1}{k+1}}{\sum\limits_{k=0}^{N} \frac{1}{\sqrt{k+1}}} e \right)$$

Plugging in Lemma 24 and bounding the sums (e.g., by using Beck, 2017, Lemma 8.27(a)) yields,

$$v_\lambda^{\pi_N} - v^* \leq O\left(\frac{h_\omega}{1-\gamma} \frac{D_\omega + \log N}{\sqrt{N}} e \right).$$

Plugging the expressions for h_ω, D_ω in Lemma 21 and Lemma 24 we conclude the proof. □

The Regularized case

Proof. Applying Lemma 9 with $\pi = \pi^*$ and $\lambda > 0$,

$$\begin{aligned}
&t_k (I - \gamma P^{\pi^*}) \left(v_\lambda^{\pi_k} - v_\lambda^*\right) \\
&\leq (1 - \lambda t_k) B_\omega(\pi^*, \pi_k) - B_\omega(\pi^*, \pi_{k+1}) + \lambda t_k (\omega(\pi_k) - \omega(\pi_{k+1})) + \frac{t_k^2 h_\omega^2}{2(1 - \lambda t_k)} e.
\end{aligned}$$

Plugging $t_k = \frac{1}{\lambda(k+2)}$ and multiplying by $\lambda(k+2)$,

$$\begin{aligned}
&(I - \gamma P^{\pi^*}) \left(v_\lambda^{\pi_k} - v_\lambda^*\right) \\
&\leq \lambda(k+1) B_\omega(\pi^*, \pi_k) - \lambda(k+2) B_\omega(\pi^*, \pi_{k+1}) + \lambda \omega(\pi_k) - \lambda \omega(\pi_{k+1}) + \frac{h_\omega^2}{2\lambda} \frac{1}{k+1} e.
\end{aligned}$$

Summing the above inequality over $k = 0, ..., N$ yields

$$\begin{aligned}
&\sum_{k=0}^{N} (I - \gamma P^{\pi^*}) \left(v_\lambda^{\pi_k} - v_\lambda^*\right) \\
&\leq \lambda B_\omega\left(\pi^*, \pi_0\right) - \lambda(N+3) B_\omega\left(\pi^*, \pi_{N+1}\right) + \lambda\omega(\pi_2) - \lambda\omega(\pi_{N+1}) + \frac{h_\omega^2}{2\lambda} e \sum_{k=0}^{N} \frac{1}{k+1},
\end{aligned}$$

as the summation results in a telescopic sum.

Observe that for any π, π' and both our choices of ω, $\omega(\pi) - \omega(\pi') \leq \max_\pi |\omega(\pi)|$. For the euclidean case $\max_\pi |\omega(\pi)| < 1$ and for the non euclidean case $\max_\pi |\omega(\pi)| \leq \log A$. These bounds are the same bounds as the bound for the Bregman distance, D_ω (see Lemma 24). Thus, for both our choices of ω we can bound $\omega(\pi) - \omega(\pi') < D_\omega$.

Furthermore, since $B_\omega\left(\pi^*, \pi_{N+1}\right) \geq 0$ the following bound holds:

$$\begin{aligned}
&\sum_{k=0}^{N} (I - \gamma P^{\pi^*}) \left(v_\lambda^{\pi_k} - v_\lambda^*\right) \leq 2\lambda D_\omega e + \frac{h_\omega^2}{2\lambda} e \sum_{k=1}^{N} \frac{1}{k+1} \\
\iff & (I - \gamma P^{\pi^*}) \sum_{k=0}^{N} \left(v_\lambda^{\pi_k} - v_\lambda^*\right) \leq 2\lambda D_\omega e + \frac{h_\omega^2}{2\lambda} e \sum_{k=1}^{N} \frac{1}{k+1} \\
\iff & \sum_{k=0}^{N} \left(v_\lambda^{\pi_k} - v_\lambda^*\right) \leq \frac{2\lambda D_\omega}{1-\gamma} e + \frac{h_\omega^2}{2\lambda(1-\gamma)} e \sum_{k=1}^{N} \frac{1}{k+1},
\end{aligned} \tag{35}$$

and in the third relation we multiplied both side by $(I - \gamma P^{\pi^*})^{-1} \geq 0$ component-wise and used $(I - \gamma P^{\pi})^{-1} e = \frac{1}{1-\gamma} e$ for any π.

By Lemma 10 the value $v_\lambda^{\pi_k}$ decreases in k, and, thus,

$$(N+1)(v_\lambda^{\pi_N} - v_\lambda^*) \leq \sum_{k=0}^{N} \left(v_\lambda^{\pi_k} - v_\lambda^*\right). \tag{36}$$

Combining (35), (36), and dividing by $N+1$ we get the following component-wise inequality,

$$v_\lambda^{\pi_N} - v_\lambda^* \leq \left(\frac{2\lambda D_\omega}{(1-\gamma)(N+1)} + \frac{h_\omega^2}{2\lambda(1-\gamma)(N+1)} \sum_{k=1}^{N+1} \frac{1}{k} \right) e$$

Using the fact that $\sum\limits_{k=1}^{N+1} \frac{1}{k} \in O(\log n)$, we get

$$v_\lambda^{\pi_N} - v_\lambda^* \leq O\left(\frac{\lambda^2 D_\omega + h_\omega^2 \log N}{\lambda(1-\gamma)N} e \right).$$

Plugging the expressions for h_ω, D_ω in Lemma 21 and Lemma 24 we conclude the proof. □

F Sample-Based Trust Region Policy Optimization

Sample-Based TRPO is a sample-based version of Uniform TRPO (see Appendix E). Unlike Uniform TRPO which have access to the entire state space and to a model of the environment, by which it computes $v^\pi \in \mathbb{R}^S$ in each iteration, Sample-Based TRPO solely requires the ability to sample from an MDP using a ν-restart model. Similarly to [Kakade et al., 2003], it requires Assumption 1 to be satisfied, which we restate here.

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

Thus, Sample-Based TRPO operates under much more realistic assumptions, and, more importantly, puts formal grounds to first-order gradient based methods such as NE-TRPO [Schulman et al., 2015], which was so far considered a heuristic method motivated by CPI [Kakade and Langford, 2002].

In this section we prove Sample-Based TRPO (Section 5, Theorem 4) converges to an approximately optimal solution with high probability. We now describe the content of each of the following subsections: First, in Appendix F.1, we prove Proposition 3 which shows the connection between Sample-Based TRPO (using unbiased estimation) and Uniform TRPO. In Appendix F.2, we analyze the Sample-Based TRPO update rule and formalize the truncated sampling process. In Appendix F.3, in order to ease readability, we give a detailed proof sketch of the convergence theorem for Sample-Based TRPO. Then, we derive a fundamental inequality that will be used to prove the convergence of both the unregularized and regularized versions (Appendix F.4). This inequality is a *scalar* version of the fundamental inequality for Uniform TRPO (Lemma 9), but with an additional term which arises due to the approximation error. In Appendix F.5, we analyze the sample complexity needed to bound this approximation error. We go on to prove the convergence rates of Sample-Based TRPO for both the unregularized and regularized version (Appendix F.6). Finally, in Appendix F.7, we calculate the overall sample complexity of both the unregularized and regularized Sample-Based TRPO and compare it to CPI.

F.1 Relation Between Uniform and Sample-Based TRPO

Before diving into the proof of Sample-Based TRPO, we prove Proposition 3, which connects the update rules of Uniform TRPO and Sample-Based TRPO (in case of an unbiased estimator for $q_\lambda^{\pi_k}$):

Proposition 3 (Uniform 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,*

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

Proof. For any $m = 1,...,M$, we take expectation over the sampling process given the filtration $\mathcal{F}_k$, i.e., $s_m \sim d_{\nu,\pi_k}, a_m \sim U(\mathcal{A}), \hat{q}_\lambda^{\pi_k} \sim q_\lambda^{\pi_k}$ (we assume here an unbiased estimation process where we do not truncate the sample trajectories),

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

where first transition is by the definition of $\hat{\nabla}\nu v_\lambda^{\pi_k}[m]$, the second by the smoothing theorem, the third transition is due to the linearity of expectation and the fourth transition is by taking the expectation and due to the fact that $\mathbb{1}\{a = a_m\}$ is zero for any $a \neq a_m$.

$$
\begin{aligned}
(*) &= \frac{1}{1-\gamma}\mathbb{E}\Bigg[\langle A q_\lambda^{\pi_k}(s_m, \cdot)\mathbb{1}\{\cdot = a_m\} + \nabla\omega\left(s_m;\pi_k\right), \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m)\rangle + \frac{1}{t_k}B_\omega\left(s_m;\pi,\pi_k\right) \mid \mathcal{F}_k\Bigg] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s_m}\Bigg[\sum_{a_m\in\mathcal{A}}\frac{1}{A}\langle A q_\lambda^{\pi_k}(s_m, \cdot)\mathbb{1}\{\cdot = a_m\} + \nabla\omega\left(s_m;\pi_k\right), \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m)\rangle + \frac{1}{t_k}B_\omega\left(s_m;\pi,\pi_k\right) \mid \mathcal{F}_k\Bigg] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s_m}\Bigg[\langle \sum_{a_m\in\mathcal{A}}\frac{1}{A} A q_\lambda^{\pi_k}(s_m, \cdot)\mathbb{1}\{\cdot = a_m\} + \nabla\omega\left(s_m;\pi_k\right), \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m)\rangle + \frac{1}{t_k}B_\omega\left(s_m;\pi,\pi_k\right) \mid \mathcal{F}_k\Bigg] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s_m}\Bigg[\langle q_\lambda^{\pi_k}(s_m, \cdot)\sum_{a_m\in\mathcal{A}}\mathbb{1}\{\cdot = a_m\} + \nabla\omega\left(s_m;\pi_k\right), \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m)\rangle + \frac{1}{t_k}B_\omega\left(s_m;\pi,\pi_k\right) \mid \mathcal{F}_k\Bigg] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s_m}\Bigg[\langle q_\lambda^{\pi_k}(s_m, \cdot) + \nabla\omega\left(s_m;\pi_k\right), \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m)\rangle + \frac{1}{t_k}B_\omega\left(s_m;\pi,\pi_k\right) \mid \mathcal{F}_k\Bigg] \\
&= (**).
\end{aligned}
$$

where the second transition is by taking the expectation over a_m, the third transition is by the linearity of the inner product and due to the fact that $\langle\nabla\omega\left(s_m;\pi_k\right), \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m)\rangle$ and $B_\omega\left(s_m;\pi,\pi_k\right)$ are independent of a_m.

Now, taking the expectation over $s_m \sim d_{\nu,\pi_k}$,

$$
\begin{aligned}
(**) &= \frac{1}{1-\gamma}\mathbb{E}_{s_m}\Bigg[\langle q_\lambda^{\pi_k}(s_m, \cdot) + \nabla\omega\left(s_m;\pi_k\right), \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m)\rangle + \frac{1}{t_k}B_\omega\left(s_m;\pi,\pi_k\right) \mid \mathcal{F}_k\Bigg] \\
&= \frac{1}{1-\gamma}\sum_s d_{\nu,\pi_k}(s)\bigg(\langle q_\lambda^{\pi_k}(s, \cdot) + \nabla\omega\left(s;\pi_k\right), \pi(\cdot \mid s) - \pi_k(\cdot \mid s)\rangle + \frac{1}{t_k}B_\omega\left(s;\pi,\pi_k\right)\bigg) \\
&= \frac{1}{1-\gamma}d_{\nu,\pi_k}\langle q_\lambda^{\pi_k} + \nabla\omega(\pi_k), \pi - \pi_k\rangle + \frac{1}{t_k}\frac{1}{1-\gamma}d_{\nu,\pi_k}B_\omega\left(\pi,\pi_k\right) \\
&= \frac{1}{1-\gamma}d_{\nu,\pi_k}(T_\lambda^{\pi}v_\lambda^{\pi_k} - v_\lambda^{\pi_k} - \lambda B_\omega\left(\pi,\pi_k\right)) + \frac{1}{t_k}\frac{1}{1-\gamma}d_{\nu,\pi_k}B_\omega\left(\pi,\pi_k\right) \\
&= \langle\nabla\nu v_\lambda^{\pi_k}, \pi - \pi_k\rangle + \frac{1}{t_k(1-\gamma)}d_{\nu,\pi_k}B_\omega\left(\pi,\pi_k\right),
\end{aligned}
$$

where the second transition is by taking the expectation w.r.t. to s_m, the the fourth is by using the lemma 20 which connects the bellman operator and the q-functions, and the last transition is due to (4) in Proposition 1.

So far, we proved that

$$
\begin{aligned}
&\langle\nabla\nu v_\lambda^{\pi_k}, \pi - \pi_k\rangle + \frac{1}{t_k(1-\gamma)}d_{\nu,\pi_k}B_\omega\left(\pi,\pi_k\right) \\
&\mathbb{E}\Bigg[\langle\hat{\nabla}\nu v_\lambda^{\pi_k}[m], \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m)\rangle + \frac{1}{t_k(1-\gamma)}B_\omega\left(s_m;\pi,\pi_k\right) \mid \mathcal{F}_k\Bigg] \qquad (37)
\end{aligned}
$$

It is left to prove that,

$$
\begin{aligned}
&\nu\bigg(\langle\nabla v_\lambda^{\pi_k}, \pi - \pi_k\rangle + \frac{1}{t_k}(I - \gamma P^{\pi_k})^{-1}B_\omega\left(\pi,\pi_k\right)\bigg) \\
&\quad = \langle\nabla\nu v_\lambda^{\pi_k}, \pi - \pi_k\rangle + \frac{1}{t_k(1-\gamma)}d_{\nu,\pi_k}B_\omega\left(\pi,\pi_k\right) \qquad (38)
\end{aligned}
$$

First, notice that for every s'

$$\begin{aligned}
\nu\big\langle \nabla_{\pi_k(\cdot|s')} v_\lambda^{\pi_k}, \pi - \pi_k \big\rangle &= \sum_s \nu(s) \big\langle \nabla_{\pi_k(\cdot|s')} v_\lambda^{\pi_k}(s), \pi(\cdot \mid s') - \pi_k(\cdot \mid s') \big\rangle \\
&= \left\langle \sum_s \nabla_{\pi_k(\cdot|s')} \nu(s) v_\lambda^{\pi_k}(s), \pi(\cdot \mid s') - \pi_k(\cdot \mid s') \right\rangle \\
&= \left\langle \nabla_{\pi_k(\cdot|s')} \sum_s \nu(s) v_\lambda^{\pi_k}(s), \pi(\cdot \mid s') - \pi_k(\cdot \mid s') \right\rangle \\
&= \big\langle \nabla_{\pi_k(\cdot|s')} \nu v_\lambda^{\pi_k}, \pi(\cdot \mid s') - \pi_k(\cdot \mid s') \big\rangle,
\end{aligned}$$

where in the second and third transition we used the linearity of the inner product and the derivative, and in the last transition we used the definition of $\nu v_\lambda^{\pi_k}$.

Thus, we have,

$$\nu\langle \nabla v_\lambda^{\pi_k}, \pi - \pi_k \rangle = \langle \nabla \nu v_\lambda^{\pi_k}, \pi - \pi_k \rangle. \tag{39}$$

Now,

$$\begin{aligned}
&\nu\bigg(\langle \nabla v_\lambda^{\pi_k}, \pi - \pi_k \rangle + \frac{1}{t_k}(I - \gamma P^{\pi_k})^{-1} B_\omega(\pi, \pi_k) \bigg) \\
&= \bigg(\nu\langle \nabla v_\lambda^{\pi_k}, \pi - \pi_k \rangle + \frac{1}{t_k} \nu (I - \gamma P^{\pi_k})^{-1} B_\omega(\pi, \pi_k) \bigg) \\
&= \bigg(\langle \nabla \nu v_\lambda^{\pi_k}, \pi - \pi_k \rangle + \frac{1}{t_k} \nu (I - \gamma P^{\pi_k})^{-1} B_\omega(\pi, \pi_k) \bigg) \\
&= \langle \nabla \nu v_\lambda^{\pi_k}, \pi - \pi_k \rangle + \frac{1}{t_k(1-\gamma)} d_{\nu,\pi_k} B_\omega(\pi, \pi_k),
\end{aligned}$$

where the second transition is by plugging in (39) and the last transition is by the definition of the stationary distribution d_{ν,π_k}.

By combining (37) and (38) we conclude the proof. □

F.2 Sample-Based TRPO Update Rule and the Sampling Process

In each step, we solve the following optimization problem (8):

$$\begin{aligned}
\pi_{k+1} &\in \arg\min_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} \left\{ \frac{1}{M} \sum_{m=1}^{M} \langle \hat{\nabla} \nu v_\lambda^{\pi_k}[m], \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m) \rangle + \frac{1}{t_k(1-\gamma)} B_\omega(s_m; \pi, \pi_k) \right\} \\
&\in \arg\min_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} \left\{ \frac{1}{M} \sum_{m=1}^{M} \left(\langle A \hat{q}_\lambda^{\pi_k}(s_m, \cdot, m) \mathbb{1}\{\cdot = a_m\} + \lambda \nabla \omega(s_m; \pi_k), \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m) \rangle + \tfrac{1}{t_k} B_\omega(s_m; \pi, \pi_k) \right) \right\} \\
&\in \arg\min_{\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}} \left\{ \sum_{s \in \mathcal{S}} \sum_{m=1}^{M} \mathbb{1}\{s = s_m\} \begin{pmatrix} \langle A \hat{q}_\lambda^{\pi_k}(s_m, \cdot, m) \mathbb{1}\{\cdot = a_m\} + \lambda \nabla \omega(s_m; \pi_k), \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m) \rangle \\ + \tfrac{1}{t_k} B_\omega(s_m; \pi, \pi_k) \end{pmatrix} \right\},
\end{aligned}$$

where $s_m \sim d_{\nu,\pi_k}(\cdot)$, $a_m \sim U(\mathcal{A})$, and $\hat{q}_\lambda^{\pi_k}(s_m, a_m, m)$ is the truncated Monte Carlo estimator of $q_\lambda^{\pi_k}(s_m, a_m)$ in the m-th trajectory. The notation $\hat{q}_\lambda^{\pi_k}(s_m, \cdot, m)\mathbb{1}\{\cdot = a_m\}$ is a vector with the estimator value at the index a_m, and zero elsewhere. Also, we remind the reader we use the notation $A := |\mathcal{A}|$. We can obtain a sample $s_m \sim d_{\nu,\pi_k}(\cdot)$ by a similar process as described in [Kakade and Langford, 2002, Kakade et al., 2003]. Draw a start state s from the ν-restart distribution. Then, $s_m = s$ is chosen w.p. γ. Otherwise, w.p. $1-\gamma$, an action is sampled according to $a \sim \pi_k(s)$ to receive the next state s. This process is repeated until s_m is chosen. If the time $T = \frac{1}{1-\gamma} \log \frac{\epsilon}{8 r_\omega(k,\lambda)}$ is reached, we accept the current state as s_m. Note that $r_\omega(k, \lambda)$ is defined in Lemma 16, and ϵ is the required final error. Finally, when s_m is chosen, an action a_m is drawn from the uniform distribution, and then the trajectory is unrolled using the current policy π_k for $T = \frac{1}{1-\gamma} \log \frac{\epsilon}{8 r_\omega(k,\lambda)}$ time-steps, to calculate $\hat{q}_\lambda^{\pi_k}(s_m, a_m, m)$. Note that this introduces a bias into the estimation of $q_\lambda^{\pi_k}$ [Kakade et al., 2003][Sections 2.3.3 and 7.3.4]. Lastly, note that the A factor in the estimator is due to importance sampling.

