On the Convergence Rates of Policy Gradient Methods

Lin Xiao　　　　　　　　　　　　　　　　　　　　　　　　　　　　LINX@FB.COM
Meta AI Research
Seattle, WA 98109, USA

Editor: Alekh Agarwal

Abstract

We consider infinite-horizon discounted Markov decision problems with finite state and action spaces and study the convergence rates of the projected policy gradient method and a general class of policy mirror descent methods, all with direct parametrization in the policy space. First, we develop a theory of weak gradient-mapping dominance and use it to prove sharp sublinear convergence rate of the projected policy gradient method. Then we show that with geometrically increasing step sizes, a general class of policy mirror descent methods, including the natural policy gradient method and a projected Q-descent method, all enjoy a linear rate of convergence without relying on entropy or other strongly convex regularization. Finally, we also analyze the convergence rate of an inexact policy mirror descent method and estimate its sample complexity under a simple generative model.

Keywords: discounted Markov decision problem, policy gradient, gradient domination, policy mirror descent, sample complexity.

1. Introduction

Markov decision process (MDP) is a fundamental model for sequential decision-making and serves as the mathematical framework for stochastic control and reinforcement learning. In this paper, we consider infinite-horizon, discounted Markov decision processes with finite state and action spaces. They are specified as a 5-tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is a finite state space with cardinality $|\mathcal{S}|$, $\mathcal{A}$ is a finite action space with cardinality $|\mathcal{A}|$, P is a transition probability function with $P(s'|s, a)$ denoting the probability of transitioning to s' when taking action a from state s, $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a reward function with $R_{s,a}$ or $R(s, a)$ being the (expected) reward of taking action a from state s, and finally $\gamma \in [0, 1)$ is a discount factor applied to the reward one-step in the future.

Starting from an initial state $s_0 \in \mathcal{S}$, an agent takes an action $a_t \in \mathcal{A}$ at each time step $t = 0, 1, 2, \ldots$, which leads to the next state s_{t+1} with probability $P(s_{t+1}|s_t, a_t)$ and obtains the immediate reward $r_t = R(s_t, a_t)$. Such interactions generate a *trajectory*

$$(s_0, a_0, r_0), \ (s_1, a_1, r_1), \ (s_2, a_2, r_2), \ \ldots.$$

The goal of the agent is to find a *policy* of choosing the actions $a_0, a_1, a_2, \ldots$ that maximizes the discounted cumulative reward $\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$. Here the expectation is taken with respect to the possible randomness in s_0, any randomness in choosing the actions a_t, and the randomness of state transitions prescribed by P.

In general, a policy that determines the action at time t may depends on the whole history of the trajectory up to time t. A *stationary policy* π specifies a decision rule that depends only on the current state. Specifically, we let $\pi_s \in \Delta(\mathcal{A})$ be the decision rule at state s, where $\Delta(\mathcal{A})$ denotes the probability simplex supported on $\mathcal{A}$, and $\pi_{s,a}$ denotes the probability of taking action a at state s. The value of a stationary policy $\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}$ starting from an arbitrary state s is defined as

$$V_s(\pi) := \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\,\Big|\, s_0 = s\right], \tag{1}$$

where the expectation is taken with respect to $a_t \sim \pi_{s_t}$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$ for all $t \geq 0$. We define $V : \Delta(\mathcal{A})^{|\mathcal{S}|} \to \mathbf{R}^{|\mathcal{S}|}$ as a vector-valued function with components $V_s(\pi)$. By the assumption that $R(s, a) \in [0, 1]$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we immediately have

$$0 \leq V_s(\pi) \leq \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma}, \qquad \forall\, s \in \mathcal{S}, \quad \forall\, \pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}. \tag{2}$$

The conventional formulation of Markov decision problems is about maximizing the discounted total reward. In this paper, we adopt a minimization formulation in order to better align with conventions in the optimization literature. To this end, we regard each $R(s, a) \in [0, 1]$ as a value measuring *regret* rather than reward. Given a reward matrix R, we can reset $R(s, a) \leftarrow 1 - R(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ to turn it into a regret matrix. Suppose $\rho \in \Delta(\mathcal{S})$ is an arbitrary initial state distribution. We consider the problem of minimizing

$$V_\rho(\pi) := \mathop{\mathbf{E}}_{s \sim \rho} V_s(\pi) = \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\,\Big|\, s_0 \sim \rho\right]. \tag{3}$$

For infinite-horizon MDPs with finite state and actions spaces, there exists a (deterministic) stationary policy $\pi^\star$ that is simultaneously optimal in minimizing $V_s(\cdot)$ for all $s \in \mathcal{S}$ (e.g., Puterman, 1994, Section 6.2.4). Such a solution is insensitive to the choice of ρ.

In this paper, we focus on *policy gradient methods* for minimizing the weighted value function V_ρ. These methods generate a sequence of policies $\{\pi^{(k)}\}$ through repeated evaluation of the policy gradient ∇V_μ, where $\mu \in \Delta(\mathcal{S})$ is not necessarily equal to ρ. The most straightforward variant is the projected policy gradient method,

$$\pi^{(k+1)} = \mathbf{proj}_\Pi\left(\pi^{(k)} - \eta_k \nabla V_\mu(\pi^{(k)})\right), \tag{4}$$

where η_k is the step size, $\Pi := \Delta(\mathcal{A})^{|\mathcal{S}|}$ is the set of feasible policies, and $\mathbf{proj}_\Pi(\cdot)$ denotes projection onto Π in the Euclidean norm. More generally, policy gradient methods can be derived from the mirror-descent form

$$\pi^{(k+1)} = \arg\min_{\pi \in \Pi}\left\{\eta_k\langle\nabla V_\mu(\pi^{(k)}), \pi\rangle + D_k(\pi, \pi^{(k)})\right\}, \tag{5}$$

where $D_k(\cdot, \cdot)$ is a distance-like function that may depend on $\pi^{(k)}$. For example, setting D_k as the squared Euclidean distance yields the projected policy gradient method (4). Shani et al. (2020) showed that by setting D_k as an appropriately weighted Kullback-Leibler (KL) divergence, one recovers the natural policy gradient (NPG) method of Kakade (2001). In general, we can think of (5) as a class of preconditioned policy gradient methods. The main results of this paper concern the convergence rates of such methods.

1.1 Previous Work

Many classical algorithms for MDPs are based on dynamical programming (Bellman, 1957), including value iteration, policy iteration, temporal difference learning and Q-learning (see, e.g., Puterman, 1994; Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 2018). Analyses of these methods in the tabular case mostly rely on the contraction property of the Bellman operator, which are difficult to extend with nonlinear function approximation and policy parametrization. In contrast, policy gradient methods (Williams, 1992; Sutton et al., 2000; Konda and Tsitsiklis, 2000; Kakade, 2001) aim to find a local minimum of an expected value function, thus are applicable to any differentiable policy parametrization and admit easy extensions to function approximation. In particular, they appear to work well when parametrized with modern deep neural networks (Schulman et al., 2015, 2017).

Despite the long history and empirical successes of policy gradient methods, their convergence properties are not well understood until recently. For example, it was widely accepted that they converge asymptotically to a stationary point or a local minimum because the objective function is nonconvex in general. However, Fazel et al. (2018) show that for linear quadratic control problems, policy gradient methods converge to the global optimal solution despite the nonconvex cost function, thanks to a gradient dominance property (Polyak, 1963). Agarwal et al. (2021) derive a variational gradient-dominance property and use it to obtain global convergence of the projected policy gradient method (4). Bhandari and Russo (2019) identify more general structural properties of policy gradient methods to ensure gradient domination and hence convergence to global optimum.

Using direct policy parametrization (over $\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}$), Agarwal et al. (2021) show that the projected policy gradient method (4) converges to a global optimum at an $O(1/\sqrt{k})$ sublinear rate. Specifically, the number of iterations to obtain $V_\rho(\pi^{(k)}) - V_\rho^\star \le \epsilon$ is

$$O\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^6\,\epsilon^2}\left\|\frac{d_\rho(\pi^\star)}{\mu}\right\|_\infty^2\right), \tag{6}$$

where $d_\rho(\pi^\star) \in \Delta(\mathcal{S})$ is a *discounted state-visitation distribution* and $\left\|d_\rho(\pi^\star)/\mu\right\|_\infty$ is a *distribution mismatch coefficient* (see Section 2.1 for definition and explanation). Zhang et al. (2020) develop a variational policy gradient framework and use it to show that the projected policy gradient method converges to global optimum at a faster $O(1/k)$ rate. In both cases, the constants in the iteration complexity are very large and depend on the Lipschitz constant characterizing the smoothness of the objective function.

Shani et al. (2020) show that the natural policy gradient (NPG) method (Kakade, 2001) can be cast as a special case of policy mirror descent method (5) and has an $O(1/\sqrt{k})$ convergence rate. Agarwal et al. (2021) improve the convergence rate of NPG to $O(1/k)$; more concretely, the number of iterations to obtain $V_\rho(\pi^{(k)}) - V_\rho^\star \le \epsilon$ is

$$\frac{2}{(1-\gamma)^2\epsilon}, \tag{7}$$

which is independent of the dimensions $|\mathcal{S}|$ and $|\mathcal{A}|$ or any distribution mismatch coefficient. Interestingly, the step sizes that guarantee such a rate can be chosen arbitrarily large, regardless of the Lipschitz constant of the policy gradient.

With entropy regularization, Cen et al. (2020) show that the NPG method has linear (geometric) convergence. Their approach relies on the contraction property of a generalized Bellman operator and the convergence guarantees are in terms of the infinity norm of the "soft" Q-functions. With appropriate choice of the regularization parameter and step size, they obtain iteration complexity on the order of

$$\frac{1}{1-\gamma} \log \frac{1}{(1-\gamma)\epsilon}. \tag{8}$$

Lan (2021) proposes a general policy mirror descent method that is similar to (5) with either convex or strongly convex regularizations. He focuses on the case of minimizing $V_{\rho^\star}$ where $\rho^\star$ is the stationary distribution of the MDP under the optimal policy $\pi^\star$, which avoids any distribution mismatch coefficient in the analysis. In order to guarantee $V_{\rho^\star}(\pi^{(k)}) - V_{\rho^\star}^\star \leq \epsilon$, Lan (2021) obtains iteration complexity on the orders of (7) and (8) for the settings without and with entropy regularization, respectively. Lan (2021) also obtains linear convergence for the un-regularized Markov decision problems using diminishing regularization combined with increasing step sizes (while maintaining a constant product of the two).

More recently, Zhan et al. (2021) extend the framework of Lan (2021) to accommodate a broader class of convex regularizers including those that are nonsmooth. For un-regularized MDPs, Khodadadian et al. (2021) show that the NPG method can obtain linear convergence with an adaptive step-size rule, and Bhandari and Russo (2021) show that several variants of policy gradient methods has linear convergence with exact line search.

For the exact policy gradient method with softmax parametrization, Agarwal et al. (2021) show that it converges asymptotically to a global optimum, and attains an $O(1/\sqrt{k})$ rate with log barrier regularization. Mei et al. (2020) derive an $O(1/k)$ convergence rate and Mei et al. (2021) further improve it to linear convergence by exploiting non-uniform variants of the smoothness and gradient dominance properties. However, these fast rates are associated with problem-dependent constants that can be very large (Li et al., 2021).

1.2 Contributions and Outline

In this paper, we present a systematic study of policy gradient methods with direct policy parametrization, focusing on their convergence rates for minimizing V_ρ over $\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}$.

Section 2 contains an overview of structural properties of discounted MDPs that are well-known but essential for the main results of the paper.

In Section 3, we develop a theory of *weak gradient-mapping domination* for general nonconvex composite optimization, and use it to obtain an $O(1/k)$ convergence rate for the projected policy gradient method. Concretely, our result on iteration complexity replaces ϵ^{-2} in (6) with ϵ^{-1} and $(1-\gamma)^{-6}$ with $(1-\gamma)^{-5}$. Although this result is the same as the one obtained by Zhang et al. (2020), our analysis are quite different. Zhang et al. (2020) exploit the bijection structure of the primal-dual MDP formulations, while we derive this result as a special case of nonconvex optimization with weak gradient-mapping domination which, to our best knowledge, is new and of independent interest.

In Section 4, we study exact policy mirror descent methods of the form (5). First, we show that with a constant step size (which can be arbitrarily large), they obtain the same dimension-free iteration complexity (7). This result extend the one of Agarwal et al. (2021) on NPG with KL-divergence to a general class of Bregman divergences, including a

projected Q-descent method derived with squared Euclidean distance. Second, we show that with geometrically increasing step sizes, as simple as $\eta_{k+1} = \eta_k/\gamma$, policy mirror descent methods enjoy linear convergence *without* relying on any regularization. Specifically, their iteration complexity for reaching $V_\rho(\pi^{(k)}) - V_\rho^\star \le \epsilon$ is

$$\frac{1}{1-\gamma} \left\| \frac{d_\rho(\pi^\star)}{\rho} \right\|_\infty \log \frac{2}{(1-\gamma)\epsilon}.$$

If ρ is set to be the stationary distribution under the optimal policy $\pi^\star$, then the distribution mismatch coefficient $\left\| d_\rho(\pi^\star)/\rho \right\|_\infty = 1$ and we recover (8). In addition, we discuss conditions for superlinear convergence and make connections with the classical Policy Iteration method.

In Section 5, we investigate the iteration complexity of inexact policy mirror descent methods and show that the geometrically increasing step sizes do not cause instability even with errors in evaluating the policy gradients or Q-functions. They converge with the same linear rate up to an asymptotic error floor. With a simple Q-estimator by repeated simulation of truncated trajectories, we obtain a sample complexity of

$$\widetilde{O} \left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^8 \, \epsilon^2} \left\| \frac{d_\rho(\pi^\star)}{\rho} \right\|_\infty^3 \right),$$

where the notation $\widetilde{O}(\cdot)$ hides poly-logarithmic factors of $|\mathcal{S}||\mathcal{A}|$, $1/(1-\gamma)$ and $1/\epsilon$.

Finally, in Section 6, we discuss the limitations of our work and possible extensions.

