A Novel Framework for Policy Mirror Descent with General Parametrization and Linear Convergence

Anonymous Authors[1]

Abstract

Modern policy optimization methods in applied reinforcement learning, such as Trust Region Policy Optimization and Policy Mirror Descent, are often based on the policy gradient framework. While theoretical guarantees have been established for this class of algorithms, particularly in the tabular setting, the use of a general parametrization scheme remains mostly unjustified. In this work, we introduce a novel framework for policy optimization based on mirror descent that naturally accommodates general parametrizations. The policy class induced by our scheme recovers known classes, e.g. softmax, and it generates new ones, depending on the choice of the mirror map. For a general mirror map and parametrization class, we establish the quasi-monotonicity of the updates in value function, global linear convergence rates, and we bound the total expected Bregman divergence of the algorithm along its path. To showcase the ability of our framework to accommodate general parametrization schemes, we present a case study involving shallow neural networks.

1. Introduction

Policy optimization represents one of the most widely-used classes of algorithms for reinforcement learning (RL). Among policy optimization techniques, policy gradient (PG) methods (Williams and Peng, 1991; Sutton et al., 1999; Konda and Tsitsiklis, 2000; Baxter and Bartlett, 2001) are gradient-based algorithms that optimize the policy over a parametrized policy class and have emerged as a popular class of algorithms for RL (Kakade, 2002; Peters and Schaal, 2008; Bhatnagar et al., 2009; Mnih et al., 2016; Lan, 2022).

The design of gradient-based policy updates has been key in achieving empirical successes in many settings, such as games (Berner et al., 2019) and autonomous driving (Shalev-Shwartz et al., 2016). In particular, a class of PG algorithms that has proven successful in practice consists in building updates that include a hard constraint (e.g. trust region constraint) or a penalty term ensuring that the updated policy does not move too far away from the previous policy. Two main examples of algorithms belonging to this category are trust region policy optimization (TRPO) (Schulman et al., 2015a), which imposes a Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) constraint on its updates, and policy mirror descent (PMD) (Tomar et al., 2022; Lan, 2022; Xiao, 2022; Kuba et al., 2022; Vaswani et al., 2022), which applies mirror descent (MD) (Nemirovski and Yudin, 1983) to RL. Shani et al. (2020) propose a variant of TRPO which is in fact a special case of PMD, thus draw the connection between TRPO and PMD.

From a theoretical perspective, motivated by the empirical success of PMD, there is now a concerted effort to develop convergence theories for PMD methods. For instance, it has been established that PMD converges linearly to the global optimum in the tabular setting by using a geometrically increasing step size (Lan, 2022; Xiao, 2022), by adding entropy regularization (Cen et al., 2021) and more generally by adding convex regularization (Zhan et al., 2021). The linear convergence of PMD has also been established for the negative entropy mirror map in the linear function approximation regime, i.e. for log-linear policies, either by adding entropy regularization (Cayci et al., 2021), or by using a geometrically increasing step size (Chen and Theja Maguluri, 2022; Alfano and Rebeschini, 2022; Yuan et al., 2023). The proof of those results rely on specific policy parametrizations, that is tabular and log-linear, while PMD remains mostly unjustified for general policy parametrizations, which leaves out important practical cases such as neural networks. In particular, it remains to see if the theoretical results

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

obtained for tabular policy classes transfer to this more general setting. We refer to Appendix A.2 for a thorough review of the PMD literature. In particular, Table 1 and 2 provide a detailed overview of our results.

In this work, we provide an affirmative answer to that question by proposing a novel framework based on the MD algorithm which recovers PMD in the tabular setting, is capable of generating new algorithms and is amenable to theoretical analysis for any parametrization class. Since the MD update can be viewed as a two-step procedure, that is an update on the dual space and a mapping onto the probability simplex, our starting point is to define the policy class based on this second MD step. This policy class recovers the softmax policy class as a special case (Example 3.2) and accommodates any parametrization class, such as tabular, linear or neural network parametrizations. We then develop a new Approximate Mirror Policy Optimization (AMPO) framework for this policy class, based on the actor-critic family (Konda and Tsitsiklis, 2000) and on the mirror descent methodology. We illustrate the versatility of our framework by instantiating our algorithm with different choices of mirror map in Examples 3.3, 3.4 and 3.5.

In addition, we provide a theoretical analysis of AMPO. The key point in our analysis is Lemma 4.1, which is an extension of the three-point descent lemma given by Chen and Teboulle (1993, Lemma 3.2) to the function approximation setting, thus greatly expanding the scope of applications of the lemma. Such extension is highly nontrivial and is established by exploiting the non-expansivity property of the Bregman projection (Lemma D.2) to account for general parametrization classes. This result, together with the formulation of AMPO, permits us to keep track of errors induced by our choice of policy class.

Building on Lemma 4.1 and leveraging the PMD proof techniques from Xiao (2022), we establish theoretical guarantees for AMPO that hold for any parametrization class and any mirror map. More precisely, we show that our algorithm enjoys quasi-monotonic improvements (Proposition 4.2), sublinear convergence when the step size is non-decreasing and linear convergence when the step size is increasing geometrically (Theorem 4.3). Additionally, we give a bound on the total expected Bregman divergence between consequent policies along the path of the algorithm (Corollary 4.4), which sheds light on how the choice of mirror map plays a role in the regularization of the algorithm. To the best of our knowledge, AMPO is the first gradient-based policy optimization algorithm with linear convergence that can accommodate any parametrization class and choice of mirror map.

Lastly, we apply our results to the important case of shallow neural networks, where further theoretical derivations enable us to show that the error coming from our choice of policy class can be made arbitrarily small by increasing the width of the network (Theorem 4.5).

2. Preliminaries

In this section, we first present the main RL setting before reviewing the mirror descent methodology.

2.1. RL setting

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$ be a discounted Markov Decision Process (MDP), where $\mathcal{S}$ is a possibly infinite state space, $\mathcal{A}$ is a finite action space, $P(s'|s,a)$ is the transition probability from state s to s' under action a, $r(s,a) \in [0,1]$ is a reward function, γ is a discount factor, and μ is a target state distribution.

The behaviour of an agent on an MDP is then modelled by a *policy* $\pi \in (\Delta(\mathcal{A}))^{\mathcal{S}}$, where $a \sim \pi(\cdot \mid s)$ is the density of the distribution over actions at state $s \in \mathcal{S}$ and $\Delta(\mathcal{A})$ is the probability simplex over $\mathcal{A}$.

Given a policy π, let $V^\pi : \mathcal{S} \to \mathbb{R}$ denote the associated *value function*. Letting s_t and a_t be the current state and action at time t, the value function V^π is defined as the expected discounted cumulative reward with starting state $s_0 = s$, namely,

$$V^\pi(s) := \mathop{\mathbb{E}}_{\substack{a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi, s_0 = s \right].$$

Now letting $V^\pi(\mu) := \mathbb{E}_{s \sim \mu}[V^\pi(s)]$, one of the main objectives in RL is for the agent to find an optimal policy

$$\pi^\star \in \mathop{\mathrm{argmax}}_{\pi \in (\Delta(\mathcal{A}))^{\mathcal{S}}} V^\pi(\mu). \tag{1}$$

Similarly to the value function, for each pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, the state-action value function, or *Q-function*, associated to the policy π is defined as

$$Q^\pi(s,a) := \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid \pi, s_0 = s, a_0 = a \right].$$

where $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$. We also define the discounted state visitation distribution by

$$d_\mu^\pi(s) := (1-\gamma)\mathbb{E}_{s_0 \sim \mu} \left[\sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \pi, s_0) \right],$$

where $P(s_t = s \mid \pi, s_0)$ represents the probability of the agent being in state s at time t when following policy π

and starting from s_0. The probability $d_\mu^\pi(s)$ represents the time spent on state s when following policy π.

The gradient of the value function $V^\pi(\mu)$ with respect to the policy can be easily expressed by the policy gradient theorem (Sutton et al., 1999):

$$\nabla_s V^\pi(\mu) := \frac{\partial V^\pi(\mu)}{\partial \pi(\cdot|s)} = \frac{1}{1-\gamma} d_\mu^\pi(s) Q^\pi(s,\cdot). \quad (2)$$

We next introduce the mirror descent framework which we will use to define and motivate a novel framework.

2.2. Mirror descent

The first tools we recall from the MD framework are mirror maps and Bregman divergences (Bubeck, 2015, Chapter 4). Let $\mathcal{Y} \subseteq \mathbb{R}^{|\mathcal{A}|}$ be a convex set. A *mirror map* $h : \mathcal{Y} \to \mathbb{R}$ is a strictly convex, continuously differentiable and essentially smooth function[1] such that $\nabla h(\mathcal{Y}) = \mathbb{R}^{|\mathcal{A}|}$. The convex conjugate of h, denoted by h^*, is given by

$$h^*(x^*) := \sup_{x \in \mathcal{Y}} \langle x^*, x \rangle - h(x), \quad x^* \in \mathbb{R}^{|\mathcal{A}|}.$$

The gradient of the mirror map $\nabla h : \mathcal{Y} \to \mathbb{R}^{|\mathcal{A}|}$ allows to map objects from the primal space $\mathcal{Y}$ to its dual space $\mathbb{R}^{|\mathcal{A}|}$, $x \mapsto \nabla h(x)$, and viceversa for ∇h^*, i.e. $x^* \mapsto \nabla h^*(x^*)$. In particular, from $\nabla h(\mathcal{Y}) = \mathbb{R}^{|\mathcal{A}|}$, we have: for all $(x, x^*) \in \mathcal{Y} \times \mathbb{R}^{|\mathcal{A}|}$,

$$x = \nabla h^*(\nabla h(x)) \quad \text{and} \quad x^* = \nabla h(\nabla h^*(x^*)). \quad (3)$$

Furthermore, the mirror map h induces a Bregman divergence (Bregman, 1967; Censor and Zenios, 1997), defined as

$$\mathcal{D}_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle,$$

where $\mathcal{D}_h(x, y) \geq 0$ for all $x, y \in \mathcal{Y}$. We can now present the standard MD algorithm (Nemirovski and Yudin, 1983; Bubeck, 2015). Let $\mathcal{X} \subseteq \mathcal{Y}$ be a convex set and $V : \mathcal{X} \to \mathbb{R}$ be a differentiable function[2]. The MD algorithm can be formalized[3] as the following iterative procedure in order to solve the minimization problem $\min_{x \in \mathcal{X}} V(x)$: for all $t \geq 0$,

$$y^{t+1} = \nabla h(x^t) - \eta_t \nabla V(x)|_{x=x^t}, \quad (4)$$
$$x^{t+1} = \mathrm{Proj}_{\mathcal{X}}^h(\nabla h^*(y^{t+1})), \quad (5)$$

where η_t is set according to a step size schedule $(\eta_t)_{t\geq 0}$ and $\mathrm{Proj}_{\mathcal{X}}^h(\cdot)$ is the Bregman projection

$$\mathrm{Proj}_{\mathcal{X}}^h(y) := \operatorname*{argmin}_{x \in \mathcal{X}} \mathcal{D}_h(x, y). \quad (6)$$

Precisely, at time t, $x^t \in \mathcal{X}$ is mapped to the dual space through $\nabla h(\cdot)$, where a gradient step is performed as in (4) to obtain y^{t+1}. The next step is to project y^{t+1} back in the primal space using $\nabla h^*(\cdot)$. In case $\nabla h^*(y^{t+1})$ does not belong to $\mathcal{X}$, it is projected as in (5).

In the next section, we explore how the MD setup we just presented can be applied in the context of RL to define a novel parametrized policy class alongside with an efficient update rule. We will then show how this framework can be applied for different parametrization schemes, including neural networks as special case.

3. Approximate Mirror Policy Optimization

The starting point of our framework is the introduction of a novel parametrized policy class, which relies on the Bregman projection expression recalled in (6).

Definition 3.1. Given a parametrized function class $\mathcal{F}^\Theta = \{f^\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}, \theta \in \Theta\}$, a mirror map $h : \mathcal{Y} \to \mathbb{R}$, where $\mathcal{Y} \subseteq \mathbb{R}^{|\mathcal{A}|}$ is a convex set with $\Delta(\mathcal{A}) \subseteq \mathcal{Y}$, and $\eta > 0$, the *Bregman projected policy* class associated to $\mathcal{F}^\Theta$ and h consists of all the policies of the form:

$$\left\{\pi^\theta : \pi_s^\theta = \mathrm{Proj}_{\Delta(\mathcal{A})}^h(\nabla h^*(\eta f_s^\theta)), \; s \in \mathcal{S}; \; \theta \in \Theta\right\},$$

where, for all $s \in \mathcal{S}$, $\pi_s^\theta := \pi^\theta(\cdot \mid s)$ and $f_s^\theta := f^\theta(s, \cdot)$.

In this definition, the policy is induced by a mirror map h and a parametrized function f^θ and is obtained by mapping f^θ to $\mathcal{Y}$ with the operator $\nabla h^*(\cdot)$, which may not be a well-defined probability distribution and is thus projected on the convex probability simplex $\Delta(\mathcal{A})$. Note that the choice of h will be key in deriving convenient expressions for the policy π^θ. The Bregman projected policy class contains large families of policy classes. We provide below an example of h that recovers widely used policy classes (Beck, 2017, Example 9.10).

Example 3.2 (Negative entropy mirror map)**.** If $\mathcal{Y} = \mathbb{R}_+^{|\mathcal{A}|}$ and h is the negative entropy mirror map, i.e. $h(\pi(\cdot|s)) = \sum_{a \in \mathcal{A}} \pi(a|s) \log(\pi(a|s))$, $\mathrm{Proj}_{\Delta(\mathcal{A})}^h(\nabla h^*(\cdot))$ is equivalent to the following common policy class

$$\left\{\pi^\theta : \pi_s^\theta = \frac{\exp(\eta f_s^\theta)}{\|\exp(\eta f_s^\theta)\|_1}, \; s \in \mathcal{S}; \; \theta \in \Theta\right\}, \quad (7)$$

where the exponential and the fraction are element-wise and $\|\cdot\|_1$ is ℓ_1 norm. In particular, when $f^\theta(s, a) = \theta_{s,a}$,

[1] h is essentially smooth if $\lim_{x \to \partial\mathcal{Y}} \|\nabla h(x)\|_2 = +\infty$, where $\partial\mathcal{Y}$ denotes the boundary of $\mathcal{Y}$.

[2] The results in our paper can be easily extended for subdifferentiable functions. Here we consider differentiable functions for simplicity.

[3] See a different formulation of MD in (10) or further in Appendix B (Lemma B.1).

the policy class (7) becomes tabular softmax policy; when f^θ is a linear function, (7) becomes the log-linear policy; and when f^θ is a neural network, (7) becomes the neural policy class defined by Agarwal et al. (2021). We refer to Appendix C.1 for proofs and more details.

We now construct a policy mirror descent type algorithm to optimize V^{π^θ} over the Bregman projected policy class associated to a mirror map h and a parametrization class $\mathcal{F}^\Theta$ by adapting Section 2.2 to our setting. First, we use the following shorthand: at each time t, let $\pi^t := \pi^{\theta_t}$, $f^t := f^{\theta_t}$, $V^t := V^{\pi^t}$, $Q^t := Q^{\pi^t}$, and $d^t_\mu := d^{\pi^t}_\mu$. Further, for any function $y : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and distribution v over $\mathcal{S} \times \mathcal{A}$, let $y_s := y(s, \cdot) \in \mathbb{R}^{|\mathcal{A}|}$ and $\|y\|^2_{L_2(v)} = \mathbb{E}_v[(y(s,a))^2]$. Ideally, we would like to execute the exact MD-based algorithm: for all $t \geq 0$ and for all $s \in \mathcal{S}$,

$$\begin{aligned} f^{t+1}_s &= \nabla h(\pi^t_s) + \eta_t (1-\gamma) \nabla_s V^t(\mu) \text{ [4]} \\ &\overset{(2)}{=} \nabla h(\pi^t_s) + \eta_t Q^t_s, \qquad (8) \\ \pi^{t+1}_s &= \mathrm{Proj}^h_{\Delta(\mathcal{A})}(\nabla h^*(\eta_t f^{t+1}_s)). \qquad (9) \end{aligned}$$

Here, (9) reflects our Bregman projected policy class 3.1. Once f^{t+1}_s is provided, π^{t+1}_s can be obtained with a computational cost of $\widetilde{O}(|\mathcal{A}|)$ (Krichene et al., 2015). However, we usually cannot perform the update (8) exactly. In general, we do not have access to the exact gradient of the value function and especially, when f^θ belongs to a parametrized class $\mathcal{F}^\Theta$, there might not exist any $\theta_{t+1} \in \Theta$ such that (8) is satisfied for all $s \in \mathcal{S}$.

To remedy these issues, we propose Approximate Mirror Policy Optimization (AMPO), described in Algorithm 1, which lies within the actor-critic algorithmic family (Konda and Tsitsiklis, 2000). More specifically, at each time t, our algorithm involves a critic $\mathcal{C}$ evaluating the policy and an actor which updates the policy, following an approximate version of updates (8) and (9).

• On the one hand, in line 2 the critic $\mathcal{C}$ returns a function $\widehat{Q}^t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ that estimates the Q-function Q^t associated to the current policy π^t and serves as an approximation of the gradient of the value function at time t (Agarwal et al., 2021; Xiao, 2022), hence solving the first issue. Since in this work we focus on the policy optimization side of RL, we will assume that we have access to the critic $\mathcal{C}$, yet keeping in mind that $\mathcal{C}$ will typically estimate the Q-function through rollout (Schulman et al., 2015b) or temporal difference learning (Tesauro, 1995).