First, the update rule of Sample-Based TRPO can be written as a state-wise update rule for any $s \in \mathcal{S}$. Observe that,

$$\begin{aligned}
\pi_{k+1} &\in \underset{\pi\in\Delta_{\mathcal{A}}^{\mathrm{S}}}{\arg\min} \left\{ \sum_{m=1}^{M} \langle A\hat{q}_{\lambda}^{\pi_k}(s_m,\cdot,m)\mathbb{1}\{\cdot = a_m\} + \lambda\nabla\omega\left(s_m;\pi_k\right), \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m)\rangle + \frac{1}{t_k}B_\omega\left(s_m;\pi,\pi_k\right) \right\} \\
&= \underset{\pi\in\Delta_{\mathcal{A}}^{\mathrm{S}}}{\arg\min} \left\{ \sum_{s\in\mathcal{S}} \sum_{m=1}^{M} \mathbb{1}\{s = s_m\} \begin{pmatrix} \langle A\hat{q}_{\lambda}^{\pi_k}(s_m,\cdot,m)\mathbb{1}\{\cdot = a_m\} + \lambda\nabla\omega\left(s_m;\pi_k\right), \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m)\rangle \\ +\frac{1}{t_k}B_\omega\left(s_m;\pi,\pi_k\right) \end{pmatrix} \right\},
\end{aligned} \tag{40}$$

The first relation is the definition of the update rule (8) without the constant factor $\frac{1}{M}$. See that multiplying the optimization problem by the constant M does not change the minimizer. In the second relation we used the fact that summation on $\sum_s \mathbb{1}\{s = s_m\}$ leaves the optimization problem unchanged (as the indicator function is 0 for all states that are not s_m).

Thus, using this update rule we can solve the optimization problem individually per $s \in \mathcal{S}$,

$$\pi_{k+1}(\cdot|s) = \underset{\pi\in\Delta_{\mathcal{A}}}{\arg\min} \left\{ \sum_{m=1}^{M} \mathbb{1}\{s = s_m\} \left(\langle A\hat{q}_{\lambda}^{\pi_k}(s,\cdot,m)\mathbb{1}\{\cdot = a_m\} + \lambda\nabla\omega\left(s;\pi_k\right), \pi - \pi_k(\cdot \mid s)\rangle + \tfrac{1}{t_k}B_\omega\left(s;\pi,\pi_k\right) \right) \right\}. \tag{41}$$

Note that using this representation optimization problem, the solution for states which were not encountered in the k-th iteration, $s \notin \mathcal{S}_M^k$, is arbitrary. To be consistent, we always choose to keep the current policy, $\pi_{k+1}(\cdot \mid s) = \pi_k(\cdot \mid s)$.

Now, similarly to Uniform TRPO, the update rule of Sample-Based TRPO can be written such that the optimization problem is solved individually per visited state $s \in \mathcal{S}_M^k$. This results in the final update rule used in Algorithm 2.

To prove this, let $n(s) = \sum_a n(s,a)$ be the number of times the state s was observed at the k-th episode. Using this notation and (40), the update rule has the following equivalent forms,

$$\begin{aligned}
\pi_{k+1} &\in \underset{\pi\in\Delta_{\mathcal{A}}^{\mathrm{S}}}{\arg\min} \left\{ \sum_{m=1}^{M} \langle A\hat{q}_{\lambda}^{\pi_k}(s_m,\cdot,m)\mathbb{1}\{\cdot = a_m\} + \lambda\nabla\omega\left(s_m;\pi_k\right), \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m)\rangle + \frac{1}{t_k}B_\omega\left(s_m;\pi,\pi_k\right) \right\} \\
&= \underset{\pi\in\Delta_{\mathcal{A}}^{\mathrm{S}}}{\arg\min} \left\{ \sum_{s\in\mathcal{S}} \sum_{m=1}^{M} \mathbb{1}\{s = s_m\} \begin{pmatrix} \langle A\hat{q}_{\lambda}^{\pi_k}(s_m,\cdot,m)\mathbb{1}\{\cdot = a_m\} + \lambda\nabla\omega\left(s_m;\pi_k\right), \pi(\cdot \mid s_m) - \pi_k(\cdot \mid s_m)\rangle \\ +\frac{1}{t_k}B_\omega\left(s_m;\pi,\pi_k\right) \end{pmatrix} \right\} \\
&= \underset{\pi\in\Delta_{\mathcal{A}}^{\mathrm{S}}}{\arg\min} \left\{ \sum_{s\in\mathcal{S}} \begin{pmatrix} \left\langle \sum_{m=1}^{M} \mathbb{1}\{s = s_m\} A\hat{q}_{\lambda}^{\pi_k}(s_m,\cdot,m)\mathbb{1}\{\cdot = a_m\} + n(s)\lambda\nabla\omega\left(s;\pi_k\right), \pi(\cdot \mid s) - \pi_k(\cdot \mid s)\right\rangle \\ +n(s)\frac{1}{t_k}B_\omega\left(s;\pi,\pi_k\right) \end{pmatrix} \right\} \\
&= \underset{\pi\in\Delta_{\mathcal{A}}^{\mathrm{S}}}{\arg\min} \left\{ \sum_{s\in\mathcal{S}_M^k} \begin{pmatrix} \left\langle \sum_{m=1}^{M} \mathbb{1}\{s = s_m\} A\hat{q}_{\lambda}^{\pi_k}(s_m,\cdot,m)\mathbb{1}\{\cdot = a_m\} + n(s)\lambda\nabla\omega\left(s;\pi_k\right), \pi(\cdot \mid s) - \pi_k(\cdot \mid s)\right\rangle \\ +n(s)\frac{1}{t_k}B_\omega\left(s;\pi,\pi_k\right) \end{pmatrix} \right\} \\
&= \underset{\pi\in\Delta_{\mathcal{A}}^{\mathrm{S}}}{\arg\min} \left\{ \sum_{s\in\mathcal{S}_M^k} \begin{pmatrix} \left\langle \frac{1}{n(s)} \sum_{m=1}^{M} \mathbb{1}\{s = s_m\} A\hat{q}_{\lambda}^{\pi_k}(s_m,\cdot,m)\mathbb{1}\{\cdot = a_m\} + \lambda\nabla\omega\left(s;\pi_k\right), \pi(\cdot \mid s) - \pi_k(\cdot \mid s)\right\rangle \\ +\frac{1}{t_k}B_\omega\left(s;\pi,\pi_k\right) \end{pmatrix} \right\}.
\end{aligned} \tag{42}$$

In the third relation we used the fact for any π, π_k

$$\sum_s \sum_{m=1}^{M} B_\omega\left(s_m;\pi,\pi_k\right)\mathbb{1}\{s = s_m\} = \sum_s B_\omega\left(s;\pi,\pi_k\right) \sum_{m=1}^{M} \mathbb{1}\{s = s_m\} = \sum_s B_\omega\left(s;\pi,\pi_k\right) n(s).$$

The fourth relation holds as the optimization problem is not affected by $s \notin \mathcal{S}_M^k$, and the last relation holds by dividing by $n(s) > 0$ as $s \in \mathcal{S}_M^k$ and using linearity of inner product.

Lastly, we observe that (42) is a sum of functions of $\pi(\cdot \mid s)$, i.e.,

$$\pi_{k+1} \in \arg \min_{\pi\in\Delta_{\mathcal{A}}^{\mathrm{S}}} \left\{ \sum_{s\in\mathcal{S}_M^k} f(\pi(\cdot \mid s)) \right\},$$

where $f = \langle g_s, \pi(\cdot \mid s)\rangle + \frac{1}{t_k} B_\omega\left(s; \pi, \pi_k\right)$, $g_s \in \mathbb{R}^A$ is the vector inside the inner product of (42). Meaning, the minimization problem is a sum of independent summands. Thus, in order to minimize the function on $\Delta_{\mathcal{A}}^{\mathrm{S}}$ it is enough to minimize independently each one of the summands. From this observation, we conclude that the update rule (8) is equivalent to update the policy for all $s \in \mathcal{S}_M^k$ by

$$\pi_{k+1}(\cdot \mid s) \in \arg\min_{\pi \in \Delta_{\mathcal{A}}} \left\{ \left(\begin{array}{c} \left\langle \frac{1}{n(s)} \sum_{m=1}^{M} \mathbb{1}\{s = s_m\} A \hat{q}_\lambda^{\pi_k}(s_m, \cdot, m) \mathbb{1}\{\cdot = a_m\} + \lambda \nabla \omega\left(s; \pi_k\right), \pi - \pi_k(\cdot \mid s) \right\rangle \\ + \frac{1}{t_k} B_\omega\left(s; \pi, \pi_k\right) \end{array} \right) \right\}, \tag{43}$$

Finally, by plugging in $\hat{q}_\lambda^{\pi_k}(s,a) = \frac{1}{n(s)} \sum_{i=1}^{n(s,a)} \hat{q}_\lambda^{\pi_k}(s, a, m_i)$, we get

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

where m_i is the trajectory index of the i-th occurrence of the state s.

F.3 Proof Sketch of Theorem 4

In order to keep things organized for an easy reading, we first go through the proof sketch in high level, which serves as map for reading the proof of Theorem 4 in the following sections:

1. In Appendix F.4, we use the Sample-Based TRPO optimization problem described in F.2, to derive a fundamental inequality in Lemma 15 for the sample-based case.:
   (a) We derive a state-wise inequality for the sample-based case. By adding and subtracting a term which relates to the expectation of the state-wise inequality, we write this inequality as a sum between the expected error and an approximation error term.
   (b) For each state, we employ importance sampling of $\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}$ to relate the derived state-wise inequality, to a global guarantee w.r.t. the optimal policy π^* and measure μ. This importance sampling procedure is allowed by assumption 1, which states that for any s such that $d_{\mu,\pi^*}(s) > 0$ it also holds that $\nu(s) > 0$, and thus $d_{\nu,\pi_k}(s) > 0$ since $d_{\nu,\pi_k}(s) \geq (1-\gamma)\nu(s)$.
   (c) By summing over all states we get the required fundamental inequality, which is a sum between the expected error and an approximation error term.
2. In Appendix F.5, we show that the approximation error term is made of two sources of errors: (a) a sampling error due to the finite number of trajectories in each iteration; (b) a truncation error due to the finite length of each trajectory, even in the infinite-horizon case.
   (a) In Lemma 16 we deal with the sampling error. We show that this error is due to the difference between an empirical mean of i.i.d. random variables and their expected value. Using Lemma 22 and Lemma 23, we show that these random variables are bounded, and also that they are proportional to the step size t_k. Then, similarly to [Kakade et al., 2003], we use Hoeffding's inequality and the union bound over the policy space (in our case, the space of deterministic policies), in order to bound this error term uniformly. This enables us to find the number of trajectories needed in the k-th iteration to reach an error proportional to $C^{\pi^*} t_k \epsilon = \left\| \frac{d_{\mu,\pi^*}}{\nu} \right\|_\infty t_k \epsilon$ with high probability. The common concentration efficient C^{π^*}, arises due to $\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}$, the importance sampling ratio used for the global convergence guarantee.
   (b) In Lemma 17 we deal with the truncation error. We show that we can bound this error to be proportional to $C^{\pi^*} t_k \epsilon$, by using $O\left(\frac{1}{1-\gamma}\right)$ samples in each trajectory.

   Finally, in Lemma 19, we use the union bound over all $k \in \mathbb{N}$ in order to uniformly bound the error propagation over N iterations of Sample-Based TRPO.
3. In Appendix F.6, we use the above results to prove Theorem 4 using a similar analysis to the one used for the rates guarantees of Uniform TRPO (Appendix E.4). The main difference is the additional approximation term which we bound in F.5. There, we make use of the fact that the approximation term is proportional to the step size t_k and thus decreasing with the number of iterations, to prove a bounded approximation error for any N. Moreover, differently than in Uniform TRPO, in the sample-based case we don't have improvement guarantees, and therefore the convergence is proved for the best policy in hindsight.
4. Lastly, in Appendix F.7, we calculate the overall sample complexity – previously we bounded the number of needed iterations and the number of samples needed in every iteartion – for each of the four cases of Sample-Based TRPO (euclidean vs. non-euclidean, unregularized vs. regularized).

F.4 Fundamental Inequality of Sample-Based TRPO

First, we will prove a lemma which will be used in the proof of Lemma 13.

Lemma 12 (helper state-wise inequality)**.** *For all states s the following inequality holds:*

$$\begin{aligned}&\left\langle t_k\left(q_\lambda^{\pi_k}(s,\cdot)+\lambda\nabla\omega\left(s;\pi_k\right)\right)+\nabla_{\pi_{k+1}}B_\omega\left(s;\pi_{k+1},\pi_k\right),\pi-\pi_{k+1}(\cdot\mid s)\right\rangle\\&\leq t_k(T_\lambda^\pi v_\lambda^{\pi_k}(s)-v_\lambda^{\pi_k}(s))+\frac{t_k^2h_\omega^2(k;\lambda)}{2}+(1-\lambda t_k)B_\omega\left(s;\pi,\pi_k\right)-B_\omega\left(s;\pi,\pi_{k+1}\right).\end{aligned}$$

where h_ω is defined at the third claim of Lemma 21.

Proof. First, consider the LHS of the inequality,

$$\begin{aligned}&\left\langle t_k\left(q_\lambda^{\pi_k}(s,\cdot)+\lambda\nabla\omega\left(s;\pi_k\right)\right)+\nabla_{\pi_{k+1}}B_\omega\left(s;\pi_{k+1},\pi_k\right),\pi-\pi_{k+1}(\cdot\mid s)\right\rangle\\&=\underbrace{t_k\langle q_\lambda^{\pi_k}(s,\cdot)+\lambda\nabla\omega\left(s;\pi_k\right),\pi-\pi_{k+1}(\cdot\mid s)\rangle}_{(1)}+\underbrace{\left\langle\nabla_{\pi_{k+1}}B_\omega\left(s;\pi_{k+1},\pi_k\right),\pi-\pi_{k+1}(\cdot\mid s)\right\rangle}_{(2)}\end{aligned}\tag{44}$$

The first term can be bounded as follows.

$$\begin{aligned}(1)&=t_k\langle q_\lambda^{\pi_k}(s,\cdot)+\lambda\nabla\omega\left(s;\pi_k\right),\pi-\pi_{k+1}(\cdot\mid s)\rangle\\&=t_k\langle q_\lambda^{\pi_k}(s,\cdot)+\lambda\nabla\omega\left(s;\pi_k\right),\pi-\pi_k(\cdot\mid s)\rangle\\&+t_k\langle q_\lambda^{\pi_k}(s,\cdot)+\lambda\nabla\omega\left(s;\pi_k\right),\pi_k(\cdot\mid s)-\pi_{k+1}(\cdot\mid s)\rangle\\&\leq t_k\langle q_\lambda^{\pi_k}(s,\cdot)+\lambda\nabla\omega\left(s;\pi_k\right),\pi-\pi_k(\cdot\mid s)\rangle\\&+|\langle t_kq_\lambda^{\pi_k}(s,\cdot)+t_k\lambda\nabla\omega\left(s;\pi_k\right),\pi_k(\cdot\mid s)-\pi_{k+1}(\cdot\mid s)\rangle|\\&\leq t_k\langle q_\lambda^{\pi_k}(s,\cdot)+\lambda\nabla\omega\left(s;\pi_k\right),\pi-\pi_k(\cdot\mid s)\rangle\\&+\frac{t_k^2\left\|q_\lambda^{\pi_k}(s,\cdot)+\lambda\nabla\omega\left(s;\pi_k\right)\right\|_*^2}{2}+\frac{1}{2}\left\|\pi_k(\cdot\mid s)-\pi_{k+1}(\cdot\mid s)\right\|^2,\end{aligned}$$

where the last relation follows from Fenchel's inequality using the euclidean or non-euclidean norm $\|\cdot\|$, and where $\|\cdot\|_*$ is its dual norm, which is L_2 in the euclidean case, and L_∞ in the non-euclidean case. Note that the norms are applied over the action space. Furthermore, by adding and subtracting $\lambda\omega\left(s;\pi\right)$,

$$\begin{aligned}&\langle q_\lambda^{\pi_k}(s,\cdot)+\lambda\nabla\omega\left(s;\pi_k\right),\pi-\pi_k\rangle\\&=\langle q_\lambda^{\pi_k}(s,\cdot),\pi-\pi_k(\cdot\mid s)\rangle+\lambda\langle\nabla\omega\left(s;\pi_k\right),\pi-\pi_k(\cdot\mid s)\rangle\\&=T^\pi v_\lambda^{\pi_k}(s)-T^{\pi_k}v_\lambda^{\pi_k}(s)-\lambda\omega\left(s;\pi\right)+\lambda\omega\left(s;\pi_k\right)+\lambda\langle\nabla\omega\left(s;\pi_k\right),\pi-\pi_k(\cdot\mid s)\rangle\\&=T_\lambda^\pi v_\lambda^{\pi_k}(s)-T_\lambda^{\pi_k}v_\lambda^{\pi_k}(s)-\lambda B_\omega\left(s;\pi,\pi_k\right)\\&=T_\lambda^\pi v_\lambda^{\pi_k}(s)-v_\lambda^{\pi_k}(s)-\lambda B_\omega\left(s;\pi,\pi_k\right),\end{aligned}\tag{45}$$

where the second transition follows the same steps as in equation (19) in the proof of Proposition 1, and the third transition is by the definition of the Bregman distance of ω. Note that (45) is actually given in Lemma 20, but is re-derived here for readability.