2. Preliminaries on Discounted MDPs

In this section, we overview the structural properties of infinite-horizon discounted MDPs that are essential for the developments in later sections. We start with a few definitions. Let $\Delta(\mathcal{S})$ denote the probability simplex defined over the state space $\mathcal{S}$, i.e.,

$$\Delta(\mathcal{S}) = \left\{ p \in \mathbf{R}^{|\mathcal{S}|} \mid \sum_{s \in \mathcal{S}} p_s = 1, \; p_s \ge 0 \text{ for all } s \in \mathcal{S} \right\}.$$

Similarly, $\Delta(\mathcal{A})$ denotes the probability simplex over the action space $\mathcal{A}$. The set of admissible policies is defined as

$$\Pi := \Delta(\mathcal{A})^{|\mathcal{S}|} = \left\{ \pi = \{\pi_s\}_{s \in \mathcal{S}} \mid \pi_s \in \Delta(\mathcal{A}) \text{ for all } s \in \mathcal{S} \right\}.$$

With slight abuse of notation, we define the following functions of $\pi \in \Pi$:

- $P : \Pi \to \mathbf{R}^{|\mathcal{S}| \times |\mathcal{S}|}$: a matrix function with entries $P_{s,s'}(\pi) = \sum_{a \in \mathcal{A}} \pi_{s,a} P(s'|s,a)$;

- $r : \Pi \to \mathbf{R}^{|\mathcal{S}|}$: a vector function with components $r_s(\pi) = \sum_{a \in \mathcal{A}} \pi_{s,a} R_{s,a}$.

Using the definitions above, the value function $V : \Pi \to \mathbf{R}^{|\mathcal{S}|}$, whose components V_s are defined in (1), admits the following analytic form (see, e.g., Puterman, 1994, Section 6.1)

$$V(\pi) = \sum_{t=0}^{\infty} \gamma^t P(\pi)^t \, r(\pi) = \left(I - \gamma P(\pi) \right)^{-1} r(\pi). \tag{9}$$

5

Since $P(\pi)$ is a row stochastic matrix and $0 \le \gamma < 1$, the spectral radius of $\gamma P(\pi)$ is strictly less than one (by the Perron-Frobenius theorem) and thus $I - \gamma P(\pi)$ is always invertible. Given $\rho \in \Delta(\mathcal{S})$, the weighted value function V_ρ defined in (3) can be written as

$$V_\rho(\pi) = \rho^T V(\pi) = \rho^T (I - \gamma P(\pi))^{-1} r(\pi). \tag{10}$$

Here we treat ρ as a column vector and use matrix multiplication conventions.

2.1 Distribution Mismatch Coefficient

Starting from $s \in \mathcal{S}$, the discounted state-visitation distribution under a policy π is a vector $d_s(\pi) \in \Delta(\mathcal{S})$ whose components are defined as

$$d_{s,s'}(\pi) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr^\pi \big(s_t = s' \mid s_0 = s \big), \qquad \forall \, s' \in \mathcal{S}. \tag{11}$$

The coefficient $1 - \gamma$ ensures that $\sum_{s' \in \mathcal{S}} d_{s,s'}(\pi) = 1$. In fact, $d_{s,s'}(\pi)$ is the (s, s') entry of the matrix $(1 - \gamma)(I - \gamma P(\pi))^{-1}$. In other words, if we define $e_s \in \mathbf{R}^{|\mathcal{S}|}$ with components $e_{s,s'} = 1$ if $s = s'$ and 0 otherwise, then we have

$$d_{s,s'}(\pi) = (1 - \gamma) e_s^T \big(I - \gamma P(\pi) \big)^{-1} e_{s'}, \qquad \forall \, s, s' \in \mathcal{S}. \tag{12}$$

Given an initial state distribution $\rho \in \Delta(\mathcal{S})$, we define $d_\rho(\pi) \in \Delta(\mathcal{S})$ with components

$$d_{\rho,s'}(\pi) = \mathop{\mathbf{E}}_{s \sim \rho} d_{s,s'}(\pi) = \sum_{s \in \mathcal{S}} \rho_s d_{s,s'}(\pi).$$

Some useful facts from the above definitions are:

$$d_{s,s}(\pi) \ge 1 - \gamma \qquad \text{and} \qquad d_{\rho,s}(\pi) \ge (1 - \gamma) \rho_s, \qquad \forall \, s \in \mathcal{S}. \tag{13}$$

For any $\rho, \mu \in \Delta(\mathcal{S})$, we define the distribution mismatch of ρ from μ as

$$\left\| \frac{\rho}{\mu} \right\|_\infty := \max_{s \in \mathcal{S}} \frac{\rho_s}{\mu_s},$$

with the convention $0/0 = 1$. This is an asymmetric measure of mismatch and it is finite if and only if the support (set of indices with nonzero entries) of μ contains that of ρ. If μ is the uniform distribution, then the mismatch is bounded by $|\mathcal{S}|$.

The convergence properties of policy gradient methods often depend on the distribution mismatch coefficients between two discounted state-visitation distributions (e.g., Kakade and Langford, 2002; Agarwal et al., 2021). According to (13), we have for any $\rho, \mu \in \Delta(\mathcal{S})$ and $\pi, \pi' \in \Pi$,

$$\left\| \frac{d_\rho(\pi)}{d_\mu(\pi')} \right\|_\infty \le \frac{1}{1 - \gamma} \left\| \frac{d_\rho(\pi)}{\mu} \right\|_\infty. \tag{14}$$

Our results in this paper mostly concern the case with $\rho = \mu$ and $\pi = \pi^\star$, i.e.,

$$C_\rho^\star := \left\| \frac{d_\rho(\pi^\star)}{\rho} \right\|_\infty.$$

6

In order for $C_\rho^\star$ to be finite, it suffices to assume $\rho > 0$, which means $\rho_s > 0$ for all $s \in \mathcal{S}$.

The distribution mismatch coefficient is closely related to the *concentrability coefficients* in the analysis of approximate dynamic programming algorithms (Munos, 2003, 2005; Szepesvári and Munos, 2008). In fact, $C_\rho^\star$ is considered the "best" one among all concentrability coefficients in the sense that it does not impose any restrictions on the MDP dynamics and it can be finite when other concentrability coefficients are infinite (Scherrer, 2014). See Agarwal et al. (2021, Section 2) for further discussions.

If ρ is chosen as the stationary distribution of the MDP under the optimal policy $\pi^\star$, denoted as $\rho^\star$, then we have $d_{\rho^\star}(\pi^\star) = \rho^\star$ and hence $C_{\rho^\star}^\star = 1$. This is the setting adopted by Liu et al. (2019) and Lan (2021), which leads to simplified analysis for minimizing $V_{\rho^\star}$. For discounted MDPs with entropy regularization (Lan, 2021; Cen et al., 2020), the resulting $\rho^\star$ always have full support over $\mathcal{S}$. However, in general $\rho^\star$ may not have full support over $\mathcal{S}$ unless the underlying MDP is *ergodic* (Puterman, 1994, Section A.2).

2.2 Q-functions, Policy Gradient and Performance Difference Lemma

For each pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, the state-action value function $Q_{s,a} : \Pi \to \mathbf{R}$ is defined as

$$Q_{s,a}(\pi) := \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \,\Big|\, s_0 = s, \, a_0 = a\right], \tag{15}$$

where the expectation is taken with respect to $a_t \sim \pi_{s_t}$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$ for all $t \geq 0$. It is straightforward to verify that

$$Q_{s,a}(\pi) = R_{s,a} + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) V_{s'}(\pi). \tag{16}$$

Let $Q_s(\pi) \in \mathbf{R}^{|\mathcal{A}|}$ denote the vector with components $Q_{s,a}(\pi)$ for all $a \in \mathcal{A}$. Then,

$$V_s(\pi) = \sum_{a \in \mathcal{A}} \pi_{s,a} Q_{s,a}(\pi) = \langle \pi_s, Q_s(\pi) \rangle, \tag{17}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors.

Policy gradients refer to the gradients of the value functions $V_s(\pi)$ and $V_\rho(\pi)$. We can obtain their expressions as special cases of the policy gradient theorem (Sutton et al., 2000) which covers the general case with policy parametrization. For easy reference, we give a simple, self-contained derivation in the Appendix (Section A.1). Specifically, we have

$$\nabla_s V_\rho(\pi) := \frac{\partial V_\rho(\pi)}{\partial \pi_s} = \frac{1}{1 - \gamma} d_{\rho,s}(\pi) Q_s(\pi), \tag{18}$$

and ∇V_ρ is the concatenation of $\nabla_s V_\rho$ for all $s \in \mathcal{S}$. In other words, policy gradients are weighted Q-functions where the weights are block-diagonal and proportional to the discounted state-visitation probabilities.

A fundamental result for analyzing discounted MDPs and related algorithms is the *performance difference lemma* of Kakade and Langford (2002). In this paper, we mostly rely on the following variant, which has appeared in Liu et al. (2019) and Lan (2021). For completeness, here we provide an alternative proof.

Lemma 1 (Performance difference lemma) *For any $\pi, \tilde{\pi} \in \Pi$, it holds that*

$$V_s(\pi) - V_s(\tilde{\pi}) = \frac{1}{1-\gamma} \mathop{\mathbf{E}}_{s' \sim d_s(\pi)} \langle Q_{s'}(\tilde{\pi}), \pi_{s'} - \tilde{\pi}_{s'} \rangle, \qquad \forall\, s \in \mathcal{S}. \qquad (19)$$

Proof Using the relation (17) on both π and $\tilde{\pi}$ and the definition of Q-functions, we obtain

$$\begin{aligned}
V_s(\pi) - V_s(\tilde{\pi}) &= \langle Q_s(\pi), \pi_s \rangle - \langle Q_s(\tilde{\pi}), \tilde{\pi}_s \rangle \\
&= \langle Q_s(\tilde{\pi}), \pi_s - \tilde{\pi}_s \rangle + \langle Q_s(\pi) - Q_s(\tilde{\pi}), \pi_s \rangle \\
&= \langle Q_s(\tilde{\pi}), \pi_s - \tilde{\pi}_s \rangle + \gamma \sum_{s \in \mathcal{S}} \pi_{s,a} \sum_{s' \in \mathcal{S}} P(s'|s,a)\big(V_{s'}(\pi) - V_{s'}(\tilde{\pi})\big), \qquad \forall\, s \in \mathcal{S}.
\end{aligned}$$

Define $u \in \mathbf{R}^{|\mathcal{S}|}$ with components $u_s = \langle Q_s(\tilde{\pi}), \pi_s - \tilde{\pi}_s \rangle$. Then the above result leads to

$$V(\pi) - V(\tilde{\pi}) = u + \gamma P(\pi)\big(V(\pi) - V(\tilde{\pi})\big),$$

which further implies

$$V(\pi) - V(\tilde{\pi}) = \big(I - \gamma P(\pi)\big)^{-1} u.$$

Using the expression of $d_{s,s'}(\pi)$ in (12), we write the above equality component-wise as

$$V_s(\pi) - V_s(\tilde{\pi}) = e_s^T \big(I - \gamma P(\pi)\big)^{-1} u = \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} d_{s,s'}(\pi)\, u_{s'},$$

which is the same as (19). ∎

The weighted version of the performance difference lemma is,

$$V_\rho(\pi) - V_\rho(\tilde{\pi}) = \frac{1}{1-\gamma} \mathop{\mathbf{E}}_{s' \sim d_\rho(\pi)} \langle Q_{s'}(\tilde{\pi}), \pi_{s'} - \tilde{\pi}_{s'} \rangle, \qquad \forall\, \pi, \tilde{\pi} \in \Pi. \qquad (20)$$

Considering the expression of policy gradient in (18), the above characterization resembles that of a linear function (precisely so if the expectation were taken with respect to $s' \sim d_\rho(\tilde{\pi})$ instead of $s' \sim d_\rho(\pi)$). The performance difference lemma is directly responsible for variational gradient domination (Agarwal et al., 2021, Lemma 4), which is a convexity-like property, and also for descent with arbitrarily large step sizes (see Section 4), which is a concavity-like property.

3. Projected Policy Gradient Method

In this section, we analyze the projected policy gradient method for solving the problem

$$\operatorname*{minimize}_{\pi \in \Pi} V_\rho(\pi), \qquad (21)$$

where $V_\rho(\pi)$ is defined in (3) or equivalently (10). We assume that the policy gradients are computed with respect to an initial state distribution μ, which may be different from the performance evaluation distribution ρ.

Starting from an initial policy $\pi^{(0)} \in \Pi$, the projected policy gradient method generates a sequence $\pi^{(k)}$ for $k = 1, 2, \ldots$ as follows:

$$\pi^{(k+1)} = \mathbf{proj}_\Pi \left(\pi^{(k)} - \eta_k \nabla V_\mu(\pi^{(k)}) \right), \tag{22}$$

where η_k is the step size and $\mathbf{proj}_\Pi(\cdot)$ denotes projection onto Π in Euclidean norm, i.e., $\mathbf{proj}_\Pi(\pi) = \arg\min_{\pi' \in \Pi} \|\pi' - \pi\|_2^2$. Since $\Pi = \Delta(\mathcal{A})^{|\mathcal{S}|}$ is a Cartesian product, the projections associated with different states can be done separately:

$$\pi_s^{(k+1)} = \mathbf{proj}_{\Delta(\mathcal{A})} \left(\pi_s^{(k)} - \eta_k \nabla_s V_\mu(\pi^{(k)}) \right), \qquad s \in \mathcal{S}, \tag{23}$$

where $\nabla_s V_\mu$ is given by (18) with ρ replaced by μ.

Agarwal et al. (2021, Theorem 5) show that with a constant step size $\eta_k = \frac{(1-\gamma)^3}{2\gamma|\mathcal{A}|}$ for all $k \geq 0$, the projected policy gradient method converges at an $O(1/\sqrt{k})$ rate. More precisely,

$$\min_{0 < k \leq K} \left\{ V_\rho(\pi^{(k)}) - V_\rho^\star \right\} \leq \epsilon \quad \text{whenever} \quad K > \frac{64\gamma|\mathcal{S}||\mathcal{A}|}{\epsilon^2 (1-\gamma)^6} \left\| \frac{d_\rho(\pi^\star)}{\mu} \right\|_\infty^2. \tag{24}$$

The following two ingredients are key to their analysis.

- **Smoothness** (Agarwal et al., 2021, Lemma 54): For any $\pi, \pi' \in \Pi$, it holds that

$$\left\| \nabla V_\rho(\pi) - \nabla V_\rho(\pi') \right\|_2 \leq \frac{2\gamma|\mathcal{A}|}{(1-\gamma)^3} \left\| \pi - \pi' \right\|_2. \tag{25}$$

- **Variational gradient domination** (Agarwal et al., 2021, Lemma 4): For any $\pi \in \Pi$,

$$V_\rho(\pi) - V_\rho^\star \leq \frac{1}{1-\gamma} \left\| \frac{d_\rho(\pi^\star)}{\mu} \right\|_\infty \max_{\pi' \in \Pi} \langle \nabla V_\mu(\pi), \pi - \pi' \rangle. \tag{26}$$

  Here "variational" refers to the term $\max_{\pi' \in \Pi} \langle \nabla V_\mu(\pi), \pi - \pi' \rangle$, which is different from $\|\nabla V_\mu(\pi)\|$ as in gradient dominance conditions for unconstrained optimization.

