POLICY OPTIMIZATION OVER GENERAL STATE AND ACTION SPACES [*]

GUANGHUI LAN [†]

Abstract.

Reinforcement learning (RL) problems over general state and action spaces are notoriously challenging. In contrast to the tableau setting, one cannot enumerate all the states and then iteratively update the policies for each state. This prevents the application of many well-studied RL methods especially those with provable convergence guarantees. In this paper, we first present a substantial generalization of the recently developed policy mirror descent method to deal with general state and action spaces. We introduce new approaches to incorporate function approximation into this method, so that we do not need to use explicit policy parameterization at all. Moreover, we present a novel policy dual averaging method for which possibly simpler function approximation techniques can be applied. We establish linear convergence rate to global optimality or sublinear convergence to stationarity for these methods applied to solve different classes of RL problems under exact policy evaluation. We then define proper notions of approximation errors for policy evaluation and investigate their impact on the convergence of these methods applied to general-state RL problems with either finite-action or continuous-action spaces. To the best of our knowledge, the development of these algorithmic frameworks as well as their convergence analysis appear to be new in the literature.

1. Introduction. Stochastic dynamic programming provides a general framework to model the interactions between an agent and its environment, and to improve the agent's decisions through such interactions. More specifically, the status of the environment is described by either discrete or continuous state variables, while the agent's behavior is described by actions. Upon the agent's action, the system's state gets updated, and the agent receives some reward or pays some cost. The goal of stochastic dynamic programming is to find the optimal policy which specifies the agent's best action at a given state.

Markov decision process (MDP) is among the most widely used stochastic dynamic programming models. Consider the Markov decision process $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, c, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ denotes the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is transition model, $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the cost function, and $\gamma \in [0, 1)$ is the discount factor. A policy $\pi : \mathcal{S} \to \mathcal{A}$ determines a particular action to be chosen at a given state. Since MDP can be used to model many important sequential decision making problems (e.g., inventory control, resource allocation, and organ donation and transplantation), it had been a classic topic in the area of operations research and stochastic control [22, 3]. While in the classic MDP literature, the transition model $\mathcal{P}$ is assumed to be given when searching for the optimal policy π^*, we often have no access to exact information about $\mathcal{P}$ in the more recent reinforcement (RL) literature [26]. This can be viewed as the main difference between MDP and RL. During the past few years, RL has received widespread attention partly due to its success in computer games such as AlphaZero [25, 24].

In general, RL methods can be divided into model-based and model-free approaches. In model-based methods, one seeks to approximate the transition model $\mathcal{P}$ by using various statistical learning techniques and then to solve an approximate problem to the original MDP. In model-free methods, one attempts to search improved policies without formulating an approximate transition model first. The difference between model-based and model-free RL is similar to the one between sample average approximation (SAA) and stochastic approximation (SA) in the area of stochastic optimization. Model-based methods separate statistical analysis from the design of optimization algorithms, while model-free methods intend to carefully trade-off the effort needed for data collection and computation through a combined design and analysis. Model-free methods appear to be more applicable to online reinforcement learning where the policies need to be improved as streaming data is being collected. These types of methods are the main subject of this paper.

In spite of the popularity of RL, the development of efficient RL methods with guaranteed convergence is still lacking in this area. Most existing theoretical studies focused only on problems with finite state and finite action spaces. In this setting, a rich set of model-free methods have been developed by incorporating different sampling schemes into the classic methods for Markov decision processes, including linear programming, value iteration (or Q-learning), and policy iteration. In addition, there has been considerable interest in the development of nonlinear optimization based methods that utilize the gradient information (i.e., Q function) for solving RL problems (e.g., [27, 9, 8, 7, 2, 1, 23, 6, 31, 28, 17]). While these methods are very popular

in practice, it was only until recently that we know these stochastic gradient type methods can exhibit comparable and even superior performance than those based on classic dynamic programming and/or linear programming approaches. In particular, we show in [12] that a general class of policy mirror descent method can achieve linear convergence for solving both regularized and unregularized MDP problems. When applied to RL problems with unknown $\mathcal{P}$, it can achieve the optimal $\mathcal{O}(1/\epsilon^2)$ and $\mathcal{O}(1/\epsilon)$ sample complexity for solving unregularized and regularized problems, respectively. It is worth noting the latter sample complexity has not been achieved even for model-based approaches. Further developments of PMD type methods can be found in some more recent works (e.g., [10, 5, 32, 30, 16, 13, 14]).

RL problems with general state and action spaces are notoriously more challenging than the finite-state finite-action setting mentioned above. One significant difficulty associated with a general state space exists in that one cannot enumerate all the states and properly define the policy found for each state. In order to address this issue, the current practice is to apply function approximation, e.g., by using a neural network, for policy parameterization (see, e.g., [4, 5]). Then one can possibly apply stochastic gradient type methods restricted to this policy class to search for the policy parameters. However, the incorporation of policy parameterization would destroy the possible convexity of action space, and make it difficult to satisfy the constraints imposed on policies. In many applications, to meet these constraints might be even more important than to find an optimal policy, since the latter goal is usually not achievable for problems with general state and action spaces due to their high nonconvexity. Moreover, with policy parameterization, it is often difficult to assess the quality of the solutions since the stationarity is defined w.r.t. the specific policy class rather than the original policy optimization problem. In addition, by using policy parameterization, we often end up with an optimization problem of very high dimension in the space of policy parameters, even if the dimension of the original action space is rather small. Finally, even with policy parameterization, another type of function approximation is also needed to evaluate a policy and compute the gradients, since one cannot save such gradient information in a tableau form. In general, there are very little theoretical guarantees (e.g., feasibility, optimality, stationarity) on the convergence of existing RL methods dealing with general state and action spaces.

This paper attempts to address some of the aforementioned issues by presenting novel algorithmic frameworks to solve RL problems over general state and action spaces. The main contributions of this paper mainly consist of the following two aspects. The first aspect of our contribution exists in a substantial generalization of the policy mirror descent (PMD) method. More specifically, we start with an idealized situation for policy optimization, where exact gradient information can be computed over general state and action spaces. We generalize the PMD method and show its linear convergence to the global optimality and sublinear convergence to a stationary point for solving different classes of RL problems depending on the curvature of the action-value function. We then discuss how function approximation can be incorporated into the PMD method. In contrast to existing methods, we will not incorporate policy parameterization into this algorithmic scheme at all. Instead, we apply function approximation only to the policy evaluation step applied to an augmented action-value function. The policies will be represented by a simple optimization problem, and will be computed, whenever needed, based on the parameters obtained during the policy evaluation step. For RL problems with general state and finite action spaces, this method converges to the global optimality up to function approximation errors. Moreover, its subproblems to compute the policy can have explicit solutions in this setting. Furthermore, for RL problems with general state and continuous action spaces, it converges to the global optimality up to function approximation errors if certain curvature condition is satisfied for the approximation functions. Otherwise, it will converge to a stationary point for the original policy optimization problem under certain conditions for the approximation error.

The second aspect of our contribution exists in the development of a novel class of policy dual averaging (PDA) methods for solving RL problems. The dual averaging method was originally developed for convex optimization [19] and later further studied for machine learning (e.g., [29]). Similar to the original dual averaging method, the PDA method developed herein minimizes a weighted summation of the advantage functions plus a certain regularization term at each iteration. The PDA method is equivalent to PMD under the tabular setting with specific selection of distance generating functions, but these two methods are different in general. In comparison with PMD, the PDA method appears to be more amenable to function approximation as it only requires function approximation for the original action-value function, and the stepsize parameters will not impact function approximation. However, other than the tableau setting, the convergence of the PDA method has not been studied before in the literature. Through a novel analysis,

we show that PDA also exhibits linear convergence to the global optimality and sublinear convergence to a stationary point (in a slightly different sense than PMD) depending on the curvature of the action-value function. We also study the impact of stochastic policy evaluation and function approximation on the convergence of the PDA method for solving RL problems with general state and finite action spaces, as well as general state and continuous action spaces. In general, PDA exhibits comparable convergence properties to PMD. In some cases it does have some advantages over PMD by either allowing more adaptive stepsize policy, e.g., when applied to general un-regularized RL problems with finite action spaces, or requiring less restrictive assumptions on the function approximation error, e.g., for solving RL problems with continuous action spaces.

This paper is organized as follows. We first formally introduce the problems of interest in Section 2. Section 3 is devoted to the development of the PMD method for solving RL problems with general state and action spaces. The development of the PDA method is detailed in Section 4, whose structure is parallel to that of Section 3. Finally some concluding remarks are made in Section 5.

1.1. Notation and terminology. Let $\| \cdot \|$ be given norm in action space $\mathcal{A}$ and $\omega : \mathcal{A} \to \mathbb{R}$ be a strongly convex function w.r.t. to given $\| \cdot \|$ s.t.

$$D(a_2, a_1) := \omega(a_1) - [\omega(a_2) + \langle \omega'(a_2), a_1 - a_2 \rangle] \geq \tfrac{1}{2}\|a_1 - a_2\|^2, \forall a_1, a_2 \in \mathcal{A}. \tag{1.1}$$

For general-state and finite-action MDPs (see (2.1)-(2.2)), a common selection of ω for RL is the entropy function $\omega(a) = \sum_{i=1}^m a_i \log a_i$. With this selection, the Bregman distance becomes the KL divergence (see [12]) given by $D(a', a) = \mathrm{KL}(a \parallel a') = \sum_{i=1}^m a_i \log \frac{a_i}{a_i'}$. For general-state and continonus-action MDPs, the selection of ω will be problem-dependent. A common section would be the Euclidean norm $\omega(a) = \sum_{i=1}^m a_i^2/2$, with the corresponding Bregrman distance given by $D(a', a) = \tfrac{1}{2}\|a - a'\|_2^2$.

2. Problems of Interest. For a space Ω with a σ-algebra σ_Ω, we define $\mathcal{M}(\Omega)$ as the set of all probability measures over σ_Ω. In this paper, the term "measurable" refers to a Borel measurable set or function depending on the context.

2.1. Markov Decision Processes. A Markov decision process (MDP) is a five-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, c, \gamma)$, where $\mathcal{S} \subseteq \mathbb{R}^n$ is a measurable state space, $\mathcal{A} \subseteq \mathbb{R}^n$ is a measurable action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{M}(\mathcal{S})$ is a mapping with domain $\mathcal{S} \times \mathcal{A}$, $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denote the cost function, and $\gamma \in [0, 1)$ is a discount factor. The transition kernel $\mathcal{P}(\cdot|s, a)$ is a probability measure on $\mathcal{S}$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. The measurable cost function $c(s, a)$ specifies the instantaneous cost associated with the selection of action a in state s.

A policy $\pi : \mathcal{S} \to \mathcal{A}$ is a measurable mapping that determines a feasible action $\pi(s) \in \mathcal{A}$ for each state $s \in \mathcal{S}$. In this paper, we focus on the case when the action space $\mathcal{A} \subset \mathbb{R}^n$ a compact convex set, and intend to find a deterministic policy that maps a given state s to a specific selection of $a \in \mathcal{A}$. It should be noted, however, that our development covers the finite set of actions $\{\mathcal{A}_1, \ldots, \mathcal{A}_m\}$ as an important special case. Indeed, when the set of possible actions is finite, a widely used approach in policy optimization is to seek a randomized policy that determines the probability of selecting a specific action. This is equivalent to specify a continuous action a over the simplex set

$$\mathcal{A} \equiv \Delta_m := \{p \in \mathbb{R}^m : \textstyle\sum_{i=1}^m p_i = 1, p_i \geq 0, i = 1, \ldots, m\}, \tag{2.1}$$

with each of its extreme point corresponding to an action $\mathcal{A}_i$, $i = 1, \ldots, m$. Accordingly, the cost function and the transition kernel, respectively, are given by

$$c(s, a) = \textstyle\sum_{i=1}^m a_i c_0(s, \mathcal{A}_i) \quad \text{and} \quad \mathcal{P}(s, a) = \textstyle\sum_{i=1}^m a_i \mathcal{P}_0(\cdot|s, \mathcal{A}_i), \tag{2.2}$$

Here $c_0(s, \mathcal{A}_i)$ and $\mathcal{P}_0(\cdot|s, \mathcal{A}_i)$ denote the cost and transition probability when the action $\mathcal{A}_i$ is chosen. In this rest of the paper, we will refer to the aforementioned situation as general-state and finite-action MDP even though the action space is given by the simplex as in (2.1). More generally, our development also covers an action space with mixed discrete and continuous actions by employing randomization over the discrete ones.

Let Π denote the set of all measurable stationary policies that are invariant of time epoch t. For any $\pi \in \Pi$, we measure its performance by the action-value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defined as

$$Q^\pi(s, a) := \mathbb{E}\left[\textstyle\sum_{t=0}^\infty \gamma^t [c(s_t, a_t) + h^{\pi(s_t)}(s_t)] \mid s_0 = s, a_0 = a, a_t = \pi(s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)\right],$$

3

where $h^a(s)$ is a regularizer which is strongly convex w.r.t. a with modulus $\mu_h \geq 0$, i.e.,

$$h^{a_1}(s) - h^{a_2}(s) - \langle (h^{a_2})'(s), a_1 - a_2 \rangle \geq \mu_h D(a_2, a_1).$$

Moreover, we define the state-value function $V^\pi : \mathcal{S} \to \mathbb{R}$ associated with π as

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t [c(s_t, a_t) + h^\pi(s_t)] \mid s_0 = s, a_t = \pi(s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)\right]. \tag{2.3}$$

Throughout the paper, we assume that $c(s,a) + h^a(s) \in [0, \bar{c}]$ so that $V^\pi(s) \in [0, \frac{\bar{c}}{1-\gamma}]$. It can be easily seen from the definitions of Q^π and V^π that

$$V^\pi(s) = Q^\pi(s, \pi(s)), \tag{2.4}$$

$$Q^\pi(s,a) = c(s,a) + h^{\pi(s)}(s) + \gamma \int \mathcal{P}(ds'|s,a) V^\pi(s'). \tag{2.5}$$

Without specific mention, we assume throughout the paper that the integral is taken over the state space $\mathcal{S}$. The main objective in MDP is to find an optimal policy $\pi^* : \mathcal{S} \to \mathcal{A}$ s.t.

$$V^{\pi^*}(s) \leq V^\pi(s), \forall \pi \in \Pi, \forall s \in \mathcal{S}. \tag{2.6}$$

Sufficient conditions that guarantee the existence of π^* have been intensively studied (e.g., [3, 22]). For example, one such sufficient condition is given by the existence of a measurable policy $\pi \in \Pi$ such that

$$c(s, \pi(s)) + h^{\pi(s)}(s) + \gamma \int V(s') P(ds'|s, \pi(s)) = \min_{a \in \mathcal{A}} \left[c(s,a) + h^a(s) + \gamma \int V(s') P(ds'|s,a) \right].$$

Under these conditions, (2.6) can be formulated as a nonlinear optimization problem with a single objective function. More specifically, given an initial state distribution ρ over $\mathcal{S}$, let f_ρ be defined as

$$f_\rho(\pi) := \int V^\pi(s) \rho(ds). \tag{2.7}$$

Our goal is to solve the following policy optimization problem $\min_{\pi(s) \in \mathcal{A}, \forall s \in \mathcal{S}} f_\rho(\pi)$.

For a given policy π, let $\mathcal{P}^\pi$ be its associated transition kernel, i.e., $\mathcal{P}^\pi(\cdot|s) = \mathcal{P}(\cdot|s, \pi(s))$. A probability distribution ν^π is said to be a stationary distribution of $\mathcal{P}^\pi$ if for all $B \in \sigma_\mathcal{S}$, $\nu^\pi(B) = \int \mathcal{P}^\pi(B|s) \nu^\pi(ds)$. The existence and uniqueness of a stationary distribution ν^π, and the convergence of the Markov chain to ν^π is ensured by ergodicity conditions. A sufficient condition is that the Markov chain is Harris ergodic, i.e., it has a regeneration set, and is aperiodic and positively recurrent (see [18]). For discrete-state Markov chains, it is well-known that any aperiodic and positive recurrent Markov chain has a unique invariant stationary distribution. Throughout this paper, we assume that ν^π is the unique stationary distribution associated with $\mathcal{P}^\pi$.

In particular, we use $\nu^* \equiv \nu^{\pi^*}$ to denote the stationary distribution induced by the optimal policy π^*. To simplify our discussion, we choose ρ in (2.7) as the stationary distribution ν^*. In this case, our problem of interest becomes

$$f^* := \min_{\pi(s) \in \mathcal{A}, \forall s \in \mathcal{S}} \{ f(\pi) := f_{\nu^*}(\pi) \}. \tag{2.8}$$

It is clear that π^* must be an optimal solution of problem (2.8). Moreover, an optimal solution of (2.8) must be an optimal policy for our MDP if ν^* is sufficiently random (i.e., supported over $\mathcal{S}$).

2.2. Performance Difference and Policy Gradient. Given a feasible policy $\pi \in \Pi$, we define the discounted state visitation measure by

$$\kappa_s^\pi(B) = (1 - \gamma) \sum_{t=0}^\infty \gamma^t \mathrm{Pr}^\pi(s_t \in B | s_0 = s) \tag{2.9}$$

for $B \in \mathcal{S}$. Here $\mathrm{Pr}^\pi(s_t \in B | s_0 = s)$ denotes the probability distribution of s_t at time epoch t after we follow the policy π starting from $s_0 = s$. Let $(\mathcal{P}^\pi)^t$ be the t-step transition kernel associated with $\mathcal{P}^\pi$ defined recursively according to

$$(\mathcal{P}^\pi)^t(B|s) = \int (\mathcal{P}^\pi)^{t-1}(ds'|s) \mathcal{P}^\pi(B|s'), t \geq 2.$$

Then we have $\Pr^\pi(s_t \in B \mid s_0 = s) = (\mathcal{P}^\pi)^t(B \mid s)$.

We now state an important "performance difference" lemma which tells us the difference on the value functions for two different policies.