From which, we conclude that

$$\begin{aligned}(1)&\leq t_k(T_\lambda^\pi v_\lambda^{\pi_k}(s)-v_\lambda^{\pi_k}(s)-\lambda B_\omega\left(s;\pi,\pi_k\right))\\&+\frac{t_k^2\left\|q_\lambda^{\pi_k}(s,\cdot)+\lambda\nabla\omega\left(s;\pi_k\right)\right\|_*^2}{2}+\frac{1}{2}\left\|\pi_k(\cdot\mid s)-\pi_{k+1}(\cdot\mid s)\right\|^2\\&\leq t_k(T_\lambda^\pi v_\lambda^{\pi_k}(s)-v_\lambda^{\pi_k}(s)-\lambda B_\omega\left(s;\pi,\pi_k\right))+\frac{t_k^2h_\omega^2(k;\lambda)}{2}+\frac{1}{2}\left\|\pi_k(\cdot\mid s)-\pi_{k+1}(\cdot\mid s)\right\|^2,\end{aligned}$$

where in the last transition we used the third claim of Lemma 21, We now continue analyzing (2).

$$
\begin{aligned}
(2) &= \left\langle \nabla_{\pi_{k+1}} B_\omega\left(s ; \pi_{k+1}, \pi_k\right), \pi-\pi_{k+1}(\cdot \mid s)\right\rangle \\
&= \left\langle \nabla \omega\left(s ; \pi_{k+1}\right)-\nabla \omega\left(s ; \pi_k\right), \pi-\pi_{k+1}(\cdot \mid s)\right\rangle \\
&= B_\omega\left(s ; \pi, \pi_k\right)-B_\omega\left(s ; \pi, \pi_{k+1}\right)-B_\omega\left(s ; \pi_{k+1}, \pi_k\right) \\
&\leq B_\omega\left(s ; \pi, \pi_k\right)-B_\omega\left(s ; \pi, \pi_{k+1}\right)-\frac{1}{2}\left\|\pi_k(\cdot \mid s)-\pi_{k+1}(\cdot \mid s)\right\|^2 .
\end{aligned}
$$

The first relation, $\nabla_{\pi_{k+1}} B_\omega\left(s ; \pi_{k+1}, \pi_k\right)=\nabla \omega\left(s ; \pi_{k+1}\right)-\nabla \omega\left(s ; \pi_k\right)$, holds by simply taking the derivative of any Bregman distance w.r.t. π_{k+1}. The second relation holds by the three-points lemma (Lemma 26). The third relation holds by the strong convexity of the Bregman distance, i.e., $\frac{1}{2}\|x-y\|^2 \leq B_\omega(x, y)$, which is straight forward in the euclidean case, and is the well known Pinsker's inequality in the non-euclidean case.

Plugging the above upper bounds for (1) and (2) into (44) we get,

$$
\begin{aligned}
&\left\langle t_k\left(q_\lambda^{\pi_k}(s, \cdot)+\lambda \nabla \omega\left(s ; \pi_k\right)\right)+\nabla_{\pi_{k+1}} B_\omega\left(s ; \pi_{k+1}, \pi_k\right), \pi-\pi_{k+1}(\cdot \mid s)\right\rangle \\
&\leq t_k(T_\lambda^\pi v_\lambda^{\pi_k}(s)-v_\lambda^{\pi_k}(s))+\frac{t_k^2(h_\omega^2(k ; \lambda))}{2}+(1-\lambda t_k) B_\omega\left(s ; \pi, \pi_k\right)-B_\omega\left(s ; \pi, \pi_{k+1}\right),
\end{aligned}
$$

which concludes the proof. □

Now, we are ready to derive the fundamental inequality of Sample-Based TRPO. We first derive the following state-wise inequality:

Lemma 13 (sample-based state-wise inequality)**.** *Let $\{\pi_k\}_{k\geq 0}$ be the sequence generated by Aproximate TRPO using stepsizes $\{t_k\}_{k\geq 0}$. Then, for all states s for which $d_{\nu,\pi_k}(s)>0$ the following inequality holds for all $\pi \in \Delta_{\mathcal{A}}^{\mathrm{S}}$,*

$$
0 \leq t_k(T_\lambda^\pi v_\lambda^{\pi_k}(s)-v_\lambda^{\pi_k}(s))+\frac{t_k^2 h_\omega^2(k ; \lambda)}{2}+(1-\lambda t_k) B_\omega\left(s ; \pi, \pi_k\right)-B_\omega\left(s ; \pi, \pi_{k+1}\right)+\epsilon_k(s, \pi).
$$

where h_ω is defined at the third claim of Lemma 22.

Proof. Using the first order optimality condition for the update rule (41), the following holds for any $s \in \mathcal{S}$ and thus for any $s \in \{s' : d_{\nu,\pi_k}(s)>0\}$,

$$
0 \leq \frac{1}{M} \sum_{m=1}^M \mathbb{1}\{s=s_m\}\left\langle t_k\left(A \hat{q}_\lambda^{\pi_k}(s_m, \cdot, m) \mathbb{1}\{\cdot=a_m\}+\lambda \nabla \omega\left(s_m ; \pi_k\right)\right)+\nabla_{\pi_{k+1}} B_\omega\left(s_m ; \pi_{k+1}, \pi_k\right), \pi-\pi_{k+1}(\cdot \mid s_m)\right\rangle .
$$

Dividing by $d_{\nu,\pi_k}(s)$ which is strictly positive for all s such that $\mathbb{1}\{s=s_m\}=1$ and adding and subtracting the term

$$
\left\langle t_k\left(q_\lambda^{\pi_k}(s, \cdot)+\lambda \nabla \omega\left(s ; \pi_k\right)\right)+\nabla_{\pi_{k+1}} B_\omega\left(s ; \pi_{k+1}, \pi_k\right), \pi-\pi_{k+1}(\cdot \mid s)\right\rangle,
$$

we get

$$
0 \leq \underbrace{\left\langle t_k\left(q_\lambda^{\pi_k}(s, \cdot)+\lambda \nabla \omega\left(s ; \pi_k\right)\right)+\nabla_{\pi_{k+1}} B_\omega\left(s ; \pi_{k+1}, \pi_k\right), \pi-\pi_{k+1}(\cdot \mid s)\right\rangle}_{(*)}+\epsilon_k(s, \pi), \tag{46}
$$

where we defined $\epsilon_k(s,\pi)$,

$$
\begin{aligned}
&\epsilon_k(s, \pi) \\
&:= \frac{1}{d_{\nu,\pi_k}(s)} \frac{1}{M} \sum_{m=1}^M \mathbb{1}\{s=s_m\}\left\langle t_k\left(A \hat{q}_\lambda^{\pi_k}(s_m, \cdot, m) \mathbb{1}\{\cdot=a_m\}+\lambda \nabla \omega\left(s_m ; \pi_k\right)\right)+\nabla_{\pi_{k+1}} B_\omega\left(s_m ; \pi_{k+1}, \pi_k\right), \pi-\pi_{k+1}(\cdot \,|\, s_m)\right\rangle \\
&-\left\langle t_k\left(q_\lambda^{\pi_k}(s, \cdot)+\lambda \nabla \omega\left(s ; \pi_k\right)\right)+\nabla_{\pi_{k+1}} B_\omega\left(s ; \pi_{k+1}, \pi_k\right), \pi-\pi_{k+1}(\cdot \mid s)\right\rangle \\
&= \frac{1}{d_{\nu,\pi_k}(s)} \frac{1}{M} \sum_{m=1}^M \mathbb{1}\{s=s_m\}\langle t_k\left(A \hat{q}_\lambda^{\pi_k}(s_m, \cdot, m) \mathbb{1}\{\cdot=a_m\}+\lambda \nabla \omega\left(s_m ; \pi_k\right)\right)+\nabla \omega\left(s ; \pi_{k+1}\right)-\nabla \omega\left(s ; \pi_k\right), \pi-\pi_{k+1}(\cdot \,|\, s_m)\rangle \\
&-\langle t_k\left(q_\lambda^{\pi_k}(s, \cdot)+\lambda \nabla \omega\left(s ; \pi_k\right)\right)+\nabla \omega\left(s ; \pi_{k+1}\right)-\nabla \omega\left(s ; \pi_k\right), \pi-\pi_{k+1}(\cdot \mid s)\rangle .
\end{aligned} \tag{47}
$$

By bounding $(*)$ in (46) using Lemma 12 we conclude the proof. □

Now, we state another lemma which connects the state-wise inequality using the discounted stationary distribution of the optimal policy d_{μ,π^*}.

Lemma 14. *Let Assumption 1 hold and let $\{\pi_k\}_{k\geq 0}$ be the sequence generated by Aproximate TRPO using stepsizes $\{t_k\}_{k\geq 0}$. Then, for all $k \geq 0$ Then, the following inequality holds for all π,*

$$0 \leq t_k d_{\mu,\pi^*}(T^{\pi}_{\lambda} v^{\pi_k}_{\lambda} - v^{\pi_k}_{\lambda}) + \frac{t_k^2 h_{\omega}^2(k;\lambda)}{2} + (1-\lambda t_k) d_{\mu,\pi^*} B_{\omega}\left(\pi, \pi_k\right) - d_{\mu,\pi^*} B_{\omega}\left(\pi, \pi_{k+1}\right) + d_{\mu,\pi^*}\epsilon_k(\cdot, \pi).$$

where h_ω is defined in the third claim of Lemma 22.

Proof. By Assumption 1, for all s for which $d_{\mu,\pi^*}(s) > 0$ it also holds that $d_{\nu,\pi_k}(s) > 0$. Thus, for all s for which $d_{\mu,\pi^*}(s) > 0$ the component-wise relation in Lemma 13 holds. By multiplying each inequality by the positive number $d_{\mu,\pi^*}(s)$ and summing over all s we get,

$$0 \leq t_k d_{\mu,\pi^*}(T^{\pi}_{\lambda} v^{\pi_k}_{\lambda} - v^{\pi_k}_{\lambda}) + \frac{t_k^2 h_{\omega}^2(k;\lambda)}{2} + (1-\lambda t_k) d_{\mu,\pi^*} B_{\omega}\left(\pi, \pi_k\right) - d_{\mu,\pi^*} B_{\omega}\left(\pi, \pi_{k+1}\right) + d_{\mu,\pi^*}\epsilon_k(\cdot, \pi),$$

which concludes the proof.

□

Lemma 15 (fundamental inequality of Sample-Based TRPO.)**.** *Let $\{\pi_k\}_{k\geq 0}$ be the sequence generated by Aproximate TRPO using stepsizes $\{t_k\}_{k\geq 0}$. Then, for all $k \geq 0$*

$$t_k(1-\gamma)(\mu v^{\pi_k}_{\lambda} - \mu v^{\pi^*}_{\lambda}) \leq d_{\mu,\pi^*}((1-\lambda t_k) B_{\omega}\left(\pi^*, \pi_k\right) - B_{\omega}\left(\pi^*, \pi_{k+1}\right)) + \frac{t_k^2 h_{\omega}^2(k;\lambda)}{2} + d_{\mu,\pi^*}\epsilon_k,$$

where $h_\omega(k;\lambda)$ is defined in Lemma 22 and $\epsilon_k := \epsilon_k(\cdot,\pi^*)$ where the latter defined in (47).

Proof. Setting $\pi = \pi^*$ in Lemma 14 and denoting $\epsilon_k := \epsilon_k(\cdot,\pi^*)$, we get that for any k,

$$\begin{aligned} &- t_k d_{\mu,\pi^*}\left(T^{\pi^*}_{\lambda} v^{\pi_k}_{\lambda} - v^{\pi_k}_{\lambda}\right) \\ &\leq d_{\mu,\pi^*}((1-\lambda t_k) B_{\omega}\left(\pi^*, \pi_k\right) - B_{\omega}\left(\pi^*, \pi_{k+1}\right)) + \frac{t_k^2 h_{\omega}^2(k;\lambda)}{2} + d_{\mu,\pi^*}\epsilon_k. \end{aligned}$$

Furthermore, by the third claim of Lemma 25,

$$(1-\gamma)\mu(v^*_{\lambda} - v^{\pi_k}_{\lambda}) = d_{\mu,\pi^*}\left(T^{\pi^*}_{\lambda} v^{\pi_k}_{\lambda} - v^{\pi_k}_{\lambda}\right).$$

Combining the two relations on both sides we concludes the proof.

□

F.5 Approximation Error Bound

In this section we deal with the approximation error, the term $d_{\mu,\pi^*}\epsilon_k$ in Lemma 15. Two factors effects $d_{\mu,\pi^*}\epsilon_k$: (1) the error due to Monte-Carlo sampling, which we bound using Hoeffding's inequality and the union bound; (2) the error due to the truncation in the sampling process (see Appendix F.2). The next two lemmas bound these two sources of error. We first discuss the analysis of using an unbiased sampling process (Lemma 16), i.e., when no truncation is taking place, and then move to discuss the use of the truncated trajectories (Lemma 17). Finally, in Lemma 18 we combine the two results to bound $d_{\mu,\pi^*}\epsilon_k$ in the case of the full truncated sampling process discussed in Appendix F.2.

The unbiased q-function estimator uses a full unrolling of a trajectory, i.e., calculates the (possibly infinite) sum of retrieved costs following the policy π_k in the m-th trajecotry of the k-th iteration,

$$\hat{q}^{\pi_k}_{\lambda}(s_m, a_m, m) := \sum_{t=0}^{\infty} \gamma^t \Big(c\Big(s^{k,m}_t, a^{k,m}_t\Big) + \lambda\omega\left(s^{k,m}_t; \pi_k\right)\Big),$$

where the notation $s_t^{k,m}$ refer to the state encountered in the m-th trajectory of the k-th iteration, at the t step of estimating the $q_\lambda^{\pi_k}$ function. Moreover, $(s_m, a_m) = (s_0^{k,m}, a_0^{k,m})$ and $\hat{q}_\lambda^{\pi_k}(s, a, m) = 0$ for any $(s, a) \neq (s_m, a_m)$.

The truncated biased q-function estimator, truncates the trajectory after T interactions with the MDP, where T is predefined:

$$\hat{q}_{\lambda,\text{trunc}}^{\pi_k}(s, a, m) := \sum_{t=0}^{T-1} \gamma^t \Big(c\Big(s_t^{k,m}, a_t^{k,m}\Big) + \lambda\omega\left(s_t^{k,m}; \pi_k\right)\Big)$$

The following lemma describes the number of trajectories needed in the k-th update, in order to bound the error to be proportional to ϵ w.p. $1 - \delta'$, using an unbiased estimator.

Lemma 16 (Approximation error bound with unbiased sampling). *For any $\epsilon, \tilde{\delta} > 0$, if the number of trajectories in the k-th iteration is*

$$M_k \geq \frac{2 r_\omega(k, \lambda)^2}{\epsilon^2}\Big(S \log 2A + \log 1/\tilde{\delta}\Big),$$

then with probability of $1 - \tilde{\delta}$,

$$d_{\mu,\pi^*}\epsilon_k \leq t_k \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty \frac{\epsilon}{2},$$

where $r_\omega(k, \lambda) = \frac{4A\,\mathrm{C}_{max,\lambda}}{1-\gamma}$ and $r_\omega(k, \lambda) = \frac{4A\,\mathrm{C}_{max,\lambda}}{1-\gamma}(1 + \mathbb{1}\{\lambda \neq 0\}\log k)$ in the euclidean and non-euclidean settings respectively.

Proof. Plugging the definition of $\epsilon_k := \epsilon_k(\cdot, \pi^*)$ in (47), we get,

$$\begin{aligned}
&d_{\mu,\pi^*}\epsilon_k \\
&= \sum_s \frac{d_{\mu,\pi^*}(s)}{M_k d_{\nu,\pi_k}(s)} \sum_{m=1}^{M_k} \mathbb{1}\{s = s_m\} \langle t_k\left(A\hat{q}_\lambda^{\pi_k}(s_m, \cdot, m) + \lambda\nabla\omega\left(s_m; \pi_k\right)\right) + \nabla\omega\left(s; \pi_{k+1}\right) - \nabla\omega\left(s; \pi_k\right), \pi^*(\cdot\,|\,s) - \pi_{k+1}(\cdot\,|\,s_m)\rangle \\
&- \sum_s d_{\mu,\pi^*}(s) \langle t_k\left(q_\lambda^{\pi_k}(s, \cdot) + \lambda\nabla\omega\left(s; \pi_k\right)\right) + \nabla\omega\left(s; \pi_{k+1}\right) - \nabla\omega\left(s; \pi_k\right), \pi^*(\cdot \mid s) - \pi_{k+1}(\cdot \mid s)\rangle \\
&= \frac{1}{M_k} \sum_{m=1}^{M_k} \sum_s \mathbb{1}\{s = s_m\} \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \langle t_k\left(A\hat{q}_\lambda^{\pi_k}(s_m, \cdot, m) + \lambda\nabla\omega\left(s_m; \pi_k\right)\right) + \nabla\omega\left(s; \pi_{k+1}\right) - \nabla\omega\left(s; \pi_k\right), \pi^*(\cdot\,|\,s) - \pi_{k+1}(\cdot\,|\,s_m)\rangle \\
&- \sum_s d_{\nu,\pi_k}(s) \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \langle t_k\left(q_\lambda^{\pi_k}(s, \cdot) + \lambda\nabla\omega\left(s; \pi_k\right)\right) + \nabla\omega\left(s; \pi_{k+1}\right) - \nabla\omega\left(s; \pi_k\right), \pi^*(\cdot \mid s) - \pi_{k+1}(\cdot \mid s)\rangle,
\end{aligned}$$

where in the last transition we used the fact that for every $s \neq s_m$ the identity function $\mathbb{1}\{s = s_m\} = 0$.

We define,

$$\hat{X}_k(s_m, \cdot, m) := t_k\left(A\hat{q}_\lambda^{\pi_k}(s_m, \cdot, m) + \lambda\nabla\omega\left(s_m; \pi_k\right)\right) + \nabla\omega\left(s_m; \pi_{k+1}\right) - \nabla\omega\left(s_m; \pi_k\right), \tag{48}$$

$$X_k(s, \cdot) := t_k\left(q_\lambda^{\pi_k}(s, \cdot) + \lambda\nabla\omega\left(s; \pi_k\right)\right) + \nabla\omega\left(s; \pi_{k+1}\right) - \nabla\omega\left(s; \pi_k\right). \tag{49}$$

Using this definition, we have,

$$\begin{aligned}
d_{\mu,\pi^*}\epsilon_k =& \frac{1}{M_k} \sum_{m=1}^{M_k} \sum_s \mathbb{1}\{s = s_m\} \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \Big\langle \hat{X}_k(s_m, \cdot, m), \pi^*(\cdot \mid s_m) - \pi_{k+1}(\cdot \mid s_m) \Big\rangle \\
&- \sum_s d_{\mu,\pi^*}(s) \langle X_k(s, \cdot), \pi^*(\cdot \mid s) - \pi_{k+1}(\cdot \mid s)\rangle.
\end{aligned} \tag{50}$$

In order to remove the dependency on the randomness of π_{k+1}, we can bound this term in a uniform way:

$$\begin{aligned}
d_{\mu,\pi^*}\epsilon_k \leq& \max_{\pi'} \Big\{ \frac{1}{M_k} \sum_{m=1}^{M_k} \sum_s \mathbb{1}\{s = s_m\} \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \Big\langle \hat{X}_k(s_m, \cdot, m), \pi^*(\cdot \mid s_m) - \pi'(\cdot \mid s_m) \Big\rangle \\
&- \sum_s d_{\mu,\pi^*}(s) \langle X_k(s_m, \cdot), \pi^*(\cdot \mid s) - \pi'(\cdot \mid s)\rangle \Big\}.
\end{aligned} \tag{51}$$

In this lemma, we analyze the case where no truncation is taken into account. In this case we, we will now show that for any π'

$$\mathbb{E}\left[\sum_s \mathbb{1}\{s = s_m\}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\left\langle \hat{X}_k(s,\cdot,m), \pi^*(\cdot \mid s_m) - \pi'(\cdot \mid s_m)\right\rangle\right] = \sum_s d_{\mu,\pi^*}(s)\langle X_k(s,\cdot), \pi^*(\cdot \mid s) - \pi'(\cdot \mid s)\rangle,$$

which means that $\frac{1}{M_k}\sum_{m=1}^{M_k}\sum_s \mathbb{1}\{s = s_m\}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\left\langle \hat{X}_k(s,\cdot,m), \pi^*(\cdot \mid s_m) - \pi'(\cdot \mid s_m)\right\rangle$ is an unbiased estimator.