Based on the same two results above, we show that the projected policy gradient method enjoys a faster $O(1/k)$ convergence rate. The following theorem holds for the case $\rho = \mu$.

Theorem 2 *Suppose $\rho = \mu$ and the step size $\eta_k = \frac{(1-\gamma)^3}{2\gamma|\mathcal{A}|}$ for all $k \geq 0$. Then the projected policy gradient method (22) generates a sequence of policies $\pi^{(k)}$ satisfying*

$$V_\rho(\pi^{(k)}) - V_\rho^\star \leq \frac{256|\mathcal{S}||\mathcal{A}|}{k(1-\gamma)^5} \left\| \frac{d_\rho(\pi^\star)}{\rho} \right\|_\infty^2.$$

The general case with $\rho \neq \mu$ can be handled with an additional distribution mismatch coefficient. Concretely,

$$\begin{aligned}
V_\rho(\pi^{(k)}) - V_\rho^\star &= \sum_{s \in \mathcal{S}} \rho_s \left(V_s(\pi^{(k)}) - V_s^\star \right) \\
&= \sum_{s \in \mathcal{S}} \frac{\rho_s}{\mu_s} \mu_s \left(V_s(\pi^{(k)}) - V_s^\star \right) \\
&\leq \left\| \frac{\rho}{\mu} \right\|_\infty \left(V_\mu(\pi^{(k)}) - V_\mu^\star \right).
\end{aligned}$$

Then applying Theorem 2 with ρ replaced by μ yields

$$V_\rho(\pi^{(k)}) - V_\rho^\star \leq \frac{256|\mathcal{S}||\mathcal{A}|}{k(1-\gamma)^5} \left\|\frac{d_\mu(\pi^\star)}{\mu}\right\|_\infty^2 \left\|\frac{\rho}{\mu}\right\|_\infty.$$

Comparing with (24), in addition to improving the rate from $O(1/\sqrt{k})$ to $O(1/k)$, our bound also has better dependence on the discount factor: $1/(1-\gamma)^5$ as opposed to $1/(1-\gamma)^6$. However, our bound uses a different distribution mismatch coefficient and has an additional factor of $\|\rho/\mu\|_\infty$. We prove Theorem 2 as a special case of a more general result on *gradient-mapping domination*, which we present next.

3.1 An Interlude on Gradient-Mapping Domination

In this section, we consider the following composite optimization problem

$$\underset{x\in\mathbf{R}^n}{\text{minimize}} \ F(x) := f(x) + \Psi(x), \tag{27}$$

where f is smooth and Ψ is convex and lower semi-continuous (Rockafellar, 1970, Section 7). Specifically, we assume that there exists a constant $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2, \qquad \forall\, x, y \in \text{dom}\,\Psi. \tag{28}$$

The MDP formulation in (21) is a special case of (27) with the mappings $x \leftarrow \pi$, $f \leftarrow V_\rho$ and Ψ as the indicator function of Π, i.e., $\Psi(\pi) = 0$ if $\pi \in \Pi$ and ∞ otherwise.

For any convex function ϕ, the **prox** operator is defined as

$$\mathbf{prox}_\phi(x) = \arg\min_y \left\{\phi(y) + \frac{1}{2}\|y-x\|_2^2\right\}.$$

A generic algorithm for solving problem (27) is the *proximal gradient method*:

$$\begin{aligned}
x^{(k+1)} &= \arg\min_x \left\{\left\langle\nabla f(x^{(k)}), x-x^{(k)}\right\rangle + \frac{1}{2\eta_k}\|x-x^{(k)}\|^2 + \Psi(x)\right\} \\
&= \mathbf{prox}_{\eta_k\Psi}\left(x^{(k)} - \eta_k\nabla f(x^{(k)})\right),
\end{aligned} \tag{29}$$

where η_k is the step size. If Ψ is the indicator function of Π, then $\mathbf{prox}_{\eta_k\Psi}$ becomes the projection operator $\mathbf{proj}_\Pi$ for any $\eta_k > 0$. In this section, we focus on the proximal gradient method with the constant step size $\eta_k = 1/L$. To simplify presentation, we define

$$T_L(x) := \mathbf{prox}_{\frac{1}{L}\Psi}\left(x - \frac{1}{L}\nabla f(x)\right), \tag{30}$$

thus the proximal gradient method (29) can be written simply as $x^{(k+1)} = T_L(x^{(k)})$. The *gradient mapping* associated with problem (27) is defined as

$$G_L(x) := L\big(x - T_L(x)\big). \tag{31}$$

In the special case $\Psi \equiv 0$, we have $G_L(x) = \nabla f(x)$ for any $L > 0$. The norm of the gradient mapping, $\|G_L(x^{(k)})\|$, can serve as a measure of closeness to a first-order stationary point.

Key to the convergence analysis of the proximal gradient method is the following descent property (Nesterov, 2013, Theorem 1),

$$F(x^{(k)}) - F(x^{(k+1)}) \geq \frac{1}{2L} \big\| G_L(x^{(k)}) \big\|_2^2. \tag{32}$$

Summing up over $k = 0, 1, \ldots, K$, we obtain

$$F(x^{(0)}) - F(x^{(K+1)}) \geq \frac{1}{2L} \sum_{k=0}^{K} \big\| G_L(x^{(k)}) \big\|_2^2,$$

which, together with the fact $F(x^{(K+1)}) \geq F^\star$, implies (Beck, 2017, Theorem 10.15)

$$\min_{0 \leq k \leq K} \big\| G_L(x^{(k)}) \big\|_2 \leq \sqrt{\frac{2L \left(F(x^{(0)}) - F^\star \right)}{K + 1}}. \tag{33}$$

The $O(1/\sqrt{k})$ convergence rate stated in (24) is obtained by combining (33) with (26), which is the approach taken by Agarwal et al. (2021) and Bhandari and Russo (2019).

We show that under a *weak gradient mapping dominance condition*, which we define next, the proximal gradient method enjoys a much faster sublinear convergence rate.

Definition 3 (weak gradient-mapping domination) *Suppose $F := f + \Psi$ where f is L-smooth and Ψ is proper, convex and closed. We say that F satisfies a weak gradient-mapping dominance condition, of exponent $\alpha \in (1/2, 1]$, if there exists $\omega > 0$ such that*

$$\|G_L(x)\|_2 \geq \sqrt{2\omega} \left(F\big(T_L(x)\big) - F^\star \right)^\alpha, \qquad \forall\, x \in \operatorname{dom} \Psi, \tag{34}$$

where $F^\star = \min_x F(x)$ and T_L and G_L are defined in (30) and (31) respectively.

This *weak* version of gradient-mapping domination corresponds to the Kurdyka-Łojasiewicz (KŁ) condition with exponent $\alpha \in (1/2, 1]$ (Kurdyka, 1998). We note that classical KŁ theory mostly considers the range $\alpha \in [0, 1)$, where $\alpha = 0$ implies convergence in finite number of steps, $\alpha \in (0, 1/2]$ implies linear convergence, and $\alpha \in (1/2, 1)$ leads to sublinear convergence (e.g., Li and Pong, 2018; Cui and Pang, 2022, Section 8.4.2). We consider extensions of classical results to *gradient-mapping* domination and include the case $\alpha = 1$, which is what we need for the analysis of the policy gradient method. We discuss the case $\alpha = 1/2$, which corresponds to the strongly convex case, in Appendix A.2.

Theorem 4 *Consider the problem of minimizing $F := f + \Psi$ where f is L-smooth and Ψ is proper, convex and closed. Let $F^\star = \min_x F(x)$. Suppose F is weakly gradient-mapping dominant with exponent $\alpha \in (1/2, 1]$ and parameter ω. Then the proximal gradient method (29) with step size $\eta_k = 1/L$ generates a sequence $\{x^{(k)}\}$ that satisfies, for all $k \geq 0$,*

$$F(x^{(k)}) - F^\star \leq \max\left\{ \left(\frac{2^{2\alpha+1} L}{(2\alpha - 1)\omega} \right)^{\frac{1}{2\alpha-1}} \left(\frac{1}{k} \right)^{\frac{1}{2\alpha-1}}, \left(\frac{\sqrt{2}}{2} \right)^k \left(F(x^{(0)}) - F^\star \right) \right\}. \tag{35}$$

In particular, for $\alpha = 1$, we have

$$F(x^{(k)}) - F^\star \leq \max\left\{ \frac{8L}{\omega k}, \left(\frac{\sqrt{2}}{2} \right)^k \left(F(x^{(0)}) - F^\star \right) \right\}. \tag{36}$$

Proof Combining the descent property (32) with the inequality (34) yields

$$F(x^{(k)}) - F(x^{(k+1)}) \geq \frac{1}{2L} \left\| G_L(x^{(k)}) \right\|_2^2 \geq \frac{\omega}{L} \big(F(x^{(k+1)}) - F^\star \big)^{2\alpha}.$$

Let $\delta_k = F(x^{(k)}) - F^\star \geq 0$. Then we have

$$\delta_k - \delta_{k+1} \geq \frac{\omega}{L} \delta_{k+1}^{2\alpha}. \tag{37}$$

Given that $\alpha \in (1/2, 1]$, we have $2\alpha - 1 \in (0, 1]$. Expanding a telescoping sum, we get

$$\frac{1}{\delta_k^{2\alpha-1}} - \frac{1}{\delta_0^{2\alpha-1}} = \sum_{i=0}^{k-1} \left(\frac{1}{\delta_{i+1}^{2\alpha-1}} - \frac{1}{\delta_i^{2\alpha-1}} \right) = \sum_{i=0}^{k-1} \frac{2\alpha-1}{\delta_{i+1}^{2\alpha-1}} \frac{1}{2\alpha-1} \left(1 - \left(\frac{\delta_{i+1}}{\delta_i} \right)^{2\alpha-1} \right)$$

Notice that due to the descent property, we always have $\delta_{i+1} \leq \delta_i$ and thus $\delta_{i+1}/\delta_i \leq 1$. Mei et al. (2021, Lemma 13) proved that for any $\beta > 0$ and $\rho \in [0, 1]$, it holds that

$$\frac{1}{\beta}(1 - \rho^\beta) \geq \rho^\beta (1 - \rho).$$

Apply this result with $\beta = 2\alpha - 1$ and $\rho = \delta_{i+1}/\delta_i$, we obtain

$$\frac{1}{\delta_k^{2\alpha-1}} - \frac{1}{\delta_0^{2\alpha-1}} \geq \sum_{i=0}^{k-1} \frac{2\alpha-1}{\delta_{t+1}^{2\alpha-1}} \left(\frac{\delta_{i+1}}{\delta_i} \right)^{2\alpha-1} \left(1 - \frac{\delta_{i+1}}{\delta_i} \right) = (2\alpha-1) \sum_{i=0}^{k-1} \frac{\delta_i - \delta_{i+1}}{\delta_i^{2\alpha}}.$$

Now invoking (37) leads to

$$\frac{1}{\delta_k^{2\alpha-1}} - \frac{1}{\delta_0^{2\alpha-1}} \geq (2\alpha-1) \frac{\omega}{L} \sum_{i=0}^{k-1} \left(\frac{\delta_{i+1}}{\delta_i} \right)^{2\alpha}.$$

For any two constant $r, c \in (0, 1)$, let's define $n(k, r)$ to be the number of times that the ratio δ_{i+1}/δ_i is at least r among the first k iterations. If $n(k, r) \geq ck$, then $\delta_{i+1}/\delta_i \geq r$ at least $\lceil ck \rceil$ times, thus

$$\frac{1}{\delta_k^{2\alpha-1}} - \frac{1}{\delta_0^{2\alpha-1}} \geq (2\alpha-1) \frac{\omega}{L} r^{2\alpha} ck,$$

which implies

$$\delta_k \leq \left(\frac{1}{1/\delta_0^{2\alpha-1} + (2\alpha-1)\frac{\omega}{L} r^{2\alpha} ck} \right)^{\frac{1}{2\alpha-1}} \leq \left(\frac{L}{(2\alpha-1)\omega r^{2\alpha} c} \right)^{\frac{1}{2\alpha-1}} \left(\frac{1}{k} \right)^{\frac{1}{2\alpha-1}}$$

Otherwise, we must have $n(k, r) < ck$, which means that $\delta_{i+1}/\delta_i < r$ at least $\lceil (1-c)k \rceil$ times. Noticing that $\delta_{i+1}/\delta_i \leq 1$ for all i, we arrive at

$$\delta_k \leq \delta_0 r^{(1-c)k} = \delta_0 \left(r^{1-c} \right)^k.$$

Combining the above two cases and using the fact that $r, c \in (0, 1)$ can be chosen arbitrarily, we conclude that

$$\delta_k \leq \min_{0 < r, c < 1} \max \left\{ \left(\frac{L}{(2\alpha-1)\omega r^{2\alpha} c} \right)^{\frac{1}{2\alpha-1}} \left(\frac{1}{k} \right)^{\frac{1}{2\alpha-1}}, \ \left(r^{1-c} \right)^k \delta_0 \right\}.$$

Setting $r = c = 1/2$ gives the desired result (35), and further letting $\alpha = 1$ yields (36). ∎

3.2 Proof of Theorem 2

In order to prove Theorem 2, we only need to verify that the weak gradient-mapping domination holds for the weighted value function V_ρ. This is the result of the next lemma.

Lemma 5 *Consider the problem of minimizing V_ρ over Π and suppose that V_ρ is L-smooth. We have*

$$V_\rho\big(T_L(\pi)\big) - V_\rho^\star \leq \frac{2\sqrt{2|\mathcal{S}|}}{1-\gamma} \left\| \frac{d_\rho(\pi^\star)}{\mu} \right\|_\infty \|G_L(\pi)\|_2 , \tag{38}$$

where

$$T_L(\pi) := \mathbf{proj}_\Pi \left(\pi - \frac{1}{L}\nabla V_\mu(\pi) \right), \qquad G_L(\pi) := L\big(\pi - T_L(\pi)\big).$$

Proof Applying a result of Nesterov (2013, Theorem 1) to our setting yields

$$\big\langle \nabla V_\mu(T_L(\pi)), T_L(\pi) - \pi' \big\rangle \leq 2\|G_L(\pi)\|_2 \cdot \|T_L(\pi) - \pi'\|_2, \qquad \forall\, \pi, \pi' \in \Pi.$$

Using the facts $T_L(\pi) \in \Pi$ and $\|\pi'' - \pi'\|_2 \leq \sqrt{2|\mathcal{S}|}$ for any $\pi'', \pi' \in \Pi$, we obtain

$$\max_{\pi' \in \Pi} \big\langle \nabla V_\mu(T_L(\pi)), T_L(\pi) - \pi' \big\rangle \leq 2\sqrt{2|\mathcal{S}|}\, \|G_L(\pi)\|_2.$$

Combining the above bound with (26) yields the desired result. ∎

An argument equivalent to Lemma 5 was used by Agarwal et al. (2021) which relies on a result of Ghadimi and Lan (2016). Combining (38) with (33) gives the result in (24). On the other hand, we recognize that with $\rho = \mu$, inequality (38) implies (34) with

$$\alpha = 1, \qquad \omega = \frac{(1-\gamma)^2}{16|\mathcal{S}|} \left\| \frac{d_\rho(\pi^\star)}{\rho} \right\|_\infty^{-2}, \qquad L = \frac{2\gamma|\mathcal{A}|}{(1-\gamma)^3}.$$

Now we can apply Theorem 4. Notice that in this case, the exponential decay part in (35) is always smaller than the sublinear part, which leads to $V_\rho(\pi^{(k)}) - V_\rho^\star \leq 8L/(\omega k)$, i.e.,

$$V_\rho(\pi^{(k)}) - V_\rho^\star \leq \frac{256|\mathcal{S}||\mathcal{A}|}{k(1-\gamma)^5} \left\| \frac{d_\rho(\pi^\star)}{\rho} \right\|_\infty^2.$$

This finishes the proof of Theorem 2.