LEMMA 2.1. *Let π and π' be two feasible policies. Then we have*

$$V^{\pi'}(s) - V^\pi(s) = \tfrac{1}{1-\gamma} \int \psi^\pi(q, \pi'(q))\, \kappa_s^{\pi'}(dq), \forall s \in \mathcal{S}, \tag{2.10}$$

where

$$\psi^\pi(q, a) := Q^\pi(q, a) - V^\pi(q) + h^a(q) - h^{\pi(q)}(q). \tag{2.11}$$

Proof. By the definition of $V^{\pi'}$ in (2.3), we have

$$
\begin{aligned}
&V^{\pi'}(s) - V^\pi(s) \\
&= \mathbb{E}\left[\textstyle\sum_{t=0}^\infty \gamma^t[c(s_t, a_t) + h^{\pi'(s_t)}(s_t)] \mid s_0 = s, a_t = \pi'(s_t), s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)\right] - V^\pi(s) \\
&= \mathbb{E}\left[\textstyle\sum_{t=0}^\infty \gamma^t[c(s_t, a_t) + h^{\pi'(s_t)}(s_t) + \gamma V^\pi(s_{t+1}) - V^\pi(s_t)]\right.\\
&\qquad\left.\mid s_0 = s, a_t = \pi'(s_t), s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)\right] \\
&\quad + \mathbb{E}\left[V^\pi(s_0) \mid s_0 = s, a_t = \pi'(s_t), s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)\right] - V^\pi(s) \\
&= \mathbb{E}\left[\textstyle\sum_{t=0}^\infty \gamma^t[c(s_t, a_t) + h^{\pi(s_t)}(s_t) + \gamma V^\pi(s_{t+1}) - V^\pi(s_t) + h^{\pi'(s_t)}(s_t) - h^{\pi(s_t)}(s_t)]\right.\\
&\qquad\left.\mid s_0 = s, a_t = \pi'(s_t), s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)\right],
\end{aligned}
$$

where the second identity follows from the cancelation of the terms by taking telescoping sum, and the last identity follows from

$$\mathbb{E}\left[V^\pi(s_0) \mid s_0 = s, a_t = \pi'(s_t), s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)\right] = V^\pi(s).$$

Now using the above identity, (2.5) and (2.9), we have

$$
\begin{aligned}
&V^{\pi'}(s) - V^\pi(s) \\
&= \mathbb{E}\left[\textstyle\sum_{t=0}^\infty \gamma^t[Q^\pi(s_t, a_t) - V^\pi(s_t) + h^{\pi'(s_t)}(s_t) - h^{\pi(s_t)}(s_t)]\right.\\
&\qquad\left.\mid s_0 = s, a_t = \pi'(s_t), s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)\right] \\
&= \tfrac{1}{1-\gamma} \int [Q^\pi(q, \pi'(q)) - V^\pi(q) + h^{\pi'(q)}(q) - h^{\pi(q)}(q)]\, d_s^{\pi'}(dq).
\end{aligned}
$$

∎

We call ψ^π in (2.11) the advantage function. It is different than the advantage function used in the RL literature in that it deals with deterministic policies and involves the term h^a.

LEMMA 2.2. *For any feasible policy π, we have*

$$\mathbb{E}_{s\sim\nu^*}\left[-\psi^\pi(s, \pi^*(s))\right] = (1-\gamma)\mathbb{E}_{s\sim\nu^*}[V^\pi(s) - V^{\pi^*}(s)]$$

Proof. By Lemma 2.1 (with $\pi' = \pi^*$),

$$(1-\gamma)[V^{\pi^*}(s) - V^\pi(s)] = \mathbb{E}_{q\sim\kappa_s^{\pi^*}}\left[\psi^\pi(q, \pi^*(q))\right].$$

Taking expectation for w.r.t. ν^* and noticing the stationarity of ν^*, we have

$$(1-\gamma)\mathbb{E}_{s\sim\nu^*}[V^{\pi^*}(s) - V^\pi(s)] = \mathbb{E}_{s\sim\nu^*, q\sim\kappa_s^{\pi^*}}\left[\psi^\pi(q, \pi^*(q))\right] = \mathbb{E}_{s\sim\nu^*}\left[\psi^\pi(s, \pi^*(s))\right].$$

∎

It is known that one can sometimes derive from Lemma 2.1 the formula for the gradients of the objective function f in (2.8). Indeed, for simplicity let us assume that the state space is finite and denote $\delta = \pi' - \pi$. Then by (2.10), we have

$$
\begin{aligned}
V^{\pi'}(s) - V^{\pi}(s) &= \int [Q^{\pi}(q, \pi'(q)) - Q^{\pi}(q, \pi(q)) + h^{\pi'(q)}(q) - h^{\pi(q)}(q)] \, \kappa_s^{\pi'}(dq) \\
&= \int [Q^{\pi}(q, \pi'(q)) - Q^{\pi}(q, \pi(q)) + h^{\pi'(q)}(q) - h^{\pi(q)}(q)] \, \kappa_s^{\pi}(dq) \\
&\quad + \int [Q^{\pi}(q, \pi'(q)) - Q^{\pi}(q, \pi(q)) + h^{\pi'(q)}(q) - h^{\pi(q)}(q)] \, [\kappa_s^{\pi'}(dq) - \kappa_s^{\pi}(dq)]
\end{aligned}
$$

Suppose that $Q^{\pi}(s, a)$ is differentiable w.r.t. its second argument a, i.e.,

$$
Q^{\pi}(q, \pi(q) + \delta) - Q^{\pi}(q, \pi(q)) = \langle \nabla_a Q^{\pi}(q, \pi(q)), \delta \rangle + o(\|\delta\|^2)
$$

Also assume that h^{π} is also differentiable, i.e.,

$$
h^{\pi'(q)}(q) - h^{\pi(q)}(q) = \langle \nabla h^{\pi(q)}(q), \delta \rangle + o(\|\delta\|^2),
$$

and that κ_s^{π} is locally Lipschitz continuous, i.e.,

$$
\|\kappa_s^{\pi'}(dq) - \kappa_s^{\pi}(dq)\| \le L_{\pi} \|\delta\|.
$$

Sufficient conditions to guarantee some of these assumptions have been developed in the literature (e.g., [20, 21, 4, 15]). Then for any δ s.t. $\pi' = \pi + \delta \in \mathcal{A}$

$$
\begin{aligned}
&\lim_{\|\delta\| \to 0} \tfrac{1}{\|\delta\|} \left[V^{\pi'}(s) - V^{\pi}(s) - \left\langle \int \nabla_a Q^{\pi}(q, \pi(q)) \kappa_s^{\pi}(dq), \delta \right\rangle \right] \\
&= \lim_{\|\delta\| \to 0} \tfrac{1}{\|\delta\|} \left[V^{\pi'}(s) - V^{\pi}(s) - \int \langle \nabla_a Q^{\pi}(q, \pi(q)), \delta \rangle \, \kappa_s^{\pi}(dq) \right] \\
&= 0.
\end{aligned}
$$

Therefore, whenever its exists, we can use the gradient of V^{π} denoted by

$$
[\nabla V^{\pi}(s)](q) := [\nabla_a Q^{\pi}(q, \pi(q)) + \nabla h^{\pi}(q)] \kappa_s^{\pi}(q) \tag{2.12}
$$

as the termination criterion to evaluate the efficiency of policy optimization algorithms.

3. Policy Mirror Descent. The policy mirror descent method was initially designed for solving RL problems over finite state and finite action spaces. Our goal in this section is to generalize this algorithm for RL over general state and action spaces.

3.1. The Generic Algorithmic Scheme. For a given policy π_k, the PMD method minimizes the advantage function $\psi^{\pi_k}(s, a)$, plus the regularization term $D(\pi_k(s), a)$ with a certain stepsize parameter $\eta_k \ge 0$. In order to guarantee that each step of PMD is well-defined, we need to assume through this section that the policy value function $Q^{\pi}(s, a)$ satisfies certain curvature conditions. In particular, we assume that for some finite $\mu_Q \ge 0$, the function

$$
Q^{\pi}(s, a) + \mu_Q D(\pi_0(s), a) \tag{3.1}
$$

is convex w.r.t. a for any $s \in \mathcal{S}$ and any $\pi \in \Pi$. $Q^{\pi}(s, \cdot)$ is convex if $\mu_Q = 0$, and strongly convex if $\mu_Q < 0$. Otherwise, if $\mu_Q > 0$, $Q^{\pi}(s, \cdot)$ is not necessarily convex and function of this type is also referred to as being weakly convex. It is well-known that any differential function with Lipschitz continuous gradients are weakly convex. Let us denote

$$
\mu_d := \mu_h - \mu_Q. \tag{3.2}
$$

The advantage function $\psi^{\pi}(s, \cdot)$ defined in (2.11) is convex if $\mu_d \ge 0$, and weakly convex otherwise.

Algorithm 1 The policy mirror descent method

Input: $\eta_k \geq 0$ and π_0.
for $k = 0, 1, \ldots,$ **do**

$$\pi_{k+1}(s) = \arg\min_{a \in \mathcal{A}} \{\psi^{\pi_k}(s, a) + \tfrac{1}{\eta_k} D(\pi_k(s), a)\} \tag{3.3}$$

$$= \arg\min_{a \in \mathcal{A}} \{Q^{\pi_k}(s, a) + h^a(s) + \tfrac{1}{\eta_k} D(\pi_k(s), a)\}, \forall s \in \mathcal{S}. \tag{3.4}$$

end for

With a properly chosen η_k, the objective function of the subproblem in (3.3) is strongly convex, and thus the new policy π_{k+1} is well-defined. The following result is well-known for mirror descent methods.

LEMMA 3.1. *If η_k in (3.3) satisfies*

$$\mu_d + \tfrac{1}{\eta_k} \geq 0, \tag{3.5}$$

then for any $a \in \mathcal{A}$,

$$\psi^{\pi_k}(s, \pi_{k+1}(s)) + \tfrac{1}{\eta_k} D(\pi_k(s), \pi_{k+1}(s)) + (\mu_d + \tfrac{1}{\eta_k}) D(\pi_{k+1}(s), a) \leq \psi^{\pi_k}(s, a) + \tfrac{1}{\eta_k} D(\pi_k(s), a). \tag{3.6}$$

Proof. Observe that the objective function of (3.3) is strongly convex w.r.t. $a \in \mathcal{A}$ with modulus $\mu_d + 1/\eta_k$. The result then follows immediately from the optimality condition of (3.3) (see, e.g., Lemma 3.5 of [11]). ∎

We now show the progress made by each iteration of the PMD method.
PROPOSITION 3.2. *Suppose η_k in the PMD method satisfies (3.5). Then for any $s \in S$,*

$$V^{\pi_{k+1}}(s) - V^{\pi_k}(s) \leq \psi^{\pi_k}(s, \pi_{k+1}(s)) \leq -[\tfrac{1}{\eta_k} D(\pi_k(s), \pi_{k+1}(s)) + (\mu_d + \tfrac{1}{\eta_k}) D(\pi_{k+1}(s), \pi_k(s))]. \tag{3.7}$$

Proof. By (3.6) with $a = \pi_k(s)$, we have

$$\psi^{\pi_k}(s, \pi_{k+1}(s)) + \tfrac{1}{\eta_k} D(\pi_k(s), \pi_{k+1}(s)) + (\mu_d + \tfrac{1}{\eta_k}) D(\pi_{k+1}(s), \pi_k(s))$$
$$\leq \psi^{\pi_k}(s, \pi_k(s)) + \tfrac{1}{\eta_k} D(\pi_k(s), \pi_k(s)) = 0, \tag{3.8}$$

where the last identity follows from the fact that $\psi^{\pi_k}(s, \pi_k(s)) = 0$ due to (2.4) and (2.11). By Lemma 2.1, (3.8) and the fact that $\kappa_s^{\pi_{k+1}}(\{s\}) \geq 1 - \gamma$ due to (2.9), we have

$$V^{\pi_{k+1}}(s) - V^{\pi_k}(s) = \tfrac{1}{1-\gamma} \int \psi^{\pi_k}(q, \pi_{k+1}(q)) \kappa_s^{\pi_{k+1}}(dq)$$
$$\leq \tfrac{1}{1-\gamma} \psi^{\pi_k}(s, \pi_{k+1}(s)) \kappa_s^{\pi_{k+1}}(\{s\})$$
$$\leq \psi^{\pi_k}(s, \pi_{k+1}(s)). \tag{3.9}$$

The result then follows by combining (3.8) with (3.9). ∎

In the following two results, we establish the convergence of the PMD method under the assumption $\mu_d \geq 0$ and $\mu_d < 0$, respectively.
THEOREM 3.3. *Suppose that $\mu_d \geq 0$ and η_k in the PMD method satisfies (3.5). Then for any $k \geq 1$,*

$$\sum_{t=0}^{k-1} \{\eta_t[f(\pi_{t+1}) - f(\pi^*)] - \gamma\eta_t[f(\pi_t) - f(\pi^*)]\} + (\mu_d\eta_{k-1} + 1)\mathcal{D}(\pi_k, \pi^*) \leq \mathcal{D}(\pi_0, \pi^*), \tag{3.10}$$

where

$$\mathcal{D}(\pi, \pi') := \mathbb{E}_{s \sim \nu^*}[D(\pi(s), \pi(s'))]. \tag{3.11}$$

In particular, if $\eta_k = \gamma^{-k}$, then

$$f(\pi_k) - f(\pi^*) + (\mu_d + \gamma^{k-1})\mathcal{D}(\pi_k, \pi^*) \le \gamma^k[f(\pi_0) - f(\pi^*)] + \gamma^{k-1}\mathcal{D}(\pi_0, \pi^*).$$

Proof. Multiplying (3.6) by η_k and then taking the telescopic sum (in view of $\mu_d \ge 0$), we obtain

$$\sum_{t=0}^{k-1} \eta_t \psi^{\pi_t}(s, \pi_{t+1}(s)) + \sum_{t=0}^{k-1} D(\pi_t(s), \pi_{t+1}(s)) + (\mu_d\eta_{k-1} + 1)D(\pi_k(s), a) \le \sum_{t=0}^{k-1} \eta_t \psi^{\pi_t}(s, a) + D(\pi_0(s), a).$$

Setting $a = \pi^*(s)$ in the above inequality, and using the first relation in (3.7), we then have

$$\sum_{t=0}^{k-1} \eta_t[V^{\pi_{t+1}}(s) - V^{\pi_t}(s)] + (\mu_d\eta_{k-1} + 1)D(\pi_k(s), \pi^*(s)) \le \sum_{t=0}^{k-1} \eta_t \psi^{\pi_t}(s, \pi^*(s)) + D(\pi_0(s), \pi^*(s)).$$

Now taking expectation of the above inequality w.r.t. $s \sim \nu^*$, applying Lemma 2.2, and using the definition of f and $\mathcal{D}$ in (2.8) and (3.11), we then obtain

$$\sum_{t=0}^{k-1} \eta_t[f(\pi_{t+1}) - f(\pi_t)] + (\mu_d\eta_{k-1} + 1)\mathcal{D}(\pi_k, \pi^*) \le (1 - \gamma)\sum_{t=0}^{k-1} \eta_t[f(\pi^*) - f(\pi_t)] + \mathcal{D}(\pi_0, \pi^*),$$

which clearly implies (3.10) after rearranging the terms. ∎

In view of Theorem 3.3, the PMD method will exhibit a linear rate of convergence if the advantage function is convex w.r.t. $a \in \mathcal{A}$. This result generalizes the previous linear rate of convergence for the PMD method in the tabular setting to more general problems. The following result further shows that this algorithm converges to a stationary point of problem (2.8) if $\mu_d < 0$.

THEOREM 3.4. *Suppose that $\mu_d < 0$ and that $\eta_k = |\mu_d|/2$ for any $k \ge 0$. Then for any $s \in \mathcal{S}$, there exist iteration indices $\bar{k}(s)$ found by running k iterations of the PMD method s.t.*

$$0 \le -\psi^{\pi_{\bar{k}(s)}}(s, \pi_{\bar{k}(s)+1}(s))] \le \tfrac{1}{k}[V^{\pi_0}(s) - V^{\pi^*}(s)], \tag{3.12}$$

$$2D(\pi_{\bar{k}(s)}, \pi_{\bar{k}(s)+1}(s)) + D(\pi_{\bar{k}(s)+1}(s), \pi_{\bar{k}(s)}(s)) \le \tfrac{1}{k|\mu_d|}[V^{\pi_0}(s) - V^{\pi^*}(s)]. \tag{3.13}$$

Proof. Taking telescopic sum of (3.7), we obtain

$$\sum_{t=0}^{k-1}[-\psi^{\pi_t}(s, \pi_{t+1}(s))] \le \sum_{t=0}^{k-1}[-\psi^{\pi_t}(s, \pi_{t+1}(s))] \le V^{\pi_0}(s) - V^{\pi_k}(s) \le V^{\pi_0}(s) - V^{\pi^*}(s), \forall s \in \mathcal{S}. \tag{3.14}$$

The iteration indices $\bar{k}(s)$ in (3.12) are given by the ones with the smallest value of $[-\psi^{\pi_t}(s, \pi_{t+1}(s))]$, $t = 0, \ldots, k-1$. The relation in (3.13) follows directly from (3.7) and (3.12). ∎

Observe that the relation in (3.13) tells us some information about the stationarity of the policy $\pi_{\bar{j}(s)}$ for any $s \in \mathcal{S}$. For simplicity, let us assume that $\omega(a) = \|a\|_2^2/2$, $\mathcal{A} = \mathbb{R}^m$, and that $Q^\pi(s, \cdot)$ and $h^a(s)$ have Lipschitz continuous gradient w.r.t. a. Then the optimality condition of problem (3.4) given by

$$\nabla_a Q^{\pi_{\bar{j}(s)}}(s, \pi_{\bar{j}(s)+1}(s)) + \nabla_a h^{\pi_{\bar{j}(s)+1}}(s) + 2|\mu_d|[\pi_{\bar{j}(s)+1}(s) - \pi_{\bar{j}(s)}(s)] = 0, \forall s \in \mathcal{S},$$

implies that

$$\|\nabla_a Q^{\pi_{\bar{j}(s)}}(s, \pi_{\bar{j}(s)}(s)) + \nabla_a h^{\pi_{\bar{j}(s)}}(s)\| \le (L_Q + L_h + 2|\mu_d|)\|\pi_{\bar{j}(s)+1}(s) - \pi_{\bar{j}(s)}(s)\|] = \mathcal{O}(1/\sqrt{k}), \forall s \in \mathcal{S}.$$

Hence, by (2.12), the gradient ∇V^π is also $\mathcal{O}(1/\sqrt{k})$ with the selection of the policy $\pi(s) = \pi_{\bar{j}(s)}(s)$, $s \in \mathcal{S}$. It should be noted, however, that it might be difficult to implement this strategy in practice, since it requires us to specify a different index $\bar{j}(s)$ for a different state $s \in \mathcal{S}$. In practice, one can possibly select one random iterate for all different states. In addition, the strategy of choosing a different iteration index for a different state can possibly be implemented when function approximation is used in PMD as shown in the next subsection.