• On the other hand, in line 3 the actor returns a function $f^{t+1} \in \mathcal{F}^\Theta$, such that the expected actor error is bounded by $\varepsilon_{\text{actor}}$ for a generic probability distribution v_t over $\mathcal{S} \times \mathcal{A}$. Line 3 is an instance of function approximation where we try to approximate $\widehat{Q}^t + \eta_t^{-1} \nabla h(\pi^t)$ with f^{t+1} and which has been extensively studied when f^θ is a neural network (Ji et al., 2019). At the same time, f^{t+1} is not constrained to a specific explicit update. If f^θ is differentiable with respect to θ, one can obtain f^{t+1} by minimizing the expression in line 3 with gradient descent methods. We can then readily use (9) to update the policy π^{t+1} within our defined policy class in line 4.

[4]The update is (4) up to a constant scaling $(1-\gamma)$ of η_t.

Algorithm 1: Approximate Mirror Policy Optimization

Input: Initial policy π^0, mirror map h, parametrization class $\mathcal{F}^\Theta$, actor error $\varepsilon_{\text{actor}} > 0$, critic $\mathcal{C}$, iteration number T, step size schedule $(\eta_t)_{t\geq 0}$, state-action distribution sequence $(v_t)_{t\geq 0}$.

1 **for** $t = 0$ ***to*** $T-1$ **do**
2   Obtain $\widehat{Q}^t$ from the critic $\mathcal{C}$.
3   Find $\theta_{t+1} \in \Theta$ such that
$$\|f^{\theta_{t+1}} - \widehat{Q}^t - \eta_t^{-1} \nabla h(\pi^t)\|^2_{L_2(v_t)} \leq \varepsilon_{\text{actor}}.$$
4     $\pi^{t+1}_s = \underset{\pi' \in \Delta(\mathcal{A})}{\operatorname{argmin}} \, \mathcal{D}_h(\pi', \nabla h^*(\eta_t f^{\theta_{t+1}}_s)), \ \forall s \in \mathcal{S}.$

Remark. Line 3 associates AMPO with the *compatible function approximation* framework developed by Sutton et al. (1999); Kakade (2002); Agarwal et al. (2021), as both frameworks aim to find the parameters θ^{t+1} that minimize an actor error defined as a regression problem. Differently from Agarwal et al. (2021), the regression problem in line 3 is not constrained to be linear or to depend on the distrubution d^t_μ.

To better illustrate the novelty of our framework, we now give a comparison between AMPO and previous approximations of PMD (Vaswani et al., 2022; Tomar et al., 2022). In both approaches, the algorithm provides an expression to optimize. For AMPO this expression is the one in line 3 of Algorithm 1, while, for instance, Vaswani et al. (2022) aim to maximize an expression that is similar to

$$\pi^{t+1} = \underset{\pi \in \Pi}{\operatorname{argmin}} \, \mathbb{E}_{s \sim d^t_\mu} [\langle Q^t_s, \pi_s \rangle + D_h(\pi_s, \pi^t_s)], \qquad (10)$$

where Π is a given policy class. When the policy class Π is the entire policy space $\Delta(\mathcal{A})^{\mathcal{S}}$, this is equivalent to the two step procedure (8)-(9), thanks to the policy gradient theorem (2). A derivation of this observation is provided in Appendix B (Lemma B.1) for completeness. However, in practice the parametrized policy class Π is often non-convex with respect to θ, which prevents the application of modern PMD proof techniques (Xiao, 2022) relying on the convexity of the tabular parametrization. On the contrary, AMPO side-steps this problem thanks to the definition of the Bregman projected policy class and the update in line 4 of Algorithm 1, as we will see in the theoretical analysis. We refer to Appendix A.1 for a thorough review of the PMD framework literature.

AMPO provides a flexible framework that accommodates any parametrization class $\mathcal{F}^\Theta$, as we highlight in the following examples where we instantiate AMPO for specific choices of mirror map and show how it can recover existing approaches to policy optimization.

Example 3.3 (Squared ℓ_2-norm)**.** If $\mathcal{Y} = \mathbb{R}^{|\mathcal{A}|}$ and h is the squared ℓ_2-norm, that is $h(\pi_s) = \|\pi_s\|^2/2$, line 3 in Algorithm 1 becomes

$$\left\| f^{t+1}(s,a) - \widehat{Q}^t(s,a) - \eta_t^{-1}\pi^t(a|s) \right\|_{L_2(v_t)}^2 \le \varepsilon_{\text{actor}},$$

and the policy update is given for all $s \in \mathcal{S}$ by

$$\pi_s^{t+1} = \text{Proj}_{\Delta(\mathcal{A})}^{l_2}(\eta_t f_s^{t+1}), \tag{11}$$

where $\text{Proj}_{\Delta(\mathcal{A})}^{l_2}$ represents the Euclidean projection on the probability simplex. In the tabular setting where $\mathcal{S}$ and $\mathcal{A}$ are finite and $f^\theta(s,a) = \theta_{s,a}$, $\varepsilon_{\text{actor}}$ can be set to 0 (Agarwal et al., 2021) and (11) recovers the projected Q-descent algorithm (Xiao, 2022). We refer to Appendix C.2 for detailed derivations. As a byproduct, we generalize the projected Q-descent algorithm from the tabular setting to a general parametrization class $\mathcal{F}^\Theta$, which is a novel algorithm in the RL literature.

Example 3.4 (Negative entropy)**.** If h is the negative entropy mirror map from Example 3.2, line 3 in Algorithm 1 becomes

$$\left\| f^{t+1} - \widehat{Q}^t - \frac{\eta_{t-1}}{\eta_t} f^t \right\|_{L_2(v_t)}^2 \le \varepsilon_{\text{actor}}. \tag{12}$$

Consequently, based on Example 3.2, we have

$$\pi_s^{t+1} = \frac{\exp(\eta_t f_s^{t+1})}{\left\|\exp(\eta_t f_s^{t+1})\right\|_1}, \tag{13}$$

for all $s \in \mathcal{S}$. In this example, AMPO recovers tabular NPG (Shani et al., 2020) when $f^\theta(s,a) = \theta_{s,a}$, and NPG with log-linear polices (Yuan et al., 2023) when f^θ and $\widehat{Q}^t$ are linear functions for all $t \ge 0$. We refer to Appendix C.1 for detailed derivations and an extension to Tsallis entropy.

We next present an example for a mirror map that has not yet been considered in the RL literature.

Example 3.5 (Hyperbolic entropy)**.** If $\mathcal{Y} = \mathbb{R}^{|\mathcal{A}|}$ and h_b is the hyperbolic entropy mirror map with a scalar parameter $b > 0$, that is

$$h_b(\pi_s) = \sum_{a \in \mathcal{A}} \pi(a|s)\,\text{arcsinh}(\pi(a|s)/b) - \sqrt{\pi(a|s)^2 + b^2},$$

the update in line 3 of Algorithm 1 with $h = h_b$ becomes

$$\left\| f^{t+1} - \widehat{Q}^t - \frac{\eta_{t-1}}{\eta_t}(f^t - \text{arcsinh}\, c_s) \right\|_{L_2(v_t)}^2 \le \varepsilon_{\text{actor}},$$

and the policy satisfies

$$\pi_s^{t+1} = b\sinh(f_s^{t+1})\sqrt{1+c_s^2} - b\cosh(f_s^{t+1})c_s, \quad \forall\, s \in \mathcal{S},$$

where $c_s \in \mathbb{R}$ has an explicit expression for all times t. We refer to Appendix C.3 for detailed derivations.

Given the extensive literature on mirror maps (Orabona, 2020; Vaškevičius et al., 2020; Ghai et al., 2020), we expect our approach to pave the way for the use of mirror maps tailored for MDP structures. We leave this direction as future work but outline in the example below how to instantiate AMPO for a wide class of mirror maps, which includes the negative entropy and the hyperentropy as particular cases.

Example 3.6 (ω-potential mirror map)**.** Let the mirror map h be defined as

$$h(\pi_s) = \sum_{a \in \mathcal{A}} \int_1^{\pi(a|s)} \phi^{-1}(u)du$$

where ϕ is a ω-potential[5]. Using a result by Krichene et al. (2015, Proposition 2), the policy π^t obtained by Algorithm 1 can be written as

$$\pi^t(a|s) = [\phi(\eta_{t-1} f^t(s,a) + \lambda_s)]_+ \quad \forall s \in \mathcal{S}, a \in \mathcal{A},$$

where $\lambda_s \in \mathbb{R}$ for all $s \in \mathcal{S}$ and $[z]_+ = \max(z, 0)$ for $z \in \mathbb{R}$. We refer to this class as the *ω-potential policy class*. While there are some cases where λ_s has an explicit expression, as we have shown in Examples 3.4 and 3.5, this does not appear possible for all mirror maps. We refer to Appendix C.4 for more details.

4. Theoretical analysis

This section is devoted to the theoretical analysis of AMPO, which will rely on the following lemma. For convenience, denote $\mathcal{D}_{\bar\pi}^{\pi}(s) = \mathcal{D}_h(\pi_s, \bar\pi_s)$ for all $s \in \mathcal{S}$.

[5]For $a \in (-\infty, +\infty]$ and $\omega \le 0$, an increasing C^1-diffeomorphism $\phi : (-\infty, a) \to (\omega, +\infty)$ is called an ω-potential if

$$\lim_{u \to -\infty} \phi(u) = \omega, \lim_{u \to a} \phi(u) = +\infty, \int_0^1 \phi^{-1}(u)du \le \infty.$$

Lemma 4.1. *For any policies π and $\bar{\pi}$, for any function $f^\theta \in \mathcal{F}^\Theta$ and for $\eta > 0$, we have for all $s \in \mathcal{S}$*

$$\langle \eta f_s^\theta - \nabla h(\bar{\pi}_s), \pi_s - \tilde{\pi}_s \rangle \leq \mathcal{D}_{\bar{\pi}}^{\pi}(s) - \mathcal{D}_{\bar{\pi}}^{\tilde{\pi}}(s) - \mathcal{D}_{\tilde{\pi}}^{\pi}(s),$$

where $\tilde{\pi}$ is the Bregman projected policy induced by f^θ and h according to Definition 3.1, that is $\tilde{\pi}_s = \operatorname{argmin}_{\pi' \in \Delta(\mathcal{A})} \mathcal{D}_h(\pi', \nabla h^*(\eta f_s^\theta))$ for all $s \in \mathcal{S}$.

The proof of Lemma 4.1 is presented in Appendix D.1. Lemma 4.1 describes a relationship between any two policies and a policy belonging to the Bregman projected policy class associated to $\mathcal{F}^\Theta$ and h. While similar results have been obtained and exploited for the tabular setting through convexity (Xiao, 2022) and for the negative entropy mirror map (Liu et al., 2019; Hu et al., 2022; Yuan et al., 2023), Lemma 4.1 is the first to allow any parametrization class $\mathcal{F}^\Theta$ and any choice of mirror map, thus greatly expanding the scope of applications of the lemma. Since Lemma 4.1 does not depend on Algorithm 1, we expect it to be helpful in contexts outside this work.

Lemma 4.1 becomes useful when we set $\bar{\pi} = \pi^t$, $f^\theta = f^{t+1}$, $\eta = \eta_t$ and $\pi = \pi^t$ or $\pi = \pi^\star$. In particular, when $\eta_t f_s^{t+1} - \nabla h(\pi_s^t) \approx \eta_t Q_s^\pi$, which is the case in the proof of Lemma 7 in Xiao (2022), Lemma 4.1 enables us to obtain telescopic sums and recursive relationships and to handle error terms efficiently, as we show in Appendix D. This is possible thanks to our two-step formulation (4)-(5) applied in Algorithm 1, whereas Lemma 4.1 cannot be applied to algorithms based on the update in (10) (Tomar et al., 2022; Vaswani et al., 2022), due to the non-convexity of the optimization problem.

In the next section, we consider the parametrization class $\mathcal{F}^\Theta$ and the mirror map h fixed but arbitrary.

4.1. General policy parametrization

We show that AMPO enjoys (i) quasi-monotonic improvements, (ii) sublinear or linear convergence depending on the step size schedule and (iii) upper-bounded total expected Bregman divergence of the algorithm along its path. The first step to do so is to control the approximation error of AMPO by Assumptions (A1) and (A2) below.

(A1) (*Critic error*). There exists $\varepsilon_{\text{critic}} \geq 0$ such that, for all times $t \geq 0$, the critic $\mathcal{C}$ returns a function $\widehat{Q}^t$ that satisfies

$$\left\| Q^t - \widehat{Q}^t \right\|_{L_2(\rho_t)}^2 \leq \varepsilon_{\text{critic}},$$

where $(\rho^t)_{t\geq 0}$ is a sequence of distributions over states and actions.

(A2) (*Actor error*). There exists $\varepsilon_{\text{actor}} \geq 0$ such that, for all times $t \geq 0$, the actor returns a function f^{t+1} that satisfies

$$\left\| f^{t+1} - \widehat{Q}^t - \eta_t^{-1} \nabla h(\pi^t) \right\|_{L_2(v_t)}^2 \leq \varepsilon_{\text{actor}},$$

where $(v^t)_{t\geq 0}$ is a sequence of distributions over states and actions.

Assumption (A1) requires the critic to provide an estimate of the Q-function with a bounded expected square error, where the expectation is taken over some arbitrary distribution. Several works focus on designing methods to provide such estimates (Sutton et al., 1998; Schulman et al., 2015b; Espeholt et al., 2018). It is weaker than the standard critic error assumption in the RL literature (Agarwal et al., 2021; Cayci et al., 2021; Yuan et al., 2023), as v_t can be any distribution, not necessarily dependent to π^t.

Assumption (A2) ensures the update line 3 of Algorithm 1 holds so that AMPO is well defined. In particular, we show later in Section 4.2 that, when $\mathcal{F}^\Theta$ is a class of shallow neural networks, it is always possible to find f^{t+1} such that Assumption (A2) holds, provided that a modulus of continuity condition is respected. When the step size η_t is sufficiently large, Assumption (A2) measures how well f^{t+1} approximates the Q-function estimate $\widehat{Q}^t$. We reflect this observation again later in our analysis.

We highlight that in both assumptions, the distributions ρ^t and v^t do not depend on the current policy π^t for all times $t \geq 0$. Therefore, Assumptions (A1) and (A2) allow off-policy policy evaluation techniques (Thomas and Brunskill, 2016) and policy updates, enabling the use of replay buffers (Mnih et al., 2015). To quantify how the choice of these distributions affect the error terms in the convergence rates, we introduce the following two coefficients.

(A3) (*Concentrability coefficients*). There exist $C_1 \geq 0$ and $C_2 \geq 0$ such that, for all times t, the distributions ρ^t and v^t satisfy

$$\mathbb{E}_{(s,a)\sim\rho^t}\left[\left(\frac{d_\mu^\pi(s)\pi(a|s)}{\rho^t(s,a)} \right)^2 \right] \leq C_1$$

and

$$\mathbb{E}_{(s,a)\sim v^t}\left[\left(\frac{d_\mu^\pi(s)\pi(a|s)}{v^t(s,a)} \right)^2 \right] \leq C_2,$$

whenever (d_μ^π, π) is either $(d_\mu^\star, \pi^\star)$, (d_μ^{t+1}, π^{t+1}), $(d_\mu^\star, \pi^t)$ or (d_μ^{t+1}, π^t).

The concentrability coefficients C_1 and C_2 describe how much the distributions ρ^t and v^t overlap with the distributions $(d_\mu^\star, \pi^\star)$, (d_μ^{t+1}, π^{t+1}), $(d_\mu^\star, \pi^t)$ and (d_μ^{t+1}, π^t). Our (A3) is weaker than the previous best known concentrability coefficients in Yuan et al. (2023, Assumption 9) in the sense that we have the full control of (ρ^t, v^t). We refer to Appendix D.8 for a more detailed discussion on the concentrability coefficients.

The quantities we have introduced so far determine the main error term in the algorithm, which we denote by

$$\tau = \frac{2}{1-\gamma}(\sqrt{C_1 \varepsilon_{\text{critic}}} + \sqrt{C_2 \varepsilon_{\text{actor}}}).$$

We can now present Proposition 4.2, which shows that Algorithm 1 enjoys quasi-monotonic updates.

Proposition 4.2. *Let (A1), (A2) and (A3) be true. Then, for all times $t \geq 0$, Algorithm 1 enjoys quasi-monotonic updates, that is*

$$V^{t+1}(\mu) - V^t(\mu) \geq -\tau.$$

We refer to Appendix D.3 for a proof and a tighter bound. Proposition 4.2 ensures that an update of Algorithm 1 cannot lead to a decrease in performance larger than τ.

We next present our main convergence results. To do so, we need the following assumption on the state space coverage for the agent at each time t.

(A4) (*Distribution mismatch coefficient*). Let $d_\mu^\star := d_\mu^{\pi^\star}$. There exists $C_3 \geq 0$ such that, for all times $t \geq 0$,

$$\max_{s \in \mathcal{S}} \frac{d_\mu^\star(s)}{d_\mu^t(s)} \leq C_3.$$

Since $d_\mu^t(s) \geq (1-\gamma)\mu(s)$ for all $s \in \mathcal{S}$, we have that

$$\max_{s \in \mathcal{S}} \frac{d_\mu^\star(s)}{d_\mu^t(s)} \leq \frac{1}{1-\gamma} \max_{s \in \mathcal{S}} \frac{d_\mu^\star}{\mu},$$

where assuming boundedness for the term on the right-hand side is standard in the literature on the PG convergence analysis (e.g., Zhang et al., 2020; Wang et al., 2020) and the NPG convergence analysis (e.g., Agarwal et al., 2021; Cayci et al., 2021; Xiao, 2022).