This fact comes from the from the following relations:

$$\begin{aligned}
&\mathbb{E}[\sum_s \mathbb{1}\{s = s_m\}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\left\langle \hat{X}_k(s,\cdot,m), \pi^*(\cdot \mid s_m) - \pi'(\cdot \mid s_m)\right\rangle] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_s \mathbb{1}\{s = s_m\}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\left\langle \hat{X}_k(s,\cdot,m), \pi^*(\cdot \mid s) - \pi'(\cdot \mid s)\right\rangle \mid s_m\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{d_{\mu,\pi^*}(s_m)}{d_{\nu,\pi_k}(s_m)}\left\langle \hat{X}_k(s_m,\cdot,m), \pi^*(\cdot \mid s_m) - \pi'(\cdot \mid s_m)\right\rangle \mid s_m\right]\right] \\
&= \mathbb{E}\left[\frac{d_{\mu,\pi^*}(s_m)}{d_{\nu,\pi_k}(s_m)}\mathbb{E}\left[\left\langle \hat{X}_k(s_m,\cdot,m), \pi^*(\cdot \mid s_m) - \pi'(\cdot \mid s_m)\right\rangle \mid s_m\right]\right] \\
&= \mathbb{E}\left[\frac{d_{\mu,\pi^*}(s_m)}{d_{\nu,\pi_k}(s_m)}\left\langle \mathbb{E}\left[\hat{X}_k(s_m,\cdot,m) \mid s_m\right], \pi^*(\cdot \mid s_m) - \pi'(\cdot \mid s_m)\right\rangle\right] \\
&= \mathbb{E}\left[\frac{d_{\mu,\pi^*}(s_m)}{d_{\nu,\pi_k}(s_m)}\langle X_k(s_m,\cdot), \pi^*(\cdot \mid s_m) - \pi'(\cdot \mid s_m)\rangle\right] \\
&= \sum_s d_{\nu,\pi_k}(s)\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\langle X_k(s,\cdot), \pi^*(\cdot \mid s) - \pi'(\cdot \mid s)\rangle \\
&= \sum_s d_{\mu,\pi^*}(s)\langle X_k(s,\cdot), \pi^*(\cdot \mid s) - \pi'(\cdot \mid s)\rangle, \qquad (52)
\end{aligned}$$

where the first transition is by law of total expectation; the second transition is by the fact the indicator function is zero for every $s \neq s_m$; the third transition is by the fact s_m is not random given s_m; the fourth transition is by the linearity of expectation and the fact that $\pi^*(\cdot \mid s_m) - \pi'(\cdot \mid s_m)$ is not random given s_m; the fifth transition is by taking the expectation of $\hat{X}$ in the state s_m; finally, the sixth transition is by explicitly taking the expectation over the probability that s_m is drawn from d_{ν,π_k} in the m-th trajectory (by following π_k from the restart distribution ν).

Meaning, (51) is a difference between an empirical mean of M_k random variables and their mean for a the fixed policy π', which maximizes the following expression

$$\begin{aligned}
d_{\mu,\pi^*}\epsilon_k \leq &\max_{\pi'}\left\{\frac{1}{M_k}\sum_{m=1}^{M_k}\sum_s \mathbb{1}\{s = s_m\}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\left\langle \hat{X}_k(s,\cdot,m), \pi^*(\cdot \mid s) - \pi'(\cdot \mid s)\right\rangle\right. \\
&\left. - \mathbb{E}\left[\sum_s \mathbb{1}\{s = s_m\}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\left\langle \hat{X}_k(s,\cdot,m), \pi^*(\cdot \mid s) - \pi'(\cdot \mid s)\right\rangle\right]\right\}. \qquad (53)
\end{aligned}$$

As we wish to obtain a uniform bound on π', we can use the common approach of bounding (53) uniformly for all $\pi' \in \Delta_{\mathcal{A}}^{\mathrm{S}}$ using the union bound. Note that the above optimization problem is a linear programming optimization problem in π', where $\pi' \in \Delta_{\mathcal{A}}^{\mathrm{S}}$. It is a well known fact that for linear programming, there is an extreme point which is the optimal solution of the problem [Bertsimas and Tsitsiklis, 1997][Theorem 2.7]. The set of extreme points of $\Delta_{\mathcal{A}}^{\mathrm{S}}$ is the set of all deterministic policies denoted by Π^{det}. Therefore, in order to bound the maximum in (53), it suffices to uniformly bound all policies $\pi' \in \Pi^{\mathrm{det}}$.

Now, notice that $\frac{d_{\mu,\pi^*}(s_m)}{d_{\nu,\pi_k}(s_m)}\left\langle \hat{X}_k(s_m,\cdot,m), \pi^*(\cdot \mid s_m) - \pi'(\cdot \mid s_m)\right\rangle$ is bounded for all s_m and π',

$$
\begin{aligned}
&\frac{d_{\mu,\pi^*}(s_m)}{d_{\nu,\pi_k}(s_m)}\left\langle \hat{X}_k(s_m,\cdot,m), \pi^*(\cdot \mid s_m) - \pi'(\cdot \mid s_m)\right\rangle \\
&= \left\langle \frac{d_{\mu,\pi^*}(s_m)}{d_{\nu,\pi_k}(s_m)}\hat{X}_k(s_m,\cdot,m), \pi^*(\cdot \mid s_m) - \pi'(\cdot \mid s_m)\right\rangle \\
&\le \left\| \frac{d_{\mu,\pi^*}(s_m)}{d_{\nu,\pi_k}(s_m)}\hat{X}_k(s_m,\cdot,m)\right\|_\infty \left\|\pi^*(\cdot \mid s_m) - \pi'(\cdot \mid s_m)\right\|_1 \\
&\le 2\frac{d_{\mu,\pi^*}(s_m)}{d_{\nu,\pi_k}(s_m)}\left\|\hat{X}_k(s_m,\cdot,m)\right\|_\infty \\
&\le 2\left\|\frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}}\right\|_\infty \left\|\hat{X}_k(s_m,\cdot,m)\right\|_\infty \\
&= 2\left\|\frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}}\right\|_\infty \left\|t_k\left(A\hat{q}_\lambda^{\pi_k}(s_m,\cdot,m) + \lambda\nabla\omega\left(s_m;\pi_k\right)\right) + \nabla\omega\left(s_m;\pi_{k+1}\right) - \nabla\omega\left(s_m;\pi_k\right)\right\|_\infty \\
&\le 2\left\|\frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}}\right\|_\infty \left(t_k\left\|A\hat{q}_\lambda^{\pi_k}(s_m,\cdot,m) + \lambda\nabla\omega\left(s_m;\pi_k\right)\right\|_\infty + \left\|\nabla\omega\left(s_m;\pi_{k+1}\right) - \nabla\omega\left(s_m;\pi_k\right)\right\|_\infty\right) \\
&\le 2t_k\left\|\frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}}\right\|_\infty \left(\hat{h}_\omega(k;\lambda) + 2A_\omega(k)\right) \\
&= 2\left\|\frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}}\right\|_\infty \left(t_k\hat{h}_\omega(k;\lambda) + A_\omega(k)\right) \\
&:= t_k\left\|\frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}}\right\|_\infty r_\omega(k,\lambda), \qquad (54)
\end{aligned}
$$

where the second transition is due to Hölder's inequality; the third transition is due to the bound of the TV distance between two random variables; the sixth transition is due to the triangle inequality; finally, the seventh transition is by plugging in the bounds in Lemma 22 and Lemma 23. Also, we defined $r_\omega(k,\lambda) = \frac{4A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}$ and $r_\omega(k,\lambda) = \frac{4A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}(1 + \mathbb{1}\{\lambda \neq 0\}\log k)$ in the euclidean and non-euclidean cases respectively.

Thus, by Hoeffding and the union bound over the set of deterministic policies,

$$P\left(d_{\mu,\pi^*}\epsilon_k \ge t_k\left\|\frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}}\right\|_\infty \frac{\epsilon}{2}\right) \le 2|\Pi^{\text{det}}|\exp\left(-\frac{M_k\epsilon^2}{2r_\omega(k,\lambda)^2}\right) = \tilde{\delta}.$$

In other words, in order to guarantee that

$$d_{\mu,\pi^*}\epsilon_k \le t_k\left\|\frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}}\right\|_\infty \frac{\epsilon}{2},$$

we need the number of trajectories M_k to be at least

$$M_k \ge \frac{2r_\omega(k,\lambda)^2}{\epsilon^2}\left(S\log 2A + \log 1/\tilde{\delta}\right),$$

where we used the fact that there are $|\Pi^{\text{det}}| = A^S$ deterministic policies.

which concludes the result.

□

The following lemma described with error due to the use of truncated trajectories:

Lemma 17 (Truncation error bound)**.** *The bias of the truncated sampling process in the k-th iteration, with maximal trajectory length of $T = \frac{1}{1-\gamma}\log\frac{\epsilon}{8r_\omega(k,\lambda)}$ is $t_k\left\|\frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}}\right\|_\infty \frac{\epsilon}{4}$, where $r_\omega(k,\lambda) = \frac{4A\,\mathrm{C}_{max,\lambda}}{1-\gamma}$ and $r_\omega(k,\lambda) = \frac{2A\,\mathrm{C}_{max,\lambda}}{1-\gamma}\left(\frac{1}{1-\lambda t_k} + 1 + \lambda\log k\right)$ in the euclidean and non-euclidean settings respectively.*

Proof. We start this proof by defining notation related to the truncated sampling process. First, denote $d^{\text{trunc}}_{\nu,\pi_k}(s)$, the probability to choose a state s, using the truncated biased sampling process of length T, as described in Appendix F.2. Observe that

$$d^{\text{trunc}}_{\nu,\pi_k}(s) = (1-\gamma)\sum_{t=0}^{T-1}\gamma^t p(s_t = s \mid \nu, \pi_k) + \gamma^T p(s_T = s \mid \nu, \pi_k)$$

We also make use in this proof in the following definitions (see (48) and (49)),

$$\begin{aligned}\hat{X}_k(s_m,\cdot,m) &:= t_k\left(A\hat{q}^{\pi_k}_{\lambda,\text{trunc}}(s_m,\cdot,m) + \lambda\nabla\omega\left(s_m;\pi_k\right)\right) + \nabla\omega\left(s_m;\pi_{k+1}\right) - \nabla\omega\left(s_m;\pi_k\right),\\ X_k(s,\cdot) &:= t_k\left(q^{\pi_k}_\lambda(s,\cdot) + \lambda\nabla\omega\left(s;\pi_k\right)\right) + \nabla\omega\left(s;\pi_{k+1}\right) - \nabla\omega\left(s;\pi_k\right).\end{aligned}$$

Lastly, we denote the expectation of $\hat{X}_k(s,\cdot,m)$ using the truncated sampling process as $X^{\text{trunc}}_k(s,\cdot)$,

$$X^{\text{trunc}}_k(s,a) = \mathbb{E}\hat{X}_k(s,a,m)$$

Now, we move on to the proof. We first split the bias to two different sources of bias:

$$\begin{aligned}&\mathbb{E}_{s\sim d^{\text{trunc}}_{\nu,\pi_k}}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\left\langle X^{\text{trunc}}_k(s,\cdot),\pi(\cdot\mid s)-\pi'(\cdot\mid s)\right\rangle - \mathbb{E}_{s\sim d_{\nu,\pi_k}}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\langle X_k(s,\cdot),\pi(\cdot\mid s)-\pi'(\cdot\mid s)\rangle\\ &= \left(\mathbb{E}_{s\sim d^{\text{trunc}}_{\nu,\pi_k}}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\left\langle X^{\text{trunc}}_k(s,\cdot),\pi(\cdot\mid s)-\pi'(\cdot\mid s)\right\rangle - \mathbb{E}_{s\sim d_{\nu,\pi_k}}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\left\langle X^{\text{trunc}}_k(s,\cdot),\pi(\cdot\mid s)-\pi'(\cdot\mid s)\right\rangle\right)\\ &+\left(\mathbb{E}_{s\sim d_{\nu,\pi_k}}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\left\langle X^{\text{trunc}}_k(s,\cdot),\pi(\cdot\mid s)-\pi'(\cdot\mid s)\right\rangle - \mathbb{E}_{s\sim d_{\nu,\pi_k}}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\langle X_k(s,\cdot),\pi(\cdot\mid s)-\pi'(\cdot\mid s)\rangle\right).\end{aligned}$$

The first source of bias is due to the truncation of the state sampling after T iterations, and the second source of bias is due to the truncation done in the estimation of $q^{\pi_k}_\lambda(s,a)$, for the chosen state s and action a.

First, we bound the first error term. Observe that for any s,

$$\begin{aligned}\sum_s\left|d^{\text{trunc}}_{\nu,\pi_k}(s)-d_{\nu,\pi_k}(s)\right| &= \sum_s\left|(1-\gamma)\sum_{t=0}^{T-1}\gamma^t p(s_t=s\mid\nu,\pi_k)+\gamma^T p(s_T=s\mid\nu,\pi_k)-(1-\gamma)\sum_{t=0}^{\infty}\gamma^t p(s_t=s\mid\nu,\pi_k)\right|\\ &= \sum_s\left|\gamma^T p(s_T=s\mid\nu,\pi_k)-(1-\gamma)\sum_{t=T}^{\infty}\gamma^t p(s_t=s\mid\nu,\pi_k)\right|\\ &\le \sum_s\left|\gamma^T p(s_T=s\mid\nu,\pi_k)\right|+\sum_s\left|(1-\gamma)\sum_{t=T}^{\infty}\gamma^t p(s_t=s\mid\nu,\pi_k)\right|\\ &= \sum_s\gamma^T p(s_T=s\mid\nu,\pi_k)+\sum_s(1-\gamma)\sum_{t=T}^{\infty}\gamma^t p(s_t=s\mid\nu,\pi_k)\\ &= \gamma^T\sum_s p(s_T=s\mid\nu,\pi_k)+\ (1-\gamma)\sum_{t=T}^{\infty}\gamma^t\sum_s p(s_t=s\mid\nu,\pi_k)\\ &\le \gamma^T+(1-\gamma)\sum_{t=T}^{\infty}\gamma^t\\ &= 2\gamma^T \qquad (55)\end{aligned}$$

where the third transition is due to the triangle inequality, the fourth transition is due to the fact that for any t, $\gamma^t p(s_t\mid\nu,\pi_k)\ge 0$ and the sixth transition is by the fact that $\sum_s p(s_t=s|\nu,\pi_k)\le 1$ for any t as a probability distribution.

Thus,

$$
\begin{aligned}
&\mathbb{E}_{s\sim d^{\text{trunc}}_{\nu,\pi_k}} \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \left\langle X^{\text{trunc}}_k(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s) \right\rangle - \mathbb{E}_{s\sim d_{\nu,\pi_k}} \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \left\langle X^{\text{trunc}}_k(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s) \right\rangle \\
&= \sum_s \left(d^{\text{trunc}}_{\nu,\pi_k}(s) - d_{\nu,\pi_k}(s) \right) \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \left\langle X^{\text{trunc}}_k(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s) \right\rangle \\
&\le \sum_s \left| d^{\text{trunc}}_{\nu,\pi_k}(s) - d_{\nu,\pi_k}(s) \right| \left| \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \left\langle X^{\text{trunc}}_k(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s) \right\rangle \right| \\
&\le \max_s \left| \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \left\langle X^{\text{trunc}}_k(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s) \right\rangle \right| \sum_s \left| d^{\text{trunc}}_{\nu,\pi_k}(s) - d_{\nu,\pi_k}(s) \right| \\
&\le 2\gamma^T \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty \max_s \left| \left\langle X^{\text{trunc}}_k(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s) \right\rangle \right| \\
&\le \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty t_k r_\omega(k,\lambda) 2\gamma^T,
\end{aligned}
$$

where the fourth transition is by plugging in (55) and the last transition is by repeating similar analysis to (54).

Now, by simple arithmetic, for any $\epsilon > 0$, if the trajectory length $T > \frac{1}{1-\gamma} \log \frac{\epsilon}{16 r_\omega(k,\lambda)}$, we get that the first bias term is bounded,

$$
\begin{aligned}
&\mathbb{E}_{s\sim d^{\text{trunc}}_{\nu,\pi_k}} \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \left\langle X^{\text{trunc}}_k(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s) \right\rangle - \mathbb{E}_{s\sim d_{\nu,\pi_k}} \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \left\langle X^{\text{trunc}}_k(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s) \right\rangle \\
&\le \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty t_k \frac{\epsilon}{8} \qquad (56)
\end{aligned}
$$

Next, we bound the second error term.

First, observe that for any s, a,

$$
\begin{aligned}
&\left| \mathbb{E}\hat{q}^{\pi_k}_{\lambda,\text{trunc}}(s,a,m) - q^{\pi_k}_\lambda(s,a) \right| \\
&= \left| \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t (c_t(s_t,a_t) + \lambda\omega(s_t;\pi_k)) \mid s_0 = s, a_0 = a \right] - \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (c_t(s_t,a_t) + \lambda\omega(s_t;\pi_k)) \mid s_0 = s, a_0 = a \right] \right| \\
&= \left| \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t (c_t(s_t,a_t) + \lambda\omega(s_t;\pi_k)) - \sum_{t=0}^{\infty} \gamma^t (c_t(s_t,a_t) + \lambda\omega(s_t;\pi_k)) \mid s_0 = s, a_0 = a \right] \right| \\
&= \left| \mathbb{E}\left[\sum_{t=T}^{\infty} \gamma^t (c_t(s_t,a_t) + \lambda\omega(s_t;\pi_k)) \mid s_0 = s, a_0 = a \right] \right| \\
&\le \gamma^T \frac{\mathrm{C}_{\max,\lambda}}{1-\gamma} \qquad (57)
\end{aligned}
$$

Now,

$$
\begin{aligned}
&\mathbb{E}_{s\sim d_{\nu,\pi_k}} \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \big\langle X_k^{\text{trunc}}(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s)\big\rangle - \mathbb{E}_{s\sim d_{\nu,\pi_k}} \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \langle X_k(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s)\rangle \\
&= \mathbb{E}_{s\sim d_{\nu,\pi_k}} \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \big\langle X_k^{\text{trunc}}(s,\cdot) - X_k(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s)\big\rangle \\
&= \sum_s d_{\nu,\pi_k}(s) \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \big\langle X_k^{\text{trunc}}(s,\cdot) - X_k(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s)\big\rangle \\
&\le \max_s \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \big\langle X_k^{\text{trunc}}(s,\cdot) - X_k(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s)\big\rangle \\
&\le t_k \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty \max_s \big\langle X_k^{\text{trunc}}(s,\cdot) - X_k(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s)\big\rangle \\
&= t_k \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty \max_s \Big\langle \mathbb{E}\hat{q}_{\lambda,\text{trunc}}^{\pi_k}(s,\cdot,m) - q_\lambda^{\pi_k}(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s)\Big\rangle \\
&\le t_k \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty \max_s \left\| \mathbb{E}\hat{q}_{\lambda,\text{trunc}}^{\pi_k}(s,\cdot,m) - q_\lambda^{\pi_k}(s,\cdot) \right\|_\infty \|\pi(\cdot \mid s) - \pi'(\cdot \mid s)\|_1 \\
&\le 2 \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty t_k \frac{\mathrm{C}_{\max,\lambda}}{1-\gamma} \gamma^T,
\end{aligned}
$$

where the first transition is due to the linearity of expectation, the third transition is by the fact the summation of d_{ν,π_k} is convex, the fourth transition is by the fact $\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}$ is non-negative for any s and by maximizing each term separately, the fifth transition is by using the definitions of X_k and X_k^{trunc}, the sixth is using Hölder's inequality and the last transition is due to (57).