8

3.2. Function Approximation in PMD. In an ideal situation for the PMD method, we update a given policy π_k by

$$\pi_{k+1}(s) = \underset{a \in \mathcal{A}}{\arg\min} \{\underbrace{Q^{\pi_k}(s,a) - \tfrac{1}{\eta_k}\langle \nabla\omega(\pi_k(s)), a\rangle}_{L^{\pi_k}(s,a)} + h^a(s) + \tfrac{1}{\eta_k}\omega(a)\}, \forall s \in \mathcal{S} \qquad (3.15)$$

for some properly chosen stepsize $\eta_k \geq 0$. However, for the general state space $\mathcal{S}$, it is impossible to perform these updates since $\mathcal{S}$ is possibly infinite. Current practice is to apply different function approximations for both estimating $Q^{\pi_k}(s,a)$ and computing a parameterized form of $\pi_k(s)$. While the idea of using a parameterized form of π_k appears to be conceptually simple, it often results in a few significant limitations. First, it introduces additional approximation error due to the difficulty to identify the optimal policy class. Second, it will destroy the convexity of the action space $\mathcal{A}$, and make it difficult to satisfy these constraints.

In order to deal with these issues for RL with general state spaces, we adopt the idea of using function approximation for the value functions. However, different from existing methods that require function approximation for both value function and policy parameterization, we will not use policy parameterization in the policy optimization step. Instead, as shown in Algorithm 2, we suggest to compute a function approximation for an augmented action-value function $L^{\pi_k}(s,a)$ (rather than $Q^{\pi_k}(s,a)$) used to define the objective function in (3.15).

Algorithm 2 The policy mirror descent method with function approximation

Input: $\eta_k \geq 0$ and π_0.
for $k = 0, 1, \ldots,$ **do**
   Set

$$L^{\pi_k}(s,a) = Q^{\pi_k}(s,a) + \tfrac{1}{\eta_k}\langle \nabla\omega(\pi_k(s)), a\rangle. \qquad (3.16)$$

   Run a policy evaluation procedure to estimate $L^{\pi_k}(s,a)$ s.t.

$$\tilde{L}(s,a;\theta_k) \approx L^{\pi_k}(s,a), \forall s \in \mathcal{S}, \forall a \in \mathcal{A}. \qquad (3.17)$$

   Update the policy to

$$\pi_{k+1}(s) = \underset{a \in \mathcal{A}}{\arg\min} \{\tilde{L}(s,a;\theta_k) + h^a(s) + \tfrac{1}{\eta_k}\omega(a)\}, \forall s \in \mathcal{S}. \qquad (3.18)$$

end for

In (3.17), $\tilde{L}(s,a;\theta_k)$ represents a function approximation, parameterized by θ_k, of the augmented action-value function $L^{\pi_k}(s,a)$. In the simplest form of function approximation, we are given a set of linearly independent basis functions $l_1(s,a), l_2(s,a), \ldots, l_d(s,a)$, and the goal is to determine the coefficients $\theta_k \in \mathbb{R}^d$ so that $\tilde{L}(s,a;\theta_k) = \sum_{i=1}^{d} \theta_k^i l_i(s,a)$ approximates $L^{\pi_k}(s,a)$. More powerful and sophisticated nonlinear function approximation methods, such as Reproducing Kernel Hilbert Space (RKHS) and Neural network, can also be used to estimate $L^{\pi_k}(s,a)$. We will study impact of the error of approximating $L_k(s,a)$ on the convergence of the PMD method in the next few subsections.

The equation in (3.18) tells us how to compute the new policy π_{k+1}. In some cases, the solution of π_{k+1} is given in an explicit form (see Section 3.3). In a more general setting, π_{k+1} is represented as an optimal solution of problem (3.18). With a properly choosen η_k, the objective function of (3.18) is strongly convex, and hence the policy π_{k+1} is uniquely defined. Moreover, any efficient numerical procedures (e.g., the Ellipsoid method and accelerated gradient method) can be used to solve (3.18). For the sake of simplicity, we assume throughout this paper that exact solutions can be computed for (3.18). However, our analysis can be extended to the case when only approximate solutions of (3.18) are available (see Section 6 of [12]) for an illustration. Observe that, different from existing policy gradient methods, we do not actually compute $\pi_{k+1}(s)$ for any $s \in \mathcal{S}$. Since the function approximation $L(s,a;\theta_k)$ is given in a parametric form (through θ_k's), we can compute $\pi_{k+1}(s)$ using (3.18) whenever needed, e.g., inside the policy evaluation procedure. The idea of representing a policy by an optimization problem has been used in the area of multi-stage stochastic optimization.

3.3. PMD for general state and finite action spaces. In this subsection, we focus on RL with a fine set of actions $\{\mathcal{A}_1, \ldots, \mathcal{A}_m\}$. As mentioned earlier, in this case we search for an optimal randomized policy in the probability simplex $\mathcal{A} = \Delta_m$ in (2.1), with the cost and transition probability defined in (2.2). For a given iterate $\pi_k(s) \in \mathcal{A}$, we use

$$Q_0^{\pi_k}(s, \mathcal{A}_i) := c_0(s, \mathcal{A}_i) + h^{\pi_k(s)}(s) + \gamma \int \mathcal{P}_0(ds'|s, \mathcal{A}_i) V^{\pi_k}(s'), i = 1, \ldots, m, \tag{3.19}$$

$$L_0^{\pi_k}(s, \mathcal{A}_i) := Q_0^{\pi_k}(s, \mathcal{A}_i) + \tfrac{1}{\eta_k} \nabla_i \omega(\pi_k(s)), i = 1, \ldots, m, \tag{3.20}$$

respectively, to denote the value function and augmented value function associated with each action $\mathcal{A}_i$. For notational convenience, we also denote $Q_0^{\pi_k}(s) = (Q_0^{\pi_k}(s, \mathcal{A}_1), \ldots, Q_0^{\pi_k}(s, \mathcal{A}_m))$ and $L_0^{\pi_k}(s) = (L_0^{\pi_k}(s, \mathcal{A}_1), \ldots, L_0^{\pi_k}(s, \mathcal{A}_m))$. It then follows that $Q^{\pi_k}(s, a) = \langle Q_0^{\pi_k}(s), a \rangle$ and

$$L^{\pi_k}(s, a) = Q^{\pi_k}(s, a) + \langle \nabla \omega(\pi_k(s)), a \rangle / \eta_k = \langle L_0^{\pi_k}(s), a \rangle.$$

We suggest to approximate $L_0^{\pi_k}(s, \mathcal{A}_i)$ by $\tilde{L}_0(s, \mathcal{A}_i; \theta_k)$, $i = 1, \ldots, m$, where $\tilde{L}_0(s, \mathcal{A}_i; \theta)$ denotes some function class parameterized by θ. Then the subproblem in (3.4) reduces

$$\pi_{k+1}(s) = \arg\min_{a \in \Delta_m} \{ \langle \tilde{L}_0(s; \theta_k), a \rangle + h^a(s) + \tfrac{1}{\eta_k} \omega(a) \}, \forall s \in \mathcal{S}, \tag{3.21}$$

where $\tilde{L}_0(s; \theta_k) := (\tilde{L}_0(s, \mathcal{A}_1; \theta_k), \ldots, \tilde{L}_0(s, \mathcal{A}_m; \theta_k))$. The objective function of (3.21) is strongly convex with modulus $\mu_h + 1/\eta_k$. It is well-known that the above problem has an explicit solution, for example, when $\omega(a) = \sum_{i=1}^m a_i \log a_i$ and $h^a = \lambda \mathrm{KL}(a||\pi_0(s))$ for some $\lambda \geq 0$.

In order to find the parameter θ_k, we need to solve the following statistical learning/estimation problem

$$\min_{\theta \in \Theta} \int \tfrac{1}{2} \|\tilde{L}_0(s; \theta) - L_0^{\pi_k}(s)\|_\infty^2 \rho(ds)$$

$$= \min_{\theta \in \Theta} \mathbb{E}_{s \sim \rho} \left\{ \max_{y \geq 0} \max_{\|z\|_1 \leq 1} y \left[\langle \tilde{L}_0(s; \theta) - L_0^{\pi_k}(s), z \rangle \right] - \tfrac{y^2}{2} \right\}, \tag{3.22}$$

for a given distribution ρ over $\mathcal{S}$. Observe that there exist two types of uncertainty for the nested stochastic composite optimization (NSCO) problem in (3.22), including: a) the sampling of s from $\mathcal{S}$ according to ρ in the outer minimization problem, and b) the estimation of $L_0^{\pi_k}(s, \mathcal{A}_i)$ through the sampling of the Markov Chain induced by π_k for the inner maximization problem. We use ξ_k to denote these random samples generated at iteration k to solve problem (3.22), and $F_k := \{\xi_1, \ldots, \xi_k\}$ denote its associated filtration. Let θ_k^* be an optimal solution of (3.22) and θ_k be an approximate solution found by using some policy evaluation algorithms. Note that problem (3.22) is convex when linear function approximation is used, and we can apply either sample averaging approximation or stochastic gradient type methods to solve it.

It is worth noting that in the PMD method the stepsize η_k appears in the definition of L^{π_k}, which may impact the magnitude of the stochastic first-order information of the policy evaluation problem (see (3.22)). As a result, the sample complexity of solving this statistical learning problem might increase as η_k decreases. One can possibly alleviate this issue by using constant stepsize η_k, which might worsen the dependence on $1 - \gamma$ as noticed in [12]. Another possible solution is to use a larger sample size to approximate the term $\nabla \omega(\pi_k(s))$ than that for the action-value function Q^{π_k}, since the former sampling scheme does not involve the Markov chain and can be performed faster than collecting samples from the Markov chain associated with π_k. Yet we will propose another different approach that can possibly address this issue by developing the policy dual averaging method in the next section.

In this section we assume that the approximation error

$$\delta_k(s) := \tilde{L}_0(s; \theta_k) - L_0^{\pi_k}(s) \tag{3.23}$$

can be decomposed into the following components:

$$\delta_k(s) = \underbrace{\mathbb{E}_{\xi_k}[\tilde{L}_0(s; \theta_k)] - L_0^{\pi_k}(s)}_{\delta_k^{det}(s)} + \underbrace{\tilde{L}_0(s; \theta_k) - \mathbb{E}_{\xi_k}[\tilde{L}_0(s; \theta_k)]}_{\delta_k^{sto}(s)}$$

$$= \underbrace{\tilde{L}_0(s; \theta_k^*) - L_0^{\pi_k}(s)}_{\delta_k^{app}(s)} + \underbrace{\mathbb{E}_{\xi_k}[\tilde{L}_0(s; \theta_k)] - \tilde{L}_0(s; \theta_k^*)}_{\delta_k^{bia}(s)} + \underbrace{\tilde{L}_0(s; \theta_k) - \mathbb{E}_{\xi_k}[\tilde{L}_0(s; \theta_k)]}_{\delta_k^{sto}(s)}, \tag{3.24}$$

10

where the expectation $\mathbb{E}_{\xi_k}$ denote the expectation w.r.t. ξ_k conditional on F_{k-1}. Note that $\delta_k^{det}(s)$ and $\delta_k^{sto}(s)$ denote the deterministic and stochastic errors, respectively. Moreover, the deterministic error $\delta_k^{det}(s)$ can be further decomposed into the error caused by function approximation $\delta_k^{app}(s)$ and the bias term $\delta_k^{bia}(s)$ when computing the estimation. One can possibly reduce the $\delta_k^{app}(s)$ by using more complex function class. In particular, we have $\delta_k^{app}(s) = 0$ in the tabular case, and for some specific kernels in RKHS. Moreover, the bias term $\delta_k^{bia}(s)$ can be usually reduced much faster than the stochastic error $\delta_k^{sto}(s)$ as observed previously in [12]. While these two deterministic error terms are caused by different sources, their impact on the overall convergence of the PMD method is similar. We make the following assumptions about these error terms:

$$\mathbb{E}_{s \sim \nu^*}[\|\delta_k^{det}(s)\|_\infty] \leq \varsigma, \tag{3.25}$$

$$\mathbb{E}_{\xi_k}\left\{\mathbb{E}_{s \sim \nu^*}[\|\delta_k^{sto}(s)\|_\infty^2]\right\} \leq \sigma^2, \tag{3.26}$$

for some $\varsigma \geq 0$ and $\sigma \geq 0$.

The following result states an important recursion about each PMD iteration for solving RL problems with finite action space.

PROPOSITION 3.5. *For any $s \in \mathcal{S}$, we have*

$$- \psi^{\pi_k}(s, a) + h^{\pi_k(s)}(s) - h^a(s) + (\mu_h + \tfrac{1}{\eta_k})D(\pi_{k+1}(s), a)$$
$$\leq \tfrac{1}{\eta_k}D(\pi_k(s), a) + \tfrac{\eta_k}{2}[\tfrac{\bar{c}}{1-\gamma} + M_h + \|\delta_k^{sto}(s)\|_\infty]^2 + \|\delta_k^{det}(s)\|_\infty$$
$$- \langle \delta_k^{sto}(s), \pi_k(s) - a \rangle, \forall a \in \mathcal{A}. \tag{3.27}$$

Proof. By the optimality condition of (3.21), we have

$$\langle L_0^{\pi_k}(s), \pi_{k+1}(s) - a \rangle + h^{\pi_{k+1}(s)}(s) - h^a(s) + \tfrac{1}{\eta_k}[\omega(\pi_{k+1}(s)) - \omega(a)] + (\mu_h + \tfrac{1}{\eta_k})D(\pi_{k+1}(s), a)$$
$$+ \langle \delta_k(s), \pi_{k+1}(s) - a \rangle \leq 0. \tag{3.28}$$

Using the facts $\langle L_0^{\pi_k}(s), a \rangle = L^{\pi_k}(s, a) = Q^{\pi_k}(s, a) + \tfrac{1}{\eta_k}\langle \nabla\omega(\pi_k), a \rangle$ in the above relation, we obtain

$$Q^{\pi_k}(s, \pi_{k+1}(s)) - Q^{\pi_k}(s, a) + h^{\pi_{k+1}(s)}(s) - h^a(s) + \tfrac{1}{\eta_k}[\omega(\pi_{k+1}(s)) - \omega(a)] + (\mu_h + \tfrac{1}{\eta_k})D(\pi_{k+1}(s), a)$$
$$+ \tfrac{1}{\eta_k}\langle \nabla\omega(\pi_k(s)), \pi_{k+1}(s) - a \rangle + \langle \delta_k(s), \pi_{k+1}(s) - a \rangle \leq 0.$$

Noticing that

$$\omega(\pi_{k+1}(s)) - \omega(a) + \langle \nabla\omega(\pi_k(s)), \pi_{k+1}(s) - a \rangle = D(\pi_k(s), \pi_{k+1}(s)) - D(\pi_k(s), a),$$

we have

$$Q^{\pi_k}(s, \pi_{k+1}(s)) - Q^{\pi_k}(s, a) + h^{\pi_{k+1}(s)}(s) - h^a(s) + \tfrac{1}{\eta_k}[D(\pi_k(s), \pi_{k+1}(s)) - D(\pi_k(s), a)]$$
$$+ (\mu_h + \tfrac{1}{\eta_k})D(\pi_{k+1}(s), a) + \langle \delta_k(s), \pi_{k+1}(s) - a \rangle \leq 0,$$

which, in view of (2.11), then implies that

$$\psi^{\pi_k}(s, \pi_{k+1}(s)) - \psi^{\pi_k}(s, a) + h^{\pi_{k+1}(s)}(s) - h^{\pi_k}(s) + \tfrac{1}{\eta_k}[D(\pi_k(s), \pi_{k+1}(s)) - D(\pi_k(s), a)]$$
$$+ (\mu_h + \tfrac{1}{\eta_k})D(\pi_{k+1}(s), a) + \langle \delta_k(s), \pi_{k+1}(s) - a \rangle \leq 0. \tag{3.29}$$

Noticing that

$$\psi^{\pi_k}(s, \pi_{k+1}(s)) = Q^{\pi_k}(s, \pi_{k+1}(s)) - V^{\pi_k}(s) = \langle Q_0^{\pi_k}(s), \pi_{k+1}(s) - \pi_k(s) \rangle, \tag{3.30}$$

$$|h^{\pi_{k+1}(s)}(s) - h^{\pi_k(s)}(s)| \leq M_h \|\pi_{k+1}(s) - \pi_k(s)\|_\infty, \tag{3.31}$$

$$\langle \delta_k(s), \pi_{k+1}(s) - a \rangle = \langle \delta_k^{det}(s) + \delta_k^{sto}(s), \pi_{k+1}(s) - a \rangle$$
$$= \langle \delta_k^{det}(s), \pi_{k+1}(s) - \pi^*(s) \rangle + \langle \delta_k^{sto}(s), \pi_{k+1}(s) - \pi_k(s) \rangle$$
$$+ \langle \delta_k^{sto}(s), \pi_k(s) - a \rangle, \tag{3.32}$$

11

we have

$$\psi^{\pi_k}(s, \pi_{k+1}(s)) + h^{\pi_{k+1}(s)}(s) - h^{\pi_k(s)}(s) + \langle \delta_k(s), \pi_{k+1}(s) - a \rangle + \frac{1}{\eta_k} D(\pi_k(s), \pi_{k+1}(s))$$

$$\geq -[\|Q_0^{\pi_k}(s)\|_\infty + M_h) + \|\delta_k^{sto}(s)\|_\infty] \|\pi_{k+1}(s) - \pi_k(s)\|_1 + \frac{1}{2\eta_k} \|\pi_k(s) - \pi_{k+1}(s)\|_1^2$$

$$\quad - \|\delta_k^{det}(s)\|_\infty + \langle \delta_k^{sto}(s), \pi_k(s) - a \rangle$$

$$\geq -\frac{\eta_k}{2} [\|Q_0^{\pi_k}(s)\|_\infty + M_h + \|\delta_k^{sto}(s)\|_\infty]^2 - \|\delta_k^{det}(s)\|_\infty + \langle \delta_k^{sto}(s), \pi_k(s) - a \rangle$$

$$\geq -\frac{\eta_k}{2} [\frac{\bar{c}}{1-\gamma} + M_h + \|\delta_k^{sto}(s)\|_\infty]^2 - \|\delta_k^{det}(s)\|_\infty + \langle \delta_k^{sto}(s), \pi_k(s) - a \rangle.$$

The result then follows by using the above relation in (3.29) and rearranging the terms. ∎

Observe that in the proof of Proposition 3.27, we bound the term $\psi^{\pi_k}(s, \pi_{k+1}(s))$ and $h^{\pi_{k+1}(s)}(s) - h^{\pi_k(s)}(s)$ in (3.30) and (3.31) simply by the Cauchy Schwarz inequality and Lipschtiz continuity of h^a, respectively. These relations are sufficient to prove the sublinear rate of convergence for the PMD method. In order to prove the linear rate of convergence of the PMD method in terms of number of iterations for the stochastic setting, more sophisticated techniques to bound this term are needed as initially shown in [12].