We also denote the expected Bregman divergence between the optimal policy and the initial policy π^0 by $\mathcal{D}_0^\star = \mathbb{E}_{s \sim d_\mu^\star}[\mathcal{D}_h(\pi_s^\star, \pi_s^0)]$, where the expectation is taken over the discounted state visitation distribution associated to the optimal policy. We then have our main results below.

Theorem 4.3 (Convergence rates)**.** *Let (A1), (A2), (A3) and (A4) hold. If the step size schedule is non-decreasing, i.e. $\eta_t \leq \eta_{t+1}$ for all $t \geq 0$, we have that the iterates of Algorithm 1 satisfy: for every $T \geq 0$,*

$$V^\star(\mu) - \frac{1}{T}\sum_{t<T} V^t(\mu) \leq \frac{1}{T}\left(\frac{\mathcal{D}_0^\star}{(1-\gamma)\eta_0} + \frac{C_3}{1-\gamma}\right) + (1+C_3)\tau.$$

Furthermore, if the step size schedule is geometrically increasing, i.e. satisfies

$$\eta_{t+1} \geq \frac{C_3}{C_3 - 1}\eta_t \qquad \forall t \geq 0, \tag{14}$$

we have: for every $T \geq 0$,

$$V^\star(\mu) - V^T(\mu) \leq \frac{1}{1-\gamma}\left(1 - \frac{1}{C_3}\right)^T \left(1 + \frac{\mathcal{D}_0^\star}{\eta_0(C_3-1)}\right) + (1+C_3)\tau.$$

Theorem 4.3 is, to the best of our knowledge, the first result that establishes both linear and sublinear convergence for a gradient based method in a setting with general policy parametrization. In particular, Theorem 4.3 attests how AMPO allows to use parametrization classes such as neural networks while still retaining theoretical guarantees for any mirror map. We refer to Appendix D.7 for a stronger version of Theorem 4.3 where the statements hold in expected value.

In terms of theoretical guarantees, Theorem 4.3 improves upon previous works on policy gradient methods for non-tabular policy parametrizations (Liu et al., 2019; Shani et al., 2020; Liu et al., 2020; Wang et al., 2020; Agarwal et al., 2021; Vaswani et al., 2022; Cayci et al., 2022) by having faster convergence rates. As special cases, it recovers the best known linear convergence results for PMD in the tabular setting (Xiao, 2022) and for NPG in the setting of linear function approximation (Alfano and Rebeschini, 2022; Yuan et al., 2023).

In terms of tools necessary for the analysis, we highlight that the guarantees in Theorem 4.3 hold without need to implement regularization (Cen et al., 2021; Zhan et al., 2021; Cayci et al., 2021; 2022; Lan, 2022), to impose bounded updates or smoothness of the policy (Agarwal et al., 2021; Liu et al., 2020; Vaswani et al., 2022), nor to restrict the analysis for the case where h is the negative entropy (Liu et al., 2019; Hu et al., 2022).

When the step size is infinitely large, AMPO can be interpreted as an approximation of the generalized policy iteration algorithm (see Appendix A.1), which also enjoys a linear convergence rate. This observation is already detailed in Xiao (2022). We also recover the linear converge speed of previous results on Q-learning with neural networks (Fan et al., 2020).

When $\mathcal{S}$ is a finite state space, a sufficient condition for C_3 in (A4) to be bounded is requiring μ to have full support on $\mathcal{S}$. When μ does not have full support, one can still obtain a linear convergence rate for $V^\star(\mu') - V^T(\mu')$, for an arbitrary state distribution μ' with full support, and relate this quantity to $V^\star(\mu) - V^T(\mu)$. We refer to Appendix D.8 for a more detailed discussion on the distribution mismatch coefficient.

Next, we give a result regarding the distance between policy updates, showing that it is always bounded.

Corollary 4.4 (Total expected Bregman divergence)**.** *Let (A1), (A2) and (A3) hold. If the step size is non-decreasing, i.e. $\eta_t \leq \eta_{t+1}$ for all $t \geq 0$, we have that the iterates of Algorithm 1 satisfy, for all $T \geq 0$,*

$$\sum_{t<T} \frac{\mathbb{E}_{s\sim d_\mu^\star}[\mathcal{D}_h(\pi_s^{t+1}, \pi_s^t)]}{\eta_t} \leq \left(\frac{\mathcal{D}_0^\star}{\eta_0} + C_3 - 1\right) + T(1-\gamma)(1+C_3)\tau.$$

In particular, if $\eta_t \equiv \eta$, we have, for all $T \geq 0$,

$$\sum_{t<T} \mathbb{E}_{s\sim d_\mu^\star}[\mathcal{D}_h(\pi_s^{t+1}, \pi_s^t)] \leq (\mathcal{D}_0^\star + \eta(C_3 - 1)) + \eta T(1-\gamma)(1+C_3)\tau.$$

Corollary 4.4 illustrates how the total expected Bregman divergence between subsequent iterates along the path of AMPO is upper-bounded by a constant term and a term that is proportional to η and T. When $\eta = c/T$, for some constant $c > 0$, the total expected Bregman divergence of the path of the algorithm is always upper bounded by the expected Bregman divergence between the optimal and the initial policy $\mathcal{D}_0^\star$ and an additive constant error term proportional to c.

We emphasize how the choice of the mirror map h in the definition of Algorithm 1 influences these results through a Bregman divergence term. First, the expected Bregman divergence $\mathcal{D}_0^\star$ between the optimal and the starting policy appears explicitly in both statements of Theorem 4.3 and Corollary 4.4, which motivates to find a starting policy π_0 and a mirror map h such that $\mathcal{D}_0^\star$ is small. However, as already highlighted in Xiao (2022), the convergence rate in Theorem 4.3 is not influenced by the choice of the mirror map. On the other hand, Corollary 4.4 provides a step further in understanding the importance of the mirror map. In particular, Corollary 4.4 encourages to choose a mirror map such that the associated Bregman divergence represents some form of update cost, as the total expected update cost of the algorithm would have an explicit bound at each iteration.

In the next section, we show that $\varepsilon_{\text{actor}}$ can be made arbitrarily small when f^θ is parametrized by a shallow neural network.

4.2. Neural network parametrization

Neural networks are popular choices for the parametrization class $\mathcal{F}^\Theta$ due to their empirical successes in RL applications (Mnih et al., 2013; 2015; Silver et al., 2017). Yet, few theoretical guarantees exist for this parametrization class. We therefore examine here how we can use our framework to fill this gap and build novel theoretical results. We will consider the case where, for each action $a \in \mathcal{A}$, $f^\theta(\cdot, a)$ belongs to the family of shallow ReLU networks, which has been shown to be a universal approximator (Jacot et al., 2018; Allen-Zhu et al., 2019; Du et al., 2019b; Ji et al., 2019). That is, for $s \in \mathcal{S} \subseteq \mathbb{R}^n$ and $a \in \mathcal{A}$, we define

$$f^\theta(\cdot, a) : s \mapsto \sum_{j=1}^m x_j^a \sigma(\langle w_j^a, s\rangle + b_j^a),$$

where $\sigma(y) = \max(y, 0)$ for all $y \in \mathbb{R}$ is the ReLU activation function, m represents the width of the network, $w_j^a \in \mathbb{R}^n$ and $b_j^a, x_j^a \in \mathbb{R}$ for all $j = 1, \dots, m$.

In the context of Algorithm 1, we want to show that $\varepsilon_{\text{actor}}$ can be made arbitrarily small by choosing a wide enough shallow ReLU network for the parametrization of $f^\theta(\cdot, a)$, for all $a \in \mathcal{A}$. In other words, at all times t, we want to find an approximation f^{t+1} of $g^t := \widehat{Q}^t + \eta_t^{-1}\nabla h(\pi^t)$. We achieve this by extending the framework developed by Ji et al. (2019) to our setting, but we highlight that other function approximation results, such as Ji and Telgarsky (2020), can be applied to this problem.

In order to do so, we first introduce the modulus of continuity ω_{g^t} for function g^t (Ji et al., 2019, Definition 4.1), which is defined for $\delta > 0$ as

$$\omega_{g^t}(\delta) := \sup_{s,s'\in\mathcal{S};\ a\in\mathcal{A}} \{g^t(s,a) - g^t(s',a) : \max(\|s\|_2, \|s'\|_2) \leq 1+\delta,\ \ \|s - s'\|_2 \leq \delta\}.$$

Denote $\|p\|_{L_1} = \int |p(w)|dw$. We additionally define a sample from a signed density $p^a : \mathbb{R}^{n+1} \to \mathbb{R}$ with $\|p\|_{L_1} < \infty$ as (w^a, b^a, x^a), where (w^a, b^a) is sampled from the probability density $|p|/\|p\|_{L_1}$ and $x^a = \text{sign}(p^a(w^a, b^a))$ (Ji et al., 2019, Definition 4.4). We can now state the following result adapted from Ji et al. (2019, Theorem E.1), which gives an explicit bound on $\varepsilon_{\text{actor}}$. Without loss of generality, assume that $\|s\|_2 \leq 1$ for all $s \in \mathcal{S}$.

Theorem 4.5. *There exist a set of signed densities $(p^a)_{a\in\mathcal{A}}$ and a set of parameters $((w_j^a, b_j^a, x_j^a))_{a\in\mathcal{A}}$ for*

$j \in \{m+1, m+2, m+3\}$ *such that, if*

$$f^{t+1}(s,a) = \frac{1}{m}\sum_{j=1}^{m+3} x_j^a \sigma(\langle w_j^a, s\rangle + b_j^a),$$

where $((w_j^a, b_j^a, x_j^a))_{j=1}^m$ *are sampled from* p^a*, for all* $a \in \mathcal{A}$*, with probability at least* $1-3\lambda$*, we have: for all times* $0 \le t < T$*,*

$$\sqrt{\varepsilon_{\text{actor}}} \le 3 \max_{t<T} \omega_{g^t}(\delta) + \mathcal{O}\left(\sqrt{\frac{\log(T|\mathcal{A}|/\lambda)}{m}}\right).$$

We refer to Appendix E for a proof and an explicit expression for the bound. Theorem 4.5 shows that $\varepsilon_{\text{actor}}$ can be made arbitrarily small when f^θ is parametrized by shallow ReLU neural networks. Additionally, while Ji et al. (2019) provide a method based on Fourier transforms and sampling to obtain a set of parameters such that the bound in Theorem 4.5 is satisfied, Theorem 4.5 does not exclude the possibility of using existing optimizers (Kingma and Ba, 2015) to solve $\min_\theta \left\| f^\theta - g^t \right\|^2_{L_2(v_t)}$. Combined with Theorem 4.3, Theorem 4.5 shows that our framework allows the exploitation of existing theoretical results on function approximation, thanks to the formulation of AMPO.

5. Conclusion

We have introduced a novel framework for RL which, given a mirror map and any parametrization class, induces a policy class and an update rule. We have proven that this framework enjoys sublinear and linear convergence for non-decreasing and geometrically increasing step size, respectively. Furthermore, for a proper choice of step size, we have shown that the total expected Bregman divergence between subsequent iterates of the algorithm is always upper bounded. Future venues of investigation include studying the sample complexity of AMPO in on-policy and off-policy settings , exploiting the properties of specific mirror maps to take advantage of the structure of the MDP and efficiently including representation learning in the algorithm. Additionally, finding parametrization classes and mirror maps that allow a small actor error and a simple Bregman projection step is of great interest. We refer to Appendix A.3 for a thorough discussion of future work. We believe that the main contribution of AMPO is to provide a general framework with theoretical guarantees that can help the analysis of specific algorithms and MDP structures. Furthermore, while AMPO recovers and improves several convergence rate guarantees in the literature, it is important to keep in consideration how previous works have exploited particular settings, as AMPO only tackles the most general case. It will be promising to see whether these previous works combined with our fast linear convergence result can derive new efficient sample complexity results.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021. (Cited on pages 4, 5, 6, 7, 16, 17, and 19.)

Carlo Alfano and Patrick Rebeschini. Linear convergence for natural policy gradient with log-linear policy parametrization, 2022. (Cited on pages 1, 7, 17, 18, 20, and 31.)

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 09–15 Jun 2019. (Cited on page 8.)

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 02 1998. ISSN 0899-7667. (Cited on page 16.)

Jonathan Baxter and Peter L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, Nov 2001. ISSN 1076-9757. doi: 10.1613/jair.806. (Cited on page 1.)

Amir Beck. *First-Order Methods in Optimization*. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2017. ISBN 1611974984. (Cited on pages 3 and 18.)

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3): 167–175, 2003. (Cited on pages 16 and 21.)

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019. (Cited on page 1.)

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods, 2019. (Cited on page 17.)

Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite MDPs.

In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2386–2394. PMLR, 13–15 Apr 2021. (Cited on pages 17 and 20.)

Shalabh Bhatnagar, Richard S. Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, pages 2471–2482, 2009. (Cited on pages 1 and 16.)

Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967. ISSN 0041-5553. (Cited on page 3.)

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 2015. (Cited on pages 3, 21, 24, and 25.)

Semih Cayci, Niao He, and R Srikant. Linear convergence of entropy-regularized natural policy gradient with linear function approximation. *arXiv preprint arXiv:2106.04096*, 2021. (Cited on pages 1, 6, 7, 17, and 20.)

Semih Cayci, Niao He, and R Srikant. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *arXiv preprint arXiv:2206.00833*, 2022. (Cited on pages 7, 16, 17, 19, and 31.)

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021. (Cited on pages 1, 7, 17, 20, and 31.)

Yair Censor and Stavros A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, USA, 1997. (Cited on page 3.)

Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993. (Cited on pages 2 and 24.)

Zaiwei Chen and Siva Theja Maguluri. Sample complexity of policy-based methods under off-policy sampling and linear function approximation. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 11195–11214. PMLR, 28–30 Mar 2022. (Cited on pages 1, 17, 20, and 31.)

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. (Cited on pages 16 and 18.)

Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global optimum convergence of momentum-based policy gradient. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 1910–1934. PMLR, 28–30 Mar 2022. (Cited on page 17.)

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1665–1674. PMLR, 09–15 Jun 2019a. (Cited on page 18.)

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in RL. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2826–2836. PMLR, 18–24 Jul 2021. (Cited on page 18.)

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b. (Cited on page 8.)

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*, pages 1407–1416. PMLR, 2018. (Cited on page 6.)

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR, 2020. (Cited on page 7.)

Ilyas Fatkhullin, Jalal Etesami, Niao He, and Negar Kiyavash. Sharp analysis of stochastic optimization under global Kurdyka-łojasiewicz inequality. In *Advances in Neural Information Processing Systems*, 2022. (Cited on page 17.)

Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies, 2023. (Cited on page 17.)

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1467–1476. PMLR, 10–15 Jul 2018. (Cited on page 17.)

Udaya Ghai, Elad Hazan, and Yoram Singer. Exponentiated gradient meets gradient descent. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 386–407. PMLR, 08 Feb–11 Feb 2020. (Cited on page 5.)

Yuzheng Hu, Ziwei Ji, and Matus Telgarsky. Actor-critic is implicitly biased towards high entropy optimal policies. In *International Conference on Learning Representations*, 2022. (Cited on pages 6, 7, 16, 17, and 19.)

Feihu Huang, Shangqian Gao, and Heng Huang. Bregman gradient policy optimization. In *International Conference on Learning Representations*, 2022. (Cited on pages 16 and 19.)

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. (Cited on page 8.)

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2020. (Cited on page 8.)

Ziwei Ji, Matus Telgarsky, and Ruicheng Xian. Neural tangent kernels, transportation mappings, and universal approximation. In *International Conference on Learning Representations*, 2019. (Cited on pages 4, 8, 9, 31, and 32.)

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020. (Cited on pages 16 and 18.)

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of 19th International Conference on Machine Learning*, pages 267–274, 2002. (Cited on page 27.)

Sham M. Kakade. A natural policy gradient. *Advances in Neural Information Processing Systems*, 2002. (Cited on pages 1, 4, and 16.)

William Karush. Minima of functions of several variables with inequalities as side conditions. Master's thesis, Department of Mathematics, University of Chicago, Chicago, IL, USA, 1939. (Cited on page 21.)

Michael J. Kearns and Daphne Koller. Efficient reinforcement learning in factored mdps. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, IJCAI '99, page 740–747, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606130. (Cited on page 18.)

Sajad Khodadadian, Zaiwei Chen, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor-critic algorithm. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5420–5431. PMLR, 18–24 Jul 2021a. (Cited on pages 16 and 19.)

Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, page 3794–3799. IEEE Press, 2021b. (Cited on pages 17 and 20.)

Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On linear and super-linear convergence of natural policy gradient algorithm. *Systems & Control Letters*, 164:105214, 2022. ISSN 0167-6911. (Cited on pages 17 and 20.)

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. (Cited on page 9.)

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, volume 12, pages 1008–1014. MIT Press, 2000. (Cited on pages 1, 2, and 4.)

Walid Krichene, Syrine Krichene, and Alexandre Bayen. Efficient bregman projections onto the simplex. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 3291–3298, 2015. doi: 10.1109/CDC.2015.7402714. (Cited on pages 4, 5, and 23.)