Zhang et al. (2020, Theorem 5) have also established the $O(1/k)$ rate of the projected policy gradient method with direct parametrization. However, their proof appears to be quite different from ours, which leverages the dual linear programming parametrization (Puterman, 1994, Section 6.9). In contrast, our approach is based on a novel notion of weak gradient-mapping domination and applies to general nonconvex composite optimization problems.

4. Exact Policy Mirror Descent Methods

Mirror descent (Nemirovski and Yudin, 1983) is a general framework for the construction and analysis of optimization algorithms, which covers the projected gradient method as a special case. Here we adopt the form of mirror descent based on proximal minimization with respect to a Bregman divergence (Beck and Teboulle, 2003).

Let $h : \Delta(\mathcal{A}) \to \mathbf{R}$ be a strictly convex function and continuously differentiable on the relative interior of $\Delta(\mathcal{A})$, denoted as rint $\Delta(\mathcal{A})$. The Bregman divergence generated by h is a distance-like function defined for any $p \in \Delta(\mathcal{A})$ and $p' \in \text{rint}\,\Delta(\mathcal{A})$ as

$$D(p, p') := h(p) - h(p') - \langle \nabla h(p'), p - p' \rangle.$$

Two most popular examples of Bregman divergence are:

- Squared Euclidean distance, generated by the squared 2-norm:

$$h(p) = (1/2)\|p\|_2^2, \qquad D(p, p') = (1/2)\|p - p'\|_2^2.$$

- Kullback-Leibler (KL) divergence, generated by the negative entropy:

$$h(p) = \sum_{a \in \mathcal{A}} p_a \log p_a, \qquad D(p, p') = \sum_{a \in \mathcal{A}} p_a \log \frac{p_a}{p'_a}.$$

  Notice that the gradient of negative entropy vanishes on the boundary of the simplex. Therefore we need to restrict the second argument p' to lie within the relative interior of $\Delta(\mathcal{A})$. We shall address such subtleties later in the convergence analysis.

Recall that the set of feasible policies is $\Pi = \Delta(\mathcal{A})^{|\mathcal{S}|}$, which is a Cartesian product of $|\mathcal{S}|$ copies of $\Delta(\mathcal{A})$. For any $\rho \in \Delta(\mathcal{S})$, we define a weighted divergence function

$$D_\rho(\pi, \pi') := \mathop{\mathbf{E}}_{s \sim \rho} \big[D(\pi_s, \pi'_s) \big] = \sum_{s \in \mathcal{S}} \rho_s D(\pi_s, \pi'_s).$$

This function satisfies the basic properties of a Bregman divergence; in particular, it is nonnegative and for $\rho > 0$ it equals to 0 if and only if $\pi = \pi'$.

Following the derivations of Shani et al. (2020), we consider *policy mirror descent* (PMD) methods with dynamically weighted divergences:

$$\pi^{(k+1)} = \arg\min_{\pi \in \Pi} \left\{ \eta_k \big\langle \nabla V_\mu(\pi^{(k)}), \pi \big\rangle + \frac{1}{1-\gamma} D_{d_\mu(\pi^{(k)})}(\pi, \pi^{(k)}) \right\},$$

where η_k is the step size, $\mu \in \Delta(\mathcal{S})$ is an arbitrary state distribution and $d_\mu(\pi^{(k)})$ is the discounted state-visitation distribution under the policy $\pi^{(k)}$. Using the fact

$$\big\langle \nabla V_\mu(\pi^{(k)}), \pi \big\rangle = \sum_{s \in \mathcal{S}} \big\langle \nabla_s V_\mu(\pi^{(k)}), \pi_s \big\rangle,$$

and plugging in the policy gradient formula (18), we obtain

$$\pi^{(k+1)} = \arg\min_{\pi \in \Pi} \left\{ \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} d_{\mu,s}(\pi^{(k)}) \Big(\eta_k \big\langle Q_s(\pi^{(k)}), \pi_s \big\rangle + D(\pi_s, \pi_s^{(k)}) \Big) \right\}$$

$$= \arg\min_{\pi \in \Pi} \left\{ \sum_{s \in \mathcal{S}} \Big(\eta_k \big\langle Q_s(\pi^{(k)}), \pi_s \big\rangle + D(\pi_s, \pi_s^{(k)}) \Big) \right\},$$

where the second equality, strictly speaking, requires $d_{\mu,s}(\pi^{(k)}) > 0$ for all $s \in \mathcal{S}$. The above update rule can be written separately for each state as

$$\pi_s^{(k+1)} = \underset{p \in \Delta(\mathcal{A})}{\arg\min} \left\{ \eta_k \langle Q_s(\pi^{(k)}), p \rangle + D(p, \pi_s^{(k)}) \right\}, \qquad \forall\, s \in \mathcal{S}. \tag{39}$$

Clearly, this update rule is independent of the choice of μ, which is the result of *adaptive preconditioning* with a dynamically weighted divergence: the weight for each state in $D_{d_\mu(\pi^{(k)})}$ matches the coefficient of $Q_s(\pi^{(k)})$ in the policy gradient $\nabla V_\mu(\pi^{(k)})$.

For the two prominent examples of Bregman divergence listed before, the corresponding PMD methods have closed-form update rules:

- **Projected Q-descent.** If $D(\cdot, \cdot)$ is the squared Euclidean distance, then (39) becomes

$$\pi_s^{(k+1)} = \mathbf{proj}_{\Delta(\mathcal{A})} \left(\pi_s^{(k)} - \eta_k Q_s(\pi^{(k)}) \right), \qquad \forall\, s \in \mathcal{S}. \tag{40}$$

  Compared with the projected policy gradient method (23), we replaced the policy gradient $\nabla_s V_\mu(\pi^{(k)})$ by $Q_s(\pi^{(k)})$ as the result of adaptive preconditioning.

- **Exponentiated Q-descent.** If $D(\cdot, \cdot)$ is the KL-divergence, then (39) takes the form

$$\pi_{s,a}^{(k+1)} = \pi_{s,a}^{(k)} \frac{\exp\left(-\eta_k Q_{s,a}(\pi^{(k)})\right)}{z_s^{(k)}}, \qquad \forall\, a \in \mathcal{A},\ s \in \mathcal{S}, \tag{41}$$

  where

$$z_s^{(k)} = \sum_{a \in \mathcal{A}} \exp\left(-\eta_k Q_{s,a}(\pi^{(k)})\right), \qquad \forall\, s \in \mathcal{S}.$$

  This is exactly the Natural Policy Gradient (NPG) method (Kakade, 2001) expressed in the policy space (Agarwal et al., 2021). For the derivation of (41), see, for example, Beck (2017, Section 9.1).

In the rest of this section, we investigate the convergence rate of the PMD method (39). We show that with a constant step size, it has $O(1/k)$ convergence rate. When the step size increases exponentially as $\eta_k = \eta_0/\gamma^k$, we have linear convergence and the convergence rate depends on the distribution mismatch coefficient $\|d_\rho(\pi^\star)/\rho\|_\infty$. In addition, we discuss situations of super-linear convergence and connections to policy iteration.

Our results hold for PMD methods constructed with general Bregman divergences, matching or improving over the best known convergence rates. In particular, the projected Q-descent method has the same rate of convergence as NPG. We show that the key ingredient for fast convergence of the PMD method is the adaptive preconditioning using weighted divergence functions. The adopted local Bregman divergence, being KL-divergence or squared Euclidean distance, does not make much difference.

4.1 Sublinear Convergence

Our analysis is based on two key ingredients: the performance difference lemma (Lemma 1) and a three-point descent lemma on proximal optimization with Bregman divergences.

In order to cover both the squared Euclidean distance and KL-divergence without loss of rigor, we need some technical conditions. Specifically, we say a function h is of Legendre

type (Rockafellar, 1970, Section 26) if it is essentially smooth and strictly convex in the relative interior of $\operatorname{dom} h$, denoted as $\operatorname{rint} \operatorname{dom} h$. Essential smoothness means that h is differentiable and $\|\nabla h(x_k)\| \to \infty$ for every sequence $\{x_k\}$ converging to a boundary point of $\operatorname{dom} h$. The following result is a slight variation of Chen and Teboulle (1993, Lemma 3.2), where we replaced the original assumption of h being a *Bregman function* with h being of Legendre type. The proof essentially follows the same arguments and thus is omitted here.

Lemma 6 (Three-point descent lemma) *Suppose that $\mathcal{C} \subset \mathbf{R}^n$ is a closed convex set, $\phi : \mathcal{C} \to \mathbf{R}$ is a proper, closed convex function, $D(\cdot, \cdot)$ is the Bregman divergence generated by a function h of Legendre type and $\operatorname{rint} \operatorname{dom} h \cap \mathcal{C} \neq \varnothing$. For any $x \in \operatorname{rint} \operatorname{dom} h$, let*

$$x^+ = \arg\min_{u \in \mathcal{C}} \{ \phi(u) + D(u, x) \}.$$

Then $x^+ \in \operatorname{rint} \operatorname{dom} h \cap \mathcal{C}$ and for any $u \in \mathcal{C}$,

$$\phi(x^+) + D(x^+, x) \leq \phi(u) + D(u, x) - D(u, x^+).$$

In the context of the PMD method (39), $\mathcal{C} = \Delta(\mathcal{A})$ and ϕ is the linear function $\eta_k \langle Q_s(\pi^{(k)}), \cdot \rangle$. There are some subtle differences between the two Bregman divergences we consider, as explained below.

- For the squared Euclidean distance, $h(\cdot) = (1/2)\| \cdot \|_2^2$ is of Legendre type with $\operatorname{rint} \operatorname{dom} h = \mathbf{R}^{|\mathcal{A}|}$ and thus $\operatorname{rint} \operatorname{dom} h \cap \mathcal{C} = \Delta(\mathcal{A})$. Therefore each iterate generated by the PMD method, specifically (40), can be on the boundary of $\Delta(\mathcal{A})$.

- For the KL divergence, h is the negative entropy function, which is also of Legendre type, but with $\operatorname{rint} \operatorname{dom} h \cap \mathcal{C} = \operatorname{rint} \operatorname{dom} h = \operatorname{rint} \Delta(\mathcal{A})$. Therefore, if we start with an initial point in $\operatorname{rint} \Delta(\mathcal{A})$, then every iterates will stay in $\operatorname{rint} \Delta(\mathcal{A})$.

We first use Lemma 6 to prove a descent property of PMD. This result is elementary and has appeared in various forms before (e.g., Liu et al., 2019; Lan, 2021). We present the proof for completeness as we will need to refer to some intermediate steps in it later.

Lemma 7 (Descent property of PMD) *Suppose the initial point $\pi^{(0)} \in \operatorname{rint} \Pi$. Then the sequences generated by the PMD method (39) satisfy*

$$\left\langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \leq 0, \qquad \forall\, s \in \mathcal{S}, \tag{42}$$

and for any $\rho \in \Delta(s)$,

$$V_\rho(\pi^{(k+1)}) \leq V_\rho(\pi^{(k)}), \qquad \forall\, k \geq 0. \tag{43}$$

Proof Applying Lemma 6 to the update rule (39) with $\mathcal{C} = \Delta(\mathcal{A})$ and $\phi(\cdot) = \eta_k \langle Q_s(\pi^k), \cdot \rangle$, we obtain that for any $p \in \Delta(\mathcal{A})$,

$$\eta_k \left\langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} \right\rangle + D(\pi_s^{(k+1)}, \pi_s^{(k)}) \leq \eta_k \left\langle Q_s(\pi^{(k)}), p \right\rangle + D(p, \pi_s^{(k)}) - D(p, \pi_s^{(k+1)}).$$

Rearranging terms and dividing both sides by η_k, we get

$$\langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - p \rangle + \frac{1}{\eta_k} D(\pi_s^{(k+1)}, \pi_s^{(k)}) \leq \frac{1}{\eta_k} D(p, \pi_s^{(k)}) - \frac{1}{\eta_k} D(p, \pi_s^{(k+1)}). \tag{44}$$

Letting $p = \pi_s^{(k)}$ in (44) yields

$$\langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \rangle \leq -\frac{1}{\eta_k} D(\pi_s^{(k+1)}, \pi_s^{(k)}) - \frac{1}{\eta_k} D(\pi_s^{(k)}, \pi_s^{(k+1)}),$$

which implies (42) since the Bregman divergence $D(\cdot, \cdot)$ is always nonnegative. By the performance difference lemma, specifically the weighted version (20), we have

$$V_\rho(\pi^{(k+1)}) - V_\rho(\pi^{(k)}) = \frac{1}{1 - \gamma} \mathop{\mathbf{E}}_{s \sim d_\rho(\pi^{(k+1)})} \langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \rangle \leq 0,$$

which is the same as (43). ∎

The next result is a generalization of the $O(1/k)$ convergence rate of the NPG method obtained by Agarwal et al. (2021, Theorem 16), where they focused on the setting of KL-divergence and their proof also relies on specific properties of the KL-divergence. Here we extend it to more general Bregman divergence. Lan (2021, Theorem 2) derived a similar result using techniques that work for general Bregman divergence. However, he worked with the special objective function $V_{\rho^\star}$ where $\rho^\star$ is the stationary distribution of the optimal policy $\pi^\star$. As a result, the proof of Lan (2021, Theorem 2) avoids some subtle arguments required for the more general objective function V_ρ where $\rho \in \Delta(\mathcal{S})$ can be arbitrary.

In order to simplify presentation, we use the following notation throughout this paper:

$$D_k^\star := D_{d_\rho(\pi^\star)}(\pi^\star, \pi^{(k)}) = \sum_{s \in \mathcal{S}} d_{\rho,s}(\pi^\star) D(\pi_s^\star, \pi_s^{(k)}), \tag{45}$$

where $d_\rho(\pi^\star)$ is the state-visitation distribution under $\pi^\star$ with initial state distribution ρ. Although ρ does not appear in the notation $D_k^\star$, we hope it is clear from the context.