THEOREM 3.6. *Assume that the stepsize η_k satisfies*

$$\frac{\beta_k}{\eta_k} \leq \beta_{k-1}(\mu_h + \frac{1}{\eta_{k-1}}), k \geq 1, \tag{3.33}$$

for some $\beta_k \geq 0$. Then

$$[\textstyle\sum_{t=0}^{k-1} \beta_t]^{-1} \left\{ (1-\gamma) \textstyle\sum_{t=0}^{k-1} \beta_t \mathbb{E}[f(\pi_t) - f^*] + \beta_{k-1}(\mu_h + \frac{1}{\eta_{k-1}}) \mathbb{E}[\mathcal{D}(\pi_k, \pi^*)] \right\}$$

$$\leq [\textstyle\sum_{t=0}^{k-1} \beta_t]^{-1} \left\{ \frac{\beta_0}{\eta_0} \mathcal{D}(\pi_0, \pi^*) + \textstyle\sum_{t=0}^{k-1} \beta_t \eta_t (\frac{\bar{c}}{1-\gamma} + M_h + \sigma^2) \right\} + \varsigma.$$

where $\mathcal{D}(\pi, \pi') := \mathbb{E}_{s \sim \nu^*}[D(\pi(s), \pi'(s))]$.

Proof. Fixing $a = \pi^*(s)$, taking expectation w.r.t. $s \sim \nu^*$ for (3.27), and using (3.25)-(3.26), we have

$$\mathbb{E}_{s \sim \nu^*} \left[-\psi^{\pi_k}(s, \pi^*(s)) + h^{\pi_k(s)}(s) - h^{\pi^*(s)}(s) + (\mu_h + \frac{1}{\eta_k}) D(\pi_{k+1}(s), \pi^*(s)) \right]$$

$$\leq \frac{1}{\eta_k} \mathbb{E}_{s \sim \nu^*} D(\pi_k(s), \pi^*(s)) + \frac{\eta_k}{2} [\frac{\bar{c}}{1-\gamma} + M_h + \mathbb{E}_{s \sim \nu^*} \|\delta_k^{sto}(s)\|_\infty]^2 + \varsigma$$

$$\quad - \mathbb{E}_{s \sim \nu^*}[\langle \delta_k^{sto}(s), \pi_k(s) - \pi^*(s) \rangle], \tag{3.34}$$

Taking conditional expectation of (3.27) w.r.t. ξ_k, it then follows from Lemma 2.2, (??) and $\mathbb{E}_{\xi_k}[\delta_k^{sto}(s)] = 0$ that

$$(1-\gamma)\mathbb{E}_{s \sim \nu^*} \left\{ [V^{\pi_k}(s) - V^{\pi^*}(s)] + (\mu_h + \frac{1}{\eta_k}) \mathbb{E}_{\xi_k}[\mathbb{E}_{s \sim \nu^*} D(\pi_{k+1}(s), \pi^*(s))] \right\}$$

$$\leq \frac{1}{\eta_k} \mathbb{E}_{s \sim \nu^*} \left\{ D(\pi_k(s), \pi^*(s)) + \eta_k[\frac{\bar{c}}{1-\gamma} + M_h + \sigma^2] + \varsigma \right\}.$$

Taking full expectation w.r.t. F_k and the β_k-weighted sum of the above inequalities, it then follows from (3.33) and the definition of f in (2.8) that

$$(1-\gamma) \textstyle\sum_{t=0}^{k-1} \beta_t \mathbb{E}[f(\pi_t) - f^*] + \beta_{k-1}(\mu_h + \frac{1}{\eta_{k-1}}) \{ \mathbb{E}[\mathbb{E}_{s \sim \nu^*} D(\pi_k(s), \pi^*(s))] \}$$

$$\leq \frac{\beta_0}{\eta_0} \mathbb{E}_{s \sim \nu^*} D(\pi_0(s), \pi^*(s)) + \textstyle\sum_{t=0}^{k-1} \beta_t \left[\eta_t (\frac{\bar{c}}{1-\gamma} + M_h + \sigma^2) + \varsigma \right],$$

which implies the result follows immediately. ∎

We specify the selection of η_k and β_k in the following result which follows immediately from Theorem 3.6 by straightforward calculations.

COROLLARY 3.7.

a) *If $\mu_h = 0$, the number of iterations k is given a priori, and*

$$\eta_t = \sqrt{\frac{\mathcal{D}(\pi_1, \pi^*)}{k[\bar{c}/(1-\gamma) + M_h + \sigma^2]}} \quad and \quad \beta_t = 1, t = 0, \ldots, k-1,$$

then

$$(1-\gamma)\frac{1}{k} \textstyle\sum_{t=0}^{k-1} [f(\pi_t) - f^*] \leq 2\sqrt{\frac{\mathcal{D}(\pi_1, \pi^*)[\bar{c}/(1-\gamma) + M_h + \sigma^2]}{k}} + \varsigma.$$

12

b) If $\mu_h > 0$ and

$$\eta_t = \frac{1}{\mu_h t} \quad and \quad \beta_t = 1, t = 0, \ldots, k-1,$$

then

$$(1-\gamma)\frac{1}{k}\sum_{t=0}^{k-1}[f(\pi_t) - f^*] + \mu_h\mathbb{E}[\mathcal{D}(\pi_{k+1}, \pi^*)]$$
$$\leq \frac{1}{k}\left\{\mu_h\mathcal{D}(\pi_1, \pi^*) + (\frac{\bar{c}}{1-\gamma} + M_h + \sigma^2)\log k\right\} + \varsigma.$$

c) If $\mu_h > 0$ and

$$\eta_t = \frac{2}{\mu_h t} \quad and \quad \beta_t = t + 1, t = 1, 2, \ldots,$$

then

$$\frac{2(1-\gamma)}{k(k+1)}\sum_{t=0}^{k}(t+1)\mathbb{E}[f(\pi_k) - f^*] + \mu_h\mathbb{E}[\mathcal{D}(\pi_k, \pi^*)]$$
$$\leq \frac{2}{k(k+1)}\mu_h\mathcal{D}(\pi_1, \pi^*) + \frac{8}{k+1}(\frac{\bar{c}}{1-\gamma} + M_h + \sigma^2) + \varsigma.$$

In view of the previous result, we can randomly select the output solution π_R from $\{\pi_0, \ldots, \pi_{k-1}\}$ with $\text{Prob}\{R = t\} = \beta_t/\sum_{t=0}^{k-1}\beta_t$, $t = 0, \ldots, k-1$. Observe that both part a) and part b) also provides a so-called regret bound, i.e., $\sum_{k=0}^{k-1}[f(\pi_t) - f^*]$, for RL problems.

3.4. PMD for general state and continuous action spaces. In this subsection, we consider RL problems with continuous action spaces. Different from finite action spaces, the augmented-value function $L^{\pi_k}(s, a)$ and its approximation $\tilde{L}(s, a; \theta_k)$ are not necessarily linear w.r.t. to $a \in \mathcal{A}$ anymore. Instead, we assume that $\tilde{L}(s, a; \theta_k)$ is given in the form of

$$\tilde{L}(s, a; \theta_k) = \tilde{Q}(s, a; \theta_k) + \langle\tilde{\nabla}\omega(s; \theta_k), a\rangle, \tag{3.35}$$

where $\tilde{Q}(s, a; \theta_k)$ and $\tilde{\nabla}\omega(s; \theta_k)$ approximate $Q^{\pi_k}(s, a)$ and $\nabla\omega(\pi_k(s))/\eta_k$, respectively. We assume that $\tilde{Q}(s, a; \theta)$ is weakly convex, i.e.,

$$\tilde{Q}(s, a; \theta) + \mu_{\tilde{Q}}D(\pi_0(s), a) \tag{3.36}$$

is convex w.r.t. $a \in \mathcal{A}$ for some $\mu_{\tilde{Q}} \geq 0$ and denote

$$\tilde{\mu}_d = \mu_h - \mu_{\tilde{Q}}.$$

Moreover, we assume that $\tilde{Q}(s, \cdot; \theta)$ and $Q^{\pi}(s, \cdot)$ are Lipschitz continuous s.t.

$$|\tilde{Q}(s, a_1; \theta) - \tilde{Q}(s, a_2; \theta)| \leq M_{\tilde{Q}}\|a_1 - a_2\|, \forall a_1, a_2 \in \mathcal{A}, \tag{3.37}$$
$$|Q^{\pi}(s, a_1) - Q^{\pi}(s, a_1)| \leq M_Q\|a_1 - a_2\|, \forall a_1, a_2 \in \mathcal{A}. \tag{3.38}$$

Note that the Lipschitz constant of $L^{\pi_k}(s, a)$ can be very large, since η_k can possibly be small. That explains why we need to handle the two terms $\tilde{Q}(s, a; \theta_k)$ and $\langle\tilde{\nabla}\omega(s; \theta_k), a\rangle$ in the definition of $\tilde{L}$ separately.

In order to find θ_k, we can possibly formulate the following statistical learning/estimation problem

$$\min_{\theta\in\Theta}\left\{\mathbb{E}_{(s,a)}[(\tilde{Q}(s, a; \theta) - Q^{\pi_k}(s, a))^2 + \|\tilde{\nabla}\omega(s; \theta) - \nabla\omega(\pi_k(s))/\eta_k\|_*^2]\right\}, \tag{3.39}$$

where the expectation is taken w.r.t. (s, a) according to a given distribution function and $\|\cdot\|_*$ is the dual norm of a given arbitrary norm $\|\cdot\|$ of the action space. This problem is also convex if $\tilde{Q}(s, a; \theta)$ and $\tilde{\nabla}\omega(s; \theta)$ are linear w.r.t. θ. Let us denote

$$\delta_k^Q(s, a) := \tilde{Q}(s, a; \theta_k) - Q^{\pi_k}(s, a), \tag{3.40}$$
$$\delta_k^\omega(s) = \tilde{\nabla}\omega(s; \theta_k) - \frac{1}{\eta_k}\nabla\omega(\pi_k(s)) \tag{3.41}$$
$$= \underbrace{\mathbb{E}[\tilde{\nabla}\omega(s; \theta_k)] - \frac{1}{\eta_k}\nabla\omega(\pi_k(s))}_{\delta_k^{\omega,det}(s)} + \underbrace{\tilde{\nabla}\omega(s; \theta_k) - \mathbb{E}[\tilde{\nabla}\omega(s; \theta_k)]}_{\delta_k^{\omega,sto}(s)}. \tag{3.42}$$

13

We start with the convex case when $\tilde{\mu}_d \geq 0$. In this case, we need to make the following assumption about the approximation error $\delta_k^Q(s, a)$ and $\delta_k^\omega(s)$.

$$\mathbb{E}_{\xi_k} \left\{ \mathbb{E}_{s \sim \nu^*} \left[|\delta_k(s, \pi_k(s))| + |\delta_k(s, \pi_*(s))| \right] \right\} \leq \varsigma^Q, \tag{3.43}$$

$$\mathbb{E}_{\xi_k} \left\{ \mathbb{E}_{s \sim \nu^*} \left[\|\delta_k^{\omega, det}(s)\|_* \right] \right\} \leq \varsigma^\omega, \tag{3.44}$$

$$\mathbb{E}_{\xi_k} \left\{ \mathbb{E}_{s \sim \nu^*} \left[\|\delta_k^{\omega, sto}\|_*^2 \right] \right\} \leq (\sigma^\omega)^2. \tag{3.45}$$

THEOREM 3.8. *Suppose (3.43)-(3.45) hold. Also assume that the stepsize η_k satisfies*

$$\frac{\beta_k}{\eta_k} \leq \beta_{k-1}(\tilde{\mu}_d + \frac{1}{\eta_{k-1}}), k \geq 1, \tag{3.46}$$

for some $\beta_k \geq 0$. Then

$$[\textstyle\sum_{t=0}^{k-1} \beta_t]^{-1} \left\{ (1-\gamma) \sum_{t=0}^{k-1} \beta_t \mathbb{E}[f(\pi_t) - f^*] + \beta_{k-1}(\tilde{\mu}_d + \frac{1}{\eta_{k-1}})\mathbb{E}[\mathcal{D}(\pi_k, \pi^*)] \right\}$$
$$\leq [\textstyle\sum_{t=0}^{k-1} \beta_t]^{-1} \left\{ \frac{\beta_0}{\eta_0}\mathcal{D}(\pi_0, \pi^*) + \sum_{t=0}^{k-1} \beta_t \eta_t [(2M_Q + M_{\tilde{Q}} + M_h)^2 + (\sigma^\omega)^2] \right\} + \varsigma^Q + \varsigma^\omega \bar{D}_{\mathcal{A}},$$

where

$$\bar{D}_{\mathcal{A}} := \max_{a_1, a_2 \in \mathcal{A}} D(a_1, a_2). \tag{3.47}$$

Proof. Denote $\delta(s, a) := \Psi_k(s, a) - L_k(s, a)$. By the optimality condition of (3.18), we have

$$\tilde{L}(s, \pi_{k+1}(s); \theta_k) - \tilde{L}(s, a; \theta_k) + h^{\pi_{k+1}(s)}(s) - h^a(s) + \frac{1}{\eta_k}[\omega(\pi_{k+1}(s)) - \omega(a)] + (\tilde{\mu}_d + \frac{1}{\eta_k})D(\pi_{k+1}(s), a) \leq 0,$$

which implies that

$$L_k(s, \pi_{k+1}(s)) - L_k(s, a) + [h^{\pi_{k+1}(s)}(s) - h^a(s)] + \frac{1}{\eta_k}[\omega(\pi_{k+1}(s)) - \omega(a)] + (\tilde{\mu}_d + \frac{1}{\eta_k})D(\pi_{k+1}(s), a)$$
$$+ \delta_k^Q(s, \pi_{k+1}(s)) - \delta_k^Q(s, a) + \langle \delta_k^\omega(s), \pi_{k+1}(s) - a \rangle \leq 0. \tag{3.48}$$

Similarly to (3.29), we can show that

$$\psi^{\pi_k}(s, \pi_{k+1}(s)) - \psi^{\pi_k}(s, a) + h^{\pi_{k+1}(s)}(s) - h^{\pi_k(s)}(s) + \frac{1}{\eta_k}[D(\pi_k(s), \pi_{k+1}(s)) - D(\pi_k(s), a)]$$
$$+ (\tilde{\mu}_d + \frac{1}{\eta_k})D(\pi_{k+1}(s), a) + \delta_k^Q(s, \pi_{k+1}(s)) - \delta_k^Q(s, a) + \langle \delta_k^\omega(s), \pi_{k+1}(s) - a \rangle \leq 0. \tag{3.49}$$

Noting that

$$\psi^{\pi_k}(s, \pi_{k+1}(s)) = Q^{\pi_k}(s, \pi_{k+1}(s)) - Q^{\pi_k}(s, \pi_k(s)) \geq -M_Q\|\pi_{k+1}(s) - \pi_k(s)\|, \tag{3.50}$$

$$h^{\pi_{k+1}(s)}(s) - h^{\pi_k(s)}(s) \geq -M_h\|\pi_{k+1}(s) - \pi_k(s)\|, \tag{3.51}$$

$$\delta_k^Q(s, \pi_{k+1}(s)) = \delta_k^Q(s, \pi_{k+1}(s)) - \delta_k^Q(s, \pi_k(s)) + \delta_k^Q(s, \pi_k(s))$$
$$\geq -(M_Q + M_{\tilde{Q}})\|\pi_{k+1}(s) - \pi_k(s)\| + \delta_k^Q(s, \pi_k(s)), \tag{3.52}$$

$$\langle \delta_k^\omega(s), \pi_{k+1}(s) - a \rangle = \langle \delta_k^{\omega, det}(s), \pi_{k+1}(s) - a \rangle + \langle \delta_k^{\omega, sto}(s), \pi_k(s) - a \rangle + \langle \delta_k^{\omega, sto}(s), \pi_{k+1}(s) - \pi_k(s) \rangle$$
$$\geq -\|\delta_k^{\omega, det}(s)\|_* \bar{D}_{\mathcal{A}} - \|\delta_k^{\omega, sto}(s)\|_* \|\pi_{k+1}(s) - \pi_k(s)\| + \langle \delta_k^{\omega, sto}(s), \pi_k(s) - a \rangle, \tag{3.53}$$

we then conclude from (3.49) and Young's inequality that

$$-\psi^{\pi_k}(s, a) + (\tilde{\mu}_d + \frac{1}{\eta_k})D(\pi_{k+1}(s), a) \leq \frac{1}{\eta_k}D(\pi_k(s), a) + \frac{\eta_k}{2}(2M_Q + M_{\tilde{Q}} + M_h + \|\delta_k^{\omega, sto}(s)\|_*)^2$$
$$- \delta_k^Q(s, \pi_k(s)) + \delta_k^Q(s, a) + \|\delta_k^{\omega, det}(s)\|_* \bar{D}_{\mathcal{A}} - \langle \delta_k^{\omega, sto}(s), \pi_k(s) - a \rangle.$$

The rest of the proof is similar to that in Theorem 3.6. ∎

The following result follows immediately from Theorem 3.6 by straightforward calculations, and it resembles Corollary 3.7 for RL with finite action spaces.