Jakub Grudzien Kuba, Christian A Schroeder De Witt, and Jakob Foerster. Mirror learning: A unifying framework of policy optimisation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7825–7844. PMLR, 17–23 Jul 2022. (Cited on pages 1, 16, and 17.)

Harold W. Kuhn and Albert W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, Berkeley and Los Angeles, 1951. (Cited on page 21.)

Solomon Kullback and Richard A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951. doi: 10.1214/aoms/1177729694. (Cited on page 1.)

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pages 1–48, 2022. ISSN 1436-4646. (Cited on pages 1, 7, 16, 17, 20, and 31.)

Yan Li, Tuo Zhao, and Guanghui Lan. Homotopic policy mirror descent: Policy convergence, implicit regularization, and improved sample complexity. *arXiv preprint arXiv:2201.09457*, 2022. (Cited on pages 17 and 20.)

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtarik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6286–6295. PMLR, 18–24 Jul 2021. (Cited on pages 17 and 18.)

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in neural information processing systems*, 32, 2019. (Cited on pages 6, 7, 16, and 19.)

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 2020. (Cited on pages 7, 16, 17, 18, and 19.)

Stanisław Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. Equ. Derivees partielles, Paris 1962, Colloques internat. Centre nat. Rech. sci. 117, 87-89 (1963)., 1963. (Cited on page 17.)

Saeed Masiha, Saber Salehkaleybar, Niao He, Negar Kiyavash, and Patrick Thiran. Stochastic second-order methods improve best-known sample complexity of SGD for gradient-dominated functions. In *Advances in Neural Information Processing Systems*, 2022. (Cited on page 17.)

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020. ISBN 2640-3498. (Cited on page 17.)

Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7555–7564. PMLR, 18–24 Jul 2021. (Cited on page 17.)

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with deep reinforcement learning, 2013. (Cited on page 8.)

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb 2015. ISSN 1476-4687. (Cited on pages 6 and 8.)

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 *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR. (Cited on page 1.)

Arkadi Nemirovski and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley Interscience, 1983. (Cited on pages 1, 3, and 16.)

Yurii E. Nesterov and Boris T. Polyak. Cubic regularization of Newton method and its global performance. *Math. Program.*, 108(1):177–205, 2006. (Cited on page 17.)

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes, 2017. (Cited on page 16.)

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2613–2621, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. (Cited on pages 16, 17, and 18.)

Francesco Orabona. A modern introduction to online learning, 2020. URL https://open.bu.edu/handle/2144/40900. (Cited on page 5.)

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, pages 1180–1190, 2008. (Cited on pages 1 and 16.)

Boris T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. ISSN 0041-5553. (Cited on page 17.)

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1889–1897, Lille, France, 07–09 Jul 2015a. PMLR. (Cited on pages 1 and 15.)

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b. (Cited on pages 4 and 6.)

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. (Cited on page 15.)

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016. (Cited on page 1.)

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 5668–5675, 2020. (Cited on pages 1, 5, 7, 16, 17, and 19.)

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550 (7676):354–359, Oct 2017. ISSN 1476-4687. doi: 10.1038/nature24270. (Cited on page 8.)

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2898–2933. PMLR, 25–28 Jun 2019. (Cited on page 18.)

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* 2018. ISBN 0262352702. (Cited on page 15.)

Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998. (Cited on page 6.)

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, 1999. (Cited on pages 1, 3, and 4.)

Gerald Tesauro. Temporal difference learning and TD-Gammon. *Journal of the International Computer Games Association*, 18(2):88, 1995. (Cited on page 4.)

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016. (Cited on page 6.)

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. In *International Conference on Learning Representations*, 2022. (Cited on pages 1, 4, 6, and 15.)

Sharan Vaswani, Olivier Bachem, Simone Totaro, Robert Müller, Shivam Garg, Matthieu Geist, Marlos C Machado, Pablo Samuel Castro, and Nicolas Le Roux. A general class of surrogate functions for stable and efficient reinforcement learning. In *AISTATS*, pages 8619–8649, 2022. (Cited on pages 1, 4, 6, 7, 16, and 17.)

Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. The statistical complexity of early-stopped mirror descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 253–264. Curran Associates, Inc., 2020. (Cited on page 5.)

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2020. (Cited on pages 7, 16, and 19.)

Ronald J. Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 1991. (Cited on page 1.)

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23 (282):1–36, 2022. (Cited on pages 1, 2, 4, 5, 6, 7, 8, 16, 17, 18, 19, 20, 22, 24, and 31.)

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. In *Advances in Neural Information Processing Systems*, volume 33, pages 4358–4369. Curran Associates, Inc., 2020. (Cited on pages 16, 17, and 19.)

Long Yang, Yu Zhang, Gang Zheng, Qian Zheng, Pengfei Li, Jianhang Huang, and Gang Pan. Policy optimization with stochastic mirror descent. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8823–8831, Jun. 2022. doi: 10.1609/aaai.v36i8.20863. (Cited on pages 16 and 19.)

Rui Yuan, Robert M. Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3332–3380. PMLR, 28–30 Mar 2022. (Cited on pages 16, 17, and 18.)

Rui Yuan, Simon Shaolei Du, Robert M. Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. In *International Conference on Learning Representations*, 2023. (Cited on pages 1, 5, 6, 7, 16, 17, 18, 19, 20, 27, and 31.)

Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4473–4525. PMLR, 15–19 Aug 2021. ISBN 2640-3498. (Cited on pages 16, 17, and 19.)

Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D. Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence, 2021. (Cited on pages 1, 7, 17, 20, and 31.)

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*, volume 33, pages 4572–4583. Curran Associates, Inc., 2020. (Cited on pages 7 and 17.)

Junyu Zhang, Chengzhuo Ni, Zheng Yu, Csaba Szepesvari, and Mengdi Wang. On the convergence and sample efficiency of variance-reduced policy gradient method. In *Advances in Neural Information Processing Systems*, 2021. (Cited on page 17.)

Appendix

Here we provide the related work discussion, the missing proofs from the main paper and some additional noteworthy observations made in the main paper.

A. Related work

We provide an extended discussion for the context of our work, including a comparison of different PMD frameworks and a comparison of the convergence theories of PMD in the literature. Furthermore, we discuss future work, such as extending our analysis to the dual averaging updates and developing sample complexity analysis of AMPO.

A.1. Comparisons with other policy optimization frameworks

In this section, we give a comparison with some of the most popular policy optimization algorithms in the literature.

Generalised Policy Iteration (Sutton and Barto, 2018). The update consists in evaluating the Q-function of the policy and obtaining the new policy by acting greedily with respect to the estimated Q-function. That is, for all $s \in \mathcal{S}$,

$$\pi_s^{t+1} \in \underset{\pi_s \in \Delta(\mathcal{A})}{\operatorname{argmax}} \langle Q_s^t, \pi_s \rangle. \tag{15}$$

AMPO recovers this algorithm when we have access to the value of Q_s^t and $\eta_t \to +\infty$ for all times t.

Trust Region Policy Optimization (Schulman et al., 2015a). The TRPO update is as follows:

$$\pi^{t+1} \in \underset{\pi \in \Pi}{\operatorname{argmax}} \mathbb{E}_{s \sim d_\mu^t} \left[\langle A_s^t, \pi_s \rangle \right], \tag{16}$$
$$\text{such that} \quad \mathbb{E}_{s \sim d_\mu^t} \left[D_h(\pi_s^t, \pi_s) \right] \leq \delta,$$

where $A_s^t = Q_s^t - V^t$ represents the advantage function, h is the negative entropy and $\delta > 0$. TRPO is equivalent to AMPO when at each time t, the admissible policy class is $\Pi^t = \{\pi \in \Delta(\mathcal{A})^{\mathcal{S}} : \mathbb{E}_{s \sim d_\mu^t} D_h(\pi_s^t, \pi_s) \leq \delta\}$, we have access to the value of Q_s^t and $\eta_t \to +\infty$.

Proximal Policy Optimization (Schulman et al., 2017). The KL variation of the Proximal Policy Optimization (PPO) update consists in maximizing a surrogate function depending on the policy gradient with respect to the new policy. Namely,

$$\pi^{t+1} \in \underset{\pi \in \Pi}{\operatorname{argmax}} \mathbb{E}_{s \sim d_\mu^t} \left[L(\pi_s, \pi_s^t) \right], \tag{17}$$

with

$$L(\pi_s, \pi_s^t) = \mathbb{E}_{a \sim \pi^t} \min \left(r^\pi(s,a) A^t(s,a), \operatorname{clip}(r^\pi(s,a), 1 \pm \epsilon) A^t(s,a) \right),$$

where $r^\pi(s,a) = \pi(s,a)/\pi^t(s,a)$, or

$$L(\pi_s, \pi_s^t) = \eta_t \langle A_s^t, \pi_s \rangle - D_h(\pi_s^t, \pi_s),$$

where h is the negative entropy. In an exact setting and when $\Pi = \Delta(\mathcal{A})^{\mathcal{S}}$, PPO differs from PMD because it inverts the terms in the Bregman divergence penalty.

Mirror Descent Policy Optimization (Tomar et al., 2022). The algorithm consists in the following update:

$$\pi^{t+1} \in \underset{\pi \in \Pi}{\operatorname{argmin}} \mathbb{E}_{s \sim d_\mu^t} [\langle A_s^t, \pi_s \rangle + D_h(\pi_s, \pi_s^t)], \tag{18}$$

where Π is a parametrized policy class. While it is equivalent to AMPO in an exact setting and when $\Pi = \Delta(\mathcal{A})^{\mathcal{S}}$, as we show in Appendix B, the difference between the two algorithms lies on the approximation of the exact algorithm.

Functional Mirror Ascent Policy Gradient (Vaswani et al., 2022). The algorithm consists in the following update:

$$\begin{aligned} \pi^{t+1} &\in \underset{\theta\in\Theta}{\operatorname{argmin}}\, \mathbb{E}_{s\sim d_\mu^t}[V^{\pi^t}(\mu) + \langle \nabla_{\pi_s} V^{\pi^t}(\mu)\big|_{\pi=\pi^t}, \pi_s^\theta - \pi_s^t\rangle + D_h(\pi_s^\theta, \pi_s^t)], \\ &\in \underset{\theta\in\Theta}{\operatorname{argmin}}\, \mathbb{E}_{s\sim d_\mu^t}[\langle Q_s^t, \pi_s\rangle + D_h(\pi_s^\theta, \pi_s^t)], \end{aligned} \tag{19}$$

where π^θ is a policy parametrized by θ. The second line is obtained by the definition of V^t and the policy gradient theorem (2). The discussion is the same as the previous algorithm.

Mirror Learning (Kuba et al., 2022). The on-policy version of the algorithm consists in the following update:

$$\pi^{t+1} = \underset{\pi\in\Pi(\pi^t)}{\operatorname{argmin}}\, \mathbb{E}_{s\sim d_\mu^t}[\langle Q_s^t, \pi_s\rangle + D(\pi_s, \pi_s^t)], \tag{20}$$

where $\Pi(\pi^t)$ is a policy class that depends on the current policy π^t and the drift functional D is defined as a map $D : \Delta(\mathcal{A}) \times \Delta(\mathcal{A}) \to \mathbb{R}$ such that $D(\pi_s, \bar{\pi}_s) \geq 0$ and $\nabla_{\pi_s} D(\pi_s, \bar{\pi}_s)\big|_{\pi_s=\bar{\pi}_s} = 0$. The drift functional D recovers the Bregman divergence as a particular case, in which case Mirror Learning is equivalent to AMPO in an exact setting and when $\Pi = \Delta(\mathcal{A})^{\mathcal{S}}$. Once more, the main difference between the two algorithms lies on the approximation of the exact algorithm.

A.2. Theoretical analyses of policy optimization methods

Lately, the impressive empirical success of policy gradient (PG) - type methods has catalyzed the development of theoretically sound gradient-based algorithms for policy optimization. In particular, there has been a lot of attention around algorithms inspired by mirror descent (MD) (Nemirovski and Yudin, 1983; Beck and Teboulle, 2003) and, more specifically, by natural gradient descent (Amari, 1998). These two approaches led to policy mirror descent (PMD) methods (Shani et al., 2020; Lan, 2022) and natural policy gradient (NPG) methods (Kakade, 2002), which as first shown by Neu et al. (2017) is a particular case of PMD. Leveraging different techniques from the MD literature, it has been established that PMD, NPG and their variants converge to the global optimum in different settings. We refer to global optimum convergence as an analysis that guarantees that $V^\star(\mu) - \mathbb{E}\left[V^T(\mu)\right] \leq \epsilon$ after T iterations with $\epsilon > 0$. As an important variant of NPG, we will also discuss the literature of the convergence analysis of natural actor-critic (NAC) (Peters and Schaal, 2008; Bhatnagar et al., 2009).

Sublinear convergence analyses of PMD, NPG and NAC. For softmax tabular policies, Shani et al. (2020) establish a $\mathcal{O}(1/\sqrt{T})$ convergence rate for unregularized NPG and $\mathcal{O}(1/T)$ for regularized NPG. Agarwal et al. (2021); Khodadadian et al. (2021a) and Xiao (2022) improve the convergence rate for unregularized NPG and NAC to $\mathcal{O}(1/T)$ and Xiao (2022) extends the same convergence rate to projected Q-descent.

In the function approximation regime, Zanette et al. (2021) and Hu et al. (2022) achieve $\mathcal{O}(1/\sqrt{T})$ convergence rate by developing variants of PMD methods for the linear MDP (Jin et al., 2020) setting. The same $\mathcal{O}(1/\sqrt{T})$ convergence rate is obtained by Agarwal et al. (2021) for both log-linear and smooth policies, while Yuan et al. (2023) improve the convergence rate to $\mathcal{O}(1/T)$ for log-linear policies. For smooth policies, the convergence rate is later improved to $\mathcal{O}(1/T)$ either by adding an extra Fisher-non-degeneracy condition on the policies (Liu et al., 2020) or by analyzing NAC under Markovian sampling (Xu et al., 2020). Yang et al. (2022) and Huang et al. (2022) consider Lipschitz and smooth policies (Yuan et al., 2022), obtain $\mathcal{O}(1/\sqrt{T})$ convergence rates for PMD-type methods and faster $\mathcal{O}(1/T)$ convergence rates by applying the variance reduction techniques SARAH (Nguyen et al., 2017) and STORM (Cutkosky and Orabona, 2019), respectively. As for neural policy parametrization, Liu et al. (2019) establish a $\mathcal{O}(1/\sqrt{T})$ convergence rate for two-layer neural PPO. The same $\mathcal{O}(1/\sqrt{T})$ convergence rate is established by Wang et al. (2020) for two-layer neural NAC, which is later improved to $\mathcal{O}(1/T)$ by Cayci et al. (2022), using entropy regularization.

We highlight that all the sublinear convergence analyses mentioned above, for both softmax tabular policies and the function approximation regime, are obtained by either using a decaying step size or a constant step size. We refer to Table 1 for an overview of recent sublinear convergence analyses of NPG/PMD.

Linear convergence analysis of PMD, NPG, NAC and other PG methods. In the softmax tabular policy settings, the linear convergence guarantees of NPG and PMD are achieved by either adding regularization (Cen et al., 2021; Zhan et al., 2021; Lan, 2022; Li et al., 2022) or by varying the step sizes (Bhandari and Russo, 2021; Khodadadian et al., 2021b; 2022; Xiao, 2022).

In the function approximation regime, the linear convergence guarantees are achieved for NPG with log-linear policies, either by adding entropy regularization (Cayci et al., 2021) or by choosing geometrically increasing step sizes (Alfano and Rebeschini, 2022; Yuan et al., 2023). It can also be achieved for NAC with log-linear policy by using adaptive increasing step sizes (Chen and Theja Maguluri, 2022).

We refer to Table 2 for an overview of recent linear convergence analyses of NPG/PMD.

Alternatively, by leveraging a Polyak-Lojasiewicz (PL) condition (Polyak, 1963; Łojasiewicz, 1963), fast linear convergence results can be achieved for PG methods under different settings, such as linear quadratic control problems (Fazel et al., 2018) and softmax tabular policies with entropy regularization (Mei et al., 2020; Yuan et al., 2022). The PL condition is widely explored by Bhandari and Russo (2019) to identify more general MDP settings. Similar to the cases of NPG and PMD, by choosing adaptive step sizes through exact line search (Bhandari and Russo, 2021) or by exploiting non-uniform smoothness (Mei et al., 2021), linear convergence of PG can also be obtained for the softmax tabular policy without regularization. When the PL condition is relaxed to other weaker conditions, PG methods combined with variance reduction methods such as SARAH (Nguyen et al., 2017) and PAGE (Li et al., 2021) can also achieve linear convergence. This is shown by Fatkhullin et al. (2022; 2023) when the PL condition is replaced by the weak PL condition (Yuan et al., 2022), which is satisfied by Fisher-non-degenerate policies (Ding et al., 2022). It is also shown by Zhang et al. (2021), where the MDP satisfies some hidden convexity property which contains a similar property to the weak PL condition studied by Zhang et al. (2020). Lastly, linear convergence is established for the cubic-regularized Newton method (Nesterov and Polyak, 2006), a second-order method, applied on Fisher-non-degenerate policies combined with variance reduction (Masiha et al., 2022).

Outside the literature focusing on finite time convergence guarantees, Vaswani et al. (2022) and Kuba et al. (2022) provide a theoretical analysis for variations of PMD, showing monotonic improvements for their frameworks. Additionally, Kuba et al. (2022) give an infinite time convergence guarantee for their framework.