Now, using the same T, by the fact $r_\omega(k,\lambda) > \frac{2\,\mathrm{C}_{\max,\lambda}}{1-\gamma}$, we have that

$$
\begin{aligned}
&\mathbb{E}_{s\sim d_{\nu,\pi_k}} \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \big\langle X_k^{\text{trunc}}(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s)\big\rangle - \mathbb{E}_{s\sim d_{\nu,\pi_k}} \frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)} \big\langle X_k^{\text{trunc}}(s,\cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s)\big\rangle \\
&\le \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty t_k \frac{\epsilon}{8}. \qquad (58)
\end{aligned}
$$

Finally, combining (56) and (58) concludes the results. □

In the next lemma we combine the results of Lemmas 16 and 17 to bound the overall approximation error due to both sampling and truncation.

Lemma 18 (Approximation error bound using truncated biased sampling)**.** *For any $\epsilon, \tilde{\delta} > 0$, if the number of trajectories in the k-th iteration is*

$$
M_k \ge \frac{8 r_\omega(k,\lambda)^2}{\epsilon^2} \Big(S \log 2A + \log 1/\tilde{\delta} \Big),
$$

and the number of samples in the truncated sampling process is of length

$$
T_k \ge \frac{1}{1-\gamma} \log \frac{\epsilon}{8 r_\omega(k,\lambda)},
$$

then with probability of $1 - \tilde{\delta}$,

$$
d_{\mu,\pi^*} \epsilon_k \le t_k \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty \frac{\epsilon}{2},
$$

and the overall number of interaction with the MDP is in the k-th iteration is

$$
O\left(\frac{r_\omega(k,\lambda)^2 \Big(S \log A + \log 1/\tilde{\delta} \Big)}{(1-\gamma)\epsilon^2} \right),
$$

where $r_\omega(k,\lambda) = \frac{4A\,\mathrm{C}_{max,\lambda}}{1-\gamma}$ *and* $r_\omega(k,\lambda) = \frac{2A\,\mathrm{C}_{max,\lambda}}{1-\gamma}\Big(\frac{1}{1-\lambda t_k} + 1 + \lambda \log k\Big)$ *in the euclidean and non-euclidean settings respectively.*

Proof. Repeating the same steps of Lemma 16, we re-derive equation (51),

$$d_{\mu,\pi^*}\epsilon_k \le \max_{\pi'}\Big\{\frac{1}{M_k}\sum_{m=1}^{M_k}\sum_s \mathbb{1}\{s=s_m\}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\Big\langle \hat{X}_k(s,\cdot,m), \pi^*(\cdot\mid s_m)-\pi'(\cdot\mid s_m)\Big\rangle \\ -\sum_s d_{\mu,\pi^*}(s)\langle X_k(s,\cdot), \pi^*(\cdot\mid s)-\pi'(\cdot\mid s)\rangle\Big\}.$$

Now, we move on to deal with a truncated trajectory: In Appendix F.2 we defined a nearly unbiased estimation process for $q_\lambda^{\pi_k}$, i.e., $\frac{1}{M_k}\sum_{m=1}^{M_k}\sum_s \mathbb{1}\{s=s_m\}\frac{d_{\mu,\pi^*}(s)}{d_{\nu,\pi_k}(s)}\Big\langle \hat{X}_k(s,\cdot,m), \pi^*(\cdot\mid s_m)-\pi'(\cdot\mid s_m)\Big\rangle$ is no longer an unbiased estimator as in Lemma 16. In what follows we divide the error to two sources of error, one due to the finite sampling error (finite number of trajectories) and the other due to the bias admitted by the truncation.

For any π', denote the following variables,

$$\hat{Y}_m(\pi') := \frac{d_{\mu,\pi^*}(s_m)}{d_{\nu,\pi_k}(s_m)}\langle \hat{X}_k(s_m,\cdot,m), \pi^*(\cdot\mid s_m)-\pi'(\cdot\mid s_m)\rangle \\ Y(\pi') := \sum_s d_{\mu,\pi^*}(s)\langle X_k(s,\cdot), \pi^*(\cdot\mid s)-\pi'(\cdot\mid s)\rangle.$$

By plugging this new notation in (51), we can write,

$$\begin{aligned} d_{\mu,\pi^*}\epsilon_k &\le \max_{\pi'}\frac{1}{M}\sum_{m=1}^{M}\hat{Y}_m(\pi') - Y(\pi') \\ &= \max_{\pi'}\frac{1}{M}\sum_{m=1}^{M}\hat{Y}_m(\pi') - \mathbb{E}\hat{Y}_m(\pi') + \mathbb{E}\hat{Y}_m(\pi') - Y(\pi') \\ &\le \underbrace{\max_{\pi'}\frac{1}{M}\sum_{m=1}^{M}\hat{Y}_m(\pi') - \mathbb{E}\hat{Y}_m(\pi')}_{(1)} + \underbrace{\max_{\pi'}\mathbb{E}\hat{Y}_m(\pi') - Y(\pi')}_{(2)}, \end{aligned} \tag{59}$$

where the first inequality is by plugging in the definition of $Y(\pi'), \hat{Y}_M(\pi')$ in (51) and the last transition is by maximizing each of the terms in the sum independently. Note that (1) describes the error due to the finite sampling and (2) describes the error due to the truncation of the trajectories. Importantly, notice that in the case where we do not truncate the trajectory, the second term (2) equals zero by (52). We will now use Lemma 16 and Lemma 17 to bound (1) and (2) respectively:

First, look at the **first term** (1). By definition it an unbiased estimation process. Furthermore, by equation (54), $\hat{Y}_m(\pi')$ is bounded for all s_m and π' by

$$\hat{Y}_m(\pi') \le t_k \left\|\frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}}\right\|_\infty r_\omega(k,\lambda),$$

Thus by applying Lemma 16 we get that in order to guarantee that

$$\max_{\pi'}\frac{1}{M}\sum_{m=1}^{M}\Big(\hat{Y}_m(\pi') - \mathbb{E}\hat{Y}_m(\pi')\Big) \le t_k\left\|\frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}}\right\|_\infty \frac{\epsilon}{4}, \tag{60}$$

we need the number of trajectories M_k to be at least

$$M_k \ge \frac{8 r_\omega(k,\lambda)^2}{\epsilon^2}\Big(S\log 2A + \log 1/\tilde{\delta}\Big).$$

Next, we bound the **second term** (2). By Lemma 17, using a trajectory of maximal length $\frac{1}{1-\gamma}\log\frac{\epsilon}{8r_\omega(k,\lambda)}$, the errors due to the truncated estimation process are bounded as follows,

$$\max_{\pi'} \mathbb{E}\hat{Y}_m(\pi') - Y(\pi') \le t_k \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty \frac{\epsilon}{4} \tag{61}$$

Bounding the two terms by (60) and (61), and plugging them back in (59), we get that using M_k trajectories, where each trajectory is of length $O(\frac{1}{1-\gamma}\log\epsilon)$, we have that w.p. $1-\tilde{\delta}$

$$d_{\mu,\pi^*}\epsilon_k \le t_k \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty \frac{\epsilon}{4} + t_k \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty \frac{\epsilon}{4} = t_k \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty \frac{\epsilon}{2},$$

which concludes the result.

□

So far, we proved the number of samples needed for a bounded error with high probability in the k-th iteration of Sample-Based TRPO. The following Lemma gives a bound for the accumulative error of Sample-Based TRPO after k iterations.

Lemma 19 (Cumulative approximation error)**.** *For any $\epsilon, \delta > 0$, if the number of trajectories in the k-th iteration is*

$$M_k \ge \frac{8r_\omega(N,\lambda)^2}{\epsilon^2}\left(S\log 2A + \log 2(k+1)^2/\delta\right),$$

and the number of samples in the truncated sampling process is of length

$$T \ge \frac{1}{1-\gamma}\log\frac{\epsilon}{8r_\omega(k,\lambda)},$$

then, with probability greater than $1-\delta$, uniformly on all $k \in \mathbb{N}$,

$$\sum_{k=0}^{N} d_{\mu,\pi^*}\epsilon_k \le \frac{\epsilon/2}{1-\gamma}\left\| \frac{d_{\mu,\pi^*}}{\nu} \right\|_\infty \sum_{k=0}^{N} t_k,$$

where $r_\omega(k,\lambda) = \frac{4A\,\mathrm{C}_{max,\lambda}}{1-\gamma}$ and $r_\omega(k,\lambda) = \frac{4A\,\mathrm{C}_{max,\lambda}}{1-\gamma}(1+\mathbb{1}\{\lambda \ne 0\}\log k)$ in the euclidean and non-euclidean settings respectively.

Proof. Using Lemma 18 with $\tilde{\delta} = \frac{6}{\pi^2}\frac{\delta}{(k+1)^2}$ and the union bound over all $k \in \mathbb{N}$, we get that w.p. bigger than

$$\sum_{k=0}^{\infty} \frac{6}{\pi^2}\frac{\delta}{(k+1)^2} = \frac{6}{\pi^2}\delta\sum_{k=0}^{\infty}\frac{1}{(k+1)^2} = \delta,$$

for any k, the following inequality holds

$$d_{\mu,\pi^*}\epsilon_k \le t_k \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty \frac{\epsilon}{2}.$$

where we used the solution to Basel's problem (the sum of reciprocals of the squares of the natural numbers) for calculating $\sum_{k=0}^{\infty}\frac{1}{(k+1)^2}$.

Thus, by summing the inequalities for $k = 0, 1, ..., N$, we obtain

$$\sum_{k=0}^{N} d_{\mu,\pi^*}\epsilon_k \le \frac{\epsilon}{2}\sum_{k=0}^{N} t_k \left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty.$$

Now, Using the fact that $\left\| \frac{d_{\mu,\pi^*}}{d_{\nu,\pi_k}} \right\|_\infty \le \frac{1}{1-\gamma}\left\| \frac{d_{\mu,\pi^*}}{\nu} \right\|_\infty$, we have that w.p. of at least δ,

$$\sum_{k=0}^{N} d_{\mu,\pi^*}\epsilon_k \le \frac{\epsilon/2}{1-\gamma}\left\| \frac{d_{\mu,\pi^*}}{\nu} \right\|_\infty \sum_{k=0}^{N} t_k.$$

Lastly, by bounding $\pi^2/6 \le 2$ we conclude the proof.

□

We are ready to prove the convergence rates for the unregularized and regularized algorithms.

F.6 Proof of Theorem 4

For the sake of completeness and readability, we restate here Theorem 4, this time including all logarithmic factors, but excluding higher orders in λ (All constants are in the proof):

Theorem (Convergence Rate: Sample-Based TRPO)**.** *Let* $\{\pi_k\}_{k\geq 0}$ *be the sequence generated by Sample-Based TRPO, using* $M_k \geq \frac{r_\omega(N,\lambda)^2}{2\epsilon^2}\left(S\log 2A + \log \pi^2(k+1)^2/6\delta\right)$ *trajectories in each iteration, and* $\{\mu v_{\textbf{\textit{best}}}^k\}_{k\geq 0}$ *be the sequence of best achieved values,* $\mu v_{\textbf{\textit{best}}}^N := \arg\min_{k=0,\dots,N} \mu v_\lambda^{\pi_k} - \mu v_\lambda^*$*. Then, with probability greater than* $1-\delta$ *for every* $\epsilon > 0$ *the following holds for all* $N \geq 1$.

1. (Unregularized) *Let* $\lambda = 0$, $t_k = \frac{(1-\gamma)^2}{C_{\omega,1}\,\mathrm{C}_{\textit{max}}\sqrt{k+1}}$ *then*

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

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

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

Where $C_{\omega,1} = \sqrt{A}, C_{\omega,2} = 1, C_{\omega,3} = 1, r_\omega(k,\lambda) = \frac{4A\,\mathrm{C}_{\textit{max},\lambda}}{1-\gamma}$ *for the euclidean case, and* $C_{\omega,1} = 1, C_{\omega,2} = A^2, C_{\omega,3} = \log A, r_\omega(k,\lambda) = \frac{4A\,\mathrm{C}_{\textit{max},\lambda}}{1-\gamma}(1 + \mathbb{1}\{\lambda \neq 0\}\log k)$ *for the non-euclidean case.*

Finally, we prove the rates for each of the cases:

The Unregularized Case

Proof. Applying Lemma 15 and $\lambda = 0$ (the unregularized case),

$$\begin{aligned}
&t_k(1-\gamma)(\mu v^{\pi_k} - \mu v^*) \\
&\leq d_{\mu,\pi^*}\left(B_\omega\left(\pi^*,\pi_k\right) - B_\omega\left(\pi^*,\pi_{k+1}\right)\right) + \frac{t_k^2 h_\omega^2}{2} + d_{\mu,\pi^*}\epsilon_k.
\end{aligned}$$

Summing the above inequality over $k = 0, 1, \dots, N$, gives

$$\begin{aligned}
&\sum_{k=0}^N t_k(1-\gamma)(\mu v^{\pi_k} - \mu v^*) \\
&\leq d_{\mu,\pi^*}B_\omega\left(\pi^*,\pi_0\right) - d_{\mu,\pi^*}B_\omega\left(\pi^*,\pi_{N+1}\right) + \sum_{k=0}^N \frac{t_k^2 h_\omega^2}{2} + \sum_{k=0}^N d_{\mu,\pi^*}\epsilon_k \\
&\leq d_{\mu,\pi^*}B_\omega\left(\pi^*,\pi_0\right) + \sum_{k=0}^N \frac{t_k^2 h_\omega^2}{2} + \sum_{k=0}^N d_{\mu,\pi^*}\epsilon_k \\
&\leq D_\omega + \sum_{k=0}^N \frac{t_k^2 h_\omega^2}{2} + \sum_{k=0}^N d_{\mu,\pi^*}\epsilon_k.
\end{aligned}$$

where in the second relation we used $B_\omega\left(\pi^*,\pi_{N+1}\right) \geq 0$ and thus $d_{\mu,\pi^*}B_\omega\left(\pi^*,\pi_{N+1}\right) \geq 0$, and in the third relation Lemma 24.

Using the definition of v^N_{best}, we have that

$$\mu(v^N_{\text{best}} - v^*) \sum_{k=0}^{N} t_k \le \sum_{k=0}^{N} t_k(\mu v^{\pi_k} - \mu v^*),$$

and by some algebraic manipulations, we get

$$\begin{aligned}\mu v^N_{\text{best}} - \mu v^* &\le \frac{1}{1-\gamma} \frac{D_\omega + \sum\limits_{k=0}^{N} \frac{t_k^2 h_\omega^2}{2} + \sum\limits_{k=0}^{N} d_{\mu,\pi^*}\epsilon_k}{\sum\limits_{k=0}^{N} t_k} \\ &= \frac{1}{1-\gamma} \frac{D_\omega + \frac{h_\omega^2}{2}\sum\limits_{k=0}^{N} t_k^2}{\sum\limits_{k=0}^{N} t_k} + \frac{1}{1-\gamma} \frac{\sum\limits_{k=0}^{N} d_{\mu,\pi^*}\epsilon_k}{\sum\limits_{k=0}^{N} t_k},\end{aligned}$$

Plugging in the stepsizes $t_k = \frac{1}{h_\omega \sqrt{k}}$, we get,

$$\mu v^N_{\textbf{best}} - \mu v^* \le \frac{h_\omega}{1-\gamma} \frac{2D_\omega + \sum\limits_{k=0}^{N} \frac{1}{k}}{2\sum\limits_{k=0}^{N} \frac{1}{\sqrt{k}}} + \frac{1}{1-\gamma} \frac{\sum\limits_{k=0}^{N} d_{\mu,\pi^*}\epsilon_k}{\sum\limits_{k=0}^{N} t_k}.$$

Bounding the sums using (Beck, 2017, Lemma 8.27(a)) yields,

$$\mu v^N_{\textbf{best}} - \mu v^* \le O\left(\frac{h_\omega}{1-\gamma} \frac{D_\omega + \log N}{\sum\limits_{k=0}^{N} \frac{1}{\sqrt{k}}} + \frac{1}{\sum_{k=0}^{N} t_k} \frac{1}{1-\gamma} \sum_{k=0}^{N} d_{\mu,\pi^*}\epsilon_k \right).$$

Plugging in Lemma 19, we get that for any (ϵ, δ), if the number of trajectories in the k-th iteration is

$$M_k \ge \frac{r_\omega(N,\lambda)^2}{2\epsilon^2}\big(S \log 2A + \log \pi^2 (k+1)^2/6\delta\big),$$

then, with probability greater than $1-\delta$,

$$\mu v^N_{\textbf{best}} - \mu v^* \le O\left(\frac{h_\omega}{1-\gamma} \frac{D_\omega + \log N}{\sum\limits_{k=0}^{N} \frac{1}{\sqrt{k}}} + \frac{1}{\sum_{k=0}^{N} t_k} \frac{\epsilon}{(1-\gamma)^2} \left\| \frac{d_{\mu,\pi^*}}{\nu} \right\|_\infty \sum_{k=0}^{N} t_k \right),$$

where $r_\omega(k,\lambda) = \frac{4A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}$ and $r_\omega(k,\lambda) = \frac{4A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}(1 + \mathbb{1}\{\lambda \neq 0\}\log k)$ in the euclidean and non-euclidean settings respectively.