COROLLARY 3.9.

a) If $\tilde{\mu}_d = 0$, the number of iterations k is given a priori, and

$$\eta_t = \sqrt{\frac{\mathcal{D}(\pi_1, \pi^*)}{k[(2M_Q + M_{\tilde{Q}} + M_h)^2 + (\sigma^\omega)^2]}} \quad and \quad \beta_t = 1, t = 0, \ldots, k-1,$$

then

$$(1-\gamma)\frac{1}{k}\sum_{t=0}^{k-1}[f(\pi_t) - f^*] \leq 2\sqrt{\frac{\mathcal{D}(\pi_0, \pi^*)[(2M_Q + M_{\tilde{Q}} + M_h)^2 + (\sigma^\omega)^2]}{k}} + \varsigma^Q + \varsigma^\omega \bar{D}_{\mathcal{A}}.$$

b) If $\tilde{\mu}_d > 0$ and

$$\eta_t = \frac{1}{\tilde{\mu}_d t} \quad and \quad \beta_k = 1, k = 1, \ldots, N,$$

then

$$(1-\gamma)\frac{1}{k}\sum_{t=0}^{k-1}[f(\pi_t) - f^*] + \tilde{\mu}_d \mathbb{E}[\mathcal{D}(\pi_k, \pi^*)]$$
$$\leq \frac{1}{k}\left\{\tilde{\mu}_d \mathcal{D}(\pi_1, \pi^*) + [(2M_Q + M_{\tilde{Q}} + M_h)^2 + (\sigma^\omega)^2]\log k\right\} + \varsigma^Q + \varsigma^\omega \bar{D}_{\mathcal{A}}.$$

c) If $\tilde{\mu}_d > 0$ and

$$\eta_k = \frac{2}{\tilde{\mu}_d t} \quad and \quad \beta_t = t+1, t = 0, 1, \ldots,$$

then

$$\frac{2(1-\gamma)}{k(k+1)}\sum_{k=0}^{k-1}(t+1)\mathbb{E}[f(\pi_t) - f^*] + \tilde{\mu}_d \mathbb{E}[\mathcal{D}(\pi_k, \pi^*)]$$
$$\leq \frac{2}{k(k+1)}\tilde{\mu}_d \mathcal{D}(\pi_1, \pi^*) + \frac{8}{k+1}[(2M_Q + M_{\tilde{Q}} + M_h)^2 + (\sigma^\omega)^2] + \varsigma^Q + \varsigma^\omega \bar{D}_{\mathcal{A}}.$$

We now consider the case when $\tilde{\mu}_d < 0$. In this case, our goal is to find an approximate stationary point and we need to assume estimation error terms $\delta_k^{\omega, det}(s)$ and $\delta_k^{\omega, sto}(s)$ have bounded support, i.e.,

$$\|\delta_k^{\omega, det}(s)\|_* \leq \bar{\varsigma}^\omega \quad and \quad \|\delta_k^{\omega, sto}(s)\|_* \leq \bar{\sigma}^\omega. \tag{3.54}$$

On the other hand, we do not need to make any assumptions about the error term $\delta_k^Q(s, a)$ as long as it is Lipschitz continuous w.r.t. a.

THEOREM 3.10. *Let the number of iterations k is fixed a priori. Suppose that (3.54) holds. Also assume that $\tilde{\mu}_d < 0$ and that*

$$\eta_t = \eta = \min\{|\tilde{\mu}_d|/2, 1/\sqrt{k}\}, t = 0, \ldots, k-1.$$

Then for any $s \in \mathcal{S}$, there exist iteration indices $\bar{k}(s)$ found by running k iterations of the PMD method s.t.

$$-\psi^{\pi_{\bar{k}(s)}}(s, \pi_{\bar{k}(s)+1}(s))] \leq \frac{1}{k}[V^{\pi_0}(s) - V^{\pi^*}(s)] + \frac{\gamma}{1-\gamma}\left[\frac{(M_h + M_Q + M_{\tilde{Q}} + \bar{\sigma}^\omega)^2}{\sqrt{k}} + \bar{\varsigma}^\omega \bar{D}_{\mathcal{A}}\right], \tag{3.55}$$

$$\frac{1}{2\eta}D(\pi_{\bar{k}(s)}, \pi_{\bar{k}(s)+1}(s)) + (\tilde{\mu}_d + \frac{1}{\eta})D(\pi_{\bar{k}(s)+1}(s), \pi_{\bar{k}(s)}(s))$$
$$\leq \frac{1}{k}[V^{\pi_0}(s) - V^{\pi^*}(s)] + \frac{1}{1-\gamma}\left[\frac{(M_h + M_Q + M_{\tilde{Q}} + \bar{\sigma}^\omega)^2}{\sqrt{k}} + \bar{\varsigma}^\omega \bar{D}_{\mathcal{A}}\right]. \tag{3.56}$$

Proof. Setting $a = \pi_k(s)$ in (3.49), and using the facts that $\psi^{\pi_k}(s, \pi_k(s)) = 0$ and $D(\pi_k(s), \pi_k(s)) = 0$, we obtain

$$\psi^{\pi_k}(s, \pi_{k+1}(s)) + h^{\pi_{k+1}(s)}(s) - h^{\pi_k}(s) + \frac{1}{\eta_k}D(\pi_k(s), \pi_{k+1}(s))$$
$$+ (\tilde{\mu}_d + \frac{1}{\eta_k})D(\pi_{k+1}(s), \pi_k(s)) + \delta_k^Q(s, \pi_{k+1}(s)) - \delta_k^Q(s, a) + \langle\delta_k^\omega(s), \pi_{k+1}(s) - a\rangle \leq 0 \leq 0.$$

15

Using the above inequality, (3.51)-(3.53), and Young's inequality, we then have

$$
\begin{aligned}
\psi^{\pi_k}&(s, \pi_{k+1}(s)) + \tfrac{1}{2\eta_k} D(\pi_k(s), \pi_{k+1}(s)) + (\tilde{\mu}_d + \tfrac{1}{\eta_k}) D(\pi_{k+1}(s), \pi_k(s)) \\
&\leq (M_h + M_Q + M_{\tilde{Q}} + \|\delta_k^{\omega, sto}(s)\|_*)\|_* \|\pi_{k+1}(s) - \pi_k(s)\| - \tfrac{1}{2\eta_k} D(\pi_k(s), \pi_{k+1}(s)) \\
&\quad + \|\delta_k^{\omega, det}(s)\|_* \bar{D}_{\mathcal{A}} \\
&\leq \eta_k (M_h + M_Q + M_{\tilde{Q}} + \|\delta_k^{\omega, sto}(s)\|_*)^2 + \|\delta_k^{\omega, det}(s)\|_* \bar{D}_{\mathcal{A}}. \\
&\leq \eta_k (M_h + M_Q + M_{\tilde{Q}} + \bar{\sigma}^\omega)^2 + \bar{\varsigma}^\omega \bar{D}_{\mathcal{A}}.
\end{aligned} \tag{3.57}
$$

Similar to (4.10), we can show that

$$
\begin{aligned}
V^{\pi_{k+1}}(s) - V^{\pi_k}(s) &= \tfrac{1}{1-\gamma} \int \psi^{\pi_k}(q, \pi_{k+1}(q))] \kappa_s^{\pi_{k+1}}(dq) \\
&= \tfrac{1}{1-\gamma} \int \psi^{\pi_k}(q, \pi_{k+1}(q)) - \left[\eta_k (M_h + M_Q + M_{\tilde{Q}} + \bar{\sigma}^\omega)^2 + \bar{\varsigma}^\omega \bar{D}_{\mathcal{A}} \right] \kappa_s^{\pi_{k+1}}(dq) \\
&\quad + \tfrac{1}{1-\gamma} \left[\eta_k (M_h + M_Q + M_{\tilde{Q}} + \bar{\sigma}^\omega)^2 + \bar{\varsigma}^\omega \bar{D}_{\mathcal{A}} \right] \\
&\leq \psi^{\pi_k}(s, \pi_{k+1}(s)) + \tfrac{\gamma}{1-\gamma} \left[\eta_k (M_h + M_Q + M_{\tilde{Q}} + \bar{\sigma}^\omega)^2 + \bar{\varsigma}^\omega \bar{D}_{\mathcal{A}} \right]
\end{aligned} \tag{3.58}
$$

The result in (3.55) follows by taking the telescopic sum of (3.58), while the one in (3.56) follows directly from (3.55) and (3.57). ∎

Observe that the assumptions in (3.54) are needed to provide a constant upper bound in (3.57). Otherwise, if we only make the weaker assumptions in (3.44) and (3.45), the relation in (3.58) does not necessarily hold since the distribution $\kappa_s^{\pi_{k+1}}$ depends on the random sample ξ_k generated in the k-th iteration. In the next section, we will show that this issue can possibly be addressed by using the policy dual averaging method.

4. Policy Dual Averaging. Inspired by the dual averaging method for convex optimization, we present in this section a new class of policy gradient type method, referred to as policy dual averaging (PDA), for reinforcement learning. Similar to the previous section, we will first present the generic algorithmic scheme in Subsection 4.1 and discuss the incorporation of function approximation into PDA in Subsection 4.2. The convergence properties of PDA under stochastic policy evaluation will be established in Subsections 4.3 and 4.4, respectively, for RL problems with finite and continuous action spaces.

4.1. The Generic Algorithmic Scheme. In this section, we assume that we have access to exact action-value function Q^π for a given policy π. We present a basic policy dual averaging method for RL and establish its convergence properties.

Let $\| \cdot \|$ be given norm in action space $\mathcal{A}$ and $\omega : \mathcal{A} \to \mathbb{R}$ be a strongly convex function w.r.t. to given$\| \cdot \|$ and $D(\cdot, \cdot)$ be its associated Bregman's distance defined in (1.1). At the k-iteration of the PDA method (see Algorithm 3), we minimize Ψ_k defined as the weighted sum of the advantage functions plus the regularization term $\lambda_k D(\pi_0(s), a)$ (see (4.1)). For notational convenience, we set $\Psi_{-1}(s, a) = \lambda_{-1} D(\pi_0(s), a)$ for some $\lambda_{-1} \geq 0$. Since $\psi^{\pi_t}(s, a)$ differs from $Q^{\pi_t}(s, a) + h^a(s)$ by a constant independent of a, the update in (4.1) is equivalent to (4.2), and we only need to evaluate the action-value function $Q^{\pi_k}(s, a)$ at each iteration. Different from the PMD method in the previous section, the prox-center $\pi_0(s)$ in the regularization term $\lambda_k D(\pi_0(s), a)$ does not change over different iterations. This fact makes it more convenient to implement function approximation techniques in PDA. While it is easy to see that PDA is equivalent to PMD under certain specific settings (e.g. in the tabular setting with KL divergence as the Bregman's distance), the algorithmic scheme of PDA is different from PMD in general, and its convergence for solving general RL problems has not been studied before in the literature.

16

Algorithm 3 The policy dual averaging (PDA) method

Input: $\beta_k \geq 0$, $\lambda_k \geq 0$ and initial policy $\pi_0(s) \in \mathcal{A}$.
for $k = 0, 1, \ldots,$ **do**

$$\pi_{k+1}(s) = \underset{a \in \mathcal{A}}{\arg\min} \left\{ \Psi_k(s, a) := \sum_{t=0}^{k} \beta_t \psi^{\pi_t}(s, a) + \lambda_k D(\pi_0(s), a) \right\} \tag{4.1}$$

$$= \underset{a \in \mathcal{A}}{\arg\min} \left\{ \sum_{t=0}^{k} \beta_t [Q^{\pi_t}(s, a) + h^a(s)] + \lambda_k D(\pi_0(s), a) \right\}, \forall s \in \mathcal{S}. \tag{4.2}$$

end for

In the remaining part of this subsection, our goal is to show that PDA achieves mostly comparable performance to the PMD method under exact policy evaluation. However, the analysis of PDA is significantly different from PMD as well as the dual averaging method for convex optimization.

Similar to the previous section, we need to assume that $Q^\pi(s, \cdot)$ is μ_Q-weakly convex and define $\mu_d := \mu_h - \mu_Q$ (see (3.2)). We say that the advantage function $\psi^\pi(s, \cdot)$ is convex if $\mu_d \geq 0$, and weakly convex otherwise. Then, for some properly chosen $\lambda_k \geq 0$, the subproblem in (4.1) is strongly convex and has a unique solution characterized by the following result.

LEMMA 4.1. *Suppose that $\{\beta_t\}$ and λ_k are chosen such that*

$$\mu_k := \mu_d \sum_{t=0}^{k} \beta_t + \lambda_k \geq 0, \forall k \geq 0. \tag{4.3}$$

Also let us denote $\mu_{-1} = \lambda_{-1}$. Then for any $k \geq -1$, we have

$$\Psi_k(s, \pi_{k+1}(s)) + \mu_k D(\pi_{k+1}(s), a) \leq \Psi_k(s, a), \forall s \in \mathcal{S}, \forall a \in \mathcal{A}. \tag{4.4}$$

Proof. Note that the result holds obviously for $k = -1$. For any $k \geq 0$, observe that by (3.1), $\Psi_k(s, \cdot)$ is strongly convex w.r.t. modulus μ_k. The result then follows immediately from the optimality condition of (4.1) (see, e.g., Lemma 3.5 of [11]). ∎

Proposition 4.2 establishes some descent property of the PDA method.

PROPOSITION 4.2. *If $\beta_k > 0$, relation (4.3) holds, and*

$$\lambda_k \geq \lambda_{k-1}, \forall k \geq 0, \tag{4.5}$$

then

$$V^{\pi_{k+1}}(s) - V^{\pi_k}(s) - \frac{\gamma(\lambda_k - \lambda_{k-1})}{(1-\gamma)\beta_k} \bar{D}_{\mathcal{A}} \leq \psi^{\pi_k}(s, \pi_{k+1}(s))$$

$$\leq -\frac{\mu_k}{\beta_k} D(\pi_{k+1}(s), \pi_k(s)) - \frac{\mu_{k-1}}{\beta_k} D(\pi_k(s), \pi_{k+1}(s)) + \frac{\lambda_k - \lambda_{k-1}}{\beta_k} \bar{D}_{\mathcal{A}} \tag{4.6}$$

for any $k \geq 0$, where $\bar{D}_{\mathcal{A}}$ is defined in (3.47)

Proof. First note that by the definition of Ψ_k in (4.1) and Lemma 4.1 (applied to the $k-1$ iteration), we have for any $k \geq 0$,

$$\beta_k \psi^{\pi_k}(s, a) = \Psi_k(s, a) - \Psi_{k-1}(s, a) - (\lambda_k - \lambda_{k-1}) D(\pi_0(s), a)$$

$$\leq \Psi_k(s, a) - \Psi_{k-1}(s, \pi_k(s)) - \mu_{k-1} D(\pi_k(s), a) - (\lambda_k - \lambda_{k-1}) D(\pi_0(s), a). \tag{4.7}$$

Setting $a = \pi_{k+1}(s)$ and using (4.5) in the above relation, we obtain

$$\beta_k \psi^{\pi_k}(s, \pi_{k+1}(s)) \leq \Psi_k(s, \pi_{k+1}(s)) - \Psi_{k-1}(s, \pi_k(s)) - \mu_{k-1} D(\pi_k(s), \pi_{k+1}(s)). \tag{4.8}$$

By the above inequality, and Lemma 4.1 (with $a = \pi_k(s)$ in (4.4)), we have

$$\beta_k \psi^{\pi_k}(s, \pi_{k+1}(s))$$

$$\leq \Psi_k(s, \pi_k(s)) - \mu_k D(\pi_{k+1}(s), \pi_k(s)) - \Psi_{k-1}(s, \pi_k(s)) - \mu_{k-1} D(\pi_k(s), \pi_{k+1}(s))$$

$$= \beta_k \psi^{\pi_k}(s, \pi_k(s)) - \mu_k D(\pi_{k+1}(s), \pi_k(s)) - \mu_{k-1} D(\pi_k(s), \pi_{k+1}(s))$$

$$\quad + (\lambda_k - \lambda_{k-1}) D(\pi_0(s), \pi_k(s))$$

$$= -\mu_k D(\pi_{k+1}(s), \pi_k(s)) - \mu_{k-1} D(\pi_k(s), \pi_{k+1}(s)) + (\lambda_k - \lambda_{k-1}) D(\pi_0(s), \pi_k(s)), \tag{4.9}$$

17

where the last two identities follows from the definition of Ψ_k and the fact that $\psi^{\pi_k}(s, \pi_k(s)) = 0$. The previous conclusion and the assumption $\beta_k > 0$ then clearly implies that

$$\psi^{\pi_k}(s, \pi_{k+1}(s)) - \tfrac{1}{\beta_k}(\lambda_k - \lambda_{k-1})\bar{D}_{\mathcal{A}} \leq -\tfrac{\mu_k}{\beta_k}D(\pi_{k+1}(s), \pi_k(s)) - \tfrac{\mu_{k-1}}{\beta_k}D(\pi_k(s), \pi_{k+1}(s)) \leq 0, \ \forall s \in \mathcal{S}. \quad (4.10)$$

By Lemma 2.1, (4.10) and the fact that $\kappa_s^{\pi_{k+1}}(\{s\}) \geq 1 - \gamma$ due to (2.9), we have

$$
\begin{aligned}
V^{\pi_{k+1}}(s) - V^{\pi_k}(s) &= \tfrac{1}{1-\gamma} \int [\psi^{\pi_k}(q, \pi_{k+1}(q)) - \tfrac{1}{\beta_k}(\lambda_k - \lambda_{k-1})\bar{D}_{\mathcal{A}}]\kappa_s^{\pi_{k+1}}(dq) + \tfrac{1}{(1-\gamma)\beta_k}(\lambda_k - \lambda_{k-1})\bar{D}_{\mathcal{A}} \\
&\leq \tfrac{1}{1-\gamma}[\psi^{\pi_k}(s, \pi_{k+1}(s)) - \tfrac{1}{\beta_k}(\lambda_k - \lambda_{k-1})\bar{D}_{\mathcal{A}}]\kappa_s^{\pi_{k+1}}(\{s\}) + \tfrac{1}{(1-\gamma)\beta_k}(\lambda_k - \lambda_{k-1})\bar{D}_{\mathcal{A}} \\
&\leq \psi^{\pi_k}(s, \pi_{k+1}(s)) + \tfrac{\gamma}{(1-\gamma)\beta_k}(\lambda_k - \lambda_{k-1})\bar{D}_{\mathcal{A}}, \quad (4.11)
\end{aligned}
$$

which together with (4.10) clearly imply the result. ∎

Below we establish the linear convergence rate of PDA for the case when the advantage function defined in (2.11) is convex w.r.t. a, i.e., $\mu_d \geq 0$ in (4.3).