Our Contributions. Our work provides a PMD framework – AMPO. For AMPO, we establish in Theorem 4.3 both $\mathcal{O}(1/T)$ convergence guarantee by using a non-decreasing step size and linear convergence guarantee by using a geometrically increasing step size. To the best of our knowledge, Theorem 4.3 is the first result that establishes $\mathcal{O}(1/T)$ convergence by using a non-decreasing step size and is also the first result that establishes linear convergence by using a geometrically increasing step size, for a gradient based method in a setting with a general function parametrization class. The improvements of Theorem 4.3 compared to all the above sublinear and linear convergence literature can be summarized as follows.

- Firstly, Theorem 4.3 matches the current state-of-the-art sublinear convergence rate of PMD methods in all the above settings with a constant step size and is relaxed to allow the use of a non-decreasing step size. Theorem 4.3 also matches the current state-of-the-art linear convergence rate of PMD in all the above settings, that is for softmax tabular and log-linear policies.

- More importantly, our result accommodates any parametrized function class $\mathcal{F}^\Theta$. That is, it does not impose any constraints on $\mathcal{F}^\Theta$ in contrast to the analysis with smooth or Fisher-non-degenerate policies (Agarwal et al., 2021; Liu et al., 2020; Xu et al., 2020); it requires neither specific structures for the MDP as for Zanette et al. (2021); Hu et al. (2022) nor regularization Shani et al. (2020); Cayci et al. (2022); Cen et al. (2021); Zhan et al. (2021); Lan (2022); Li et al. (2022); Cayci et al. (2021). Consequently, AMPO is the first framework to achieve linear convergence for general parametrization, including widely-used neural policy parametrization as special cases. The generality of the assumptions of Theorem 4.3 allows the application of our theory to specific settings, where existing sample complexity analyses could be improved thanks to the linear convergence of AMPO. For instance, since Theorem 4.3 holds with any structural MDP, AMPO could be directly applied to the linear MDP setting to derive a sample complexity analysis of AMPO which may improve that of Zanette et al. (2021) and Hu et al. (2022). As we discuss in the next section, this is a promising direction for future work.

- Lastly, the generality of our framework and of Lemma 4.1 allow Theorem 4.3 to unify the previous results in the literature and generate new theoretically sound algorithms under one guise. Indeed, the linear convergence analysis of natural policy gradient (NPG) with softmax tabular policies (Xiao, 2022) or with log-linear policies (Alfano and Rebeschini, 2022; Yuan et al., 2023) are special cases of our general analysis. AMPO also generate new algorithms by selecting mirror maps, e.g. hyperentropy mirror map in Example 3.5 and ω-potential mirror map in Example 3.6.

A.3. Future work

The results we have obtained open up several experimental questions related to the parameterization class and the choice of mirror map in APMO. We leave such questions as an important direction to further support our theoretical findings.

On the other hand, our work also opens several interesting research directions in both algorithmic and theoretical aspects.

From an algorithmic point of view, the updates in Line 3 and 4 of AMPO are not explicit. This might be an issue in practice, especially for large scale RL problems.

It will be interesting to design efficient regression solver for minimizing the actor error in Line 3 of Algorithm 1. For instance, by using the dual averaging algorithm (Beck, 2017, Chapter 4), it could be possible to replace the term $\nabla h(\pi_s^t)$ with f_s^t for all $s \in \mathcal{S}$, to make the computation of the algorithm more efficient. That is, it could be interesting to consider the following variation of Line 3 in Algortihm 1:

$$\left\| f^{t+1} - \widehat{Q}^t - \frac{\eta_{t-1}}{\eta_t} f^t \right\|^2_{L_2(v_t)} \leq \varepsilon_{\text{actor}}. \tag{21}$$

Notice that (21) has the same update as (12), however (21) is not restricted to use the negative entropy mirror map. To efficiently solve the regression problem in Line 3 of Algorithm 1, one may want to apply the modern variance reduction techniques (Nguyen et al., 2017; Cutkosky and Orabona, 2019; Li et al., 2021). This has been done by Liu et al. (2020) for NPG method.

To make the updates in Line 4 of Algorithm 1 efficient, it will be interesting to exploit the properties of specific mirror maps to derive their explicit updates. In this work, we investigated in the negative entropy, the Tsallis entropy and the hyperbolic entropy mirror maps (Examples 3.4 and 3.5). They all contain the corresponding explicit updates for the Bregman projected policies. The ω-potential mirror map from Example 3.6 indicates one way of applying other possible mirror maps for AMPO. It is of great interest to discover other mirror maps for AMPO that guarantee explicit updates for the Bregman projected policies in Line 4 of Algorithm 1.

From a theoretical point of view, by leveraging the linear convergence of AMPO, it will be promising to derive its sample complexity analysis in specific settings. As mentioned for the linear MDP (Jin et al., 2020) in Appendix A.2, one can apply the linear convergence theory of AMPO to other structural MDPs, e.g. block MDP (Du et al., 2019a), factored MDP (Kearns and Koller, 1999; Sun et al., 2019), RKHS linear MDP and RKHS linear mixture MDP (Du et al., 2021), to build new sample complexity results for these settings, since the assumptions of Theorem 4.3 do not impose any constraint on the MDP. On the other hand, it will be interesting to explore the interaction between the Bregman projected policy class and the expected Lipschitz and smooth policies (Yuan et al., 2022) and the Fish-non-degenerate policies (Liu et al., 2020) to establish new improved sample complexity results in these settings, again thanks to the linear convergence theory of AMPO.

Finally, this work theoretically indicates that, perhaps the most important future work of PMD-type algorithms is to design efficient policy evaluation algorithms to make the critic error ϵ_{critic} as small as possible, such as using offline data for training, and to construct adaptive representation learning for $\mathcal{F}^\Theta$ to make the actor error ϵ_{actor} as small as possible. This matches one of the most important research questions for deep Q-learning type algorithms for general policy optimization problems.

Table 1. Overview of sublinear convergence results for NPG and PMD methods with constant stepsize in different settings. The dark blue cells contain our new results. The light blue cells contain previously known results that we recover as particular cases of our analysis. The pink cells contain previously best known results upon which we improve by providing a faster convergence rate. White cells contain existing results that are already improved by other literature or that we could not recover under our general analysis.

Algorithm	**Rate**	**Comparisons to our works**
Setting: Softmax tabular policies		
Adaptive TRPO (Shani et al., 2020)	$\mathcal{O}(1/\sqrt{T})$	Regularization is unnecessary
Tabular off-policy NAC (Khodadadian et al., 2021a)	$\mathcal{O}(1/T)$	Weaker approximation error asm. with L_2 instead of L_∞
Tabular NPG (Agarwal et al., 2021)	$\mathcal{O}(1/T)$	
Tabular NPG/projected Q-descent (Xiao, 2022)	$\mathcal{O}(1/T)$	Recover their results with $f^\theta(s,a) = \theta_{s,a}$; Generalize the projected Q-descent from tabular to a general parametrization class $\mathcal{F}^\Theta$; Weaker approximation error asm. with L_2 instead of L_∞
Setting: Log-linear policies		
Q-NPG with log-linear policies (Agarwal et al., 2021)	$\mathcal{O}(1/\sqrt{T})$	
Q-NPG/NPG with log-linear policies (Yuan et al., 2023)	$\mathcal{O}(1/T)$	Recover their results with linear $f^\theta(s,a)$
Setting: Two-layer neural policies		
Neural PPO (Liu et al., 2019)	$\mathcal{O}(1/\sqrt{T})$	
Neural NAC (Wang et al., 2020)	$\mathcal{O}(1/\sqrt{T})$	
Regularized neural NAC (Cayci et al., 2022)	$\mathcal{O}(1/T)$	Improve the convergence rate to $\mathcal{O}(1/T)$ or linear without regularization
Setting: Linear MDP		
Variants PMD with linear MDP (Zanette et al., 2021; Hu et al., 2022)	$\mathcal{O}(1/\sqrt{T})$	Improve the convergence rate to $\mathcal{O}(1/T)$ or linear
Setting: Smooth policies		
NPG with smooth policies (Agarwal et al., 2021)	$\mathcal{O}(1/\sqrt{T})$	
NAC under Markovian sampling with smooth policies (Xu et al., 2020)	$\mathcal{O}(1/T)$	
NPG with smooth and Fisher-non-degenerate policies (Liu et al., 2020)	$\mathcal{O}(1/T)$	
Setting: Lipschitz and Smooth policies		
Variance reduced PMD (Yang et al., 2022; Huang et al., 2022)	$\mathcal{O}(1/T)$	
Setting: Bregman projected policies with general parametrization and mirror map		
AMPO (Theorem 4.3, this work)	$\mathcal{O}(1/T)$	

Table 2. Overview of linear convergence results for NPG and PMD methods in different settings. The darker cells contain our new results. The light cells contain previously known results that we recover as special cases of our analysis, and extend the permitted concentrability coefficients settings. White cells contain existing results that we could not recover under our general analysis.

Algorithm	**Reg.**	**C.S.**	**A.I.S.**	**N.I.S.***	**Comparisons to our works**
Setting: Softmax tabular policies					
Regularized tabular NPG (Cen et al., 2021)	✓	✓			Weaker approximation error assumption with L_2 instead of L_∞ norm
Regularized tabular PMD (Zhan et al., 2021)	✓	✓			Weaker approximation error assumption with L_2 instead of L_∞ norm
Regularized tabular NPG (Lan, 2022)	✓			✓	Weaker approximation error assumption with L_2 instead of L_∞ norm
Regularized tabular NPG (Li et al., 2022)	✓			✓	Weaker approximation error assumption with L_2 instead of L_∞ norm; Li et al. (2022) derive the convergence of the policy to the fixed high entropy optimal policy, which is not studied in our work
Tabular NPG (Bhandari and Russo, 2021)			✓		Non adaptive stepsize instead of adaptive stepsize or Line search
Tabular NPG (Khodadadian et al., 2021b; 2022)			✓		Weaker approximation error assumption with L_2 instead of L_∞ norm
Tabular NPG / Projected Q-descent (Xiao, 2022)				✓	Recover their results with $f^\theta(s,a) = \theta_{s,a}$; Generalize the projected Q-descent from tabular to a general parametrization class $\mathcal{F}^\Theta$; Weaker approximation error assumption with L_2 instead of L_∞ norm
Setting: Log-linear policies					
Regularized NPG with log-linear (Cayci et al., 2021)	✓	✓			Better concentrability coefficients C_1, C_2
Off-policy NAC with log-linear (Chen and Theja Maguluri, 2022)			✓		Weaker approximation error assumption with L_2 instead of L_∞ norm
Q-NPG with log-linear (Alfano and Rebeschini, 2022)				✓	Recover their results with linear f^θ; Better concentrability coefficients C_1, C_2 ; Their relative condition number depends on t, while we donot have such assumption
Q-NPG/NPG with log-linear (Yuan et al., 2023)				✓	Recover their results with linear f^θ; Better concentrability coefficients C_1, C_2
Setting: Bregman projected policies with general parametrization and mirror map					
AMPO (Theorem 4.3, this work)				✓	

* **Reg.**: regularization; **C.S.**: constant stepsize; **A.I.S.**: Adaptive increasing stepsize; **N.I.S.**: Non-adaptive increasing stepsize.

B. Equivalence of (8)-(9) and (10) in the tabular case

To show that the policy mirror descent updates between two-step procedure (8)-(9) and one step update (10) in the tabular case are equivalent, it suffices to verify the following general lemma in optimization. The proof was shown for instance in Bubeck (2015, Chapter 4.2). We provide its proof here for completeness.

Lemma B.1 (Right after Theorem 4.2 in Bubeck (2015))**.** *Consider the mirror descent in (4)-(5) for the minimization of function $V(\cdot)$. That is,*

$$y^{t+1} = \nabla h(x^t) - \eta_t \nabla V(x)|_{x=x^t}, \tag{22}$$

$$x^{t+1} = Proj_{\mathcal{X}}^h(\nabla h^*(y^{t+1})). \tag{23}$$

Then one can rewrite the mirror descent as

$$x^{t+1} = \operatorname*{argmin}_{x\in\mathcal{X}} \eta_t x^\top \nabla V(x)|_{x=x^t} + \mathcal{D}_h(x, x^t). \tag{24}$$

Proof. From definition of the Bregman projection step, starting from (22) we have

$$\begin{aligned}
x^{t+1} &= \operatorname{Proj}_{\mathcal{X}}^h(\nabla h^*(y^{t+1})) \;=\; \operatorname*{argmin}_{x\in\mathcal{X}} \mathcal{D}_h(x, \nabla h^*(y^{t+1})) \\
&= \operatorname*{argmin}_{x\in\mathcal{X}} \nabla h(x) - \nabla h(\nabla h^*(y^{t+1})) - \left\langle \nabla h(\nabla h^*(y^{t+1})), x - \nabla h^*(y^{t+1}) \right\rangle \\
&\overset{(3)}{=} \operatorname*{argmin}_{x\in\mathcal{X}} \nabla h(x) - y^{t+1} - \left\langle y^{t+1}, x - \nabla h^*(y^{t+1}) \right\rangle \\
&= \operatorname*{argmin}_{x\in\mathcal{X}} \nabla h(x) - x^\top y^{t+1} \\
&\overset{(22)}{=} \operatorname*{argmin}_{x\in\mathcal{X}} \nabla h(x) - x^\top \left(\nabla h(x^t) - \eta_t \nabla V(x)|_{x=x^t} \right) \\
&= \operatorname*{argmin}_{x\in\mathcal{X}} \eta_t x^\top \nabla V(x)|_{x=x^t} + \nabla h(x) - \nabla h(x^t) - \left\langle \nabla h(x^t), x - x^t \right\rangle \\
&= \operatorname*{argmin}_{x\in\mathcal{X}} \eta_t x^\top \nabla V(x)|_{x=x^t} + \mathcal{D}_h(x, x^t),
\end{aligned}$$

where the second and the last lines are both obtained by the definition of the Bregman divergence. □

This (24) is often taken as the definition of mirror descent (Beck and Teboulle, 2003), which provides a proximal view point of mirror descent, i.e. a gradient step in the primal space with a regularization of Bregman divergence.

C. Deferred proofs from Section 3

In this section, we give the derivations for Examples 3.2, 3.3, 3.4, 3.5 and 3.6, which are based on the Karush-Kuhn-Tucker (KKT) conditions (Karush, 1939; Kuhn and Tucker, 1951).

C.1. Examples 3.2 and 3.4

We start by instancing Algorithm 1 for the negative entropy mirror map. In this setting, it is useful to define the dual mirror map

$$h^*(\eta f_s^\theta) = \log \sum_{a\in\mathcal{A}} \exp(\eta f^\theta(s,a)),$$

so that

$$\nabla h^*(\eta f_s^\theta) = \frac{\exp(\eta f_s^\theta)}{\|\exp(\eta f_s^\theta)\|_1},$$

where $\exp(\eta f_s^\theta)$ is taken element-wise. We can then define $h(\pi_s) = \sum_{a\in\mathcal{A}} \pi(a|s)\log(\pi(a|s))$ as the primal mirror map. With this formulation, the instantiation of Algorithm 1 becomes straightforward, as

$$\nabla h^*(\eta_{t-1} f_s^t) = \operatorname*{argmin}_{\pi_s\in\Delta(\mathcal{A})} \mathcal{D}_h(\pi_s, \nabla h^*(\eta_{t-1} f_s^t)),$$

since $\nabla h^*(\eta_{t-1} f_s^t)$ lies already on the simplex.

We now give the generalization to Tsallis entropy. Let $q \geq 1$ and

$$\exp_q(x) = \begin{cases} \exp(x) & \text{if } q = 1, \\ [1 + (q-1)x]_+^{\frac{1}{q-1}} & \text{if } q > 1, \end{cases}$$

where $x \in \mathbb{R}$ and $[z]_+ = \max(z, 0)$, and

$$\log_q(y) = \begin{cases} \log(y) & \text{if } q = 1, \\ \frac{y^{q-1}-1}{q-1} & \text{if } q > 1, \end{cases}$$

where $y \geq 0$. Then the Tsallis entropy mirror map is defined as

$$h_q(\pi_s) = \frac{1}{q}\left(\sum \pi(a|s) \log_q(\pi(a|s)) - \pi(a|s)\right).$$

To solve the minimization problem

$$\pi_s^\theta = \operatorname*{argmin}_{\pi_s \in \Delta(\mathcal{A})} \mathcal{D}_{h_q}(\pi_s, \nabla h_q^*(\eta_{t-1} f_s^t))$$

we use the KKT conditions. We formalize it as

$$\begin{aligned} &\operatorname*{argmin}_{\pi_s \in \mathbb{R}^{|\mathcal{A}|}} \; \operatorname*{argmin}_{\pi_s \in \Delta(\mathcal{A})} \mathcal{D}_{h_q}(\pi_s, \nabla h_q^*(\eta_{t-1} f_s^t)) \\ &\text{subject to } \langle \pi_s, \mathbf{1} \rangle = 1 \\ &\qquad\qquad\quad \pi(a|s) \geq 0 \qquad \forall\, a \in \mathcal{A}. \end{aligned}$$

The conditions then become

$$\begin{aligned} \text{(stationarity)} \qquad & \log_q(\pi_s) - \eta_{t-1} f_s^t + \lambda_s \mathbf{1} - \sum_{a \in \mathcal{A}} c_s^a e_a = 0, \\ \text{(complementary slackness)} \qquad & c_s^a \pi(a|s) = 0 \quad \forall\, a \in \mathcal{A}, \\ \text{(primal feasibility)} \qquad & \langle \pi_s, \mathbf{1} \rangle = 1, \; \pi(a|s) \geq 0 \qquad \forall\, a \in \mathcal{A}, \\ \text{(dual feasibility)} \qquad & c_s^a \geq 0 \quad \forall\, a \in \mathcal{A}, \end{aligned}$$

where $\log_q(\pi_s)$ is applied element-wise, λ_s and $(c_s^a)_{a \in \mathcal{A}}$ are the dual variables and e_a is a vector of all 0 with a 1 in the position associated to a, for all $a \in \mathcal{A}$. If we set $c_s^a = 0$ for all $a \in \mathcal{A}$, the complementary slackness and dual feasibility conditions are satisfied and, from the stationarity condition, we obtain that

$$\pi^t(a|s) = \exp_q(\eta_{t-1} f^t(s,a) - \lambda_s).$$

Since $\sum_{a \in \mathcal{A}} \exp_q(\eta_{t-1} f^t(s,a) - \lambda_s)$ converges to 0 and $+\infty$ when λ_s goes to $+\infty$ and $-\infty$, respectively, there always exists a λ_s such that the primal feasibility condition is satisfied, thanks to the intermediate value theorem. When $q = 1$, $\lambda_s = \log \|\exp(\eta_{t-1} f_s^t)\|_1$.