By rearranging, we get,

$$\mu v^N_{\textbf{best}} - \mu v^* \le O\left(\frac{h_\omega}{1-\gamma} \frac{D_\omega + \log N}{\sum\limits_{k=0}^{N} \frac{1}{\sqrt{k}}} + \frac{\epsilon}{(1-\gamma)^2} \left\| \frac{d_{\mu,\pi^*}}{\nu} \right\|_\infty \right).$$

Thus, for the euclidean case,

$$\mu v_{\textbf{best}}^N - \mu v^* \leq O\Bigg(\frac{\mathrm{C}_{\max} \sqrt{A} \log N}{(1-\gamma)^2 \sqrt{N}} + \frac{1}{(1-\gamma)^2} \left\| \frac{d_{\mu,\pi^*}}{\nu} \right\| \epsilon \Bigg),$$

and for the non-euclidean case,

$$\mu v_{\textbf{best}}^N - \mu v^* \leq O\Big(\frac{\mathrm{C}_{\max} (\log A + \log N)}{(1-\gamma)^2 \sqrt{N}} + \frac{1}{(1-\gamma)^2} \left\| \frac{d_{\mu,\pi^*}}{\nu} \right\| \epsilon \Big).$$

□

The Regularized Case

Proof. Applying Lemma 15 and setting $t_k = \frac{1}{\lambda(k+2)}$, we get,

$$\begin{aligned}
&\frac{1-\gamma}{\lambda(k+2)} \left(\mu v_\lambda^{\pi_k} - \mu v_\lambda^{\pi^*} \right) \\
&\leq d_{\mu,\pi^*} \Bigg((1 - \frac{1}{(k+2)}) B_\omega(\pi^*, \pi_k) - B_\omega(\pi^*, \pi_{k+1}) \Bigg) + \frac{h_\omega^2(k;\lambda)}{2\lambda^2 (k+2)^2} + d_{\mu,\pi^*} \epsilon_k \\
&\leq d_{\mu,\pi^*} \Bigg(\frac{k+1}{k+2} B_\omega(\pi^*, \pi_k) - B_\omega(\pi^*, \pi_{k+1}) \Bigg) + \frac{h_\omega^2(N;\lambda)}{2\lambda^2 (k+2)^2} + d_{\mu,\pi^*} \epsilon_k,
\end{aligned}$$

where in the second relation we used that fact $h_\omega(k;\lambda)$ is a non-decreasing function of k for both the euclidean and non-euclidean cases.

Next, multiplying both sides by $\lambda(k+2)$, summing both sides from $k = 0$ to N and using the linearity of expectation, we get,

$$\begin{aligned}
\sum_{k=0}^N (1-\gamma)(\mu v_\lambda^{\pi_k} - \mu v_\lambda^*) &\leq d_{\mu,\pi^*} (B_\omega(\pi^*, \pi_0) - (N+2) B_\omega(\pi^*, \pi_{N+1})) + \sum_{k=0}^N \frac{h_\omega^2(N;\lambda)}{2\lambda(k+2)} + \sum_{k=0}^N \lambda(k+2) d_{\mu,\pi^*} \epsilon_k \\
&\leq d_{\mu,\pi^*} B_\omega(\pi^*, \pi_0) + \sum_{k=0}^N \frac{h_\omega^2(N;\lambda)}{2\lambda(k+2)} + \sum_{k=0}^N \lambda(k+2) d_{\mu,\pi^*} \epsilon_k \\
&\leq D_\omega + \sum_{k=0}^N \frac{h_\omega^2(N;\lambda)}{2\lambda(k+2)} + \sum_{k=0}^N \lambda(k+2) d_{\mu,\pi^*} \epsilon_k \\
&= D_\omega + \sum_{k=0}^N \frac{h_\omega^2(N;\lambda)}{2\lambda(k+2)} + \sum_{k=0}^N \frac{1}{t_k} d_{\mu,\pi^*} \epsilon_k,
\end{aligned}$$

where the second relation holds by the positivity of the Bregman distance, the third relation by Lemma 24 for uniformly initialized π_0, and the last relation by plugging back $t_k = \frac{1}{\lambda(k+2)}$ in the last term..

Bounding $\sum_{k=0}^N \frac{1}{k+2} \leq O(\log N)$, we get

$$\sum_{k=0}^N \mu v_\lambda^{\pi_k} - \mu v_\lambda^* \leq O\Bigg(\frac{D_\omega}{(1-\gamma)} + \frac{h_\omega^2(N;\lambda) \log N}{\lambda(1-\gamma)} + \frac{1}{1-\gamma} \sum_{k=0}^N \frac{1}{t_k} d_{\mu,\pi^*} \epsilon_k \Bigg).$$

By the definition of v_{best}^N, which gives $(N+1)\big(\mu v_{\text{best}}^N - \mu v^*\big) \leq \sum_{k=0}^N \mu v^{\pi_k} - \mu v^*$, and some algebraic manipulations, we obtain

$$\mu v^N_{\textbf{best}} - \mu v^*_\lambda \le O\left(\frac{D_\omega}{(1-\gamma)N} + \frac{h^2_\omega(N;\lambda)\log N}{\lambda(1-\gamma)N} + \frac{1}{1-\gamma}\frac{1}{N}\sum_{k=0}^{N}\frac{1}{t_k}d_{\mu,\pi^*}\epsilon_k\right).$$

Plugging in Lemma 18, we get that for any (ϵ,δ), if the number of trajectories in the k-th iteration is

$$M_k \ge \frac{r_\omega(k,\lambda)^2}{2\epsilon^2}\big(S\log 2A + \log \pi^2(k+1)^2/6\delta\big),$$

then with probability of at least $1-\delta$,

$$\mu v^N_{\textbf{best}} - \mu v^*_\lambda \le O\left(\frac{D_\omega}{(1-\gamma)N} + \frac{h^2_\omega(N;\lambda)\log N}{\lambda(1-\gamma)N} + \frac{\epsilon}{(1-\gamma)^2}\left\|\frac{d_{\mu,\pi^*}}{\nu}\right\|_\infty\right).$$

where $r_\omega(k,\lambda) = \frac{4A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}$ and $r_\omega(k,\lambda) = \frac{4A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}(1+\mathbb{1}\{\lambda\neq 0\}\log k)$ in the euclidean and non-euclidean settings respectively.

By Plugging the bounds D_ω, h_ω and $\mathrm{C}_{\max,\lambda}$, we get in the euclidean case,

$$\mu v^N_{\textbf{best}} - \mu v^*_\lambda \le O\left(\frac{(\mathrm{C}^2_{\max} + \lambda^2)A\log N}{\lambda(1-\gamma)^3 N} + \frac{1}{(1-\gamma)^2}\left\|\frac{d_{\mu,\pi^*}}{\nu}\right\|_\infty \epsilon\right),$$

and in the non-euclidean case,

$$\mu v^N_{\textbf{best}} - \mu v^*_\lambda \le O\left(\frac{(\mathrm{C}^2_{\max} + \lambda^2\log^2 A)A^2\log^3 N}{\lambda(1-\gamma)^3 N} + \frac{1}{(1-\gamma)^2}\left\|\frac{d_{\mu,\pi^*}}{\nu}\right\|_\infty \epsilon\right),$$

□

F.7 Sample Complexity of Sample-Based TRPO

In this section we calculate the overall *sample complexity* of Sample-Based TRPO, i.e., the number interactions with the MDP the algorithm does in order to reach a close to optimal solution.

By Lemma 19, in order to have $\frac{1}{(1-\gamma)^2}\left\|\frac{d_{\mu,\pi^*}}{\nu}\right\|_\infty \frac{\epsilon}{2}$ approximation error, we need $M_k \ge O\Big(\frac{r_\omega(k,\lambda)^2}{\epsilon^2}\big(S\log 2A + \log{(k+1)^2/\delta}\big)\Big)$ trajectories in each iteration, and the number of samples in each truncated trajectory is $T_k \ge O\Big(\frac{1}{1-\gamma}\log\frac{\epsilon}{r_\omega(k,\lambda)}\Big)$, where $r_\omega(k,\lambda) = \frac{4A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}(1+\mathbb{1}\{\lambda\neq 0\}\log k)$ in the euclidean and non-euclidean settings respectively.

Therefore, the number of samples in each iteration required to guarantee a $\frac{1}{(1-\gamma)^2}\left\|\frac{d_{\mu,\pi^*}}{\nu}\right\|_\infty \frac{\epsilon}{2}$ error is

$$O\left(\frac{r_\omega(k,\lambda)^2 \log\frac{\epsilon}{r_\omega(k,\lambda)}}{(1-\gamma)\epsilon^2}\big(S\log 2A + \log{(k+1)^2/\delta}\big)\right).$$

The overall sample complexity is acquired by multiplying the number of iterations N required to reach an $\frac{\epsilon/2}{(1-\gamma)^2}$ optimization error multiplied with the iteration-wise sample complexity, given above. Combining the two errors and using the fact that $C^{\pi^*} \ge 1$, we have that the overall error

$$\frac{1}{(1-\gamma)^2}\left(1 + C^{\pi^*}\right)\frac{\epsilon}{2} \le \frac{2}{(1-\gamma)^2}C^{\pi^*}\frac{\epsilon}{2} = \frac{1}{(1-\gamma)^2}C^{\pi^*}\epsilon.$$

In other words, the overall error of the algorithm is bounded by $\frac{1}{(1-\gamma)^2}C^{\pi^*}\epsilon$

Finally, the sample complexity to reach a $\frac{1}{(1-\gamma)^2}C^{\pi^*}\epsilon$ error for the different cases is arranged in the following table (the complete analysis is provided the the next section):

	Euclidean	**Non-Euclidean (KL)**
Unregularized	$\frac{A^3 C_{\text{max}}^4}{(1-\gamma)^3 \epsilon^4}\left(\log\lvert\Pi^{\text{det}}\rvert + \log\frac{1}{\delta}\right)$	$\frac{A^2 C_{\text{max}}^4}{(1-\gamma)^3 \epsilon^4}\left(\log\lvert\Pi^{\text{det}}\rvert + \log\frac{1}{\delta}\right)$
Regularized	$\frac{A^3 C_{\text{max},\lambda}^4}{\lambda(1-\gamma)^4 \epsilon^3}\left(\log\lvert\Pi^{\text{det}}\rvert + \log\frac{1}{\delta}\right)$	$\frac{A^4 C_{\text{max},\lambda}^4}{\lambda(1-\gamma)^4 \epsilon^3}\left(\log\lvert\Pi^{\text{det}}\rvert + \log\frac{1}{\delta}\right)$

The same bound for CPI as given in [Kakade et al., 2003] is

$$\frac{A^2 C_{\text{max}}^4}{(1-\gamma)^5 \epsilon^4}\left(\log\lvert\Pi^{\text{det}}\rvert + \log\frac{1}{\delta}\right),$$

where we omitted logarithmic factors in $1-\gamma$ and ϵ. Notice that this bound is similar to the bound of Sample-Based TRPO observed in this paper, as expected.

In order to translate this bound using our notation bound, we used [Kakade et al., 2003][Theorem 7.3.3] with $H = \frac{1}{1-\gamma}$, which states that in order to guarantee a bounded advantage of for any policy π', $\mathbb{A}_\pi(\nu, \pi') \le (1-\gamma)\epsilon$ we need $O\Big(\frac{\log\epsilon(\log\Pi^{\text{det}} + \log\frac{1}{\delta}}{(1-\gamma)^5\epsilon^4}\Big)$ samples. Then, by [Kakade and Langford, 2002][Corollary 4.5] with $\mathbb{A}_\pi(\nu, \pi') \le (1-\gamma)\epsilon$ we get that $(1-\gamma)(\mu v^\pi - \mu v^*) \le \frac{\epsilon}{1-\gamma}\left\lVert\frac{d_{\mu,\pi^*}}{\nu}\right\rVert_\infty$, or $\mu v^\pi - \mu v^* \le \frac{\epsilon}{(1-\gamma)^2}\left\lVert\frac{d_{\mu,\pi^*}}{\nu}\right\rVert_\infty$. Finally, the C_{max}^4 factor comes from using a non-normalized MDP, where the maximum reward is C_{max}. We get C_{max}^2 from number of iterations needed for convergence, and the number of samples in each iteration is also proportional to C_{max}^2

The Unregularized Case

The euclidean case: The error after N iterations is bounded by

$$\mu v_{\textbf{best}}^N - \mu v^* \le O\Bigg(\frac{\mathrm{C_{max}}\sqrt{A}\log N}{(1-\gamma)^2\sqrt{N}} + \frac{1}{(1-\gamma)^2}\left\lVert\frac{d_{\mu,\pi^*}}{\nu}\right\rVert\frac{\epsilon}{2}\Bigg).$$

Thus, in order to reach an error of $\frac{1}{(1-\gamma)^2}C^{\pi^*}\epsilon$ error, we need

$$N \le O\bigg(\frac{C_{\text{max}}^2 A\log\epsilon}{\epsilon^2}\bigg).$$

Thus, the sample complexity to reach $\frac{1}{(1-\gamma)^2}C^{\pi^*}\epsilon$ error when logarithmic factors are omitted is

$$\tilde{O}\Bigg(\frac{A^3 C_{\text{max}}^4}{(1-\gamma)^3\epsilon^4}\bigg(\log\lvert\Pi^{\text{det}}\rvert + \log\frac{1}{\delta}\bigg)\Bigg)$$

The non-euclidean case: The error after N iterations is bounded by

$$\mu v_{\textbf{best}}^N - \mu v^* \le O\bigg(\frac{\mathrm{C_{max}}(\log A + \log N)}{(1-\gamma)^2\sqrt{N}} + \frac{1}{(1-\gamma)^2}\left\lVert\frac{d_{\mu,\pi^*}}{\nu}\right\rVert\frac{\epsilon}{2}\bigg).$$

Thus, in order to reach an error of $\frac{1}{(1-\gamma)^2}C^{\pi^*}\epsilon$ error, we need

$$N \le O\bigg(\frac{C_{\text{max}}^2\log^2 A\log^2\epsilon}{\epsilon^2}\bigg).$$

Thus, the sample complexity to reach $\frac{1}{(1-\gamma)^2}C^{\pi^*}\epsilon$ error when logarithmic factors are omitted is

$$\tilde{O}\Bigg(\frac{A^2 C_{\text{max}}^4}{(1-\gamma)^3\epsilon^4}\bigg(\log\lvert\Pi^{\text{det}}\rvert + \log\frac{1}{\delta}\bigg)\Bigg)$$

The Regularized Case

The euclidean case: The error after N iterations is bounded by

$$\mu v_{\textbf{best}}^N - \mu v_\lambda^* \leq O\left(\frac{\mathrm{C}^2_{\text{max},\lambda} \, A \log N}{\lambda(1-\gamma)^3 N} + \frac{1}{(1-\gamma)^2} \left\| \frac{d_{\mu,\pi^*}}{\nu} \right\| \frac{\epsilon}{2} \right),$$

Thus, in order to reach an error of $\frac{1}{(1-\gamma)^2} C^{\pi^*} \epsilon$ error, we need

$$N \leq O\left(\frac{\mathrm{C}^2_{\text{max},\lambda} \, A \log \epsilon}{\lambda(1-\gamma)\epsilon} \right)$$

Thus, the sample complexity to reach $\frac{1}{(1-\gamma)^2} C^{\pi^*} \epsilon$ error when logarithmic factors are omitted is

$$\tilde{O}\left(\frac{A^3 C^4_{\text{max},\lambda}}{\lambda(1-\gamma)^4 \epsilon^3} \left(\log |\Pi^{\text{det}}| + \log \frac{1}{\delta} \right) \right)$$

The non-euclidean case: The error after N iterations is bounded by

$$\mu v_{\textbf{best}}^N - \mu v_\lambda^* \leq O\left(\frac{\log A}{(1-\gamma)N} + \frac{\mathrm{C}^2_{\text{max},\lambda} \, A^2 \log^3 N}{\lambda(1-\gamma)^3 N} \right) + \frac{1}{(1-\gamma)^2} \left\| \frac{d_{\mu,\pi^*}}{\nu} \right\|_\infty \frac{\epsilon}{2}.$$

Rearranging, we get,

$$\mu v_{\textbf{best}}^N - \mu v_\lambda^* \leq O\left(\frac{(\mathrm{C}^2_{\text{max}} + \lambda^2 \log^2 A) A^2 \log^3 N}{\lambda(1-\gamma)^3 N} \right) + \frac{1}{(1-\gamma)^2} \left\| \frac{d_{\mu,\pi^*}}{\nu} \right\|_\infty \frac{\epsilon}{2},$$

which can also be written with $\mathrm{C}^2_{\text{max},\lambda}$

$$\mu v_{\textbf{best}}^N - \mu v_\lambda^* \leq O\left(\frac{\mathrm{C}^2_{\text{max},\lambda} \, A^2 \log^3 N}{\lambda(1-\gamma)^3 N} \right) + \frac{1}{(1-\gamma)^2} \left\| \frac{d_{\mu,\pi^*}}{\nu} \right\| \epsilon.$$

Thus, in order to reach an error of $\frac{1}{(1-\gamma)^2} C^{\pi^*} \epsilon$ error, we need

$$N \leq \tilde{O}\left(\frac{\mathrm{C}^2_{\text{max},\lambda} \, A^2}{\lambda(1-\gamma)\epsilon} \right),$$

omitting logarithmic factors.

Thus, the sample complexity to reach $\frac{1}{(1-\gamma)^2} C^{\pi^*} \epsilon$ error when logarithmic factors are omitted is

$$\tilde{O}\left(\frac{A^4 C^4_{\text{max},\lambda}}{\lambda(1-\gamma)^4 \epsilon^3} \left(\log |\Pi^{\text{det}}| + \log \frac{1}{\delta} \right) \right)$$

G Useful Lemmas

The next lemmas will provide useful bounds for uniform and Sample-Based TRPO. In this section, we define $\|\cdot\|_*$ to be the dual norm of $\|\cdot\|$.