THEOREM 4.3. *Suppose that $\beta_k > 0$ and that relations (4.3) and (4.5) hold for the PDA method. Then*

$$
\begin{aligned}
&\sum_{t=0}^{k-1} \left\{ \beta_t \left\{ [f(\pi_t) - f(\pi_*)] - \gamma[f(\pi_{t-1}) - f(\pi_*)] \right\} - \tfrac{\gamma(\lambda_t - \lambda_{t-1})}{(1-\gamma)\beta_t}\bar{D}_{\mathcal{A}} \right\} + \mu_{k-1}\mathcal{D}(\pi_k, \pi_*) \\
&+ \sum_{t=0}^{k-1} [\mu_{t-1}\mathcal{D}(\pi_t, \pi_{t+1})] \leq \lambda_k D(\pi_0, \pi_*), \quad (4.12)
\end{aligned}
$$

for any $k \geq 1$, where

$$\mathcal{D}(\pi_1, \pi_2) := \int D(\pi_1(s), \pi_2(s))\nu_*(ds). \quad (4.13)$$

In particular, if $\mu_d \geq 0$, $\beta_k = \gamma^{-k}$, and $\lambda_k = \lambda$ for some $\lambda > 0$ for any $k \geq 1$, then

$$
\begin{aligned}
&[f(\pi_k) - f(\pi_*)] + [(1 - \gamma^k)/(1-\gamma)\mu_d + \lambda\gamma^{k-1}]D(\pi_k, \pi_*) \\
&\leq \gamma^{k-1}\lambda D(\pi_0, \pi_*) + \gamma^k[f(\pi_0) - f(\pi_*)], \ \forall k \geq 1. \quad (4.14)
\end{aligned}
$$

If in addition, $\lambda = 0$ in the PDA algorithm, then

$$[f(\pi_k) - f(\pi_*)] + (1 - \gamma^k)/(1-\gamma)\mu_d\mathcal{D}(\pi_k, \pi_*) \leq \gamma^k[f(\pi_0) - f(\pi_*)], \ \forall k \geq 1.$$

Proof. We first show part a). By taking the telescopic sum of (4.8), we have

$$
\begin{aligned}
\sum_{t=0}^{k-1} \beta_t\psi^{\pi_t}(s, \pi_{t+1}(s)) &\leq \Psi_{k-1}(s, \pi_k(s)) - \Psi_{-1}(s, \pi_0(s)) - \sum_{t=0}^{k-1}[\mu_{t-1}D(\pi_t(s), \pi_{t+1}(s))] \\
&= \Psi_{k-1}(s, \pi_k(s)) - \sum_{t=0}^{k-1}[\mu_{t-1}D(\pi_t(s), \pi_{t+1}(s))],
\end{aligned}
$$

where the equality follows from the fact that $\Psi_{-1}(s, \pi_0(s)) = \lambda_{-1}D(\pi_0(s), \pi_0(s)) = 0$. It then follows from the above conclusion and Lemma 4.1 that

$$
\begin{aligned}
&\sum_{t=0}^{k-1} \beta_t\psi^{\pi_t}(s, \pi_{t+1}(s)) \\
&\leq \Psi_{k-1}(s, a) - \mu_{k-1}D(\pi_k(s), a) - \sum_{t=0}^{k-1}[\mu_{t-1}D(\pi_t(s), \pi_{t+1}(s))] \\
&= \sum_{t=0}^{k-1} \beta_t\psi^{\pi_t}(s, a) + \lambda_{k-1}D(\pi_0(s), a) - \mu_{k-1}D(\pi_k(s), a) \\
&\quad - \sum_{t=0}^{k-1}[\mu_{t-1}D(\pi_t(s), \pi_{t+1}(s))], \quad (4.15)
\end{aligned}
$$

where the identity follows from the definition of Ψ_k. Using (4.6) and the above inequality, we obtain

$$
\begin{aligned}
&\sum_{t=0}^{k-1} \beta_t[V^{\pi_{t+1}}(s) - V^{\pi_t}(s) - \tfrac{\gamma(\lambda_t - \lambda_{t-1})}{(1-\gamma)\beta_t}\bar{D}_{\mathcal{A}}] + \sum_{t=0}^{k-1}[\mu_{t-1}D(\pi_t(s), \pi_{t+1}(s))] \\
&\leq \sum_{t=0}^{k-1} \beta_t\psi^{\pi_t}(s, a) + \lambda_{k-1}D(\pi_0(s), a) - \mu_{k-1}D(\pi_k(s), a) \quad (4.16)
\end{aligned}
$$

18

for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Setting $a = \pi_*(s)$ in (4.16), taking expectation w.r.t. ν_* on its both sides, and using Lemma 2.2, we then obtain

$$\sum_{t=0}^{k-1} \left\{ \beta_t \int [V^{\pi_{t+1}}(s) - V^{\pi_t}(s)] \nu_*(ds) - \frac{\gamma(\lambda_t - \lambda_{t-1})}{(1-\gamma)\beta_t} \bar{D}_{\mathcal{A}} \right\} + \sum_{t=0}^{k-1} \int [\mu_{t-1} D(\pi_t(s), \pi_{t+1}(s))] \nu_*(ds)$$

$$\leq \sum_{t=0}^{k-1} \beta_t (1-\gamma) \int [V^{\pi_*}(s) - V^{\pi_t}(s)] \nu_*(ds) + \lambda_{k-1} \int D(\pi_0(s), \pi_*(s)) \nu_*(ds) - \mu_{k-1} \int D(\pi_k(s), \pi_*(s)) \nu_*(ds).$$

Using the definition of f and $\mathcal{D}$, and rearranging the terms, we have

$$\sum_{t=0}^{k-1} \left\{ \beta_t [f(\pi_{t+1}) - f(\pi_t)] - \frac{\gamma(\lambda_t - \lambda_{t-1})}{(1-\gamma)\beta_t} \bar{D}_{\mathcal{A}} \right\} + \sum_{t=0}^{k-1} [\mu_{t-1} \mathcal{D}(\pi_t, \pi_{t+1})]$$

$$\leq \sum_{t=0}^{k-1} \beta_t (1-\gamma)[f(\pi_*) - f(\pi_t)] + \lambda_{k-1} \mathcal{D}(\pi_0, \pi_*) - \mu_{k-1} \mathcal{D}(\pi_k, \pi_*),$$

which implies (4.12) after rearranging the terms. Now using (4.12) and the selection of β_k and λ_k, we have

$$\gamma^{-(k-1)}[f(\pi_k) - f(\pi_*)] - \gamma[f(\pi_0) - f(\pi_*)] + \mu_{k-1}\mathcal{D}(\pi_k, \pi_*) \leq \lambda \mathcal{D}(\pi_0, \pi_*).$$

Also note that by (4.3),

$$\mu_k = \mu_d \sum_{t=0}^{k-1} \gamma^{-t} + \lambda = \frac{\mu_d(1-\gamma^k)}{\gamma^{k-1}(1-\gamma)} + \lambda.$$

Combining the above two relations, we obtain

$$\gamma^{-(k-1)}[f(\pi_k) - f(\pi_*)] + [(1-\gamma^k)\gamma^{-(k-1)}(1-\gamma)\mu_d + \lambda]D(\pi_k, \pi_*)$$

$$\leq \lambda D(\pi_0, \pi_*) + \gamma[f(\pi_0) - f(\pi_*)],$$

from which (4.14) follows immediately. ∎

We add some remarks about Theorem 4.3. First, the PDA method achieves a linear rate of convergence with properly chosen β_k. Since β_k in Theorem 4.3 is geometrically increasing, in the limiting case the PDA will be approaching the classic policy iteration method. Below we provide a simple and alternative convergence analysis of the policy iteration method as it might of some interest in its own right.

COROLLARY 4.4. *Suppose $\pi_{k+1} = \arg\min_{a \in \mathcal{A}} \psi^{\pi_k}(s, a)$ for any $s \in \mathcal{S}$. Then we have*

$$f(\pi_{k+1}) - f(\pi_*) \leq \gamma[f(\pi_k) - f(\pi_*)].$$

Proof. By the definition of π_{k+1}, we have

$$\psi^{\pi_k}(s, \pi_{k+1}(s)) - \psi^{\pi_k}(s, a) \leq 0, \forall a \in \mathcal{A}. \tag{4.17}$$

Setting $a = \pi_k(s)$, we have $\psi^{\pi_k}(s, \pi_{k+1}(s)) \leq \psi^{\pi_k}(s, \pi_k(s)) = 0$. Therefore, similarly to (4.11), we have $V^{\pi_{k+1}}(s) - V^{\pi_k}(s) \leq \psi^{\pi_k}(s, \pi_{k+1}(s))$. Using the previous inequality in (4.17) (with $a = \pi_*(s)$), we have

$$V^{\pi_{k+1}}(s) - V^{\pi_k}(s) \leq \psi^{\pi_k}(s, \pi_*(s)).$$

Taking expectation w.r.t. $s \sim \nu_*$, using Lemma 2.2 and the definition of f, we conclude $f(\pi_{k+1}) - f(\pi_k)] \leq (1-\gamma)[f(\pi_*) - f(\pi_k)]$, from which the result immediately follows. ∎

Second, in view of Theorem 4.3, one can set $\lambda = 0$ and so it appears that adding the strongly convex term $\lambda_k D(\pi_0(s), a)$ in the definition of Ψ_k is useless. It should be noted, however, that this term will be critical in the following sense: a) for the convex settings (i.e., $\mu_d \geq 0$), this term will help with suppressing the impact of the noisy estimation of Q^{π_k} as will be demonstrated later; and b) for the nonconvex setting (i.e., $\mu_d < 0$), adding $\lambda_k D(\pi_0(s), a)$ will make the subproblem tractable and thus the policy easily computable. The following result discusses the convergence for PDA for RL problems with nonconvex advantage functions (i.e., $\mu_d < 0$).

THEOREM 4.5. *Suppose that $\mu_d < 0$. Assume that the number of iterations k is fixed a priori, $\lambda_t = \lambda = k(k+1)|\mu_d|$ and $\beta_t = t+1$, $t = 0, \ldots, k-1$. Then for any $s \in \mathcal{S}$, there exist iteration indices $\bar{k}(s)$ s.t.*

$$0 \leq -\psi^{\pi_{\bar{k}(s)}}(s, \pi_{\bar{k}(s)+1}(s)) \leq \frac{2[V^{\pi_0}(s) - V^{\pi_*}(s)]}{(1-\gamma)(k+1)}, \tag{4.18}$$

$$\frac{\mu_{\bar{k}(s)}}{\beta_{\bar{k}(s)}} D(\pi_{\bar{k}(s)+1}(s), \pi_{\bar{k}(s)}(s)) + \frac{\mu_{\bar{k}(s)-1}}{\beta_{\bar{k}(s)}} D(\pi_{\bar{k}(s)}(s), \pi_{\bar{k}(s)+1}(s)) \leq \frac{2[V^{\pi_0}(s) - V^{\pi_*}(s)]}{(1-\gamma)(k+1)}, \tag{4.19}$$

19

where $\mu_k = |\mu_d|k(k+1)/2$.

Proof. Clearly, the relation (4.3) is satisfied. It follows from (4.6) that

$$\tfrac{\beta_k}{1-\gamma}[V^{\pi_{k+1}}(s) - V^{\pi_k}(s)] \le \beta_k \psi^{\pi_k}(s, \pi_{k+1}(s)) \le 0. \tag{4.20}$$

Taking telescopic sum, we have

$$0 \le \textstyle\sum_{t=0}^{k-1}[-\beta_t \psi^{\pi_t}(s, \pi_{t+1}(s))] \le \tfrac{1}{1-\gamma}\sum_{t=0}^{k-1}\beta_t[V^{\pi_t}(s) - V^{\pi_{t+1}}(s)].$$

The indices $\bar{k}(s)$ are given by the one which minimizes $-\beta_t \psi^{\pi_t}(s, \pi_{t+1}(s))$ for any $s \in \mathcal{S}$. Also notice that

$$\begin{aligned}
\textstyle\sum_{t=0}^{k-1}\beta_t[V^{\pi_t}(s) - V^{\pi_{t+1}}(s)] &= \beta_0 V^{\pi_0(s)} + \textstyle\sum_{t=0}^{k-2}(\beta_{t+1} - \beta_t)V^{\pi_{t+1}}(s) - \beta_{k-1}V^{\pi_k}(s) \\
&= \textstyle\sum_{t=0}^{k-1}V^{\pi_t}(s) - kV^{\pi_k}(s) \\
&\le k[V^{\pi_0}(0) - V^{\pi_*}(s)]
\end{aligned} \tag{4.21}$$

where the first inequality follows from the monotonicity of V^{π_t} due to (4.20) and the second inequality holds since $V^{\pi_k}(s) \ge V^{\pi_*}(s)$. Combining the above two relations, we obtain the result in (4.18). The result in (4.19) follows immediately from (4.6) and (4.18). ∎

Comparing Theorems 4.3 and 4.5 with Theorems 3.3 and 3.4, we can see that the PDA method exhibits similar convergence behavior to the PMD method. However, possible limitation for the PDA method exists in that the relation in (4.19) does not necessarily imply the stationarity of the policy $\pi_{\bar{k}(s)}(s), s \in \mathcal{S}$, for the original problem in (2.8) as for the PMD method (see discussion after Theorem 3.4).

4.2. Function Approximation in PDA. In this subsection, we discuss how to incorporate function approximation into the PDA method described in the previous section to solve RL problems over general state spaces.

Similar to the PMD method, we also adopt the idea of using function approximation for the value functions in the PDA method. However, different from the PMD method that requires function approximation for an augmented value function $L^{\pi_k}(s, a)$, we will compute a function approximation $\tilde{Q}(s, a; \theta_k)$, parameterized by θ_k, directly for the action value function $Q^{\pi_k}(s, a)$ in (4.2), i.e.,

$$\tilde{Q}(s, a; \theta_k) \approx Q^{\pi_k}(s, a). \tag{4.22}$$

As a result, the algorithmic parameters β_k and λ_k do not appear in the subjects that we need to approximate. On the other hand, in the PMD method the stepsize η_k does enter into the definition of L^{π_k}, which may impact the magnitude of the stochastic first-order information of the policy evaluation problem (see the discussion after (3.22)). Moreover, the appearance of η_k in L^{π_k} also resulted stronger assumptions in (3.54) to guarantee the convergence of the PMD method for RL over continuous action spaces with $\tilde{\mu}_d < 0$.

Algorithm 4 The policy dual averaging method with function approximation

---

Input: $\beta_k \ge 0$, $\lambda_k \ge 0$ and initial policy $\pi_0(s) \in \mathcal{A}$.

for $k = 0, 1, \ldots,$ **do**

  Define an approximate advantage function

$$\tilde{\psi}(s, a; \theta_k) := \tilde{Q}(s, a; \theta_k) - V^{\pi_k}(s). \tag{4.23}$$

  Replace the update formulas in (4.1)-(4.2) by

$$\pi_{k+1}(s) = \underset{a \in \mathcal{A}}{\arg\min}\left\{\tilde{\Psi}_k(s, a) := \textstyle\sum_{t=0}^{k}\beta_t \tilde{\psi}(s, a; \theta_t) + \lambda_k D(\pi_0(s), a)\right\} \tag{4.24}$$

$$= \underset{a \in \mathcal{A}}{\arg\min}\left\{\textstyle\sum_{t=0}^{k}\beta_t[\tilde{Q}(s, a; \theta_t) + h^a(s)] + \lambda_k D(\pi_0(s), a)\right\}, \forall s \in \mathcal{S}. \tag{4.25}$$

end for

In the simplest form of function approximation, we are given a set of linearly independent basis functions $[q_1(s, a), q_2(s, a), \ldots, q_d(s, a)]$, and the goal is to determine the coefficients $\theta_k \in \mathbb{R}^d$ so that $\tilde{Q}(s, a; \theta_k) =$

$\sum_{i=1}^{d} \theta_k^i q_i(s, a)$. More powerful and sophisticated nonlinear function approximation methods, such as RKHS and Neural networks, can also be used to estimate $Q^{\pi_k}(s, a)$.

The equation in (4.25) then tells us how to compute the new policy π_{k+1} from the parameters $\theta_{[k]} := \{\theta_0, \ldots, \theta_k\}$. Instead of saving every iteration of the policy $\pi_k(s)$ for all $s \in \mathcal{S}$, we only save these parameters used for function approximation. Similar to the PMD method, we suggest to compute $\pi_{k+1}(s)$ using (4.25) whenever needed, e.g., inside the policy evaluation procedure. In fact, given $\theta_{[k]}$, for any $s \in \mathcal{S}$, $\pi_{k+1}(s)$ can be computed by an explicit formula in some important cases. This happens, for example, when the approximation function $\tilde{Q}(s, a; \theta_t)$ is linear w.r.t. to a and D is given by the KL divergence for solving RL problems with finite actions (see Section 4.3). In a more general setting, π_{k+1} is represented as an optimal solution of problem (4.25). With a properly chosen λ, the objective function of (4.25) is strongly convex, and hence the policy π_{k+1} is uniquely defined. Moreover, the dimension for the action space is usually relatively small, in comparison with the dimension of state space, for RL problems with continuous action space. Hence, any efficient numerical optimization procedures can be used to solve (4.25) quickly. For the sake of simplicity, we also assume throughout this section that exact solutions can be computed for (4.25).

Observe that when evaluating the term

$$\tilde{Q}^k(s, a; \theta_{[k]}) := \sum_{t=0}^{k} \beta_t \tilde{Q}(s, a; \theta_t)$$

used in (4.25), one can possibly use the recursive formula

$$\tilde{Q}^k(s, a; \theta_{[k]}) = \tilde{Q}^{k-1}(s, a; \theta_{[k-1]}) + \beta_k \tilde{Q}(s, a; \theta_k) \tag{4.26}$$

with $\tilde{Q}^0(s, a; \theta_{[-1]}) = 0$. When $\tilde{Q}(s, a; \theta)$ is a linear w.r.t. θ, the above formula further simplifies to a recursive updating on $\sum_{t=0}^{k} \beta_t \theta_t$. In this way, we can also significantly save the memory required to store the parameters $\{\theta_1, \ldots, \theta_t\}$. For the more general case (e.g., $\tilde{Q}(s, a; \theta)$ is given by a Neural network), in order to save both computation and memory, we can possibly modify the policy evaluation step in (4.22) so that we have a single set of parameters Θ_k satisfying

$$\tilde{Q}(s, a; \Theta_k) \approx \tilde{Q}(s, a; \Theta_{k-1}) + \beta_k Q^{\pi_k}(s, a) \tag{4.27}$$

at the k-th iteration with $\tilde{Q}(s, a; \Theta_{-1}) = 0$. In other words, we would apply function approximation directly for estimating $\sum_{t=0}^{k} \beta_t Q^{\pi_t}(s, a)$ in a recursive manner. However, this would require us to design a modified policy evaluation procedure and to consider the impact of the algorithmic parameters β_k for function approximation. In this section we focus only on the policy evaluation step in (4.22) for the sake of simplicity.