C.2. Example 3.3

Let h be the squared ℓ_2-norm, then the associated Bregman divergence is the Euclidean distance, i.e. $D_h(\pi_s, \bar{\pi}_s) = \|\pi_s - \bar{\pi}_s\|_2^2$ and the convex conjugate of h is ℓ_2-norm itself. Thus, $\nabla h^*(\cdot)$ is an identity function.

The statement in Example 3.3 with $f^\theta(s,a) = \theta_{s,a}$ follows immediately. Indeed, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we choose

$$f^{t+1}(s,a) = \widehat{Q}^t(s,a) + \frac{\pi^t(a|s)}{\eta_t},$$

such that $\varepsilon_{\text{actor}} = 0$. Then, the policy update with the Euclidean distance becomes, for all $s \in \mathcal{S}$,

$$\pi_s^{t+1} = \text{Proj}_{\Delta(\mathcal{A})}^{l_2}(\nabla h^*(\eta_t f_s^{t+1})) = \text{Proj}_{\Delta(\mathcal{A})}^{l_2}(\eta_t f_s^{t+1}) = \operatorname*{argmin}_{\pi \in \Delta(\mathcal{A})} \left\| \pi - (\pi_s^t + \eta_t \widehat{Q}_s^t) \right\|^2, \tag{25}$$

which is the inexact projected-Q descent with softmax tabular policies developed in Xiao (2022).

C.3. Example 3.5

Let h_b be the hyperentropy mirror map with scale parameter $b > 0$, that is

$$h_b(\pi_s) = \sum_{a\in\mathcal{A}} \pi(a|s)\,\mathrm{arcsinh}(\pi(a|s)/b) - \sqrt{\pi(a|s)^2 + b^2}.$$

To solve the minimization problem

$$\pi_s^\theta = \operatorname*{argmin}_{\pi_s\in\Delta(\mathcal{A})} \mathcal{D}_{h_b}(\pi_s, \nabla h_b^*(\eta_{t-1} f_s^t))$$

we use the KKT conditions. We formalize it as

$$\begin{aligned} \operatorname*{argmin}_{\pi_s\in\mathbb{R}^{|\mathcal{A}|}} \quad & \operatorname*{argmin}_{\pi_s\in\Delta(\mathcal{A})} \mathcal{D}_{h_b}(\pi_s, \nabla h_b^*(\eta_{t-1} f_s^t)) \\ \text{subject to } & \langle \pi_s, \mathbf{1}\rangle = 1 \\ & \pi(a|s) \geq 0 \qquad \forall\, a\in\mathcal{A}. \end{aligned}$$

The conditions then become

$$\begin{aligned} \text{(stationarity)} \qquad & \mathrm{arcsinh}(\pi_s/b) - \eta_{t-1} f_s^t + \lambda_s \mathbf{1} - \sum_{a\in\mathcal{A}} c_s^a e_a = 0, \\ \text{(complementary slackness)} \qquad & c_s^a \pi(a|s) = 0 \quad \forall\, a\in\mathcal{A}, \\ \text{(primal feasibility)} \qquad & \langle \pi_s, \mathbf{1}\rangle = 1,\ \pi(a|s) \geq 0 \qquad \forall\, a\in\mathcal{A}, \\ \text{(dual feasibility)} \qquad & c_s^a \geq 0 \quad \forall\, a\in\mathcal{A}, \end{aligned}$$

where $\mathrm{arcsinh}(\pi_s/b)$ is applied element-wise, λ_s and $(c_s^a)_{a\in\mathcal{A}}$ are the dual variables and e_a is a vector of all 0 with a 1 in the position associated to a, for all $a \in \mathcal{A}$. If we set $c_s^a = 0$ for all $a \in \mathcal{A}$, the complementary slackness and dual feasibility conditions are satisfied and, from the stationarity condition, we obtain that

$$\begin{aligned} \pi^t(a|s) &= b\sinh(\eta_{t-1} f^t(s,a) - \lambda_s) \\ &= b\sinh(\eta_{t-1} f^t(s,a))\sqrt{1+\sinh^2(\lambda_s)} - b\cosh(\eta_{t-1} f^t(s,a))\sinh(\lambda_s). \end{aligned}$$

It remains to satisfy the primal feasibility condition, that is find λ_s such that

$$b\sum_{a\in\mathcal{A}} \sinh(\eta_{t-1} f^t(s,a) - \lambda_s) = 1.$$

Using properties of the hyperbolic sine, we have the equivalent expression

$$\sum_{a\in\mathcal{A}} \sinh(\eta_{t-1} f^t(s,a))\sqrt{1+\sinh^2(\lambda)} - \cosh(\eta_{t-1} f^t(s,a))\sinh(\lambda_s) = 1/b.$$

Solving for $\sinh(\lambda_s)$, we obtain

$$\sinh(\lambda_s) = \frac{\beta \pm \sqrt{\beta^2 + (\beta^2-\alpha^2)(\alpha^2-1)}}{\beta^2-\alpha^2},$$

where $\alpha = b\sum_{a\in\mathcal{A}} \sinh(\eta_{t-1} f^t(s,a))$ and $\beta = b\sum_{a\in\mathcal{A}} \cosh(\eta_{t-1} f^t(s,a))$. Since $\cosh(x) \geq \sinh(x)$ for any $x \in \mathbb{R}$, at least one of the two solutions satisfies the primal feasibility condition.

C.4. Example 3.6

To obtain the statement in Example 3.6, we give a slight variation of Proposition 2 in Krichene et al. (2015).

Proposition C.1. *For $a \in (-\infty, +\infty]$ and $\omega \leq 0$, an increasing C^1-diffeomorphism $\phi : (-\infty, a) \to (\omega, +\infty)$ is called an ω-potential if*

$$\lim_{u \to -\infty} \phi(u) = \omega, \quad \lim_{u \to a} \phi(u) = +\infty, \quad \int_0^1 \phi^{-1}(u) du \leq \infty.$$

Let the mirror map h be defined as

$$h(\pi_s) = \sum_{a \in \mathcal{A}} \int_1^{\pi(a|s)} \phi^{-1}(u) du.$$

We have then that π_s^t is a solution to

$$\min_{\pi \in \Delta_s} Proj^h_{\Delta(\mathcal{A})}(\nabla h^*(\eta_{t-1} f_s^t))$$

if and only if there exist λ_s such that

$$\pi^t(a|s) = [\phi(\eta_{t-1} f^t(s,a) + \lambda_s)]_+ \quad a \in \mathcal{A}$$

and $\sum_{a \in \mathcal{A}} \pi^t(a|s) = 1$, where $\lambda_s \in \mathbb{R}$ for all $s \in \mathcal{S}$ and $[z]_+ = \max(z, 0)$ for $z \in \mathbb{R}$.

D. Deferred proofs from Section 4.1

D.1. Proof of the variant of the three-point descent lemma – Lemma 4.1

Here we provide the proof of Lemma 4.1, a variant of the three-point descent lemma with the integration of an arbitrary parametrized function, which is the key tool for our AMPO analysis. It is a variation of both Xiao (2022, Lemma 6) and Chen and Teboulle (1993, Lemma 3.2). First, we recall some technical conditions of the mirror map (Bubeck, 2015, Chapter 4).

Suppose that $\mathcal{Y} \subset \mathbb{R}^{|\mathcal{A}|}$ is a closed convex set, we say a function $h : \mathcal{Y} \to \mathbb{R}$ is a *mirror map* if it satisfies the following properties:

(i) h is strictly convex and differentiable;

(ii) h is essentially smooth, i.e., the graident of h diverges on the boundary of $\mathcal{Y}$, that is $\lim_{x \to \partial \mathcal{Y}} \|\nabla h(x)\| \to \infty$;

(iii) the gradient of h takes all possible values, that is $\nabla h(\mathcal{Y}) = \mathbb{R}^{|\mathcal{A}|}$.

To prove Lemma 4.1, we also need the following rather simple properties, three-point identity and the generalized Pythagorean theorem, satisfied by the Bregman divergence. We provide their proofs for self-containment.

Lemma D.1 (Three-point identity, Lemma 3.1 in Chen and Teboulle (1993))**.** *Let h be a mirror map. Then for any a, b in the relative interior of $\mathcal{Y}$ and $c \in \mathcal{Y}$, the following identity holds:*

$$\mathcal{D}_h(c, a) + \mathcal{D}_h(a, b) - \mathcal{D}_h(c, b) = \langle \nabla h(b) - \nabla h(a), c - a \rangle. \tag{26}$$

Proof. Indeed, using the definition of the Bregman divergence $\mathcal{D}_h$, we have

$$\langle \nabla h(a), c - a \rangle = h(c) - h(a) - \mathcal{D}_h(c, a), \tag{27}$$
$$\langle \nabla h(b), a - b \rangle = h(a) - h(b) - \mathcal{D}_h(a, b), \tag{28}$$
$$\langle \nabla h(b), c - b \rangle = h(c) - h(b) - \mathcal{D}_h(c, b). \tag{29}$$

Subtracting (27) and (28) from (29) yields (26). □

Lemma D.2 (Generalized Pythagorean Theorem of Bregman divergence, Lemma 4.1 in Bubeck (2015))**.** *Let $\mathcal{X} \subseteq \mathcal{Y}$ be a closed convex set. Let h be a mirror map defined on $\mathcal{Y}$. Let $x, y \in \mathcal{X}$, then*

$$\left\langle \nabla h \left(Proj^h_{\mathcal{X}}(y) \right) - \nabla(y), Proj^h_{\mathcal{X}}(y) - x \right\rangle \leq 0,$$

which also implies

$$\mathcal{D}_h \left(x, Proj^h_{\mathcal{X}}(y) \right) + \mathcal{D}_h \left(Proj^h_{\mathcal{X}}(y), y \right) \leq \mathcal{D}_h(x, y). \tag{30}$$

Proof. From the definition of $\text{Proj}_{\mathcal{X}}^h(y)$, which is

$$\text{Proj}_{\mathcal{X}}^h(y) \in \underset{y' \in \mathcal{X}}{\text{argmin}}\, \mathcal{D}_h(y', y),$$

and from the first-order optimality condition (Bubeck, 2015, Proposition 1.3) with

$$\nabla_{y'} \mathcal{D}_h(y', y) = \nabla h(y') - \nabla h(y), \quad \text{for all } y' \in \mathcal{Y},$$

we have

$$\left\langle \nabla_{y'} \mathcal{D}_h(y', y)|_{y' = \text{Proj}_{\mathcal{X}}^h(y)}, \text{Proj}_{\mathcal{X}}^h(y) - x \right\rangle \leq 0 \quad \Longrightarrow \quad \left\langle \nabla h\left(\text{Proj}_{\mathcal{X}}^h(y)\right) - \nabla(y), \text{Proj}_{\mathcal{X}}^h(y) - x \right\rangle \leq 0,$$

which implies (30) by applying the definition of Bregman divergence and arranging terms. □

Now we are ready to prove Lemma 4.1.

Lemma D.3 (Lemma 4.1)**.** *Let $\mathcal{Y} \subset \mathbb{R}^{|\mathcal{A}|}$ be a closed convex set with $\Delta(\mathcal{A}) \subseteq \mathcal{Y}$. For any policies $\pi \in \Delta(\mathcal{A})$ and $\bar{\pi}$ in the relative interior of $\Delta(\mathcal{A})$, any function f^θ with $\theta \in \Theta$, any $s \in \mathcal{S}$ and for $\eta > 0$, we have that,*

$$\langle \eta f_s^\theta - \nabla h(\bar{\pi}_s), \pi_s - \tilde{\pi}_s \rangle \leq \mathcal{D}_h(\pi_s, \bar{\pi}_s) - \mathcal{D}_h(\tilde{\pi}_s, \bar{\pi}_s) - \mathcal{D}_h(\pi, \tilde{\pi}_s),$$

where $\tilde{\pi}$ is induced by f^θ and η according to Definition 3.1, that is, for all $s \in \mathcal{S}$,

$$\tilde{\pi}_s = \text{Proj}_{\Delta(\mathcal{A})}^h\left(\nabla h^*(\eta f_s^\theta)\right) = \underset{\pi'_s \in \Delta(\mathcal{A})}{\text{argmin}}\, \mathcal{D}_h(\pi'_s, \nabla h^*(\eta f_s^\theta)). \tag{31}$$

Proof. Since $\nabla h(\mathcal{Y}) = \mathbb{R}^{|\mathcal{A}|}$, for every $\theta \in \Theta$, there exists $p \in \mathcal{Y}^{\mathcal{S}}$ such that for every $s \in \mathcal{S}$, we have $\nabla h(p_s) = \eta f_s^\theta$ with $p_s \in \mathcal{Y}$.

So by using the property $\nabla^*(\nabla h(\cdot)) = id(\cdot)$ with id the identity function, (31) can be rewritten as, for all $s \in \mathcal{S}$,

$$\tilde{\pi}_s = \underset{\pi'_s \in \Delta(\mathcal{A})}{\text{argmin}}\, \mathcal{D}_h(\pi'_s, p_s) = \text{Proj}_{\Delta(\mathcal{A})}^h(p_s).$$

Now plugging $a = \bar{\pi}_s, b = p_s$ and $c = \pi_s$ in the three-point identity lemma D.1, we obtain

$$\mathcal{D}_h(\pi_s, \bar{\pi}_s) - \mathcal{D}_h(\pi_s, p_s) + \mathcal{D}_h(\bar{\pi}_s, p_s) = \langle \nabla h(\bar{\pi}_s) - \nabla h(p_s), \bar{\pi}_s - \pi_s \rangle. \tag{32}$$

Similarly, plugging $a = \bar{\pi}_s, b = p_s$ and $c = \tilde{\pi}_s$ in the three-point identity lemma D.1, we obtain

$$\mathcal{D}_h(\tilde{\pi}_s, \bar{\pi}_s) - \mathcal{D}_h(\tilde{\pi}_s, p_s) + \mathcal{D}_h(\bar{\pi}_s, p_s) = \langle \nabla h(\bar{\pi}_s) - \nabla h(p_s), \bar{\pi}_s - \tilde{\pi}_s \rangle. \tag{33}$$

From (32), we have

$$\begin{aligned}
\mathcal{D}_h(\pi_s, \bar{\pi}_s) - \mathcal{D}_h(\pi_s, p_s) + \mathcal{D}_h(\bar{\pi}_s, p_s) &= \langle \nabla h(\bar{\pi}_s) - \nabla h(p_s), \bar{\pi}_s - \pi_s \rangle \\
&= \langle \nabla h(\bar{\pi}_s) - \nabla h(p_s), \bar{\pi}_s - \tilde{\pi}_s \rangle + \langle \nabla h(\bar{\pi}_s) - \nabla h(p_s), \tilde{\pi}_s - \pi_s \rangle \\
&\overset{(33)}{=} \mathcal{D}_h(\tilde{\pi}_s, \bar{\pi}_s) - \mathcal{D}_h(\tilde{\pi}_s, p_s) + \mathcal{D}_h(\bar{\pi}_s, p_s) + \langle \nabla h(\bar{\pi}_s) - \nabla h(p_s), \tilde{\pi}_s - \pi_s \rangle.
\end{aligned}$$

By rearranging terms, we have

$$\mathcal{D}_h(\pi_s, \bar{\pi}_s) - \mathcal{D}_h(\tilde{\pi}_s, \bar{\pi}_s) - \mathcal{D}_h(\pi_s, p_s) + \mathcal{D}_h(\tilde{\pi}_s, p_s) = \langle \nabla h(\bar{\pi}_s) - \nabla h(p_s), \tilde{\pi}_s - \pi_s \rangle. \tag{34}$$

Furthermore, from the non-expansivity property, which is also called the Generalized Pythagorean Theorem of the Bregman divergence in Lemma D.2, and from the fact that $\tilde{\pi}_s = \text{Proj}_{\Delta(\mathcal{A})}^h(p_s)$, we know that

$$\mathcal{D}_h\left(\pi_s, \text{Proj}_{\Delta(\mathcal{A})}^h(p_s)\right) + \mathcal{D}_h\left(\text{Proj}_{\Delta(\mathcal{A})}^h(p_s), p_s\right) \leq \mathcal{D}_h(\pi_s, p_s),$$

that is

$$\mathcal{D}_h(\pi_s, \tilde{\pi}_s) + \mathcal{D}_h(\tilde{\pi}_s, p_s) \leq \mathcal{D}_h(\pi_s, p_s) \quad \Longleftrightarrow \quad -\mathcal{D}_h(\pi_s, p_s) + \mathcal{D}_h(\tilde{\pi}_s, p_s) \leq -\mathcal{D}_h(\pi_s, \tilde{\pi}_s).$$

Plugging the above inequality into the left hand side of (34) yields

$$\mathcal{D}_h(\pi_s, \bar{\pi}_s) - \mathcal{D}_h(\tilde{\pi}_s, \bar{\pi}_s) - \mathcal{D}_h(\pi_s, \tilde{\pi}_s) \geq \langle \nabla h(\bar{\pi}_s) - \nabla h(p_s), \tilde{\pi}_s - \pi_s \rangle,$$

which concludes the proof with $\nabla h(p_s) = \eta f_s^\theta$. □

D.2. Bounding errors

In this section, we will bound error terms of the type

$$\mathbb{E}_{s\sim d_\mu^\pi, a\sim\pi_s}\left[Q^t(s,a)+\eta_t^{-1}\nabla h(\pi_s^t)(a)-f^{t+1}(s,a)\right], \tag{35}$$

where $(d_\mu^\pi,\pi)\in\{(d_\mu^\star,\pi^\star),(d_\mu^{t+1},\pi^{t+1}),(d_\mu^\star,\pi^t),(d_\mu^{t+1},\pi^t)\}$. These error terms will appear in the forthcoming proofs of our theorems. They directly induce the error floors of our convergence results.