Lemma 20 (Connection between the regularized Bellman operator and the q-function)**.** *For any π, π' the following holds:*

$$\langle q_\lambda^\pi + \lambda \nabla \omega(\pi), \pi' - \pi \rangle = T_\lambda^{\pi'} v_\lambda^\pi - v_\lambda^\pi - \lambda B_\omega(\pi', \pi)$$

Proof. First, note that for any s

$$
\begin{aligned}
&\langle q_\lambda^\pi(s,\cdot), \pi'(\cdot \mid s)\rangle \\
&= \sum_a \pi'(a \mid s) q_\lambda^\pi(s,a) \\
&= \sum_a \pi'(a \mid s)\left(c_\lambda^\pi(s,a) + \gamma \sum_{s'} p(s'|s,a) v_\lambda^\pi \right) \\
&= \sum_a \pi'(a \mid s)\left(c(s,a) + \lambda\omega\left(s;\pi\right) + \gamma \sum_{s'} p(s'|s,a) v_\lambda^\pi \right) \\
&= \sum_a \pi'(a \mid s)\left(c(s,a) + \lambda\omega\left(s;\pi'\right) - \lambda\omega\left(s;\pi'\right) + \lambda\omega\left(s;\pi\right) + \gamma \sum_{s'} p(s'|s,a) v_\lambda^\pi \right) \\
&= \sum_a \pi'(a \mid s)\left(c(s,a) + \lambda\omega\left(s;\pi'\right) + \gamma \sum_{s'} p(s'|s,a) v_\lambda^\pi \right) + \lambda\omega\left(s;\pi\right) - \lambda\omega\left(s;\pi'\right) \\
&= c_\lambda^{\pi'}(s) + \gamma P^{\pi'} v_\lambda^\pi(s) + \lambda\omega\left(s;\pi\right) - \lambda\omega\left(s;\pi'\right) \\
&= T_\lambda^{\pi'} v_\lambda^\pi(s) + \lambda\omega\left(s;\pi\right) - \lambda\omega\left(s;\pi'\right),
\end{aligned}
$$

where the second transition is by the definition of q_λ^π, the third is by the definition of c_λ^π, the fourth is by adding and subtracting $\lambda\omega\left(s;\pi'\right)$, the fifth is by the fact $\lambda\omega\left(s;\pi'\right)$ is independent of a and the seventh is by the definition of the regularized Bellman operator.

Thus,

$$
\langle q_\lambda^\pi, \pi'\rangle = T_\lambda^{\pi'} v_\lambda^\pi + \lambda\omega(\pi) - \lambda\omega(\pi')
$$

Now, note that by the definition of the q-function $\langle q_\lambda^\pi, \pi\rangle = v_\lambda^\pi$ and thus,

$$
\langle q_\lambda^\pi, \pi' - \pi\rangle = T_\lambda^{\pi'} v_\lambda^\pi - v_\lambda^\pi + \lambda\omega(\pi) - \lambda\omega(\pi').
$$

Finally, by adding to both sides $\langle \lambda\nabla\omega(\pi), \pi' - \pi\rangle$, we get,

$$
\langle q_\lambda^\pi + \lambda\nabla\omega(\pi), \pi' - \pi\rangle = T_\lambda^{\pi'} v_\lambda^\pi - v_\lambda^\pi + \lambda\omega(\pi) - \lambda\omega(\pi') + \lambda\langle \nabla\omega(\pi), \pi' - \pi\rangle.
$$

To conclude the proof, note that by the definition of the Bregman distance we have,

$$
\langle q_\lambda^\pi + \lambda\nabla\omega(\pi), \pi' - \pi\rangle = T_\lambda^{\pi'} v_\lambda^\pi - v_\lambda^\pi - \lambda B_\omega\left(\pi', \pi\right).
$$

□

Lemma 21 (Bounds regarding the updates of Uniform TRPO)**.** *For any $k \geq 0$ and state s, which is updated in the k-th iteration, the following relations hold for Uniform TRPO* (21)*:*

1. $\left\|\nabla\omega(\pi_k(\cdot|s))\right\|_* \leq O(1)$ *and* $\left\|\nabla\omega(\pi_k(\cdot|s))\right\|_* \leq O(\frac{\mathrm{C}_{max,\lambda}\log k}{\lambda(1-\gamma)})$*, in the euclidean and non-euclidean cases, respectively.*

2. $\left\|q_\lambda^{\pi_k}(s,\cdot)\right\|_* \leq h_\omega$*, where* $h_\omega = O(\frac{\sqrt{A}\,\mathrm{C}_{max,\lambda}}{1-\gamma})$ *and* $h_\omega = O(\frac{\mathrm{C}_{max,\lambda}}{1-\gamma})$ *in the euclidean and non-euclidean cases, respectively.*

3. $\left\|q_\lambda^{\pi_k}(s,\cdot) + \lambda\nabla\omega(\pi_k(\cdot|s))\right\|_* \leq h_\omega(k;\lambda)$*, where* $h_\omega(k;\lambda) = O(\frac{\sqrt{A}\,\mathrm{C}_{max,\lambda}}{1-\gamma})$ *and* $h_\omega(k;\lambda) = O(\frac{\mathrm{C}_{max,\lambda}(1+\mathbb{1}\{\lambda\neq 0\}\log k)}{1-\gamma})$ *in the euclidean and non-euclidean cases, respectively, and* $\mathbb{1}\{\lambda \neq 0\} = 0$ *in the unregularized case (λ=0) and* $\mathbb{1}\{\lambda \neq 0\} = 1$ *otherwise.*

Where for every state s, $\left\|\cdot\right\|_*$ denotes the dual norm over the action space, which is L_1 in the euclidean case, and L_∞ in non-euclidean cases.

Proof. We start by proving the **first claim**:

For the **euclidean case**, $\omega(\cdot) = \frac{1}{2}\left\|\cdot\right\|_2^2$. Thus, for every state s,

$$\left\|\nabla\omega(\pi(\cdot|s))\right\|_2 = \left\|\pi(\cdot|s)\right\|_2 \le \left\|\pi(\cdot|s)\right\|_1 = 1,$$

where the inequality is due to the fact that $\left\|\cdot\right\|_2 \le \left\|\cdot\right\|_1$.

The statement holds by the properties of $\frac{1}{2}\left\|\cdot\right\|_2^2$ and thus holds for both the uniform versions.

For the **non-euclidean case**, $\omega(\cdot) = H(\cdot) + \log \mathcal{A}$. Now, consider Uniform TRPO (21). By taking the logarithm of (26), we have,

$$\begin{aligned}\log \pi_k(a \mid s) = \log \pi_{k-1}(a \mid s)& \\ - t_{k-1}\big(q_\lambda^{\pi_{k-1}}(s,a) + \lambda \log \pi_{k-1}(a \mid s)\big)& \\ - \log\left(\sum_{a'} \pi_{k-1}(a' \mid s)\exp\big(-t_{k-1}\big(q_\lambda^{\pi_{k-1}}(s,a') + \lambda \log \pi_{k-1}(a' \mid s)\big)\big)\right)&.\end{aligned} \tag{62}$$

Notice that for $k \ge 0$, for every state-action pair, $q_\lambda^{\pi_k}(a|s) \ge 0$. Thus,

$$\begin{aligned}\log\left(\sum_{a'} \pi_k(a' \mid s)\exp(-t_k(q_\lambda^{\pi_k}(s,a') + \lambda \log \pi_k(a' \mid s)))\right) &\le \log\left(\sum_{a'} \pi_k(a' \mid s)\exp(-t_k\lambda \log \pi_k(a' \mid s))\right) \\ &= \log\left(\sum_{a'} \pi_k(a' \mid s)\pi_k^{-\lambda t_k}(a' \mid s)\right).\end{aligned} \tag{63}$$

Where the first relation holds since $q_\lambda^\pi(s,a) \ge 0$. Applying Jensen's inequality we can further bound the above.

$$\begin{aligned}(63) &= \log\left(A\sum_{a'} \frac{1}{A}\pi_k^{1-\lambda t_k}(a' \mid s)\right) \\ &= \log\left(A\sum_{a'} \frac{1}{A}\pi_k^{1-\lambda t_k}(a' \mid s)\right) \\ &\le \log\left(A\left(\sum_{a'} \frac{1}{A}\pi_k(a' \mid s)\right)^{1-\lambda t_k}\right) \\ &= \log\left(A\left(\frac{1}{A}\sum_{a'} \pi_k(a' \mid s)\right)^{1-\lambda t_k}\right) \\ &= \log\left(A\left(\frac{1}{A}\right)^{1-\lambda t_k}\right) = \log\big(A^{\lambda t_k}\big) = \lambda t_k \log A.\end{aligned} \tag{64}$$

In the third relation we applied Jensen's inequality for concave functions. As $0 \le 1 - \lambda t_k \le 1$ (by the choice of the learning rate in the regularized case) we have that $X^{1-\lambda t_k}$ is a concave function in X, and thus $\sum_{a'=1}^A \frac{1}{A}\pi_k^{1-\lambda t_k}(a' \mid s) \le \left(\sum_{a'=1}^A \frac{1}{A}\pi_k(a' \mid s)\right)^{1-\lambda t_k}$ by Jensen's inequality. Combining this inequality with the fact that A is positive and $\log$ is monotonic function establishes the third relation.

Furthermore, note that for every k, and for every s, a

$$\log \pi_k(a|s) \le 0 \tag{65}$$

Plugging (64) and (65) in (62), we get,

$$
\begin{aligned}
\log \pi_k(a \mid s) &\geq \log \pi_{k-1}(a \mid s) - t_{k-1}\big(q_\lambda^{\pi_{k-1}}(s,a) + \lambda \log A\big) \\
&\geq \log \pi_0(a|s) - \sum_{i=0}^{k-1} t_k(q_\lambda^{\pi_i}(s,a) + \lambda \log A) \\
&\geq -\log A - \left(\frac{\mathrm{C}_{\max,\lambda}}{1-\gamma} + \lambda \log A\right)\sum_{i=0}^{k-1} t_i \\
&= -\log A - \left(\frac{\mathrm{C}_{\max,\lambda}}{\lambda(1-\gamma)} + \log A\right)\sum_{i=0}^{k-1} \frac{1}{i+2} \\
&\geq -\log A - \left(\frac{\mathrm{C}_{\max,\lambda}}{\lambda(1-\gamma)} + \log A\right)(1+\log k) \\
&\geq -\frac{\mathrm{C}_{\max} + 3\lambda \log A}{\lambda(1-\gamma)}(1+\log k) \\
&\geq -\frac{3\,\mathrm{C}_{\max,\lambda}}{\lambda(1-\gamma)}(1+\log k), \qquad (66)
\end{aligned}
$$

where the second relation holds by unfolding the recursive formula for each k and the fourth by plugging in the stepsizes for the regularized case, i.e. $t_k = \frac{1}{\lambda(k+2)}$. The final relation holds since $\mathrm{C}_{\max,\lambda} = \mathrm{C}_{\max} + \lambda \log A$.

To conclude, since $\log \pi_k(a \mid s) \leq 0$ and $\nabla\omega(\pi) = \nabla H(\pi) = 1 + \log \pi$, we get that for the non-euclidean case,

$$
\|\nabla\omega(\pi_k)\|_\infty \leq O\left(\frac{\mathrm{C}_{\max,\lambda}}{\lambda(1-\gamma)} \log k\right).
$$

This concludes the proof of the first claim for both the euclidean and non-euclidean cases, in both exact scenarios. Interestingly, in the non-euclidean case, the gradients can grow to infinity due to the fact that the gradient of the entropy of a deterministic policy is unbounded. However, this result shows that a deterministic policy can only be obtained after an infinite time, as the gradient is bounded by a logarithmic rate.

Next, we prove the **second claim**:

It holds that for any state-action pair $q_\lambda^{\pi_k}(s,a) \in \left[0, \frac{\mathrm{C}_{\max,\lambda}}{1-\gamma}\right]$.

For the **euclidean case**, we have that

$$
\|q_\lambda^{\pi_k}(s,\cdot)\|_* = \|q_\lambda^{\pi_k}(s,\cdot)\|_2 \leq \sqrt{\sum_{a\in\mathcal{A}} \left(\frac{\mathrm{C}_{\max,\lambda}}{1-\gamma}\right)^2} = \frac{\sqrt{A}\,\mathrm{C}_{\max,\lambda}}{1-\gamma}.
$$

For the **non-euclidean case**, we have that

$$
\|q_\lambda^{\pi_k}(s,\cdot)\|_* = \|q_\lambda^{\pi_k}(s,\cdot)\|_\infty \leq \frac{\mathrm{C}_{\max,\lambda}}{1-\gamma},
$$

which concludes the proof of the second claim.

Finally, we prove the **third claim**: For any state s, by the triangle inequality,

$$
\|q_\lambda^{\pi_k}(s,\cdot) + \lambda\nabla\omega(\pi_k(\cdot|s))\|_* \leq \|q_\lambda^{\pi_k}(s,\cdot)\|_* + \lambda \|\nabla\omega(\pi_k(\cdot|s))\|_*\,,
$$

by plugging the two former claims for the euclidean and non-euclidean cases, we get the required result. □

The next lemma follows similar derivation to Lemma 21, with small changes tailored for the sample-based case. Note that in the sample-based case, and A factor is added in claims 1,3 and 4.

Lemma 22 (Bounds regarding the updates of Sample-Based TRPO)**.** *For any $k \geq 0$ and state s, which is updated in the k-th iteration, the following relations hold for Sample-Based TRPO* (43)*:*

1. $\|\nabla\omega(\pi_k(\cdot|s))\|_* \leq O(1)$ *and* $\|\nabla\omega(\pi_k(\cdot|s))\|_* \leq O(\frac{A\,\mathrm{C}_{max,\lambda}\log k}{\lambda(1-\gamma)})$*, in the euclidean and non-euclidean cases, respectively.*

2. $\|q_\lambda^{\pi_k}(s,\cdot)\|_* \leq h_\omega$*, where* $h_\omega = O(\frac{\sqrt{A}\,\mathrm{C}_{max,\lambda}}{1-\gamma})$ *and* $h_\omega = O(\frac{\mathrm{C}_{max,\lambda}}{1-\gamma})$ *in the euclidean and non-euclidean cases, respectively.*

3. $\|q_\lambda^{\pi_k}(s,\cdot)+\lambda\nabla\omega(\pi_k(\cdot|s))\|_* \leq h_\omega(k;\lambda)$*, where* $h_\omega(k;\lambda) = O(\frac{\sqrt{A}\,\mathrm{C}_{max,\lambda}}{1-\gamma})$ *and* $h_\omega(k;\lambda) = O(\frac{\mathrm{C}_{max,\lambda}(1+\mathbb{1}\{\lambda\neq 0\}A\log k)}{1-\gamma})$ *in the euclidean and non-euclidean cases, respectively, and* $\mathbb{1}\{\lambda \neq 0\} = 0$ *in the unregularized case (*λ*=0) and* $\mathbb{1}\{\lambda \neq 0\} = 1$ *in the regularized case (*$\lambda > 0$*).*

4. $\|A\hat{q}_\lambda^{\pi_k}(s,\cdot,m)+\lambda\nabla\omega(\pi_k(\cdot|s))\|_\infty \leq \hat{h}_\omega(k;\lambda)$*, where* $\hat{h}_\omega(k;\lambda) = O(\frac{A\,\mathrm{C}_{max,\lambda}}{1-\gamma})$ *and* $\hat{h}_\omega(k;\lambda) = O(\frac{A\,\mathrm{C}_{max,\lambda}(1+\mathbb{1}\{\lambda\neq 0\}\log k)}{1-\gamma})$ *in the euclidean and non-euclidean cases, respectively, and* $\mathbb{1}\{\lambda \neq 0\} = 0$ *in the unregularized case (*λ*=0) and* $\mathbb{1}\{\lambda \neq 0\} = 1$ *in the regularized case (*$\lambda > 0$*).*

Where for every state s, $\|\cdot\|_*$ denotes the dual norm over the action space, which is L_1 in the euclidean case, and L_∞ in non-euclidean cases.

Proof. We start by proving the **first claim**:

For the **euclidean case**, in the same manner as in the Uniform case, $\omega(\cdot) = \frac{1}{2}\|\cdot\|_2^2$. Thus, for every state s,

$$\|\nabla\omega(\pi(\cdot|s))\|_2 = \|\pi(\cdot|s)\|_2 \leq \|\pi(\cdot|s)\|_1 = 1,$$

where the inequality is due to the fact that $\|\cdot\|_2 \leq \|\cdot\|_1$.

For the **non-euclidean case**, $\omega(\cdot) = H(\cdot) + \log\mathcal{A}$. The bound for the sample-based version for the non-euclidean choice of ω follows similar reasoning with mild modification. By (42), in the sample-based case, a state s is updated in the k-th iteration using the approximation of the $q_\lambda^{\pi_k}(s,a)$ in this state,

$$\hat{q}_\lambda^{\pi_k}(s,a) := \frac{A\sum_{m=1}^M \mathbb{1}\{s=s_m, a=a_m\}\hat{q}_\lambda^{\pi_k}(s_m,\cdot,m)}{n(s)} \leq \frac{A\sum_{m=1}^M \mathbb{1}\{s=s_m, a=a_m\}\frac{\mathrm{C}_{\max,\lambda}}{1-\gamma}}{n(s)} \leq \frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma},$$

where we denoted $n(s) = \sum_a n(s,a)$ the number of times the state s was observed at the k-th episode and used the fact $\hat{q}_\lambda^{\pi_k}(s_m,\cdot,m_i)$ is sampled by unrolling the MDP. Thus, it holds that

$$\hat{q}_\lambda^{\pi_k}(s,a) \leq \frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}.$$

Interestingly, because we use the importance sampling factor A in the approximation of $q_\lambda^{\pi_k}$, we obtain an additional A factor.

Thus, by repeating the analysis in Lemma 21, equation (66), we obtain,

$$\begin{aligned}\log(\pi_k(a \mid s)) &\geq \log(\pi_{k-1}(a \mid s)) - t_{k-1}\big(\hat{q}_\lambda^{\pi_{k-1}}(s,a) + \lambda \log A\big) \\ &\geq \log \pi_0(a|s) - \sum_{i=0}^{k-1} t_i(\hat{q}_\lambda^{\pi_i}(s,a) + \lambda \log A) \\ &\geq -\log A - \left(\frac{A\,\mathrm{C}_{\text{max},\lambda}}{1-\gamma} + \lambda \log A\right)\sum_{i=0}^{k-1} t_i \\ &= -\log A - \left(\frac{A\,\mathrm{C}_{\text{max},\lambda}}{\lambda(1-\gamma)} + \log A\right)\sum_{i=0}^{k-1} \frac{1}{i+2} \\ &\geq -\log A - \left(\frac{A\,\mathrm{C}_{\text{max},\lambda}}{\lambda(1-\gamma)} + \log A\right)(1+\log k) \\ &\geq -\frac{A\mathrm{C}_{\text{max}} + 3A\lambda \log A}{\lambda(1-\gamma)}(1+\log k) \\ &\geq -\frac{3A\,\mathrm{C}_{\text{max},\lambda}}{\lambda(1-\gamma)}, \end{aligned} \tag{67}$$

where the second relation holds by unfolding the recursive formula for each k and the fourth by plugging in the stepsizes for the regularized case, i.e. $t_k = \frac{1}{\lambda(k+2)}$. The final relation holds since $\mathrm{C}_{\text{max},\lambda} = \mathrm{C}_{\text{max}} + \lambda \log A$. Thus,

$$\log(\pi_k(a \mid s)) \geq -\frac{3A\,\mathrm{C}_{\text{max},\lambda}}{\lambda(1-\gamma)}(1+\log k),$$

This concludes the proof of the first claim for both the euclidean and non-euclidean cases.

As in the uniform case, in the non-euclidean case, the gradients can grow to infinity due to the fact that the gradient of the entropy of a deterministic policy is unbounded. However, this result shows that a deterministic policy can only be obtained after an infinite time, as the gradient is bounded by a logarithmic rate.