4.3. PDA for General State and Finite Action Spaces. In this subsection, we focus on RL with a finite set of actions $\{\mathcal{A}_1, \ldots, \mathcal{A}_m\}$. As mentioned earlier, in this case we search for an optimal randomized policy in the probability simplex $\mathcal{A} = \Delta_m$ in (2.1), with the cost and transition probability defined in (2.2). For a given iterate $\pi_k \in \mathcal{A}$, we use

$$Q_0^{\pi_k}(s, \mathcal{A}_i) := c_0(s, \mathcal{A}_i) + h^{\pi_k(s)}(s) + \gamma \int \mathcal{P}_0(ds'|s, \mathcal{A}_i) V^{\pi_k}(s'), i = 1, \ldots, m, \tag{4.28}$$

to denote the value function associated with each action $\mathcal{A}_i$. For notational convenience, we also denote $Q_0^{\pi_k}(s) = (Q_0^{\pi_k}(s, \mathcal{A}_1), \ldots, Q_0^{\pi_k}(s, \mathcal{A}_m))$. It then follows that $Q^{\pi_k}(s, a) = \langle Q_0^{\pi_k}(s), a \rangle$. We suggest to approximate $Q_0^{\pi_k}(s, \mathcal{A}_i)$ by $\tilde{Q}_0(s, \mathcal{A}_i; \theta_k), i = 1, \ldots, m$, where $\tilde{Q}_0(s, \mathcal{A}_i; \theta)$ denotes some function class parameterized by θ. Let us denote $\tilde{Q}_0(s; \theta_k) = (\tilde{Q}_0(s, \mathcal{A}_1; \theta_k), \ldots, \tilde{Q}_0(s, \mathcal{A}_m; \theta_k))$ and $\tilde{Q}(s, a; \theta_k) = \langle \tilde{Q}_0(s; \theta_k), a \rangle$ and also let $\tilde{\psi}(s, a; \theta_k)$ be defined in (4.23). Then the subproblem in (4.24)-(4.25) reduces to

$$\pi_{k+1}(s) = \underset{a \in \Delta_m}{\arg\min} \{\sum_{t=0}^{k} \beta_t [\langle \tilde{Q}_0(s; \theta_t), a \rangle + h^a(s)] + \lambda_k D(\pi_0(s), a)\}, \forall s \in \mathcal{S}, \tag{4.29}$$

It is well-known that the above problem has an explicit solution, for example, when $\omega(a) = \sum_{i=1}^{m} a_i \log a_i$ and $h^a = \lambda \operatorname{KL}(a \| \pi_0(s))$ for some $\lambda \geq 0$.

In order to find the parameter θ_k, we need to solve the following statistical learning/estimation problem

$$\min_{\theta \in \Theta} \int \|\tilde{Q}_0(s; \theta) - Q_0^{\pi_k}(s)\|_\infty^2 \rho(ds) \tag{4.30}$$

for a given distribution ρ over $\mathcal{S}$. We use ξ_k to denote these random samples generated at iteration k to solve problem (4.30), and $F_k := \{\xi_1, \ldots, \xi_k\}$ denote its associated filtration. Let θ_k^* be an optimal solution of (4.30) and θ_k be an approximate solution found by using some policy evaluation algorithms. In this section we assume that the approximation error

$$\delta_{0,k}(s) := \tilde{Q}_0(s; \theta_k) - Q_0^{\pi_k}(s) \tag{4.31}$$

can be decomposed into the following components:

$$\delta_{0,k}(s) = \underbrace{\mathbb{E}_{\xi_k}[\tilde{Q}_0(s; \theta_k)] - Q_0^{\pi_k}(s)}_{\delta_{0,k}^{det}(s)} + \underbrace{\tilde{Q}_0(s; \theta_k) - \mathbb{E}_{\xi_k}[\tilde{Q}_0(s; \theta_k)]}_{\delta_{0,k}^{sto}(s)} \tag{4.32}$$

$$= \underbrace{\tilde{Q}_0(s; \theta_k^*) - Q_0^{\pi_k}(s)}_{\delta_{0,k}^{app}(s)} + \underbrace{\mathbb{E}_{\xi_k}[\tilde{Q}_0(s; \theta_k)] - \tilde{Q}_0(s; \theta_k^*)}_{\delta_{0,k}^{bia}(s)} + \underbrace{\tilde{Q}_0(s; \theta_k) - \mathbb{E}_{\xi_k}[\tilde{Q}_0(s; \theta_k)]}_{\delta_{0,k}^{sto}(s)}, \tag{4.33}$$

where the expectation $\mathbb{E}_{\xi_k}$ denote the expectation w.r.t. ξ_k conditional on F_{k-1}, $\delta_{0,k}^{det}(s)$ represent the deterministic error which can be further decomposed to $\delta_{0,k}^{app}(s)$ and $\delta_{0,k}^{bia}(s)$, and $\delta_{0,k}^{sto}(s)$ represent errors caused by stochastic estimation. We make the following assumptions about these error terms:

$$\mathbb{E}_{s \sim \nu^*}[\|\delta_{0,k}^{det}(s)\|_\infty] \leq \varsigma, \tag{4.34}$$

$$\mathbb{E}_{\xi_k}\{\mathbb{E}_{s \sim \nu^*}[\|\delta_{0,k}^{sto}(s)\|_\infty^2]\} \leq \sigma^2, \tag{4.35}$$

for some $\varsigma \geq 0$ and $\sigma \geq 0$.

THEOREM 4.6. *Assume that* β_k *and* λ_k *satisfy*

$$\tilde{\mu}_k := \mu_h \sum_{t=1}^k \beta_t + \lambda_k \geq 0 \quad and \quad \lambda_k \geq \lambda_{k-1}. \tag{4.36}$$

Also let us denote $\bar{\beta}_k := \sum_{t=0}^{k-1} \beta_t$. *Then we have*

$$(1 - \gamma)\bar{\beta}_k^{-1} \sum_{t=0}^{k-1} \beta_t \mathbb{E}[f(\pi_{t-1}) - f(\pi_*)] + \bar{\beta}_k^{-1} \tilde{\mu}_{k-1} \mathbb{E}[\mathcal{D}(\pi_k, \pi_*)]$$
$$\leq \lambda_{k-1} \bar{\beta}_k^{-1} \mathcal{D}(\pi_0, \pi_*) + [(\tfrac{\bar{c}}{1-\gamma} + M_h)^2 + \sigma^2]\bar{\beta}_k^{-1} \sum_{t=0}^{k-1} (\beta_t^2 / \tilde{\mu}_{t-1}) + \varsigma. \tag{4.37}$$

In particular, if $\mu_h > 0$, $\beta_k = k + 1$ *and* $\lambda_k = \mu_h$, *then*

$$\tfrac{2(1-\gamma)}{k(k+1)} \sum_{t=0}^{k-1} \{(t+1)\mathbb{E}[f(\pi_t) - f(\pi_*)]\} + \mu_h \mathbb{E}[\mathcal{D}(\pi_k, \pi_*)]$$
$$\leq \tfrac{2\mu_h \mathcal{D}(\pi_0, \pi_*)}{k(k+1)} + \tfrac{4}{\mu_h(k+1)}[(\tfrac{\bar{c}}{1-\gamma} + M_h)^2 + \sigma^2] + \varsigma. \tag{4.38}$$

Moreover, if $\mu_h = 0$, $\beta_k = k + 1$ *and* $\lambda_k = \lambda(k+1)^{3/2}$ *with*

$$\lambda = \sqrt{2\mathcal{D}(\pi_0, \pi_*)}[(\tfrac{\bar{c}}{1-\gamma} + M_h)^2 + \sigma^2]^{-1/2},$$

then

$$\tfrac{2(1-\gamma)}{k(k+1)} \sum_{t=0}^{k-1} \{(t+1)\mathbb{E}[f(\pi_t) - f(\pi_*)]\}$$
$$\leq 4\left(\tfrac{\sqrt{k}}{k+1} + \tfrac{2\sqrt{k+1}}{k}\right)\sqrt{2\mathcal{D}(\pi_0, \pi_*)[(\tfrac{\bar{c}}{1-\gamma} + M_h)^2 + \sigma^2]} + \varsigma. \tag{4.39}$$

Proof. Similar to (4.15), we can show that

$$\sum_{t=0}^{k-1} \beta_t \tilde{\psi}(s, \pi_{t+1}(s), \theta_t) \leq \sum_{t=0}^{k-1} \beta_t \tilde{\psi}(s, a, \theta_t) + \lambda_{k-1} D(\pi_0(s), a) - \tilde{\mu}_{k-1} D(\pi_k(s), a)$$
$$- \sum_{t=0}^{k-1} \tilde{\mu}_{t-1} D(\pi_t(s), \pi_{t+1}(s)). \tag{4.40}$$

Using the relation

$$\tilde{\psi}(s, a; \theta_k) = \psi^{\pi_k}(s, a) + \delta_k(s, a) = \psi^{\pi_k}(s, a) + \langle \delta_{0,k}(s), a \rangle.$$

22

in (4.40), we have

$$\sum_{t=0}^{k-1} \beta_t \psi^{\pi_t}(s, \pi_{t+1}(s)) \leq \sum_{t=0}^{k-1} \beta_t[\psi^{\pi_t}(s, a) + \langle \delta_{0,t}(s), a - \pi_{t+1} \rangle] + \lambda_{k-1} D(\pi_0(s), a)$$
$$- \tilde{\mu}_{k-1} D(\pi_k(s), a) - \sum_{t=0}^{k-1} \tilde{\mu}_{t-1} D(\pi_t(s), \pi_{t+1}(s)). \tag{4.41}$$

Note that

$$\psi^{\pi_t}(s, \pi_{t+1}(s)) = \langle Q_0^{\pi_t}(s), \pi_{t+1}(s) - \pi_t(s) \rangle + h^{\pi_{t+1}(s)}(s) - h^{\pi_t(s)}(s)$$
$$\geq -(\|Q_0^{\pi_t}(s)\|_\infty + M_h)\|\pi_{t+1}(s) - \pi_t(s)\|_1$$
$$\geq -(\tfrac{\bar{c}}{1-\gamma} + M_h)\|\pi_{t+1}(s) - \pi_t(s)\|_1. \tag{4.42}$$

Also observe that

$$\langle \delta_{0,t}(s), a - \pi_{t+1}(s) \rangle = \langle \delta_{0,t}^{det}(s) + \delta_{0,t}^{sto}(s), a - \pi_{t+1}(s) \rangle$$
$$= \langle \delta_{0,t}^{det}(s), a - \pi_{t+1}(s) \rangle + \langle \delta_{0,t}^{sto}(s), \pi_t(s) - \pi_{t+1}(s) \rangle$$
$$+ \langle \delta_{0,t}^{sto}(s), a - \pi_t(s) \rangle. \tag{4.43}$$

Combining the above three relations, fixing $a = \pi_*(s)$ and rearranging the terms, we have

$$- \sum_{t=0}^{k-1} \beta_t \psi^{\pi_t}(s, \pi_*(s)) + \tilde{\mu}_{k-1} D(\pi_k(s), \pi_*(s))$$
$$\leq \lambda_{k-1} D(\pi_0(s), \pi_*(s)) + \sum_{t=0}^{k-1} \beta_t[\langle \delta_{0,t-1}^{det}(s), \pi_*(s) - \pi_{t+1}(s) \rangle + \langle \delta_{0,t}^{sto}(s), \pi_*(s) - \pi_t(s) \rangle]$$
$$+ \sum_{t=0}^{k-1} [\beta_t(\tfrac{\bar{c}}{1-\gamma} + M_h + \|\delta_{0,t}^{sto}(s))\|_\infty \|\pi_{t+1}(s) - \pi_t(s)\|_1 - \tfrac{\tilde{\mu}_{t-1}}{2} \|\pi_t(s) - \pi_{t+1}(s)\|_1^2]$$
$$\leq \lambda_{k-1} D(\pi_0(s), \pi_*(s)) + \sum_{t=0}^{k-1} \beta_t[\|\delta_{0,t}^{det}(s)\|_\infty + \langle \delta_{0,t}^{sto}(s), \pi_*(s) - \pi_t(s) \rangle]$$
$$+ \sum_{t=0}^{k-1} \tfrac{\beta_t^2}{2\tilde{\mu}_{t-1}}(\tfrac{\bar{c}}{1-\gamma} + M_h + \|\delta_{0,t}^{dev}(s)\|_\infty^2).$$

Taking expectation for s w.r.t. ν_* and then taking expectation w.r.t. F_k, and using the assumptions in (4.34)-(4.35), and Lemma 2.2, we have

$$(1-\gamma) \sum_{t=0}^{k-1} \beta_t \mathbb{E}[f(\pi_t) - f(\pi_*)] + \tilde{\mu}_{k-1} \mathbb{E}[\mathcal{D}(\pi_k, \pi_*)]$$
$$\leq \lambda_{k-1} \mathcal{D}(\pi_0, \pi_*) + \sum_{t=0}^{k-1} \beta_t \varsigma + \sum_{t=0}^{k-1} \tfrac{\beta_t^2}{\tilde{\mu}_{t-1}}[(\tfrac{\bar{c}}{1-\gamma} + M_h)^2 + \sigma^2],$$

which clearly implies (4.37). The result in (4.38) follows from straightforward calculations. Similar to (4.52), we can show that

$$\tfrac{2(1-\gamma)}{k(k+1)} \sum_{t=0}^{k-1} \{t\mathbb{E}[f(\pi_t) - f(\pi_*)]\} \leq \tfrac{2\lambda \mathcal{D}(\pi_0(s),a)\sqrt{k}}{k+1} + \tfrac{4\sqrt{k+1}}{\lambda k}[(\tfrac{\bar{c}}{1-\gamma} + M_h)^2 + \sigma^2] + \varsigma, \tag{4.44}$$

which implies (4.39) by the selection of λ. ∎

Notice that in comparison with the PMD method, we do not need to fixed the number of iterations k a priori when $\mu_h = 0$. As a result, the stepsize selection of the PDA method is more adaptive than PMD for this case. Note that an alternative way to bound $\psi^{\pi_t}(s, \pi_{t+1}(s))$ in (4.42) is to modify Proposition 4.2, similarly to how the linear convergence for the SPMD method was established in [12].

4.4. PDA for General State and Continuous Action Spaces. We now investigate the impact of the approximation error in (4.22) on the convergence of the PDA method for solving RL problems with continuous action spaces. Observe that when estimating Q^{π_k} there exist two possible types of stochasticity including: a) the sampling of (s, a) from $\mathcal{S} \times \mathcal{A}$ according to some given intitial distribution $\rho_\mathcal{S} \times \rho_\mathcal{A}$; and b) the estimation of $Q^{\pi_k}(s, a)$ through the sampling of the Markov Chain induced by π_k. Similar to previous discussions, we use ξ_k to denote these random samples generated at iteration k and $F_k := \{\xi_1, \ldots, \xi_k\}$ denote its associated filtration. Let

$$\delta_k(s, a) := \tilde{Q}(s, a; \theta_k) - Q^{\pi_k}(s, a). \tag{4.45}$$

We need to make the following assumptions on these error terms.

$$\mathbb{E}_{\xi_k}\left\{\mathbb{E}_{s\sim\nu^*}\left[|\delta_{k-1}(s,\pi_{k-1}(s))| + |\delta_{k-1}(s,\pi_*(s))|]\right]\right\} \leq \varsigma, \tag{4.46}$$

where $\mathbb{E}_{\xi_k}$ denotes the expectation w.r.t. ξ_k conditional on F_{k-1}. In addition, we also need to assume that

$$\tilde{Q}(s,a;\theta_k) + \tilde{\mu}_Q D(\pi_0(s),a)$$

is convex for some $\tilde{\mu}_Q \geq 0$ and denote

$$\tilde{\mu}_d = \mu_h - \tilde{\mu}_Q.$$

Note that $\tilde{\mu}_Q$ can be viewed as an estimation of μ_Q.

Moreover, we assume that $\tilde{Q}(s,\cdot;\theta)$ and $Q^\pi(s,a)$ are Lipschitz continuous s.t.

$$|\tilde{Q}(s,a_1;\theta) - \tilde{Q}(s,a_2;\theta)| \leq M_{\tilde{Q}}\|a_1 - a_2\|, \forall a_1, a_2 \in \mathcal{A}, \tag{4.47}$$

$$|Q^\pi(s,a_1) - Q^\pi(s,a_1)| \leq M_Q\|a_1 - a_2\|, \forall a_1, a_2 \in \mathcal{A}. \tag{4.48}$$

Note that if $\tilde{Q}$ is $M_{\tilde{Q}}$-Lipschitz continuous, then $\tilde{\psi}$ is also $M_{\tilde{Q}}$-Lipschitz continuous. Moreover, $\delta_t(s,a)$ is $(M_{\tilde{Q}} + M_Q)$-Lipschitz continuous w.r.t. $a \in \mathcal{A}$.