Theorem 8 *Consider the policy mirror descent method (39) with $\pi^{(0)} \in \text{rint } \Pi$ and constant step size $\eta_k = \eta$ for all $k \geq 0$. For any $\rho \in \Delta(\mathcal{S})$, we have for all $k \geq 0$,*

$$V_\rho(\pi^{(k)}) - V_\rho^\star \leq \frac{1}{k+1} \left(\frac{D_0^\star}{\eta(1 - \gamma)} + \frac{1}{(1 - \gamma)^2} \right).$$

Proof Consider the inequality (44), we let $p = \pi_s^\star$ and subtract and add $\pi_s^{(k)}$ within the inner product term, which leads to

$$\langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \rangle + \langle Q_s(\pi^{(k)}), \pi_s^{(k)} - \pi_s^\star \rangle \leq \frac{1}{\eta_k} D(\pi_s^\star, \pi_s^{(k)}) - \frac{1}{\eta_k} D(\pi_s^\star, \pi_s^{(k+1)}).$$

Notice that we dropped the nonnegative term $(1/\eta_k) D(\pi_s^{(k+1)}, \pi_s^{(k)})$ on the left side of the inequality. Taking expectation with respect to the distribution $d_\rho(\pi^\star)$ on both sides of the above inequality and using the notation in (45), we obtain

$$\mathop{\mathbf{E}}_{s \sim d_\rho(\pi^\star)} \left[\langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \rangle \right] + \mathop{\mathbf{E}}_{s \sim d_\rho(\pi^\star)} \left[\langle Q_s(\pi^{(k)}), \pi_s^{(k)} - \pi_s^\star \rangle \right] \leq \frac{1}{\eta_k} D_k^\star - \frac{1}{\eta_k} D_{k+1}^\star. \tag{46}$$

For the first expectation in (46), we have

$$\underset{s\sim d_\rho(\pi^\star)}{\mathbf{E}}\left[\langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)}\rangle\right] = \sum_{s\in\mathcal{S}} d_{\rho,s}(\pi^\star)\langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)}\rangle$$

$$\geq \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}} d_{d_\rho(\pi^\star),s}(\pi^{(k+1)})\langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)}\rangle$$

$$= V_{d_\rho(\pi^\star)}(\pi^{(k+1)}) - V_{d_\rho(\pi^\star)}(\pi^{(k)}), \tag{47}$$

where the inequality holds because of (42) and the fact, due to (13), that

$$d_{d_\rho(\pi^\star),s}(\pi^{(k+1)}) \geq (1-\gamma)d_{\rho,s}(\pi^\star), \qquad \forall s \in \Delta(\mathcal{S}).$$

The last equality in (47) is due to the performance difference lemma. For the second expectation in (46), we again use the performance difference lemma to obtain

$$\underset{s\sim d_\rho(\pi^\star)}{\mathbf{E}}\left[\langle Q_s(\pi^{(k)}), \pi_s^{(k)} - \pi_s^\star\rangle\right] = (1-\gamma)\big(V_\rho(\pi^{(k)}) - V_\rho(\pi^\star)\big). \tag{48}$$

Substituting the two results above into (46) leads to

$$(1-\gamma)\left(V_\rho(\pi^{(k)}) - V_\rho(\pi^\star)\right) \leq \frac{1}{\eta_k}D_k^\star - \frac{1}{\eta_k}D_{k+1}^\star + V_{d_\rho(\pi^\star)}(\pi^{(k)}) - V_{d_\rho(\pi^\star)}(\pi^{(k+1)}).$$

Setting $\eta_k = \eta$ for all $k \geq 0$ and summing up over k:

$$(1-\gamma)\sum_{i=0}^{k}\left(V_\rho(\pi^{(i)}) - V_\rho(\pi^\star)\right) \leq \frac{1}{\eta}D_0^\star - \frac{1}{\eta}D_{k+1}^\star + V_{d_\rho(\pi^\star)}(\pi^{(0)}) - V_{d_\rho(\pi^\star)}(\pi^{(k+1)})$$

$$\leq \frac{1}{\eta}D_0^\star + V_{d_\rho(\pi^\star)}(\pi^{(0)}).$$

Since $V_\rho(\pi^{(k)})$ is monotone non-increasing in k (see Lemma 7), we conclude that

$$V_\rho(\pi^{(k)}) - V_\rho^\star \leq \frac{1}{k+1}\sum_{i=0}^{k}\left(V_\rho(\pi^{(i)}) - V_\rho(\pi^\star)\right) \leq \frac{1}{k+1}\left(\frac{D_0^\star}{\eta(1-\gamma)} + \frac{V_{d_\rho(\pi^\star)}(\pi^{(0)})}{1-\gamma}\right),$$

Finally, bounding $V_{d_\rho(\pi^\star)}(\pi^{(0)})$ by $1/(1-\gamma)$ as in (2) gives the desired result. $\blacksquare$

As a result of Theorem 8, whenever $\eta \geq (1-\gamma)D_0^\star = (1-\gamma)D_{d_\rho(\pi^\star)}(\pi^\star, \pi^{(0)})$, we have

$$V_\rho(\pi^{(k)}) - V_\rho^\star \leq \frac{2}{(k+1)(1-\gamma)^2}, \qquad \forall k \geq 0. \tag{49}$$

In other words, the number of iterations to reach $V_\rho(\pi^{(k)}) - V_\rho^\star \leq \epsilon$ is at most

$$\frac{2}{(1-\gamma)^2\epsilon},$$

which is independent of the problem dimensions $|\mathcal{S}|$ and $|\mathcal{A}|$. More specifically,

- For the projected Q-descent method (40), since $D(\pi_s, \pi'_s) = (1/2)\|\pi_s - \pi'_s\|^2 \leq 1$ for any $\pi_s, \pi'_s \in \Delta(\mathcal{A})$, we have $D_\rho(\pi, \pi') = \sum_{s \in \mathcal{S}} \rho_s D(\pi_s, \pi'_s) \leq 1$ for any $\rho \in \Delta(\mathcal{S})$. Therefore in order for (49) to hold, it suffices to have $\eta \geq (1 - \gamma)$.

- For the exponentiated Q-descent method (41), if we choose the uniform initial policy, i.e., $\pi_{s,a}^{(0)} = 1/|\mathcal{A}|$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, then $D_\rho(\pi^\star, \pi^{(0)}) \leq \log|\mathcal{A}|$ for all $\rho \in \Delta(\mathcal{S})$. Therefore in order for (49) to hold, it suffices to have $\eta \geq (1 - \gamma)\log|\mathcal{A}|$.

The above analysis indicates that, compared with the exponentiated update variant (NPG), the projected Q-descent method has a wider range of η to enjoy the same dimensional independent convergence guarantee (49).

A more curious fact is that for both variants, the step size η does not have an upper bound and can be as large as possible. This is in contrast to the classical analysis of smooth optimization, where the step size is usually upper bounded by $2/L$ with L being the Lipschitz constant of the gradient; see, e.g., the approach taken in Section 3.1. Here the fact the step sizes can be arbitrarily large is due to the unique structure of discounted MDPs. Indeed, we show next that PMD has linear convergence if the step size grows exponentially.

4.2 Linear Convergence

Consider again the policy mirror descent algorithm (39). In order to simplify the presentation, we define two more notations: the optimality gap

$$\delta_k := V_\rho(\pi^{(k)}) - V_\rho(\pi^\star), \tag{50}$$

which is always nonnegative, and the per-iteration distribution mismatch coefficient

$$\vartheta_k := \left\| \frac{d_\rho(\pi^\star)}{d_\rho(\pi^{(k)})} \right\|_\infty. \tag{51}$$

The following result is the basis for establishing the linear convergence and also for discussions on possible superlinear convergence.

Proposition 9 *Consider the policy mirror descent method* (39) *with* $\pi^{(0)} \in \mathrm{rint}\,\Pi$ *and* $\eta_k > 0$ *for all* $k \geq 0$. *Then for any* $\rho \in \Delta(\mathcal{S})$, *we have for all* $k \geq 0$,

$$\vartheta_{k+1}(\delta_{k+1} - \delta_k) + \delta_k \leq \frac{1}{(1 - \gamma)\eta_k} D_k^\star - \frac{1}{(1 - \gamma)\eta_k} D_{k+1}^\star, \tag{52}$$

where δ_k, ϑ_k *and* $D_k^\star$ *are defined in* (50), (51) *and* (45), *respectively*.

Proof We start with the inequality (46) and bound the first expectation as follows:

$$\begin{aligned}
\mathop{\mathbf{E}}_{s \sim d_\rho(\pi^\star)} \left[\langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \rangle \right] &= \sum_{s \in \mathcal{S}} d_{\rho,s}(\pi^\star) \langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \rangle \\
&= \sum_{s \in \mathcal{S}} \frac{d_{\rho,s}(\pi^\star)}{d_{\rho,s}(\pi^{(k+1)})} d_{\rho,s}(\pi^{(k+1)}) \langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \rangle \\
&\geq \left\| \frac{d_\rho(\pi^\star)}{d_\rho(\pi^{(k+1)})} \right\|_\infty \sum_{s \in \mathcal{S}} d_{\rho,s}(\pi^{(k+1)}) \langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \rangle \\
&= \left\| \frac{d_\rho(\pi^\star)}{d_\rho(\pi^{(k+1)})} \right\|_\infty (1 - \gamma) \left(V_\rho(\pi^{(k+1)}) - V_\rho(\pi^{(k)}) \right),
\end{aligned}$$

19

where the inequality holds because of (42), and the last equality is due to the performance difference lemma, specifically (20). Substituting the above bound and (48) into (46) and dividing both sides by $1 - \gamma$ yield the desired result. ∎

The next theorem is our main result on linear convergence. The convergence rate depends on the performance evaluation distribution ρ through the following quantity:

$$\vartheta_\rho := \frac{1}{1 - \gamma} \left\| \frac{d_\rho(\pi^\star)}{\rho} \right\|_\infty, \tag{53}$$

which is an upper bound on ϑ_k for all $k \geq 0$.

Theorem 10 *Consider the policy mirror descent method (39) with $\pi^{(0)} \in \operatorname{rint} \Pi$. Suppose the step sizes satisfy $\eta_0 > 0$ and*

$$\eta_{k+1} \geq \frac{\vartheta_\rho}{\vartheta_\rho - 1} \eta_k, \qquad k = 0, 1, 2, \ldots, \tag{54}$$

then we have for each $k \geq 0$,

$$V_\rho(\pi^{(k)}) - V_\rho^\star \leq \left(1 - \frac{1}{\vartheta_\rho} \right)^k \left(V_\rho(\pi^{(0)}) - V_\rho^\star + \frac{D_0^\star}{\eta_0 \gamma} \right). \tag{55}$$

Proof Using (13), specifically $d_{\rho,s}(\pi^{(k)}) \geq (1 - \gamma)\rho_s$ for all $s \in \mathcal{S}$, we have $\vartheta_k \leq \vartheta_\rho$ for all $k \geq 0$. In addition, by Lemma 7, we have $\delta_{k+1} - \delta_k \leq 0$ for all $k \geq 0$. Therefore (52) still holds if we replace ϑ_{k+1} by its upper bound ϑ_ρ, i.e.,

$$\vartheta_\rho(\delta_{k+1} - \delta_k) + \delta_k \leq \frac{1}{(1 - \gamma)\eta_k} D_k^\star - \frac{1}{(1 - \gamma)\eta_k} D_{k+1}^\star.$$

Dividing both sides by ϑ_ρ and rearranging terms, we obtain

$$\delta_{k+1} + \frac{1}{(1 - \gamma)\eta_k \vartheta_\rho} D_{k+1}^\star \leq \left(1 - \frac{1}{\vartheta_\rho} \right) \left(\delta_k + \frac{1}{(1 - \gamma)\eta_k(\vartheta_\rho - 1)} D_k^\star \right).$$

If the step sizes satisfy (54), i.e., $\eta_{k+1}(\vartheta_\rho - 1) \geq \eta_k \vartheta_\rho$, then we have

$$\delta_{k+1} + \frac{1}{(1-\gamma)\eta_{k+1}(\vartheta_\rho-1)} D_{k+1}^\star \leq \left(1 - \frac{1}{\vartheta_\rho} \right) \left(\delta_k + \frac{1}{(1-\gamma)\eta_k(\vartheta_\rho-1)} D_k^\star \right).$$

This forms a recursion and results in

$$\delta_k + \frac{1}{(1-\gamma)\eta_k(\vartheta_\rho-1)} D_k^\star \leq \left(1 - \frac{1}{\vartheta_\rho} \right)^k \left(\delta_0 + \frac{1}{(1-\gamma)\eta_0(\vartheta_\rho-1)} D_0^\star \right). \tag{56}$$

Finally, using the fact $\vartheta_\rho \geq 1/(1 - \gamma)$, we derive

$$(1 - \gamma)(\vartheta_\rho - 1) \geq (1 - \gamma) \left(\frac{1}{1 - \gamma} - 1 \right) = \gamma, \tag{57}$$

20

Substituting the above bound into the right side of (56), and considering the nonnegativity of $D_k^\star$ on the left side, we arrive at the desired bound (55). ■

The exact value of ϑ_ρ is hard to estimate in practice, which hinders the use of the step size rule (54). However, we can replace it with the more aggressive increasing rule

$$\eta_{k+1} = \eta_k/\gamma, \qquad \forall\, k \geq 0,$$

which always implies (54). To see this, we use $\vartheta_\rho \geq 1/(1-\gamma)$ to derive

$$\frac{\vartheta_\rho}{\vartheta_\rho - 1} \leq \frac{1/(1-\gamma)}{1/(1-\gamma) - 1} = \frac{1}{\gamma}.$$

According to Theorem 10, in order to guarantee $V_\rho(\pi^{(k)}) - V_\rho^\star \leq \epsilon$, the required number of iterations of the PMD method is

$$\frac{1}{1-\gamma} \left\| \frac{d_\rho(\pi^\star)}{\rho} \right\|_\infty \log\left(\left(V_\rho(\pi^{(0)}) - V_\rho^\star + \frac{1}{\eta_0 \gamma} D_0^\star \right) \frac{1}{\epsilon} \right).$$

Using the bound (2) and assuming $\eta_0 \geq \frac{1-\gamma}{\gamma} D_0^\star$, the iteration complexity becomes

$$\frac{1}{1-\gamma} \left\| \frac{d_\rho(\pi^\star)}{\rho} \right\|_\infty \log \frac{2}{(1-\gamma)\epsilon}.$$

Next we discuss a special choice of the performance evaluation distribution ρ.