Next, we prove the **second claim**:

It holds that for any state-action pair $q_\lambda^{\pi_k}(s,a) \in \left[0, \frac{\mathrm{C}_{\text{max},\lambda}}{1-\gamma}\right]$.

For the **euclidean case**, we have that

$$\|q_\lambda^{\pi_k}(s,\cdot)\|_* = \|q_\lambda^{\pi_k}(s,\cdot)\|_2 \leq \sqrt{\sum_{a\in\mathcal{A}}\left(\frac{\mathrm{C}_{\text{max},\lambda}}{1-\gamma}\right)^2} = \frac{\sqrt{A}\,\mathrm{C}_{\text{max},\lambda}}{1-\gamma}.$$

For the **non-euclidean case**, we have that

$$\|q_\lambda^{\pi_k}(s,\cdot)\|_* = \|q_\lambda^{\pi_k}(s,\cdot)\|_\infty \leq \frac{\mathrm{C}_{\text{max},\lambda}}{1-\gamma},$$

which concludes the proof of the second claim.

Next, we prove the **third claim**: For any state s, by the triangle inequality,

$$\|q_\lambda^{\pi_k}(s,\cdot) + \lambda\nabla\omega(\pi_k(\cdot|s))\|_* \leq \|q_\lambda^{\pi_k}(s,\cdot)\|_* + \lambda\,\|\nabla\omega(\pi_k(\cdot|s))\|_*\,,$$

by plugging the two former claims for the euclidean and non-euclidean cases, we get the required result.

Finally, the **fourth claim** is the same as the third claim, but with an additional A factor due to the importance sampling factor,

$$\|A\hat{q}_\lambda^{\pi_k}(s,\cdot,m) + \lambda\nabla\omega(\pi_k(\cdot|s))\|_\infty \leq A\,\|\hat{q}_\lambda^{\pi_k}(s,\cdot,m)\|_\infty + \lambda\,\|\nabla\omega(\pi_k(\cdot|s))\|_\infty\,.$$

□

Using the same techniques of the last lemma, we prove the following technical lemma, regarding the change in the gradient of the Bregman generating function ω of two consecutive iterations of TRPO, in the sample-based case.

Lemma 23 (bound on the difference of the gradient of ω between two consecutive policies in the sample-based case)**.** *For each state-action pair, s, a, the difference between two consecutive policies of Sample-Based TRPO is bounded by:*

$$\|\nabla\omega(\pi_{k+1}) - \nabla\omega(\pi_k)\|_{\infty,\infty} \leq A_\omega(k),$$

where $A_\omega(k) = t_k \frac{A^{3/2}\,\mathrm{C}_{\mathit{max},\lambda}}{1-\gamma}$ and $A_\omega(k) = t_k \frac{A\,\mathrm{C}_{\mathit{max},\lambda}\log k}{1-\gamma}$ in the euclidean and non-euclidean cases respectively, k is the iteration number and t_k is the step size used in the update.

Proof. In both the euclidean in non-euclidean cases, we discuss optimization problem (43) for the sample-based case. Thus, for any visited state in the k-th iteration, $s \in \mathcal{S}_M^k := \left\{s' \in \mathcal{S} : \sum_{m=1}^M \mathbb{1}\{s' = s_m\} > 0\right\}$, by (42)

$$\hat{q}_\lambda^{\pi_k}(s,a) := \frac{A\sum_{m=1}^M \mathbb{1}\{s = s_m, a = a_m\}\hat{q}_\lambda^{\pi_k}(s_m,\cdot,m)}{n(s)} \leq \frac{A\sum_{m=1}^M \mathbb{1}\{s = s_m, a = a_m\}\frac{\mathrm{C}_{\max,\lambda}}{1-\gamma}}{n(s)} \leq \frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma},$$

where we denoted $n(s) = \sum_a n(s,a)$ the number of times the state s was observed at the k-th episode and used the fact $\hat{q}_\lambda^{\pi_k}(s_m,\cdot,m_i)$ is sampled by unrolling the MDP. Thus, it holds that

$$\hat{q}_\lambda^{\pi_k}(s,a) \leq \frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}.$$

Interestingly, because we use the importance sampling factor A in the approximation of $q_\lambda^{\pi_k}$, we obtain an additional A factor.

First, notice that for states which were not encountered in the k-th iteration, i.e., all states s for which $\sum_{m=1}^M \mathbb{1}\{s = s_m\} = 0$, the solution of the optimization problem is $\pi_{k+1}(\cdot \mid s) = \pi_k(\cdot \mid s)$. Thus, $\nabla\omega\left(s;\pi_{k+1}\right) = \nabla\omega\left(s;\pi_k\right)$ and the inequality trivially holds.

We now turn to discuss the case where $\sum_{m=1}^M \mathbb{1}\{s = s_m\} > 0$, i.e., $s \in \mathcal{S}_M^k$. We separate here the analysis for the euclidean and non-euclidean cases:

For the **euclidean case**, $\omega(\cdot) = \frac{1}{2}\|\cdot\|_2^2$. Thus, the derivative of ω at a state s is,

$$\nabla\omega\left(s;\pi\right) = \pi(\cdot \mid s). \tag{68}$$

By the first order optimality condition, for any state s and policy π,

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

Plugging in $\pi := \pi_k(\cdot \mid s)$, we get

$$\langle \nabla\omega\left(s;\pi_{k+1}\right) - \nabla\omega\left(s;\pi_k\right), \pi_{k+1}(\cdot \mid s) - \pi_k(\cdot \mid s)\rangle \leq t_k\langle \hat{q}_\lambda^{\pi_k}(s,\cdot) + \lambda\nabla\omega\left(s;\pi_k\right), \pi_k(\cdot \mid s) - \pi_{k+1}(\cdot \mid s)\rangle.$$

Plugging in (68), we have that

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

which can be also written as

$$\left\|\nabla\omega\left(s;\pi_{k+1}\right) - \nabla\omega\left(s;\pi_k\right)\right\|_2^2 \leq t_k\langle \hat{q}_\lambda^{\pi_k}(s,\cdot) + \lambda\nabla\omega\left(s;\pi_k\right), \pi_k(\cdot \mid s) - \pi_{k+1}(\cdot \mid s)\rangle.$$

Bounding the RHS using the Cauchy-Schwartz inequality, we get,

$$\left\|\nabla\omega\left(s;\pi_{k+1}\right) - \nabla\omega\left(s;\pi_k\right)\right\|_2^2 \leq t_k\left\|\hat{q}_\lambda^{\pi_k}(s,\cdot) + \lambda\nabla\omega\left(s;\pi_k\right)\right\|_2 \left\|\pi_k(\cdot \mid s) - \pi_{k+1}(\cdot \mid s)\right\|_2,$$

which is the same as

$$\left\|\nabla\omega\left(s;\pi_{k+1}\right) - \nabla\omega\left(s;\pi_k\right)\right\|_2^2 \leq t_k\left\|\hat{q}_\lambda^{\pi_k}(s,\cdot) + \lambda\nabla\omega\left(s;\pi_k\right)\right\|_2 \left\|\nabla\omega\left(s;\pi_{k+1}\right) - \nabla\omega\left(s;\pi_k\right)\right\|_2,$$

Dividing by $\left\|\nabla\omega\left(s;\pi_{k+1}\right) - \nabla\omega\left(s;\pi_k\right)\right\|_2 > 0$ and noticing that in case it is 0 the bound is trivially satisfied,

$$\left\|\nabla\omega\left(s;\pi_{k+1}\right) - \nabla\omega\left(s;\pi_k\right)\right\|_2 \leq t_k\left\|\hat{q}_\lambda^{\pi_k}(s,\cdot) + \lambda\nabla\omega\left(s;\pi_k\right)\right\|_2.$$

Finally, using the norm equivalence we get,

$$\|\nabla\omega(s;\pi_{k+1})-\nabla\omega(s;\pi_k)\|_\infty \le \|\nabla\omega(s;\pi_{k+1})-\nabla\omega(s;\pi_k)\|_2 \le t_k\left\|\hat{q}_\lambda^{\pi_k}(s,\cdot)+\lambda\nabla\omega(s;\pi_k)\right\|_2.$$

Using the fourth claim of Lemma 22 (in the euclidean setting), and the fact the this inequality holds uniformly for all $s\in\mathcal{S}_M^k$ concludes the result.

For the **non-euclidean case**, $\omega(s;\pi)=\sum_a \pi(a\mid s)\log\pi(a\mid s)$. Thus, the derivative at the state action pair, s,a, is

$$\nabla_{\pi(a|s)}\omega(s;\pi)=1+\log\pi(a\mid s).$$

Thus, the difference between two consecutive policies is:

$$\nabla_{\pi_{k+1}(a|s)}\omega(s;\pi_{k+1})-\nabla_{\pi_k(a|s)}\omega(s;\pi_k)=\log\pi_{k+1}(a\mid s)-\log\pi_k(a\mid s)$$

Restating (62),

$$\begin{aligned}\log\pi_{k+1}(a\mid s)&=\log\pi_k(a\mid s)\\&\quad-t_k(\hat{q}_\lambda^{\pi_k}(s,a)+\lambda\log\pi_k(a\mid s))\\&\quad-\log\left(\sum_{a'}\pi_k(a'\mid s)\exp(-t_k(\hat{q}_\lambda^{\pi_k}(s,a')+\lambda\log\pi_k(a'\mid s)))\right).\end{aligned}$$

First, we will bound $\log\pi_{k+1}(a\mid s)-\log\pi_k(a\mid s)$ from below:

Similarly to equation 64, bounding the last term in the RHS,

$$\log\left(\sum_{a'}\pi_k(a'\mid s)\exp(-t_k(\hat{q}_\lambda^{\pi_k}(s,a')+\lambda\log\pi_k(a'\mid s)))\right)\le t_k\lambda\log A.$$

Together with the fact that $\lambda t_k\log\pi_k(a\mid s)\le 0$, we obtain,

$$\log\pi_{k+1}(a\mid s)-\log\pi_k(a\mid s)\ge -t_k(\hat{q}_\lambda^{\pi_k}(s,a)+\lambda\log A)\ge -t_k\left(\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}+\lambda\log A\right)\ge -2t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma},$$

where the last relation is by the definition of $\mathrm{C}_{\max,\lambda}$

Next, it is left to bound $\log\pi_{k+1}(a\mid s)-\log\pi_k(a\mid s)$ from above. Notice that,

$$\begin{aligned}\log\left(\sum_{a'}\pi_k(a'\mid s)\exp(-t_k(\hat{q}_\lambda^{\pi_k}(s,a')+\lambda\log\pi_k(a'\mid s)))\right)&\ge\log\sum_{a'}\pi_k(a'\mid s)\exp\left(-t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}-\lambda t_k\log\pi(a'\mid s)\right)\\&\ge\log\sum_{a'}\pi_k(a'\mid s)\exp\left(-t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}\right)\\&=\log\sum_{a'}\pi_k(a'\mid s)+\log\exp\left(-t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}\right)\\&=-t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma},\end{aligned}$$

where in the first transition we used the fact that in the sample-based case $\|\hat{q}_\lambda^{\pi_k}\|_{\infty,\infty}\le\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}$ due to the importance sampling applied in the estimation process, in the second transition we used the fact that the exponent is minimized when $\lambda t_k\log\pi(a'|s)$ is maximized and the fact that $\log\pi(a'|s)\le 0$, and the last transition is by the fact $\sum_{a'}\pi_k(a'|s)=0$.

Thus, we have

$$\begin{aligned}\log\pi_{k+1}(a\mid s)-\log\pi_k(a\mid s)&\le -t_k(\hat{q}_\lambda^{\pi_k}(s,a)+\lambda\log\pi_k(a\mid s))+t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}\\&\le t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}-\lambda t_k\log\pi_k(a\mid s)\\&\le t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}+\lambda t_k\frac{A\,\mathrm{C}_{\max}+2A\lambda\log A}{\lambda(1-\gamma)}(1+\log k)\\&\le t_k\frac{4A\,\mathrm{C}_{\max,\lambda}\log k}{1-\gamma},\end{aligned}$$

where the third transition is due to (67), and the last transition is by the the definition of $\mathrm{C}_{\max,\lambda}$.

Combining the two bounds we have,

$$\begin{aligned}
&-2t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma} \leq \log\pi_{k+1}(a\mid s) - \log\pi_k(a\mid s) \leq 4t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}\log k \\
\iff\ &-2t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma} \leq 1+\log\pi_{k+1}(a\mid s) - (1-\log\pi_k(a\mid s) \leq 4t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}\log k \\
\iff\ &-2t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma} \leq \nabla_{\pi_{k+1}(a|s)}\omega\left(s;\pi_{k+1}\right) - \nabla_{\pi_k(a|s)}\omega\left(s;\pi_k\right) \leq 4t_k\frac{A\,\mathrm{C}_{\max,\lambda}}{1-\gamma}\log k,
\end{aligned}$$

which concludes the proof.

□

Lemma 24 (bounds on initial distance D_ω)**.** *Let π_0 be the uniform policy over all states, and D_ω be an upper bound on $\max_\pi \left\|B_\omega\left(\pi_0,\pi\right)\right\|_\infty$, i.e., $\max_\pi \left\|B_\omega\left(\pi_0,\pi\right)\right\| \leq D_\omega$. Then, the following claims hold.*

1. *For $\omega(\cdot)=\frac{1}{2}\left\|\cdot\right\|_2^2$, $D_\omega=1$.*
2. *For $\omega(\cdot)=H(\cdot)$, $D_\omega=\log A$.*

Proof. For brevity, without loss of generality we omit the dependency on the state s. We start by proving the first claim. For the euclidean case,

$$\begin{aligned}
B_\omega\left(\pi,\pi_0\right) &= \frac{1}{2}\left\|\pi-\pi_0\right\|_2^2 \\
&= \frac{1}{2}\sum_a(\pi(a)-\frac{1}{A})^2 \\
&\leq \frac{1}{2}\sum_a\pi^2(a)+\sum_a\frac{1}{A^2} \\
&= \frac{1}{2A}+\frac{1}{2}\sum_a\pi^2(a) \\
&\leq \frac{1}{2A}+\frac{1}{2}\sum_a\pi(a)=\frac{1}{2A}+\frac{1}{2},
\end{aligned}$$

where the fifth relation holds since $x^2\leq x$ for $x\in[0,1]$, and the sixth relation holds since π is a probability measure.

For the non-euclidean case the following relation holds.

$$\begin{aligned}
B_\omega\left(\pi,\pi_0\right) &= d_{KL}(\pi||\pi_0) \\
&= \sum_a\pi(a)\log A\pi(a) \\
&= \sum_a\pi(a)\log\pi(a)+\sum_a\pi(a)\log A \\
&= \sum_a\pi(a)\log\pi(a)+\log A\sum_a\pi(a) \\
&= H(\pi)+\log A,
\end{aligned}$$

where H is the negative entropy. Since $H(\pi)\leq 0$ we get that $B_\omega\left(\pi,\pi_0\right)\leq\log A$ and conclude the proof.

□

The following Lemma as many instances in previous literature (e.g., [Scherrer and Geist, 2014][Lemma 1]) in the unregularized case, when $\lambda=0$. Here we generalize it to the regularized case, for $\lambda>0$.

Lemma 25 (value difference to Bellman differences)**.** *For any policies π and π', the following claims hold:*

1. $v_\lambda^{\pi'} - v_\lambda^{\pi} = (I - \gamma P^{\pi'})^{-1}(T_\lambda^{\pi'} v_\lambda^{\pi} - v_\lambda^{\pi})$.
2. $T_\lambda^{\pi'} v_\lambda^{\pi} - v_\lambda^{\pi} = (I - \gamma P^{\pi'})(v_\lambda^{\pi'} - v_\lambda^{\pi})$.
3. $\mu\Big(v_\lambda^{\pi'} - v_\lambda^{\pi}\Big) = \frac{1}{1-\gamma} d_{\mu,\pi'}(T_\lambda^{\pi'} v_\lambda^{\pi} - v_\lambda^{\pi})$.

Proof. The first claim holds by the following relations.

$$\begin{aligned} v_\lambda^{\pi'} - v_\lambda^{\pi} &= (I - \gamma P^{\pi'})^{-1} c_\lambda^{\pi'} - (I - \gamma P^{\pi'})^{-1}(I - \gamma P^{\pi'}) v_\lambda^{\pi} \\ &= (I - \gamma P^{\pi'})^{-1}(c_\lambda^{\pi'} + \gamma P^{\pi'} v_\lambda^{\pi} - v_\lambda^{\pi}) \\ &= (I - \gamma P^{\pi'})^{-1}(T_\lambda^{\pi'} v_\lambda^{\pi} - v_\lambda^{\pi}). \end{aligned}$$

The second claim follows by multiplying both sides by $(I - \gamma P^{\pi'})$. The third claim holds by multiplying both sides of the first claim by μ and using the definition $d_{\mu,\pi'} = (1-\gamma)\mu(I - \gamma P^{\pi'})^{-1}$. □

H Useful Lemmas from Convex Analysis

We state two basic results which are essential to the analysis of convergence. A full proof can be found in [Beck, 2017].

Lemma 26 (Beck 2017, Lemma 9.11, three-points lemma)**.** *Suppose that $\omega : \mathbb{E} \to (-\infty, \infty]$ is proper closed and convex. Suppose in addition that ω is differentiable over dom$(\partial\omega)$. Assume that $\mathbf{a}, \mathbf{b} \in$ dom$(\partial\omega)$ and $\mathbf{c} \in$ dom(ω). Then the following equality holds:*

$$\langle \nabla\omega(\mathbf{b}) - \nabla\omega(\mathbf{a}), \mathbf{c} - \mathbf{a}\rangle = B_\omega(\mathbf{c}, \mathbf{a}) + B_\omega(\mathbf{a}, \mathbf{b}) - B_\omega(\mathbf{c}, \mathbf{b}).$$

Theorem 27 (Beck 2017, Theorem 9.12, non-euclidean second prox theorem)**.**

- $\omega : \mathbb{E} \to (-\infty, \infty]$ *be a proper closed and convex function differentiable over dom$(\partial\omega)$.*
- $\psi : \mathbb{E} \to (-\infty, \infty]$ *be a proper closed and convex function satisfying dom$(\psi) \subseteq$ dom(ω).*
- $\omega + \delta_{dom(\psi)}$ *be σ-strongly convex ($\sigma > 0$).*

Assume that $\mathbf{b} \in$ dom$(\partial\omega)$, and let $\mathbf{a}$ be defined by

$$\mathbf{a} = \arg\min_{\mathbf{x}\in\mathbb{E}}\{\psi(\mathbf{x}) + B_\omega(\mathbf{x}, \mathbf{b})\}.$$

Then $\mathbf{a} \in$ dom$(\partial\omega)$ and for all $\mathbf{u} \in$ dom(ψ),

$$\langle \nabla\omega(\mathbf{b}) - \nabla\omega(\mathbf{a}), \mathbf{u} - \mathbf{a}\rangle \le \psi(\mathbf{u}) - \psi(\mathbf{a}).$$