Since $\nabla h(\mathcal{Y})=\mathbb{R}^{|\mathcal{A}|}$, at time $t+1$, there exists $p^{t+1}\in\mathcal{Y}^{\mathcal{S}}$ such that for every $s\in\mathcal{S}$, $\nabla h(p_s^{t+1})=f_s^{t+1}$.

In the rest of Appendix D, let $q^t:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ such that, for every $s\in\mathcal{S}$,

$$q_s^t:=\eta_t^{-1}(\nabla h(p_s^{t+1})-\nabla h(\pi_s^t))=f_s^{t+1}-\eta_t^{-1}\nabla h(\pi_s^t)\in\mathbb{R}^{|\mathcal{A}|}.$$

So (35) can be rewritten as

$$\mathbb{E}_{s\sim d_\mu^\pi, a\sim\pi_s}\left[Q^t(s,a)+\eta_t^{-1}[\nabla h(\pi_s^t)]_a-f^{t+1}(s,a)\right]=\mathbb{E}_{s\sim d_\mu^\pi, a\sim\pi_s}\left[Q^t(s,a)-q^t(s,a)\right]. \tag{36}$$

To bound it, we separate the errors in critic and actor errors. That is,

$$\mathbb{E}_{s\sim d_\mu^\pi, a\sim\pi_s}\left[Q^t(s,a)-q^t(s,a)\right]=\underbrace{\mathbb{E}_{s\sim d_\mu^\pi, a\sim\pi_s}\left[Q^t(s,a)-\widehat{Q}^t(s,a)\right]}_{\text{Error for the critic}}+\underbrace{\mathbb{E}_{s\sim d_\mu^\pi, a\sim\pi_s}\left[\widehat{Q}^t(s,a)-q^t(s,a)\right]}_{\text{Error for the actor}}.$$

Critic error. The critic error measures how accurate the estimation $\widehat{Q}^t$ of the Q-function Q^t at time t.

To bound it, let $(\rho^t)_{t\geq 0}$ be a sequence of distributions over states and actions. By using Cauchy-Schwartz's inequality, we have

$$\begin{aligned}
\mathbb{E}_{s\sim d_\mu^\pi, a\sim\pi_s}\left[Q^t(s,a)-\widehat{Q}^t(s,a)\right]&=\int_{s\in\mathcal{S},a\in\mathcal{A}}\frac{d_\mu^\pi(s)\pi(a\mid s)}{\sqrt{\rho^t(s,a)}}\cdot\sqrt{\rho^t(s,a)}(Q^t(s,a)-\widehat{Q}^t(s,a))\\
&\leq\sqrt{\int_{s\in\mathcal{S},a\in\mathcal{A}}\frac{\left(d_\mu^\pi(s)\pi(a\mid s)\right)^2}{\rho^t(s,a)}\cdot\int_{s\in\mathcal{S},a\in\mathcal{A}}\rho^t(s,a)(Q^t(s,a)-\widehat{Q}^t(s,a))^2}\\
&=\sqrt{\mathbb{E}_{(s,a)\sim\rho^t}\left[\left(\frac{d_\mu^\pi(s)\pi(a\mid s)}{\rho^t(s,a)}\right)^2\right]\cdot\mathbb{E}_{(s,a)\sim\rho^t}\left[(Q^t(s,a)-\widehat{Q}^t(s,a))^2\right]}\\
&\leq\sqrt{C_1\varepsilon_{\text{critic}}},
\end{aligned} \tag{37}$$

where the last line is obtained by Assumptions (A3) and (A1).

Actor error. The actor error evaluates how well the regression solver for θ^{t+1} performs at time t.

To bound it, let $(v^t)_{t\geq 0}$ be a sequence of distributions over states and actions. Similarly, by using Cauchy-Schwartz's inequality, we have

$$\begin{aligned}
\mathbb{E}_{s\sim d_\mu^\pi, a\sim\pi_s}\left[\widehat{Q}^t(s,a)-q^t(s,a)\right]&=\int_{s\in\mathcal{S},a\in\mathcal{A}}\frac{d_\mu^\pi(s)\pi(a\mid s)}{\sqrt{v^t(s,a)}}\cdot\sqrt{v^t(s,a)}(\widehat{Q}^t(s,a)-q^t(s,a))\\
&\leq\sqrt{\int_{s\in\mathcal{S},a\in\mathcal{A}}\frac{\left(d_\mu^\pi(s)\pi(a\mid s)\right)^2}{v^t(s,a)}\cdot\int_{s\in\mathcal{S},a\in\mathcal{A}}v^t(s,a)(\widehat{Q}^t(s,a)-q^t(s,a))^2}\\
&=\sqrt{\mathbb{E}_{(s,a)\sim v^t}\left[\left(\frac{d_\mu^\pi(s)\pi(a\mid s)}{v^t(s,a)}\right)^2\right]\cdot\mathbb{E}_{(s,a)\sim v^t}\left[(\widehat{Q}^t(s,a)-q^t(s,a))^2\right]}\\
&\leq\sqrt{C_2\varepsilon_{\text{actor}}},
\end{aligned} \tag{38}$$

where the last line is obtained by Assumptions (A3) and (A2).

Plugging (37) and (38) into (36) yields the bound

$$\left|\mathbb{E}_{s\sim d_\mu^\pi, a\sim\pi_s}\left[Q^t(s,a) - q^t(s,a)\right]\right| \leq \sqrt{C_1\varepsilon_{\text{critic}}} + \sqrt{C_2\varepsilon_{\text{actor}}}, \tag{39}$$

where $(d_\mu^\pi, \pi) \in \{(d_\mu^\star, \pi^\star), (d_\mu^{t+1}, \pi^{t+1}), (d_\mu^\star, \pi^t), (d_\mu^{t+1}, \pi^t)\}$.

D.3. Quasi-monotonic updates – Proof of Proposition 4.2

In this section, we show that the AMPO updates guarantee a quasi-monotonic property, i.e. a non-decreasing property up to a certain error floor due to the actor and the critic errors, which allows us to establish an important recursion about the AMPO iterates next. First, we recall the performance difference lemma (Kakade and Langford, 2002) which is the second key tool for our analysis and a well known result in the RL literature. Here we use a particular form of the lemma presented in Yuan et al. (2023, Lemma 3).

Lemma D.4 (Performance difference lemma, Lemma 3 in (Yuan et al., 2023))**.** *For any policy $\pi, \pi' \in \Delta(\mathcal{A})^\mathcal{S}$ and $\mu \in \Delta(\mathcal{S})$,*

$$V^\pi(\mu) - V^{\pi'}(\mu) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^\pi}\left[\left\langle Q_s^{\pi'}, \pi_s - \pi'_s\right\rangle\right].$$

Recall the notation

$$\tau := \frac{2}{1-\gamma}(\sqrt{C_1\varepsilon_{\text{critic}}} + \sqrt{C_2\varepsilon_{\text{actor}}}).$$

The following result characterizes the non-decreasing property of AMPO. The bounding errors (39) in the previous appendix D.2 will be used to prove the lemma. It is a slightly stronger result than Proposition 4.2.

Lemma D.5. *Consider the iterates of Algorithm 1, at each time $t \geq 0$, we have*

$$V^{t+1}(\mu) - V^t(\mu) \geq \mathbb{E}_{s\sim d_\mu^{t+1}}\left[\frac{\mathcal{D}_h(\pi_s^{t+1}, \pi_s^t) + \mathcal{D}_h(\pi_s^t, \pi_s^{t+1})}{\eta_t(1-\gamma)}\right] - \tau.$$

Proof. Using the variant of the three-point descent lemma 4.1 with $\bar\pi = \pi^t$, $f^\theta = f^{t+1}$, $\eta = \eta_t$, thus $\tilde\pi = \pi^{t+1}$ by Definition 3.1 and Algorithm 1, and $\pi_s = \pi_s^t$, we have

$$\langle \eta_t q_s^t, \pi_s^t - \pi_s^{t+1}\rangle \leq \mathcal{D}_h(\pi_s^t, \pi_s^t) - \mathcal{D}_h(\pi_s^{t+1}, \pi_s^t) - \mathcal{D}_h(\pi_s^t, \pi_s^{t+1}). \tag{40}$$

By rearranging terms and $\mathcal{D}_h(\pi_s^t, \pi_s^t) = 0$, we have

$$\langle \eta_t q_s^t, \pi_s^{t+1} - \pi_s^t\rangle \geq \mathcal{D}_h(\pi_s^{t+1}, \pi_s^t) + \mathcal{D}_h(\pi_s^t, \pi_s^{t+1}) \geq 0. \tag{41}$$

Then, by the performance difference lemma D.4, we have

$$\begin{aligned}
(1-\gamma)(V^{t+1}(\mu) - V^t(\mu)) &= \mathbb{E}_{s\sim d_\mu^{t+1}}\left[\langle Q_s^t, \pi_s^{t+1} - \pi_s^t\rangle\right] \\
&= \mathbb{E}_{s\sim d_\mu^{t+1}}\left[\langle q_s^t, \pi_s^{t+1} - \pi_s^t\rangle\right] + \mathbb{E}_{s\sim d_\mu^{t+1}}\left[\langle Q_s^t - q_s^t, \pi_s^{t+1} - \pi_s^t\rangle\right] \\
&\overset{(40)}{\geq} \mathbb{E}_{s\sim d_\mu^{t+1}}\left[\frac{\mathcal{D}_h(\pi_s^{t+1}, \pi_s^t) + \mathcal{D}_h(\pi_s^t, \pi_s^{t+1})}{\eta_t}\right] - \left|\mathbb{E}_{s\sim d_\mu^{t+1}}\left[\langle Q_s^t - q_s^t, \pi_s^{t+1} - \pi_s^t\rangle\right]\right| \\
&\geq \mathbb{E}_{s\sim d_\mu^{t+1}}\left[\frac{\mathcal{D}_h(\pi_s^{t+1}, \pi_s^t) + \mathcal{D}_h(\pi_s^t, \pi_s^{t+1})}{\eta_t}\right] - \tau(1-\gamma),
\end{aligned}$$

which concludes the proof after dividing both sides by $(1-\gamma)$. Indeed, the last line is obtained by bounding $\left|\mathbb{E}_{s\sim d_\mu^{t+1}}\left[\langle Q_s^t - q_s^t, \pi_s^{t+1} - \pi_s^t\rangle\right]\right|$ through the following result

$$\begin{aligned}
\left|\mathbb{E}_{s\sim d_\mu^{t+1}}\left[\langle Q_s^t - q_s^t, \pi_s^{t+1} - \pi_s^t\rangle\right]\right| &\leq \left|\mathbb{E}_{s\sim d_\mu^{t+1}, a\sim\pi_s^{t+1}}\left[Q^t(s,a) - q^t(s,a)\right]\right| + \left|\mathbb{E}_{s\sim d_\mu^{t+1}, a\sim\pi_s^t}\left[Q^t(s,a) - q^t(s,a)\right]\right| \\
&\overset{(39)}{\leq} 2(\sqrt{C_1\varepsilon_{\text{critic}}} + \sqrt{C_2\varepsilon_{\text{actor}}}) = \tau(1-\gamma),
\end{aligned} \tag{42}$$

where the first term is upper bounded by $\sqrt{C_1\varepsilon_{\text{critic}}} + \sqrt{C_2\varepsilon_{\text{actor}}}$ through (39) with $(d_\mu^\pi, \pi) = (d_\mu^{t+1}, \pi^{t+1})$, and the second term is also upper bounded by $\sqrt{C_1\varepsilon_{\text{critic}}} + \sqrt{C_2\varepsilon_{\text{actor}}}$ through (39) with $(d_\mu^\pi, \pi) = (d_\mu^{t+1}, \pi^{t})$, respectively. □

D.4. Main passage – An important recursion about the AMPO method

In this section, we show an important recursion result for the AMPO updates, which will be used for both the sublinear and the linear convergence analysis of AMPO.

To simplify proofs in the rest of Appendix D, let

$$\nu_t := \left\| \frac{d_\mu^\star}{d_\mu^{t+1}} \right\|_\infty := \max_{s\in\mathcal{S}} \frac{d_\mu^\star(s)}{d_\mu^{t+1}(s)}.$$

For two different time $t, t' \geq 0$, let $\mathcal{D}_{t'}^t$ denote the expected Bregman divergence between the policy π^t and policy $\pi^{t'}$, where the expectation is taken over the discounted state visitation distribution of the optimal policy $d_\mu^\star$, that is,

$$\mathcal{D}_{t'}^t := \mathbb{E}_{s\sim d_\mu^\star}\left[\mathcal{D}_h(\pi_s^t, \pi_s^{t'})\right].$$

Similarly, let $\mathcal{D}_t^\star$ denote the expected Bregman divergence between the optimal policy $\pi^\star$ and π^t, that is,

$$\mathcal{D}_t^\star := \mathbb{E}_{s\sim d_\mu^\star}\left[\mathcal{D}_h(\pi_s^\star, \pi_s^{t})\right].$$

Let $\Delta_t := V^\star(\mu) - V^t(\mu)$ be the optimality gap.

We can now state the following important recursion result for the AMPO method.

Proposition D.6. *Consider the iterates of Algorithm 1, at each time $t \geq 0$, we have*

$$\frac{\mathcal{D}_t^{t+1}}{(1-\gamma)\eta_t} + C_3\left(\Delta_{t+1} - \Delta_t\right) + \Delta_t \leq \frac{\mathcal{D}_t^\star}{(1-\gamma)\eta_t} - \frac{\mathcal{D}_{t+1}^\star}{(1-\gamma)\eta_t} + (1+C_3)\tau.$$

Proof. Using the three-point descent lemma 4.1 with $\bar{\pi} = \pi^t$, $f^\theta = f^{t+1}$, $\eta = \eta_t$, and thus $\tilde{\pi} = \pi^{t+1}$ by Definition 3.1 and Algorithm 1, and $\pi_s = \pi_s^\star$, we have that

$$\langle \eta_t q_s^t, \pi_s^\star - \pi_s^{t+1}\rangle \leq \mathcal{D}_h(\pi^\star, \pi^t) - \mathcal{D}_h(\pi^\star, \pi^{t+1}) - \mathcal{D}_h(\pi^{t+1}, \pi^t),$$

which can be decomposed by

$$\langle \eta_t q_s^t, \pi_s^t - \pi_s^{t+1}\rangle + \langle \eta_t q_s^t, \pi_s^\star - \pi_s^{t}\rangle \leq \mathcal{D}_h(\pi^\star, \pi^t) - \mathcal{D}_h(\pi^\star, \pi^{t+1}) - \mathcal{D}_h(\pi^{t+1}, \pi^t).$$

Taking expectation with respect to the distribution $d_\mu^\star$ over states and dividing both side by η_t, we have

$$\mathbb{E}_{s\sim d_\mu^\star}\left[\langle q_s^t, \pi_s^t - \pi_s^{t+1}\rangle\right] + \mathbb{E}_{s\sim d_\mu^\star}\left[\langle q_s^t, \pi_s^\star - \pi_s^{t}\rangle\right] \leq \frac{1}{\eta_t}(\mathcal{D}_t^\star - \mathcal{D}_{t+1}^\star - \mathcal{D}_t^{t+1}). \tag{43}$$

We lower bound the two terms on the left hand side of (43) separately.