Special case of $\rho = \rho^\star$. Let $\rho^\star \in \Delta(\mathcal{S})$ be the stationary state distribution of the MDP under the optimal policy $\pi^\star$. If the MDP starts with $s \sim \rho^\star$ and following $\pi^\star$, then the visit probability at every step is $\rho^\star$ and so is the discounted sum of them. Therefore we have $d_{\rho^\star}(\pi^\star) = \rho^\star$, which implies

$$\left\| \frac{d_{\rho^\star}(\pi^\star)}{\rho^\star} \right\|_\infty = 1 \qquad \text{and} \qquad \vartheta_{\rho^\star} = \frac{1}{(1-\gamma)}.$$

In this case, with the step size rule $\eta_0 \geq \frac{1-\gamma}{\gamma} D_0^\star$ and $\eta_{k+1} \geq \eta_k/\gamma$, we have

$$V_{\rho^\star}(\pi^{(k)}) - V_{\rho^\star}(\pi^\star) \leq \gamma^k \frac{2}{1-\gamma}$$

and the iteration complexity for $V_{\rho^\star}(\pi^{(k)}) - V_{\rho^\star}(\pi^\star) \leq \epsilon$ is dimension-independent:

$$\frac{1}{1-\gamma} \log \frac{2}{(1-\gamma)\,\epsilon}. \tag{58}$$

However, unless the MDP is ergodic, the support of $\rho^\star$ may not cover the full state space $\mathcal{S}$.

Several recent work studied policy mirror descent method for entropy-regularized MDP and obtained similar linear convergence rates (Cen et al., 2020; Lan, 2021; Zhan et al., 2021). With entropy regularization, the resulting MDP is always ergodic and the support of any stationary distribution covers the full state space $\mathcal{S}$, i.e., $\rho^\star > 0$. Lan (2021) only considers $\rho^\star$ as the performance evaluation distribution; Cen et al. (2020) and Zhan et al. (2021) rely on the contraction properties of a generalized Bellman operator and obtain guarantees of the form $\|Q(\pi^{(k)}) - Q(\pi^\star)\|_\infty \leq \epsilon$ where Q is the "soft" Q-function with regularization. Our analysis closely resembles that of Lan (2021), with the following differences:

- We consider standard discounted MDPs and show that linear convergence can be obtained without entropy regularization. Since the support of $\rho^\star$ may not cover the entire state space, we give a general analysis for any $\rho \in \Delta(\mathcal{S})$ and characterize the convergence rate in terms of the distribution mismatch coefficient $\|d_\rho(\pi^\star)/\rho\|_\infty$.

- For discounted MDPs without regularization, Lan (2021) also obtains a slower linear convergence rate ($\gamma^{k/2}$ instead of γ^k), through an approximate policy mirror descent (APMD) method. This method employs exponentially diminishing regularization and exponentially increasing step sizes, and the analysis is considerably more technical.

4.3 Superlinear Convergence

Under additional conditions, the PMD method (39) may exhibit superlinear convergence. We revisit Proposition 9 and start by rewriting the inequality (52) as

$$\delta_{k+1} + \frac{D_{k+1}^\star}{(1-\gamma)\vartheta_{k+1}\eta_k} \leq \left(1 - \frac{1}{\vartheta_{k+1}}\right)\left(\delta_k + \frac{D_k^\star}{(1-\gamma)(\vartheta_{k+1}-1)\eta_k}\right).$$

If the step sizes satisfy $\eta_k \geq \frac{\vartheta_k}{\vartheta_{k+1}-1}\eta_{k-1}$ starting with some $\eta_{-1} > 0$, then we have

$$\delta_{k+1} + \frac{D_{k+1}^\star}{(1-\gamma)\vartheta_{k+1}\eta_k} \leq \left(1 - \frac{1}{\vartheta_{k+1}}\right)\left(\delta_k + \frac{D_k^\star}{(1-\gamma)\vartheta_k\eta_{k-1}}\right)$$

$$\leq \prod_{i=0}^{k}\left(1 - \frac{1}{\vartheta_{i+1}}\right)\left(\delta_0 + \frac{D_0^\star}{(1-\gamma)\vartheta_0\eta_{-1}}\right). \tag{59}$$

Therefore, we have superlinear convergence of δ_k if $\vartheta_k \to 1$.

Recalling the definition of ϑ_k in (51), we have $\vartheta_k \to 1$ if and only if $d_\rho(\pi^{(k)}) \to d_\rho(\pi^\star)$. Apparently, a sufficient condition is $\pi^{(k)} \to \pi^\star$. However, this is hard to establish without additional assumptions, e.g., by assuming that the optimal policy $\pi^\star$ is unique. Alternatively, since $D_k^\star = D_{d_\rho(\pi^\star)}(\pi^\star, \pi^k) \to 0$ implies $\pi^k \to \pi^\star$, a reasonable attempt is to show the convergence of $D_k^\star$ by further leveraging (59). In particular, we can show

$$\frac{D_k^\star}{(1-\gamma)\vartheta_k\eta_{k-1}} \to 0$$

at the same speed as $\delta_k \to 0$, which is at least linear with an uniform upper bound on ϑ_k as we have done in Section 4.2. However, the step-size condition $\eta_k \geq \frac{\vartheta_k}{\vartheta_{k+1}-1}\eta_{k-1}$ implies that the factor $1/(\vartheta_k\eta_{k-1})$ itself converges at the same rate, thus we can not guarantee $D_k^\star \to 0$.

Nevertheless, we list here two sufficient conditions for superlinear convergence that are weaker than directly assuming $\pi^{(k)} \to \pi^\star$. Both conditions have been used to establish superlinear convergence of the classical Policy Iteration algorithm (Puterman, 1994, Corollary 6.4.10 and Theorem 6.4.8, respectively).

- Convergence of the transition probability matrix $P(\pi^k)$. Specifically,

$$\lim_{k\to\infty}\|P(\pi^k) - P(\pi^\star)\| = 0,$$

where $\|\cdot\|$ is any matrix norm. Under this condition, we have $d_\rho(\pi^{(k)}) \to d_\rho(\pi^\star)$ and thus $\vartheta_k \to 1$ because $d_\rho(\pi^{(k)}) = \left(I - \gamma P(\pi^{(k)})\right)^{-T}\rho$ is a continuous function.

- There exists a finite constant $C > 0$ such that for all $k = 1, 2, \ldots$

$$\left\| P(\pi^k) - P(\pi^\star) \right\| \leq C \left(V_\rho(\pi^k) - V_\rho^\star \right).$$

  This condition is stronger than the previous one because we already established linear convergence of $V_\rho(\pi^k) - V_\rho^\star$. As a result, it leads to local quadratic convergence.

Khodadadian et al. (2021) showed that under a variant of the second condition above, the NPG method converges superlinerly. With entropy regularization, the optimal policy $\pi^\star$ is unique and Cen et al. (2020) established local quadratic convergence of the regularized PMD method.

4.4 Connection with Policy Iteration

Our analysis of the PMD method does not impose any upper bound on the step sizes: they can be either arbitrarily large constant (Section 4.1) or geometrically increasing (Section 4.2). If we allow $\eta_k \to \infty$ for all iterations, the limit of the PMD method (39) becomes

$$\pi_s^{(k+1)} = \underset{p \in \Delta(\mathcal{A})}{\arg \min} \ \left\langle Q_s(\pi^{(k)}), p \right\rangle, \qquad \forall \, s \in \mathcal{S},$$

which is precisely the classical Policy Iteration method (e.g., Puterman, 1994; Bertsekas, 2012). In fact, our analysis still holds in the limiting case and the result corresponding to Theorem 10 is

$$\delta_{k+1} \leq \left(1 - \frac{1}{\vartheta_\rho} \right)^k \delta_0,$$

where $\delta_k = V_\rho(\pi^{(k)}) - V_\rho^\star$. Recall the definition of ϑ_ρ in (53). If $\rho = \rho^\star$, then we have $\vartheta_{\rho^\star} = 1/(1-\gamma)$ and

$$\delta_{k+1} \leq \gamma^k \delta_0,$$

which has the same convergence rate as Policy Iteration (e.g., Puterman, 1994; Ye, 2011). In general, we have the trivial bound

$$\vartheta_\rho = \frac{1}{1-\gamma} \left\| \frac{d_\rho(\pi^\star)}{\rho} \right\|_\infty \leq \frac{1}{(1-\gamma) \min_{s \in \mathcal{S}} \rho_s},$$

which leads to

$$\delta_{k+1} \leq \left(1 - (1-\gamma) \min_{s \in \mathcal{S}} \rho_s \right)^k \delta_0.$$

This convergence rate is the same as that established for several variants of policy gradient methods by Bhandari and Russo (2021, Theorem 1), which requires exact line search. Khodadadian et al. (2021) show that the NPG method with an adaptive step size rule can also achieve linear convergence. In contrast, our results in Section 4.2 show that the simple, non-adaptive step size schedule of $\eta_k = \eta_0/\gamma^k$ is sufficient to obtain linear convergence of a general class of policy mirror descent methods.

5. Inexact Policy Mirror Descent Methods

For discounted MDPs with large state and action spaces, computing the exact policy gradients or Q-functions are very costly and infeasible in practice. In this section, we consider the following inexact PMD method

$$\pi_s^{(k+1)} = \underset{p \in \Delta(\mathcal{A})}{\arg\min} \left\{ \eta_k \left\langle \widehat{Q}_s(\pi^{(k)}), p \right\rangle + D(p, \pi_s^{(k)}) \right\}, \qquad \forall\, s \in \mathcal{S}. \tag{60}$$

where $\widehat{Q}_s(\pi^{(k)})$ is an inexact evaluation of $Q_s(\pi^{(k)})$. We first study the convergence properties of (60) under the following assumption on the evaluation error.

Assumption 11 *The inexact Q-function evaluations $\widehat{Q}(\pi^{(k)})$ satisfy*

$$\left\| \widehat{Q}(\pi^{(k)}) - Q(\pi^{(k)}) \right\|_\infty \leq \tau, \qquad \forall\, k \geq 0. \tag{61}$$

The following result is the counterpart of Lemma 7 for the inexact PMD method.

Lemma 12 *Consider the inexact PMD method (60) with $\pi^{(0)} \in \operatorname{rint}\Pi$ and suppose that Assumption 11 holds. Then we have for all $k \geq 0$,*

$$\left\langle \widehat{Q}_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \leq 0, \qquad \forall\, s \in \mathcal{S}, \tag{62}$$

and for any $\rho \in \Delta(\mathcal{S})$,

$$V_\rho(\pi^{(k+1)}) - V_\rho(\pi^{(k)}) \leq \frac{2}{1-\gamma}\tau. \tag{63}$$

Proof The proof of (62) follows the same arguments as in Lemma 7. However, due to the inexact Q-function evaluations, the objectives $V_\rho(\pi^{(k)})$ are no longer monotone decreasing. We use the performance difference lemma to deduct:

$$
\begin{aligned}
V_\rho(\pi^{(k+1)}) - V_\rho(\pi^{(k)}) &= \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} d_{\rho,s}(\pi^{(k+1)}) \left\langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
&= \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} d_{\rho,s}(\pi^{(k+1)}) \left\langle \widehat{Q}_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
&\quad + \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} d_{\rho,s}(\pi^{(k+1)}) \left\langle Q_s(\pi^{(k)}) - \widehat{Q}_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle.
\end{aligned}
$$

Notice that the first term on the right-hand side is non-positive due to (62). For the second term, we use Hölder's inequality to obtain, for all $s \in \mathcal{S}$,

$$
\begin{aligned}
\left\langle Q_s(\pi^{(k)}) - \widehat{Q}_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle &\leq \left\| Q_s(\pi^{(k)}) - \widehat{Q}_s(\pi^{(k)}) \right\|_\infty \left\| \pi_s^{(k+1)} - \pi_s^{(k)} \right\|_1 \\
&\leq 2\left\| \widehat{Q}_s(\pi^{(k)}) - Q_s(\pi^{(k)}) \right\|_\infty \\
&\leq 2\tau, \tag{64}
\end{aligned}
$$

where the second inequality is due to $\left\| \pi_s^{(k+1)} - \pi_s^{(k)} \right\|_1 \leq \left\| \pi_s^{(k+1)} \right\|_1 + \left\| \pi_s^{(k)} \right\|_1 \leq 2$, and the last inequality is due to Assumption 11. Combining (64) with the previous inequality yields (63). ∎

We will need the following simple fact, whose proof is straightforward and thus omitted.

Lemma 13 *Suppose $0 < \alpha < 1$, $b > 0$, and a nonnegative sequence $\{a_k\}$ satisfies*

$$a_{k+1} \leq \alpha a_k + b, \qquad \forall\, k \geq 0.$$

Then for all $k \geq 0$,

$$a_k \leq \alpha^k a_0 + \frac{b}{1 - \alpha}.$$

The following theorem characterizes the convergence of the inexact PMD method under Assumption 11. We keep using the notations $D_k^\star$ and ϑ_ρ defined in (45) and (53), respectively.

Theorem 14 *Consider the inexact PMD method (60) with $\pi^{(0)} \in \mathrm{rint}\,\Pi$ and suppose that Assumption 11 holds. If the step sizes satisfy $\eta_0 \geq \frac{1-\gamma}{\gamma} D_0^\star$ and $\eta_{k+1} \geq \eta_k/\gamma$, then we have for all $k \geq 0$,*

$$V_\rho(\pi^{(k)}) - V_\rho^\star \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^k \frac{2}{1-\gamma} + \frac{4\vartheta_\rho}{1-\gamma}\tau. \tag{65}$$

Proof Applying Lemma 6 to the update in (60) and following the same arguments in the proof of Theorem 8, we arrive at the following counterpart of (46):

$$\mathop{\mathbf{E}}_{s \sim d_\rho(\pi^\star)}\left[\langle \widehat{Q}_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)}\rangle\right] + \mathop{\mathbf{E}}_{s \sim d_\rho(\pi^\star)}\left[\langle \widehat{Q}_s(\pi^{(k)}), \pi_s^{(k)} - \pi_s^\star\rangle\right] \leq \frac{1}{\eta_k}D_k^\star - \frac{1}{\eta_k}D_{k+1}^\star. \tag{66}$$

For the first expectation in (66), we follow the proof of Proposition 9 to obtain

$$\mathop{\mathbf{E}}_{s \sim d_\rho^\star}\left[\langle \widehat{Q}_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)}\rangle\right] \geq \left\|\frac{d_\rho(\pi^\star)}{d_\rho(\pi^{(k+1)})}\right\|_\infty \sum_{s \in \mathcal{S}} d_{\rho,s}(\pi^{(k+1)})\langle \widehat{Q}_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)}\rangle$$

$$= \vartheta_{k+1}\sum_{s \in \mathcal{S}} d_{\rho,s}(\pi^{(k+1)})\langle Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)}\rangle$$

$$+ \vartheta_{k+1}\sum_{s \in \mathcal{S}} d_{\rho,s}(\pi^{(k+1)})\langle \widehat{Q}_s(\pi^{(k)}) - Q_s(\pi^{(k)}), \pi_s^{(k+1)} - \pi_s^{(k)}\rangle$$

$$\geq \vartheta_{k+1}(1 - \gamma)\big(V_\rho(\pi^{(k+1)}) - V_\rho(\pi^{(k)})\big) - 2\vartheta_{k+1}\tau,$$

where the last inequality is due to the performance difference lemma and (64). For the second expectation in (66), we again use the performance difference lemma and Hölder's inequality to obtain

$$\mathop{\mathbf{E}}_{s \sim d_\rho(\pi^\star)}\left[\langle \widehat{Q}_s(\pi^{(k)}), \pi_s^{(k)} - \pi_s^\star\rangle\right]$$

$$= \mathop{\mathbf{E}}_{s \sim d_\rho(\pi^\star)}\left[\langle Q_s(\pi^{(k)}), \pi_s^{(k)} - \pi_s^\star\rangle\right] + \mathop{\mathbf{E}}_{s \sim d_\rho(\pi^\star)}\left[\langle \widehat{Q}_s(\pi^{(k)}) - Q_s(\pi^{(k)}), \pi_s^{(k)} - \pi_s^\star\rangle\right]$$

$$\geq (1 - \gamma)\big(V_\rho(\pi^{(k)}) - V_\rho(\pi^\star)\big) - \mathop{\mathbf{E}}_{s \sim d_\rho(\pi^\star)}\left[\|\widehat{Q}_s(\pi^{(k)}) - Q_s(\pi^{(k)})\|_\infty\|\pi_s^{(k)} - \pi_s^\star\|_1\right]$$

$$\geq (1 - \gamma)\big(V_\rho(\pi^{(k)}) - V_\rho(\pi^\star)\big) - 2\tau.$$

Substituting the last two bounds into (66) and dividing both sides by $1 - \gamma$, we get

$$\vartheta_{k+1}\left(\delta_{k+1} - \delta_k - \frac{2\tau}{1-\gamma}\right) + \delta_k \leq \frac{1}{(1-\gamma)\eta_k}D_k^\star - \frac{1}{(1-\gamma)\eta_k}D_{k+1}^\star + \frac{2\tau}{1-\gamma},$$

25

where $\delta_k := V_\rho(\pi^{(k+1)}) - V_\rho^\star$. Since $\delta_{k+1} - \delta_k - \frac{2\tau}{1-\gamma} \le 0$ (Lemma 12) and $\vartheta_{k+1} \le \vartheta_\rho$, the above inequality still holds with ϑ_{k+1} replaced by ϑ_ρ, which leads to

$$\vartheta_\rho \left(\delta_{k+1} - \delta_k\right) + \delta_k \le \frac{1}{(1-\gamma)\eta_k} D_k^\star - \frac{1}{(1-\gamma)\eta_k} D_{k+1}^\star + \frac{2(1+\vartheta_\rho)\tau}{1-\gamma}.$$

Dividing both sides by ϑ_ρ and rearranging terms, we get

$$\delta_{k+1} + \frac{1}{(1-\gamma)\eta_k\vartheta_\rho} D_{k+1}^\star \le \left(1 - \frac{1}{\vartheta_\rho}\right)\left(\delta_k + \frac{1}{(1-\gamma)\eta_k(\vartheta_\rho-1)} D_k^\star\right) + \left(1 + \frac{1}{\vartheta_\rho}\right)\frac{2\tau}{1-\gamma}.$$

If the step sizes satisfy $\eta_{k+1}(\vartheta_\rho - 1) \ge \eta_k\vartheta_\rho$, which is implied by $\eta_{k+1} \ge \eta_k/\gamma$, then

$$\delta_{k+1} + \frac{1}{(1-\gamma)\eta_{k+1}(\vartheta_\rho-1)} D_{k+1}^\star \le \left(1 - \frac{1}{\vartheta_\rho}\right)\left(\delta_k + \frac{1}{(1-\gamma)\eta_k(\vartheta_\rho-1)} D_k^\star\right) + \frac{4\tau}{1-\gamma},$$

where we also used $1 + 1/\vartheta_\rho < 2$ because $\vartheta_\rho > 1$. Next we invoke Lemma 13 with

$$a_k = \delta_k + \frac{1}{(1-\gamma)\eta_k(\vartheta_\rho-1)} D_k^\star, \qquad \alpha = 1 - \frac{1}{\vartheta_\rho} \qquad \text{and} \qquad b = \frac{4\tau}{1-\gamma},$$

which leads to

$$\delta_k \le \left(1 - \frac{1}{\vartheta_\rho}\right)^k \left(\delta_0 + \frac{1}{(1-\gamma)\eta_0(\vartheta_\rho-1)} D_0^\star\right) + \frac{4\vartheta_\rho}{1-\gamma}\tau.$$

Finally applying (57) and $\eta_0 \ge \frac{1-\gamma}{\gamma} D_0^\star$ gives the desired result (65). ∎

As a result of Theorem 14, we have the following asymptotic error bound:

$$\lim_{k\to\infty} V_\rho(\pi^{(k)}) - V_\rho^\star \le \frac{4\vartheta_\rho}{1-\gamma}\tau = \frac{4\tau}{(1-\gamma)^2}\left\|\frac{d_\rho(\pi^\star)}{\rho}\right\|_\infty,$$

which agrees with that of conservative policy iteration (CPI) of Kakade and Langford (2002, Theorem 6.2). It is also similar to the asymptotic error bound of many approximate dynamical programming algorithms (e.g., Bertsekas, 2012), with the additional factor of distribution mismatch coefficient.

5.1 Sample Complexity under a Generative Model

One way to ensure Assumption 11 hold with high probability is through multiple independent simulations (rollouts) of the MDP under a fixed policy. In this section, we analyze the sample complexity of this approach.

Suppose that for a given policy $\pi^{(k)}$ and any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, we can generate a set of M_k independent, truncated trajectories of horizon H, i.e.,

$$\mathcal{T}_{s,a}^{(k,i)} = \left\{(s_0^{(i)}, a_0^{(i)}), (s_1^{(i)}, a_1^{(i)}), \ldots, (s_{H-1}^{(i)}, a_{H-1}^{(i)}) \mid s_0^{(i)} = s, \ a_0^{(i)} = a\right\}, \quad i = 1, \ldots, M_k.$$

We construct $\widehat{Q}_{s,a}(\pi^{(k)})$ with the trajectories $\mathcal{T}_{s,a}^{(k,i)}$, $i = 1, \ldots, M_k$, as follows:

$$\widehat{Q}_{s,a}(\pi^{(k)}) := \frac{1}{M_k} \sum_{i=1}^{M_k} \widehat{Q}_{s,a}^{(i)}(\pi^{(k)}), \quad \text{where} \quad \widehat{Q}_{s,a}^{(i)}(\pi^{(k)}) := \sum_{t=0}^{H-1} \gamma^t R(s_t^{(i)}, a_t^{(i)}). \tag{67}$$

The following lemma gives a high-probability bound on the error $\|\widehat{Q}(\pi^{(k)}) - Q(\pi^{(k)})\|_\infty$.

Lemma 15 *Consider the Q-estimator given in* (67)*. For any $\delta \in (0, 1)$, if M_k satisfies*

$$M_k \geq \frac{\gamma^{-2H}}{2} \log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right),$$

then we have with probability at least $1 - \delta$,

$$\left\|\widehat{Q}(\pi^{(k)}) - Q(\pi^{(k)})\right\|_\infty \leq \frac{2\gamma^H}{1-\gamma}. \tag{68}$$

Proof We first define the expectation of the Q-estimator in (67):

$$\overline{Q}_{s,a}(\pi^{(k)}) := \mathbf{E}\left[\widehat{Q}_{s,a}(\pi^{(k)})\right] = \mathbf{E}\left[\widehat{Q}_{s,a}^{(i)}(\pi^{(k)})\right] = \mathbf{E}\left[\sum_{t=0}^{H-1} \gamma^t R(s_t^{(i)}, a_t^{(i)})\right],$$

which holds for any $i = 1, \ldots, M_k$. Recall the definition of $Q_{s,a}$ in (15). Since $R(s, a) \geq 0$, we always have $Q_{s,a}(\pi^{(k)}) - \overline{Q}_{s,a}(\pi^{(k)}) \geq 0$. On the other hand,

$$Q_{s,a}(\pi^{(k)}) - \overline{Q}_{s,a}(\pi^{(k)}) = \mathbf{E}\left[\sum_{t=H}^{\infty} \gamma^t R(s_t^{(i)}, a_t^{(i)})\right] \leq \mathbf{E}\left[\sum_{t=H}^{\infty} \gamma^t\right] = \frac{\gamma^H}{1-\gamma},$$

which holds for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Therefore,

$$\left\|\overline{Q}(\pi^{(k)}) - Q(\pi^{(k)})\right\|_\infty \leq \frac{\gamma^H}{1-\gamma}. \tag{69}$$

Next that we can decompose the estimation error into two parts:

$$\left\|\widehat{Q}(\pi^{(k)}) - Q(\pi^{(k)})\right\|_\infty \leq \left\|\widehat{Q}(\pi^{(k)}) - \overline{Q}(\pi^{(k)})\right\|_\infty + \left\|\overline{Q}(\pi^{(k)}) - Q(\pi^{(k)})\right\|_\infty. \tag{70}$$

The last term is bounded by (69), so we need to bound $\left\|\widehat{Q}(\pi^{(k)}) - \overline{Q}(\pi^{(k)})\right\|_\infty$. To this end, we notice that the random variables $\widehat{Q}_{s,a}^{(i)}(\pi^{(k)})$ are bounded in the interval $[0, 1/(1-\gamma)]$. Therefore Hoeffding's inequality (Hoeffding, 1963) implies that for any $\sigma_k > 0$,

$$\Pr\left(\left|\widehat{Q}_{s,a}(\pi^{(k)}) - \overline{Q}_{s,a}(\pi^{(k)})\right| \geq \sigma_k\right) \leq 2\exp\left(-\frac{2M_k^2\sigma_k^2}{M_k/(1-\gamma)^2}\right)$$

$$= 2\exp\left(-2(1-\gamma)^2 M_k \sigma_k^2\right).$$

Applying the union bound across all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we obtain

$$\Pr\left(\left\|\widehat{Q}(\pi^{(k)}) - \overline{Q}(\pi^{(k)})\right\|_\infty \geq \sigma_k\right) \leq 2|\mathcal{S}||\mathcal{A}| \exp\left(-2(1-\gamma)^2 M_k \sigma_k^2\right). \tag{71}$$

Therefore, for any $\delta \in (0,1)$, if we choose M_k large enough, i.e.,

$$M_k \geq \frac{1}{2(1-\gamma)^2 \sigma_k^2} \log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right), \tag{72}$$

then $\left\|\widehat{Q}(\pi^{(k)}) - \overline{Q}(\pi^{(k)})\right\|_\infty < \sigma_k$ with probability at least $1 - \delta$. Combining with (69) and (70), we conclude that with probability at least $1 - \delta$,

$$\left\|\widehat{Q}(\pi^{(k)}) - Q(\pi^{(k)})\right\|_\infty \leq \frac{\gamma^H}{1-\gamma} + \sigma_k.$$

Finally setting $\sigma_k = \gamma^H/(1-\gamma)$ gives the desired result. ∎

The next theorem characterizes the sample complexity of the inexact PMD method with the simple Q-estimator.

Theorem 16 *Consider using the Q-estimator (67) in the inexact PMD method (60), with the step sizes satisfying $\eta_0 \geq \frac{1-\gamma}{\gamma} D_0^\star$ and $\eta_{k+1} \geq \frac{1}{\gamma}\eta_k$ for all $k \geq 0$. For any $\delta \in (0,1)$ and integers $H > 0$ and $K > 0$, suppose the batch sizes M_k satisfy*

$$M_k \geq \frac{\gamma^{-2H}}{2} \log\left(\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta}\right), \qquad k = 0, 1, \ldots, K-1. \tag{73}$$

Then we have with probability at least $1 - \delta$,

$$V_\rho(\pi^{(K)}) - V_\rho^\star \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^K \frac{2}{1-\gamma} + \frac{8\vartheta_\rho}{(1-\gamma)^2}\gamma^H. \tag{74}$$

In addition, for any $\epsilon > 0$, we have $V_\rho(\pi^{(K)}) - V_\rho^\star \leq \epsilon$ with probability at least $1 - \delta$ if

$$K \geq \vartheta_\rho \log \frac{4}{(1-\gamma)\epsilon} \qquad \text{and} \qquad H \geq \frac{1}{1-\gamma} \log \frac{16\vartheta_\rho}{(1-\gamma)^2\epsilon}. \tag{75}$$

The corresponding sample complexity of state-action pairs is

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^8\epsilon^2} \left\|\frac{d_\rho(\pi^\star)}{\rho}\right\|_\infty^3\right), \tag{76}$$

where the notation $\widetilde{O}(\cdot)$ hides poly-logarithmic factors of $1/(1-\gamma)$, $1/\epsilon$ and $|\mathcal{S}||\mathcal{A}|/\delta$.

Proof Suppose the total number of iterations is K. In order to have (68) hold for all $k = 0, 1, \ldots, K-1$, we need to apply the union bound across all K iterations, which imposes an additional factor K on the right-hand side of (71). Consequently, we can extend Lemma 15 to ensure that the event

$$\left\|\widehat{Q}(\pi^{(k)}) - Q(\pi^{(k)})\right\|_\infty \leq \frac{2\gamma^H}{1-\gamma}, \qquad k = 0, 1, \ldots, K-1,$$

occurs with probability at least $1 - \delta$ provided that (73) holds. Then (74) follows directly from Theorem 14 with $\tau = 2\gamma^H/(1-\gamma)$.