For the first one, we have that

$$\begin{aligned}
\mathbb{E}_{s\sim d_\mu^\star}\left[\langle q_s^t, \pi_s^t - \pi_s^{t+1}\rangle\right] &\overset{(41)}{\geq} \left\| \frac{d_\mu^\star}{d_\mu^{t+1}} \right\|_\infty \mathbb{E}_{s\sim d_\mu^{t+1}}\left[\langle q_s^t, \pi_s^t - \pi_s^{t+1}\rangle\right] \\
&= \nu_{t+1}\mathbb{E}_{s\sim d_\mu^{t+1}}\left[\langle Q_s^t, \pi_s^t - \pi_s^{t+1}\rangle\right] + \nu_{t+1}\mathbb{E}_{s\sim d_\mu^{t+1}}\left[\langle q_s^t - Q_s^t, \pi_s^t - \pi_s^{t+1}\rangle\right] \\
&\overset{\text{Lemma D.4}}{=} \nu_{t+1}(1-\gamma)\left(V^t(\mu) - V^{t+1}(\mu)\right) + \nu_{t+1}\mathbb{E}_{s\sim d_\mu^{t+1}}\left[\langle q_s^t - Q_s^t, \pi_s^t - \pi_s^{t+1}\rangle\right] \\
&\overset{(42)}{\geq} \nu_{t+1}(1-\gamma)\left(V^t(\mu) - V^{t+1}(\mu)\right) - \nu_{t+1}\tau(1-\gamma) \\
&= \nu_{t+1}(1-\gamma)\left(\Delta_{t+1} - \Delta_t\right) - \nu_{t+1}\tau(1-\gamma).
\end{aligned}$$

For the second one, we have that

$$\begin{aligned}\mathbb{E}_{s\sim d_\mu^\star}\left[\langle q_s^t, \pi_s^\star - \pi_s^t\rangle\right] &= \mathbb{E}_{s\sim d_\mu^\star}\left[\langle Q_s^t, \pi_s^\star - \pi_s^t\rangle\right] + \mathbb{E}_{s\sim d_\mu^\star}\left[\langle q_s^t - Q_s^t, \pi_s^\star - \pi_s^t\rangle\right]\\ &\overset{\text{Lemma D.4}}{=} \Delta_t(1-\gamma) + \mathbb{E}_{s\sim d_\mu^\star}\left[\langle q_s^t - Q_s^t, \pi_s^\star - \pi_s^t\rangle\right]\\ &\geq \Delta_t(1-\gamma) - \tau(1-\gamma),\end{aligned}$$

where the upper bound of $\left|\mathbb{E}_{s\sim d_\mu^\star}\left[\langle q_s^t - Q_s^t, \pi_s^\star - \pi_s^t\rangle\right]\right|$ is $\tau(1-\gamma)$, similar to the derivation of (42), by applying (39) twice with $(d_\mu^\pi, \pi) = (d_\mu^\star, \pi^\star)$ and $(d_\mu^\pi, \pi) = (d_\mu^\star, \pi^t)$, respectively.

Plugging the two bounds in (43), dividing both side by $(1-\gamma)$ and rearranging terms, we obtain

$$\frac{\mathcal{D}_t^{t+1}}{(1-\gamma)\eta_t} + \nu_{t+1}\left(\Delta_{t+1} - \Delta_t - \tau\right) + \Delta_t \leq \frac{\mathcal{D}_t^\star}{(1-\gamma)\eta_t} - \frac{\mathcal{D}_{t+1}^\star}{(1-\gamma)\eta_t} + \tau.$$

From Proposition 4.2, we have that $\Delta_{t+1} - \Delta_t - \tau \leq 0$. Consequently, since $\nu_{t+1} \leq C_3$ by the definition of C_3 in Assumption (A4), one can lower bound the left hand side of the above inequality by replacing ν_{t+1} by C_3, that is,

$$\frac{\mathcal{D}_t^{t+1}}{(1-\gamma)\eta_t} + C_3\left(\Delta_{t+1} - \Delta_t - \tau\right) + \Delta_t \leq \frac{\mathcal{D}_t^\star}{(1-\gamma)\eta_t} - \frac{\mathcal{D}_{t+1}^\star}{(1-\gamma)\eta_t} + \tau,$$

which concludes the proof. □

D.5. Proof of the sublinear convergence analysis and Corollary 4.4

In this section, we derive the sublinear convergence result of Theorem 4.3 and Corollary 4.4 with non-decreasing step size.

First, we derive the proof of the sublinear convergence result of Theorem 4.3.

Proof. Starting from Proposition D.6

$$\frac{\mathcal{D}_t^{t+1}}{(1-\gamma)\eta_t} + C_3\left(\Delta_{t+1} - \Delta_t\right) + \Delta_t \leq \frac{\mathcal{D}_t^\star}{(1-\gamma)\eta_t} - \frac{\mathcal{D}_{t+1}^\star}{(1-\gamma)\eta_t} + (1+C_3)\tau.$$

If $\eta_t \leq \eta_{t+1}$,

$$\frac{\mathcal{D}_t^{t+1}}{(1-\gamma)\eta_t} + C_3\left(\Delta_{t+1} - \Delta_t\right) + \Delta_t \leq \frac{\mathcal{D}_t^\star}{(1-\gamma)\eta_t} - \frac{\mathcal{D}_{t+1}^\star}{(1-\gamma)\eta_{t+1}} + (1+C_3)\tau. \tag{44}$$

Summing up from 0 to $T-1$ and dropping some positive terms on the left hand side and some negative terms on the right hand side, we have

$$\sum_{t<T}\Delta_t \leq \frac{\mathcal{D}_0^\star}{(1-\gamma)\eta_0} + C_3\Delta_0 + T(1+C_3)\tau \leq \frac{\mathcal{D}_0^\star}{(1-\gamma)\eta_0} + \frac{C_3}{1-\gamma} + T(1+C_3)\tau.$$

Notice that $\Delta_0 \leq \frac{1}{1-\gamma}$ as $r(s,a) \in [0,1]$. By dividing T on both side, we yield the proof of the sublinear convergence

$$V^\star(\mu) - \frac{1}{T}\sum_{t<T}V^t(\mu) \leq \frac{1}{T}\left(\frac{\mathcal{D}_0^\star}{(1-\gamma)\eta_0} + \frac{C_3}{1-\gamma}\right) + (1+C_3)\tau.$$

□

Now we derive the proof of Corollary 4.4.

Proof. We follow the proof of the sublinear convergence result of Theorem 4.3.

From (44), summing up from 0 to $T-1$ and dropping some positive terms on the left hand side and some negative terms on the right hand side, we have

$$\sum_{t<T} \frac{\mathcal{D}_t^{t+1}}{(1-\gamma)\eta_t} \le \frac{\mathcal{D}_0^\star}{(1-\gamma)\eta_0} + C_3\Delta_0 - \Delta_0 + T(1+C_3)\tau \le \frac{\mathcal{D}_0^\star}{(1-\gamma)\eta_0} + \frac{C_3-1}{1-\gamma} + T(1+C_3)\tau.$$

Then multiplying $(1-\gamma)$ on both sides yields the proof. □

D.6. Proof of the linear convergence analysis

In this section, we derive the linear convergence result of Theorem 4.3 with exponentially increasing step size.

Proof. Starting from Proposition D.6 by dropping $\frac{\mathcal{D}_t^{t+1}}{(1-\gamma)\eta_t}$ on the left hand side, we have

$$C_3\left(\Delta_{t+1} - \Delta_t\right) + \Delta_t \le \frac{\mathcal{D}_t^\star}{(1-\gamma)\eta_t} - \frac{\mathcal{D}_{t+1}^\star}{(1-\gamma)\eta_t} + (1+C_3)\tau.$$

Dividing C_3 on both side and rearranging, we obtain

$$\Delta_{t+1} + \frac{\mathcal{D}_{t+1}^\star}{(1-\gamma)C_3\eta_t} \le \left(1 - \frac{1}{C_3}\right)\left(\Delta_t + \frac{\mathcal{D}_t^\star}{(1-\gamma)\eta_t(C_3-1)}\right) + \left(1 + \frac{1}{C_3}\right)\tau.$$

If the step sizes satisfy $\eta_{t+1}(C_3-1) \ge \eta_t C_3$ with $C_3 \ge 1$, then

$$\Delta_{t+1} + \frac{\mathcal{D}_{t+1}^\star}{(1-\gamma)\eta_{t+1}(C_3-1)} \le \left(1 - \frac{1}{C_3}\right)\left(\Delta_t + \frac{\mathcal{D}_t^\star}{(1-\gamma)\eta_t(C_3-1)}\right) + \left(1 + \frac{1}{C_3}\right)\tau.$$

Now we need the following simple fact, whose proof is straightforward and thus omitted.

Suppose $0 < \alpha < 1, b > 0$ and a nonnegative sequence $\{a_t\}_{t\ge 0}$ satisfies

$$a_{t+1} \le \alpha a_t + b \qquad \forall t \ge 0.$$

Then for all $t \ge 0$,

$$a_t \le \alpha^t a_0 + \frac{b}{1-\alpha}.$$

The proof of the linear convergence analysis follows by applying this fact with $a_t = \Delta_t + \frac{\mathcal{D}_t^\star}{(1-\gamma)\eta_t(C_3-1)}$, $\alpha = 1 - \frac{1}{C_3}$ and $b = \left(1 + \frac{1}{C_3}\right)\tau$. □

D.7. Alternative statement of Theorem 4.3

Besides, our results also hold in expectation. We give its statement here.

Theorem D.7. *Let (A1), (A2) hold in expectation. That is,*

$$\mathbb{E}\left[\mathbb{E}_{(s,a)\sim\rho_t}\left[\left(Q^t(s,a) - \widehat{Q}^t(s,a)\right)^2\right]\right] \le \varepsilon_{\text{critic}}, \tag{45}$$

$$\mathbb{E}\left[\mathbb{E}_{(s,a)\sim v_t}\left[\left(f^{t+1}(s,a) - \widehat{Q}^t(s,a) - \eta_t^{-1}\nabla h(\pi_s^t)(a)\right)^2\right]\right] \le \varepsilon_{\text{actor}}. \tag{46}$$

In both conditions, the expectations are with respect to the randomness in the sequence of iterates $\theta_0, \theta_1, \ldots, \theta_t$. Let (A3), (A4) be true. If the step size schedule is non-decreasing, i.e. $\eta_t \le \eta_{t+1}$ for all $t \ge 0$, we have that the iterates of Algorithm 1 satisfy: for every $T \ge 0$,

$$V^\star(\mu) - \frac{1}{T}\sum_{t<T}\mathbb{E}\left[V^t(\mu)\right] \le \frac{1}{T}\left(\frac{\mathcal{D}_0^\star}{(1-\gamma)\eta_0} + \frac{C_3}{1-\gamma}\right) + (1+C_3)\tau.$$

Furthermore, if the step size schedule is geometrically increasing, i.e. satisfies

$$\eta_{t+1} \geq \frac{C_3}{C_3 - 1}\eta_t \qquad \forall t \geq 0, \tag{47}$$

we have: for every $T \geq 0$,

$$V^\star(\mu) - \mathbb{E}\left[V^T(\mu)\right] \leq \frac{1}{1-\gamma}\left(1 - \frac{1}{C_3}\right)^T \left(1 + \frac{\mathcal{D}_0^\star}{\eta_0(C_3 - 1)}\right) + (1 + C_3)\tau.$$

The analysis is straightforward by following the one of Theorem 4.3.

D.8. Discussion on the Distribution Mismatch Coefficients and the Concentrability Coefficients

In our convergence analysis, Assumptions (A3) and (A4) involve concentrability coefficients C_1, C_2 and the distribution mismatch coefficients C_3, which are potentially large. We give extensive discussions on them, respectively.

Distribution mismatch coefficient C_3 As mentioned right after (A4), we have that

$$\max_{s\in\mathcal{S}} \frac{d_\mu^\star(s)}{d_\mu^t(s)} \leq \frac{1}{1-\gamma} \max_{s\in\mathcal{S}} \frac{d_\mu^\star}{\mu} := C_3',$$

which is a sufficient upper bound for C_3. As discussed in Yuan et al. (2023, Section 5.2),

$$1/(1-\gamma) \leq C_3' \leq 1/((1-\gamma)\min_s \mu(s)).$$

The upper bound $1/((1-\gamma)\min_s \mu(s))$ of C_3' is very pessimistic and the lower bound $C_3' = 1/(1-\gamma)$ is often achieved by choosing $\mu = d_\mu^\star$.

Furthermore, if μ does not have full support on the state space, i.e. $1/((1-\gamma)\min_s \mu(s))$ might be infinity, one can always convert the convergence guarantees for some state distribution $\mu' \in \Delta(\mathcal{S})$ with full support such that

$$\begin{aligned} V^\star(\mu) - V^T(\mu) &= \int_{s\in\mathcal{S}} \frac{\mu(s)}{\mu'(s)}\mu'(s)\left(V^\star(s) - V^T(s)\right) \\ &\leq \left\|\frac{\mu}{\mu'}\right\|_\infty \left(V^\star(\mu') - V^T(\mu')\right). \end{aligned}$$

Then by the linear convergence result of Theorem 4.3, we only need convergence guarantee on $V^\star(\mu') - V^T(\mu')$ with an arbitrary distribution μ' such that C_3' is finite. We refer to Yuan et al. (2023, Section 5.2) for more discussions on the distribution mismatch coefficient.

Concentrability coefficients C_1 and C_2 As discussed in Yuan et al. (2023, Section 5.2), the issue of having (potentially large) concentrability coefficients is unavoidable in all the fast linear convergence analysis of approximate PMD due to the actor and critic errors (Cen et al., 2021; Zhan et al., 2021; Lan, 2022; Cayci et al., 2022; Xiao, 2022; Chen and Theja Maguluri, 2022; Alfano and Rebeschini, 2022; Yuan et al., 2023). In particular, for these coefficients to have finite upper bounds, it is important that ρ^t and v^t cover well the state and action spaces so that the upper bounds are independent to t. However, such upper bounds are very pessimistic. Indeed, when π^t and π^{t+1} converge to $\pi^\star$, one reasonable choice of (ρ^t, v^t) is to choose $(\rho^t, v^t) \in \{(d_\mu^\star, \pi^\star), (d_\mu^{t+1}, \pi^{t+1}), (d_\mu^\star, \pi^t), (d_\mu^{t+1}, \pi^t)\}$ such that C_1 and C_2 are closed to 1. We refer to Yuan et al. (2023, Section 5.2) for more discussions on the concentrability coefficients.

E. Neural network parametrization

In this appendix, we give the proof for Theorem 4.5, which is based on the following result by Ji et al. (2019, Theorem E.1). Let $g : \mathbb{R}^n \to \mathbb{R}$ be given and define the modulus of continuity ω_g as

$$\omega_g(\delta) := \sup_{s,s'\in\mathbb{R}^n} \{f(s) - f(s') : \max(\|s\|_2, \|s'\|_2) \leq 1 + \delta, \|s - s'\|_2 \leq \delta\}.$$

Theorem E.1 (Theorem E.1 in Ji et al. (2019))**.** *Let $g : \mathbb{R}^n \to \mathbb{R}$, $\delta > 0$ and $\omega_g(\delta)$ be as above and define for $s \in \mathbb{R}^n$*

$$M := \sup_{\|s\| \le 1+\delta} |g(s)|, \qquad g_{|\delta}(s) = f(s)\mathbf{1}[\|s\| \le 1+\delta], \qquad \alpha := \frac{\delta}{\sqrt{\delta} + \sqrt{2\log(2M/\omega_g(\delta))}}.$$

Let G_α be a gaussian distribution on $\mathbb{R}^n$ with mean 0 and variance $\alpha^2\mathbb{I}$. Define $l = g_{|\delta} * G_\alpha$ (gaussian convolution) with Fourier transform $\widehat{l}$ satisfying radial decomposition $\widehat{l}(w) = |\widehat{l}(w)| \exp(2\pi i \theta_h(w))$. Let P be a probability distribution supported on $\|s\| \le 1$. Additionally define

$$\begin{aligned}
c &:= g(0)g(0) \int |\widehat{l}(w)| \big[\cos(2\pi(\theta_l(w) - \|w\|_2)) - 2\pi \|w\|_2 \sin(2\pi(\theta_l(w) - \|w\|_2)) \big] dw \\
a &= \int w |\widehat{l}(w)| dw \\
r &= \sqrt{n} + 2\sqrt{\log \frac{24\pi^2(\sqrt{d}+7)^2 \left\|g_{|\delta}\right\|_{L_1}}{\omega_g(\delta)}} \\
p &:= 4\pi^2 |\widehat{l}(w)| \cos(2\pi(\|w\|_2 - b)) \mathbf{1}[|b| \le \|w\| \le r],
\end{aligned}$$

and for convenience create fake (weight, bias, sign) triples

$$(w,b,x)_{m+1} := (0, c, m\,\mathrm{sign}(c)), \quad (w,b,x)_{m+2} := (a, 0, m), \quad (w,b,x)_{m+3} := (-a, 0, -m).$$

Then

$$\begin{aligned}
|c| &\le M + 2\sqrt{n} \left\|g_{|\delta}\right\|_{L_1} (2\pi\alpha^2)^{-d/2}, \\
\|p\|_{L_1} &\le 2 \left\|g_{|\delta}\right\|_{L_1} \sqrt{\frac{(2\pi)^3 n}{(2\pi\alpha^2)^{n+1}}},
\end{aligned}$$

and with probability at least $1 - 3\lambda$ over a draw of $((s_j, w_j, b_j))_{j=1}^m$ from p (see on top how to sample from signed densities)

$$\left\| g - \frac{1}{m} \sum_{j=1}^{m+3} x_j^a \sigma(\langle w_j, s\rangle + b_j) \right\|_{L_2(P)} \le 3\omega_g(\delta) + \frac{r \|p\|_{L_1}}{\sqrt{m}} \left[1 + \sqrt{2\log(1/\lambda)} \right].$$

Theorem 4.5 is then obtained by setting $g = \widehat{Q}^t - \eta_t^{-1} \nabla h(\pi^t)$ and using a union bound over $\mathcal{A}$ and $(0, \dots, T-1)$.