In order to have $V_\rho(\pi^{(k)}) - V_\rho^\star \leq \epsilon$ within K iterations, it suffices to have each of the two terms on the right-hand side of (74) less than $\epsilon/2$, i.e.,

$$\left(1 - \frac{1}{\vartheta_\rho}\right)^K \frac{2}{1-\gamma} \leq \frac{\epsilon}{2} \qquad \text{and} \qquad \frac{8\vartheta_\rho}{(1-\gamma)^2}\gamma^H \leq \frac{\epsilon}{2}.$$

which translate into the conditions on K and H in (75). Correspondingly, the batch sizes need to satisfy $M_k \geq M$ where

$$M := \frac{\gamma^{-2H}}{2} \log\left(\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta}\right) \geq \frac{1}{2}\left(\frac{16\vartheta_\rho}{(1-\gamma)^2\epsilon}\right)^2 \log\left(\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta}\right).$$

The total number of state-action samples can be estimated as

$$|\mathcal{S}||\mathcal{A}| \cdot K \cdot H \cdot M$$

$$= |\mathcal{S}||\mathcal{A}| \cdot \vartheta_\rho \log\left(\frac{4}{(1-\gamma)\epsilon}\right) \cdot \frac{1}{1-\gamma} \log\left(\frac{16\vartheta_\rho}{(1-\gamma)^2\epsilon}\right) \cdot \frac{1}{2}\left(\frac{16\vartheta_\rho}{(1-\gamma)^2\epsilon}\right)^2 \log\left(\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta}\right)$$

$$= \frac{128|\mathcal{S}||\mathcal{A}|\vartheta_\rho^3}{(1-\gamma)^5\epsilon^2} \log\left(\frac{4}{(1-\gamma)\epsilon}\right) \log\left(\frac{16\vartheta_\rho}{(1-\gamma)^2\epsilon}\right) \log\left(\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta}\right)$$

$$= \widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|\vartheta_\rho^3}{(1-\gamma)^5\epsilon^2}\right).$$

Finally, plugging in the definition $\vartheta_\rho = \frac{1}{1-\gamma}\left\|\frac{d_\rho(\pi^\star)}{\rho}\right\|_\infty$ gives the estimate in (76). ∎

The sample complexity obtained in Theorems 16 has $O(\epsilon^{-2})$ dependence on ϵ. This is better than that of $O(\epsilon^{-4})$ obtained by Shani et al. (2020) and Agarwal et al. (2021) and $O(\epsilon^{-3})$ by Liu et al. (2020) for policy gradient type of methods (without regularization). Cen et al. (2020) remarked that $O(\epsilon^{-2})$ sample complexity can be obtained with entropy regularization. Their approach leads to a result without the factor of the distribution mismatch coefficient, but with the same $1/(1-\gamma)^8$ factor. Lazaric et al. (2016) derived an $O(\epsilon^{-2})$ sample complexity for a variant of the policy iteration method, with a factor of at least $1/(1-\gamma)^7$. Lan (2021) studies sample complexity in expectation instead of with high probability and obtains similar results with weaker dependence on $1/(1-\gamma)$. Yuan et al. (2022) characterize the sample complexity of vanilla policy gradient method (such as REINFORCE (Williams, 1992)) under a variety of different assumptions on the parametrized value function.

Much progresses have been made for understanding the sample complexity of discounted MDPs in the tabular setting. Azar et al. (2013) established a lower bound of $\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$ for discounted MDPs under a *generative model*, which allows drawing random state-transitions repeatedly under any policy. The simple Q-estimator we use in this section fits this sample oracle model, but the dependence of our results on $1/(1-\gamma)$ is much worse than the lower bound. On the other hand, this lower bound has been matched or nearly matched by several recent work based on variance-reduced Value Iteration (Sidford et al., 2018) and Q-learning (Wainwright, 2019). There is interesting work to be done for improving the sample complexity of stochastic policy gradient methods.

6. Conclusion and Discussion

We developed a general theory of weak gradient-mapping dominance and used it to obtain an improved sublinear convergence rate of the projected policy gradient method. By exploiting additional structure of discounted MDPs, we show that with a simple, non-adaptive rule of geometrically increasing the step sizes, policy mirror descent methods enjoy linear convergence without relying on entropy or other strongly convex regularizations. In fact, the convergence rates obtained with strongly convex regularizations (Cen et al., 2020; Lan, 2021; Zhan et al., 2021) are no better than γ^k regardless of the regularization strength.

Our results on policy mirror descent methods show that dynamic preconditioning using discounted state-visitation distributions is critical for obtaining fast convergence rates that are (almost) independent of problem dimensions. The adopted local Bregman divergence, being KL-divergence or squared Euclidean distance, does not make much difference. Indeed, when the step sizes grow to infinity, preconditioned policy mirror descent methods derived with different Bregman divergences all reduce to the classical Policy Iteration algorithm. Essentially, such methods with finite step sizes can be viewed as inexact Policy Iteration methods, much like many approximate dynamic programming algorithms.

The major limitation of this work is our restriction to direct policy parametrization. (We note that the NPG method with tabular softmax parametrization has an equivalent mirror-descent form expressed in the policy space, therefore is included in our study.) A natural extension is to consider general policy parametrizations of the form $\pi(\theta)$ where the dimension of θ is much smaller than $|\mathcal{S}||\mathcal{A}|$. There are two ways to proceed. The first approach is to simply treat it as a nonlinear optimization problem of minimizing the composite objective $J_\rho(\theta) = V_\rho(\pi(\theta))$. This approach may lose some important structure of discounted MDPs and it will be hard to establish convergence to global optimum. In addition, by relying on standard theory of smooth nonconvex optimization, we may have to impose bounded step sizes and settle for relatively slow convergence rates.

The second approach is to follow the framework of *compatible function approximation* (Sutton et al., 2000; Kakade, 2001), which is extensively developed by Agarwal et al. (2021). This approach facilitates the extension of our results on inexact policy mirror descent to general policy parametrization. In particular, our results in Section 5 show that geometrically increasing step sizes do not cause instability even if the Q-functions are evaluated inaccurately. In fact, inexact policy mirror descent methods converge linearly up to an asymptotic error floor, which immediately leads to an $O(\epsilon^{-2})$ sample complexity as we have shown. It is of great interest to reduce the dependence of sample complexity on $1/(1-\gamma)$ and the distribution mismatch coefficient.

and quasi-concave. They gave a simple counter-example and pointed out the mistake in the proof. Indeed, it is neither quasi-convex nor quasi-concave. Fortunately this mistake does not affect the rest of the results that are contained in this version.

The author also thanks two anonymous reviewers for their helpful comments.

Appendix A. Appendix

A.1 Derivation of Policy Gradient using Matrix Calculus

We derive the policy gradient formula (18) using simple matrix calculus. Let $e_s \in \mathbf{R}^{|S|}$ be a vector with components $e_{s,s'} = 1$ if $s = s'$ and 0 otherwise. From the expression of $V(\pi)$ in (9), we can write its components as

$$V_s(\pi) = e_s^T V(\pi) = e_s^T \big(I - \gamma P(\pi)\big)^{-1} r(\pi).$$

Using the matrix calculus formula $\frac{\partial X^{-1}}{\partial \pi} = -X^{-1}\frac{\partial X}{\partial \pi}X^{-1}$ with $X = (I - \gamma P(\pi))$, we have

$$\frac{\partial V_s(\pi)}{\partial \pi_{s',a'}} = e_s^T \left(-\big(I - \gamma P(\pi)\big)^{-1}\left(-\gamma\frac{\partial P(\pi)}{\pi_{s',a'}}\right)\big(I - \gamma P(\pi)\big)^{-1} \right) r(\pi) + e_s^T \big(I - \gamma P(\pi)\big)^{-1}\frac{\partial r(\pi)}{\partial \pi_{s',a'}}$$

$$= e_s^T \big(I - \gamma P(\pi)\big)^{-1}\left(\frac{\partial r(\pi)}{\partial \pi_{s',a'}} + \gamma\frac{\partial P(\pi)}{\pi_{s',a'}}\big(I - \gamma P(\pi)\big)^{-1} r(\pi) \right)$$

$$= e_s^T \big(I - \gamma P(\pi)\big)^{-1}\left(R_{s',a'}e_{s'} + \gamma\frac{\partial P(\pi)}{\pi_{s',a'}}V(\pi) \right)$$

where in the last equality we used $\partial r(\pi)/\partial \pi_{s',a'} = R_{s',a'}e_{s'}$ and the definition of $V(\pi)$. From the definition of $P(\pi)$, we have $\partial P(\pi)/\partial \pi_{s',a'} = e_{s'}P(\cdot|s',a')$, which is a rank-one matrix with $P(\cdot|s',a')$ acting as a row vector. Therefore,

$$\frac{\partial V_s(\pi)}{\partial \pi_{s',a'}} = e_s^T \big(I - \gamma P(\pi)\big)^{-1}e_{s'}\Big(R_{s',a'} + \gamma P(\cdot|s',a')V(\pi) \Big) = \frac{1}{1-\gamma}d_{s,s'}(\pi)Q_{s',a'}(\pi),$$

where we used the expression of $d_{s,s'}(\pi)$ in (12) and the definition of $Q_{s',a'}(\pi)$. This gives the component-wise expression for policy gradient, which leads to the aggregated form (18).

A.2 Strong Gradient-Mapping Domination

Following the setting in Section 3.1, we define a stronger notion of gradient-mapping domination and show that it leads to geometric convergence to a global optimum. It corresponds to the setting with exponent $\alpha = 1/2$, which is excluded in Definition 3.

Definition 17 (strong gradient-mapping domination) *Suppose $F := f + \Psi$ where f is L-smooth and Ψ is proper, convex and closed. We say that F satisfies a* strong gradient-mapping dominance *condition if there exists $\omega > 0$ such that*

$$\frac{1}{2}\|G_L(x)\|_2^2 \geq \omega\big(F(T_L(x)) - F^\star\big), \qquad \forall\, x \in \mathrm{dom}\,\Psi, \tag{77}$$

where $F^\star = \min_x F(x)$ and T_L and G_L are defined in (30) and (31) respectively.

Consider the composite optimization problem of minimizing $F := f + \Psi$ where f is L-smooth and Ψ is proper, convex and closed. If F satisfies the strong gradient-mapping domination condition, then the proximal gradient method (29) converges geometrically to a global minimum. To see this, we simply combine the descent property (32) with strong gradient-mapping dominance condition (77) to obtain

$$
F(x^k) - F(x^{k+1}) \;\geq\; \frac{1}{2L} \left\| G_L(x^k) \right\|_2^2 \;\geq\; \frac{\omega}{L} \bigl(F(x^{k+1}) - F^\star \bigr).
$$

Rearranging terms, we obtain

$$
\left(1 + \frac{\omega}{L} \right) \bigl(F(x^{k+1}) - F^\star \bigr) \leq F(x^k) - F^\star.
$$

This leads to a geometric recursion and we have

$$
F(x^k) - F^\star \leq \left(1 + \frac{\omega}{L} \right)^{-k} \bigl(F(x^0) - F^\star \bigr).
$$

Connections with other notions of gradient dominance. The classical Kurdyka-Łojasiewicz (KŁ) condition with exponent $1/2$ (Kurdyka, 1998) can be expressed as

$$
\min_{s \in \partial F(x)} \frac{1}{2} \|s\|^2 \geq \tilde{\omega} \bigl(F(x) - F^\star \bigr), \qquad \forall\, x \in \mathrm{dom}\, \Psi, \tag{78}
$$

where $\partial F(x)$ denotes the set of subgradients (subdifferential) of F at x and $\tilde{\omega} > 0$ is a constant. Karimi et al. (2016) derived a proximal Polyak-Łojasiewicz (PŁ) condition

$$
\frac{1}{2} P_L(x) \geq \omega \bigl(F(x) - F^\star \bigr), \qquad \forall\, x \in \mathrm{dom}\, \Psi, \tag{79}
$$

where

$$
P_L(x) := -2L \min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2 + \Psi(y) - \Psi(x) \right\},
$$

and showed that it is equivalent to the KŁ condition (78) in the sense that they imply each other albeit with different constants $\tilde{\omega}$ and ω. Interestingly, it can be shown that there is an interlacing relationship between our definition of gradient-mapping domination and the proximal PŁ condition above:

$$
\frac{1}{2} \|P_L(x)\|_2^2 \;\geq\; \frac{1}{2} \|G_L(x)\|_2^2 \;\geq\; \omega \bigl(F(x) - F^\star \bigr) \;\geq\; \omega \bigl(F(x^+) - F^\star \bigr).
$$

The proximal PŁ condition (79) takes the first and the third terms in the above inequality chain, while our gradient-mapping dominance condition (77) takes the second and the fourth. We conjecture that these two conditions also imply each other.

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.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

Amir Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. SIAM, 2017.

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

Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1957.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2: *Approximate Dynamic Programming*. Athena Scientific, 4th edition, 2012.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. arXiv preprint, arXiv:1906.01786, 2019.

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.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research* published online 2021, arXiv:2007.06558, 2020.

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.

Ying Cui and Jong-Shi Pang. *Modern Nonconvex Nondifferentiable Optimization*. MOS-SIAM Series on Optimization. SIAM, 2022.

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.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

Sham Kakade. A natural policy gradient. In *Proceedings of the 14th International Conference on Neural Information Processing Systems (NIPS'01)*, pages 1531–1538, 2001.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning (ICML)*, volume 2, pages 267–274, 2002.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2016)*, volume 9851 of *Lectur Notes in Computer Sciencce*. Springer, 2016.

Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. In *Proceedings of the 60th IEEE Conference on Decisoin and Control (CDC)*, pages 3794–3799, December 2021.

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

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783, 1998.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, published online April 2022; arXiv:2102.00135, 2021.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Analysis of classification-based policy iteration algorithms. *Journal of Machine Learning Research*, 17:1–30, 2016.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3107–3110. PMLR, 15–19 Aug 2021.

Guoyin Li and Ting Kei Pong. Calculus of the exponent of Kurdyka-Łjasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics*, 18:1199–1232, 2018.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *Advances in Neural*

Information Processing Systems, volume 33, pages 7624–7636. Curran Associates, Inc., 2020.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37 th International Conference on Machine Learning (ICML)*, 2020.

Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvári, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *Proceedings of the 38 th International Conference on Machine Learning (ICML)*, 2021.

Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the 20th International Conference on Machine Learning (ICML'03)*, pages 560–567, 2003.

Rémi Munos. Error bounds for approximate value iteration. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI'05)*, pages 1006–1011, 2005.

A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140:125–161, 2013.

Boris T. Polyak. Gradient methods for minimizing functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. John Wiley and Sons, Inc., 1994.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Bruno Scherrer. Approximate policy iteration schemes: A comparison. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1314–1322, Bejing, China, 22–24 Jun 2014.

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 2015.

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

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*, pages 5668–5675. AAAI Press, 2020.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* Adaptive Computation and Machine Learning. The MIT Press, 2nd edition, 2018.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pages 1057–1063. MIT Press, 2000.

Csaba Szepesvári and Rémi Munos. Finite time bounds for fitted value iteration. *Journal of Machine Learning Research*, pages 815–857, 2008.

Martin J. Wainwright. Variance-reduced Q-learning is minimax optimal. arXiv e-Preprint, arXiv:1906.04697, 2019.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Yinyu Ye. The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36(4): 593–603, 2011.

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.

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. Preprint, arXiv:2105.11066, 2021.

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.