THEOREM 4.7. *Suppose $\tilde{\mu}_d \geq 0$ and that β_k and λ_k satisfy*

$$\tilde{\mu}_k := \tilde{\mu}_d \sum_{t=0}^k \beta_t + \lambda_k \geq 0 \quad and \quad \lambda_k \geq \lambda_{k-1}. \tag{4.49}$$

Also let us denote $\bar{\beta}_k := \sum_{t=0}^{k-1}\beta_t$. Then we have

$$(1-\gamma)\bar{\beta}_k^{-1}\sum_{t=0}^{k-1}\beta_t\mathbb{E}[f(\pi_t) - f(\pi_*)] + \bar{\beta}_k^{-1}\tilde{\mu}_{k-1}\mathbb{E}[\mathcal{D}(\pi_k,\pi_*)]$$
$$\leq \lambda_{k-1}\bar{\beta}_k^{-1}\mathcal{D}(\pi_0,\pi_*) + M_{\tilde{Q}}^2\bar{\beta}_k^{-1}\sum_{t=0}^{k-1}(\beta_t^2/\tilde{\mu}_{t-1}) + \varsigma. \tag{4.50}$$

In particular, if $\tilde{\mu}_d > 0$, $\beta_k = k + 1$ and $\lambda_k = \tilde{\mu}_d$, then

$$\frac{2(1-\gamma)}{k(k+1)}\sum_{t=0}^{k-1}\{(t+1)\mathbb{E}[f(\pi_{t-1}) - f(\pi_*)]\} + \tilde{\mu}_d\mathbb{E}[\mathcal{D}(\pi_k,\pi_*)] \leq \frac{2\tilde{\mu}_d\mathcal{D}(\pi_0(s),\pi_*)}{k(k+1)} + \frac{4M_{\tilde{Q}}^2}{\tilde{\mu}(k+1)} + \varsigma. \tag{4.51}$$

Moreover, if $\tilde{\mu}_d = 0$, $\beta_k = k + 1$ and $\lambda_k = \lambda(k+1)^{3/2}$ with $\lambda = M_{\tilde{Q}}/\sqrt{\mathcal{D}(\pi_0(s),\pi_*)}$, then

$$\frac{2(1-\gamma)}{k(k+1)}\sum_{t=0}^{k-1}\{t\mathbb{E}[f(\pi_t) - f(\pi_*)]\} \leq \frac{2\lambda\mathcal{D}(\pi_0(s),\pi_*)\sqrt{k}}{k+1} + \frac{4M_{\tilde{Q}}^2\sqrt{k+1}}{\lambda k} + \varsigma. \tag{4.52}$$

Proof. First observe that (4.40) still holds. Also note that $\tilde{\psi}(s,a,\theta_t) = \psi^{\pi_t}(s,a) + \delta_t(s,a)$ and that

$$\tilde{\psi}(s,\pi_{t+1}(s),\theta_t) = \tilde{\psi}(s,\pi_{t+1}(s),\theta_t) - \tilde{\psi}(s,\pi_t(s),\theta_t) + \tilde{\psi}(s,\pi_t(s),\theta_t)$$
$$\geq -M_{\tilde{Q}}\|\pi_{t+1}(s) - \pi_t(s)\| + \tilde{\psi}(s,\pi_t(s),\theta_t)$$
$$= -M_{\tilde{Q}}\|\pi_{t+1}(s) - \pi_t(s)\| + \psi^{\pi_t}(s,\pi_t(s)) + \delta_t(s,\pi_t(s))$$
$$= -M_{\tilde{Q}}\|\pi_{t+1}(s) - \pi_t(s)\| + \delta_t(s,\pi_t(s)).$$

Combining these relations and rearranging the terms, we obtain

$$-\sum_{t=0}^{k-1}\beta_t\psi^{\pi_t}(s,a) + \tilde{\mu}_{k-1}D(\pi_k(s),a) \leq \lambda_{k-1}D(\pi_0(s),a) + \sum_{t=0}^{k-1}\beta_t[\delta_t(s,a) - \delta_t(s,\pi_t(s))]$$
$$+ \sum_{t=0}^{k-1}\left[\beta_t M_{\tilde{Q}}\|\pi_{t+1}(s) - \pi_t(s)\| - \tilde{\mu}_{t-1}D(\pi_t(s),\pi_{t+1}(s))\right]$$
$$\leq \lambda_{k-1}D(\pi_0(s),a) + \sum_{t=0}^{k-1}\beta_t[\delta_t(s,a) - \delta_t(s,\pi_t(s))]$$
$$+ M_{\tilde{Q}}^2\sum_{t=0}^{k-1}(\beta_t^2/\tilde{\mu}_{t-1}).$$

24

Setting $a = \pi_*(s)$, taking expectation of s w.r.t. ν_* and then w.r.t. F_k, and using the assumption in (4.46) and Lemma 2.2, we then have

$$(1-\gamma)\sum_{t=0}^{k-1}\beta_t\mathbb{E}[f(\pi_t) - f(\pi_*)] + \tilde{\mu}_{k-1}\mathbb{E}[\mathcal{D}(\pi_k, \pi_*)] \leq \lambda_{k-1}\,\mathcal{D}(\pi_0(s), a) + \sum_{t=0}^{k-1}\beta_t\varsigma + M_{\tilde{Q}}^2\sum_{t=0}^{k-1}(\beta_t^2/\tilde{\mu}_{t-1}),$$

which clearly implies (4.50). The result in (4.51) follows from straightforward calculations, and (4.52) can be verified similarly in view of the observation that $\sum_{t=0}^{k-1}(\beta_t^2/\tilde{\mu}_{t-1}) = \frac{1}{\lambda}\sum_{t=0}^{k-1}\sqrt{t+1} \leq \frac{2}{3\lambda}(k+1)^{3/2}$. ∎

We now consider the case when $\mu_d < 0$. It is interesting to note that we only need to assume the Lipschitz continuity of $\tilde{Q}(s, a; \theta)$ and $Q^\pi(s, a)$ w.r.t. $a \in \mathcal{A}$ to show the convergence of PDA to a stationary point in this case.

THEOREM 4.8. *Suppose that $\mu_d < 0$. Assume that the number of iterations k is fixed a priori, $\lambda_t = \lambda = k(k+1)|\mu_d|$, and $\beta_t = t+1$, $t = 0, \ldots, k-1$. then for any $s \in \mathcal{S}$, there exist iteration indices $\bar{k}(s)$ s.t.*

$$-\frac{(M_Q + M_{\tilde{Q}})^2}{|\tilde{\mu}_d|(k+1)} \leq -\psi^{\pi_{\bar{k}s}}(s, \pi_{\bar{k}(s)+1}(s)) \leq \frac{2[V^{\pi_0}(0) - V^{\pi_*}(s)]}{k+1} + \frac{3(M_Q + M_{\tilde{Q}})^2}{(1-\gamma)|\tilde{\mu}_d|(k+1)}. \tag{4.53}$$

Moreover, we have

$$\frac{\tilde{\mu}_{\bar{k}(s)}}{2\beta_{\bar{k}(s)}}D(\pi_{\bar{k}(s)+1}(s), \pi_{\bar{k}(s)}(s)) + \frac{\tilde{\mu}_{\bar{k}(s)-1}}{2\beta_{\bar{k}(s)}}D(\pi_{\bar{k}(s)}(s), \pi_{\bar{k}(s)+1}(s))$$
$$\leq \frac{2[V^{\pi_0}(0) - V^{\pi_*}(s)]}{k+1} + \frac{3(M_Q + M_{\tilde{Q}})^2}{(1-\gamma)|\tilde{\mu}_d|(k+1)} + \frac{2(M_Q + M_{\tilde{Q}})^2}{|\tilde{\mu}_d(k+1)}, \tag{4.54}$$

where $\tilde{\mu}_t = [t(t+1)\tilde{\mu}_d/2 + k(k+1)|\tilde{\mu}_d|]$, $t = 0, \ldots, k-1$.

Proof. Similar to (4.10), we can show that

$$\tilde{\psi}(s, \pi_{k+1}(s), \theta_k) - \frac{1}{\beta_k}|\lambda_{k+1} - \lambda_k|\bar{D}_{\mathcal{A}}$$
$$\leq \tilde{\psi}(s, \pi_k(s), \theta_k) - \frac{\tilde{\mu}_k}{\beta_k}D(\pi_{k+1}(s), \pi_k(s)) - \frac{\tilde{\mu}_{k-1}}{\beta_k}D(\pi_k(s), \pi_{k+1}(s)),$$

which in view of the definition of $\delta_k(s, a)$ in (4.45) and $\psi^{\pi_k}(s, \pi_k(s)) = 0$ then implies that

$$\psi^{\pi_k}(s, \pi_{k+1}(s)) + \delta_k(s, \pi_{k+1}(s)) - \frac{1}{\beta_k}|\lambda_{k+1} - \lambda_k|\bar{D}_{\mathcal{A}}$$
$$\leq \delta_k(s, \pi_k(s)) - \frac{\tilde{\mu}_k}{\beta_k}D(\pi_{k+1}(s), \pi_k(s)) - \frac{\tilde{\mu}_{k-1}}{\beta_k}D(\pi_k(s), \pi_{k+1}(s))$$
$$\leq \delta_k(s, \pi_k(s)) - \frac{1}{4\beta_k}(\tilde{\mu}_k + \tilde{\mu}_{k-1})\|\pi_{k+1}(s) - \pi_k(s)\|^2$$
$$- \frac{\tilde{\mu}_k}{2\beta_k}D(\pi_{k+1}(s), \pi_k(s)) - \frac{\tilde{\mu}_{k-1}}{2\beta_k}D(\pi_k(s), \pi_{k+1}(s)).$$

Also note that

$$\delta_k(s, \pi_k(s)) - \delta_k(s, \pi_{k+1}(s)) - \frac{1}{4\beta_k}(\tilde{\mu}_k + \tilde{\mu}_{k-1})\|\pi_{k+1}(s) - \pi_k(s)\|^2$$
$$\leq (M_Q + M_{\tilde{Q}})\|\pi_{k+1}(s) - \pi_k(s)\| - \frac{1}{4\beta_k}(\tilde{\mu}_k + \tilde{\mu}_{k-1})\|\pi_{k+1}(s) - \pi_k(s)\|^2$$
$$\leq \frac{\beta_k(M_Q + M_{\tilde{Q}})^2}{\tilde{\mu}_k + \tilde{\mu}_{k-1}}.$$

Combining these two relations, we have

$$\psi^{\pi_k}(s, \pi_{k+1}(s)) + \frac{\tilde{\mu}_k}{2\beta_k}D(\pi_{k+1}(s), \pi_k(s)) + \frac{\tilde{\mu}_{k-1}}{2\beta_k}D(\pi_k(s), \pi_{k+1}(s))$$
$$\leq \frac{|\lambda_{k+1} - \lambda_k|\bar{D}_{\mathcal{A}}}{\beta_k} + \frac{\beta_k(M_Q + M_{\tilde{Q}})^2}{\tilde{\mu}_k + \tilde{\mu}_{k-1}}. \tag{4.55}$$

Using the above relation, we can show similarly to (4.11) that

$$V^{\pi_{k+1}}(s) - V^{\pi_k}(s) \leq \psi^{\pi_k}(s, \pi_{k+1}(s)) + \frac{\gamma|\lambda_{k+1} - \lambda_k|\bar{D}_{\mathcal{A}}}{(1-\gamma)\beta_k} + \frac{\gamma\beta_k(M_Q + M_{\tilde{Q}})^2}{(1-\gamma)(\tilde{\mu}_k + \tilde{\mu}_{k-1})}. \tag{4.56}$$

Using the fact that $\lambda_k = \lambda_{k-1}$ in the previous two relations, we conclude that

$$\beta_k[V^{\pi_{k+1}}(s) - V^{\pi_k}(s)] - \frac{\gamma\beta_k^2(M_Q + M_{\tilde{Q}})^2}{(1-\gamma)(\tilde{\mu}_k + \tilde{\mu}_{k-1})} \leq \beta_k\psi^{\pi_k}(s, \pi_{k+1}(s)) \leq \frac{\beta_k^2(M_Q + M_{\tilde{Q}})^2}{\tilde{\mu}_k + \tilde{\mu}_{k-1}}.$$

25

Rearranging terms in the above inequality, we then have

$$0 \le -\beta_k \left[\psi^{\pi_k}(s, \pi_{k+1}(s)) - \frac{\beta_k(M_Q+M_{\tilde{Q}})^2}{\tilde{\mu}_k+\tilde{\mu}_{k-1}} \right] \le \beta_k[V^{\pi_k}(s) - V^{\pi_{k+1}}(s)] + \frac{\beta_k^2(M_Q+M_{\tilde{Q}})^2}{(1-\gamma)(\tilde{\mu}_k+\tilde{\mu}_{k-1})}. \tag{4.57}$$

Moreover, summing up (4.55) and (4.56), we have

$$V^{\pi_{k+1}}(s) - V^{\pi_k}(s) \le \frac{\beta_k(M_Q+M_{\tilde{Q}})^2}{(1-\gamma)(\tilde{\mu}_k+\tilde{\mu}_{k-1})}, \tag{4.58}$$

which implies that

$$V^{\pi_k}(s) \le V^{\pi_0}(s) + \sum_{t=0}^{k-1} \frac{\beta_t(M_Q+M_{\tilde{Q}})^2}{(1-\gamma)(\tilde{\mu}_t+\tilde{\mu}_{t-1})} \le V^{\pi_0}(s) + \frac{(M_Q+M_{\tilde{Q}})^2}{2(1-\gamma)|\tilde{\mu}_d|},$$

where the last inequality follows from $\tilde{\mu}_t = \tilde{\mu}_d \sum_{j=0}^t \beta_j + \lambda_k \ge |\tilde{\mu}_d|k(k+1)/2$. It then follows from the selection of $\beta_t = t+1$ and the previous inequality that

$$\begin{aligned} \sum_{t=0}^{k-1} \beta_t[V^{\pi_t}(s) - V^{\pi_{t+1}}(s)] &= \sum_{t=0}^{k-1} V^{\pi_t}(s) - kV^{\pi_k}(s) \\ &\le k[V^{\pi_0}(0) + \frac{(M_Q+M_{\tilde{Q}})^2}{2(1-\gamma)|\tilde{\mu}_d|} - V^{\pi_k}(s)] \\ &\le k[V^{\pi_0}(0) + \frac{(M_Q+M_{\tilde{Q}})^2}{2(1-\gamma)|\tilde{\mu}_d|} - V^{\pi_*}(s)]. \end{aligned} \tag{4.59}$$

Taking the telescopic sum of (4.57), and utilizing the above bound in (4.59), we then have

$$\begin{aligned} &- \sum_{t=0}^{k-1} \beta_k \left[\psi^{\pi_k}(s, \pi_{k+1}(s)) - \frac{\beta_k(M_Q+M_{\tilde{Q}})^2}{\tilde{\mu}_k+\tilde{\mu}_{k-1}} \right] \\ &\le \sum_{t=0}^{k-1} \beta_t[V^{\pi_t}(s) - V^{\pi_{t+1}}(s)] + \sum_{t=0}^{k-1} \frac{\beta_k^2(M_Q+M_{\tilde{Q}})^2}{(1-\gamma)(\tilde{\mu}_k+\tilde{\mu}_{k-1})} \\ &\le k[V^{\pi_0}(0) + \frac{(M_Q+M_{\tilde{Q}})^2}{2(1-\gamma)|\tilde{\mu}_d|} - V^{\pi_*}(s)] + \sum_{t=0}^{k-1} \frac{\beta_k^2(M_Q+M_{\tilde{Q}})^2}{(1-\gamma)(\tilde{\mu}_k+\tilde{\mu}_{k-1})} \\ &\le k \left[V^{\pi_0}(0) - V^{\pi_*}(s) + \frac{3(M_Q+M_{\tilde{Q}})^2}{2(1-\gamma)|\tilde{\mu}_d|} \right], \end{aligned}$$

where the last inequality follows from the fact that

$$\sum_{t=0}^{k-1} \frac{\beta_t^2}{\tilde{\mu}_t+\tilde{\mu}_{t-1}} \le \sum_{t=0}^{k-1} \frac{(t+1)^2}{|\tilde{\mu}_d|k(k+1)} \le \frac{k}{|\tilde{\mu}_d|}.$$

Noticing the first inequality in (4.57), it then follows from the above inequality by choosing $\bar{k}(s)$ as the iteration with the smallest value of $- \left[\psi^{\pi_k}(s, \pi_{k+1}(s)) - \frac{\beta_k(M_Q+M_{\tilde{Q}})^2}{\tilde{\mu}_k+\tilde{\mu}_{k-1}} \right]$ that

$$- \frac{k(k+1)}{2} \left[\psi^{\pi_{\bar{k}(s)}}(s, \pi_{\bar{k}(s)+1}(s)) - \frac{\beta_{\bar{k}(s)}(M_Q+M_{\tilde{Q}})^2}{\tilde{\mu}_{\bar{k}(s)}+\tilde{\mu}_{\bar{k}(s)-1}} \right] \le k \left[V^{\pi_0}(0) - V^{\pi_*}(s) + \frac{3(M_Q+M_{\tilde{Q}})^2}{2(1-\gamma)|\tilde{\mu}_d|} \right].$$

The result in (4.53) then clearly follows from the above inequality and the fact that

$$0 \le \frac{\beta_{\bar{k}(s)}}{\tilde{\mu}_{\bar{k}(s)}+\tilde{\mu}_{\bar{k}(s)-1}} \le \frac{1}{|\tilde{\mu}_d|(k+1)}.$$

The result in (4.54) follows immediately from (4.53) and (4.55). ∎

We make a few remarks about the above results. First, Theorem 4.8 appears to stronger than the corresponding one in Theorem 3.10 for PMD in the sense that it requires much weaker assumptions. Second, while for the case $\tilde{\mu}_d \ge 0$, the number of iterations k does not need to fixed a priori for the PDA method, Theorem 4.8 does require us to fix the number of iterations in order to reduce $\lambda_{k+1} - \lambda_k$. It will be interesting to see if more adaptive selection of λ_k is possible for the PDA method when $\tilde{\mu}_d < 0$.

5. Concluding Remarks. In this paper, we study new policy optimization methods for solving Markov decision process and reinforcement learning problems over general state and action spaces. We first present a generalization of the policy mirror descent method and suggest a novel function approximation scheme into it to deal with general state spaces. We then introduce a new class of policy dual averaging methods for reinforcement learning that might be more amenable for function approximation in that it only requires the approximation of the action-value function. We establish the convergence properties of both methods under different situations, including the linear convergence to the global optimality under exact action-value function evaluation, the sublinear convergence to the global optimality up to function approximation error under certain curvature conditions, and the sublinear convergence to a stationary point up to function approximation if such curvature conditions are violated. Future research will be directed to efficient implementation and extensive numerical experimentation of these new algorithmic frameworks and the design of better schemes to estimate problem parameters and select stepsizes.

REFERENCE

[1] Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR, 2019.

[2] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv*, pages arXiv–1908.00261, 2019.

[3] Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case.* Athena Scientific, USA, 2nd edition, 1996.

[4] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv*, page arXiv:1906.01786, 2019.

[5] Semih Cayci, Niao He, and R. Srikant. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv*, page arXiv:2106.04096, 2021.

[6] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization. *arXiv e-prints*, page arXiv:2007.06558, July 2020.

[7] Eyal Even-Dar, Sham. M. Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

[8] S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, 2002.

[9] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[10] Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. *arXiv preprint arXiv:2105.01424*, 2021.

[11] G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning.* Springer Nature, Switzerland AG, 2020.

[12] G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, 2021. accepted for publication.

[13] Guanghui Lan, Yan Li, and Tuo Zhao. Block policy mirror descent. *arXiv preprint arXiv:2201.05756*, 2022.

[14] Tianjiao Li Li, Feiyang Wu, and Guanghui Lan. Stochastic first-order methods for average-reward markov decision processes. *arXiv preprint arXiv:2205.05800*, 2022.

[15] Yan Li, Tuo Zhao, and Guanghui Lan. First-order policy optimization for robust markov decision process. *arXiv preprint arXiv:2209.10579*, 2022.

[16] Yan Li, Tuo Zhao, and Guanghui Lan. Homotopic policy mirror descent: Policy convergence, implicit regularization, and improved sample complexity. *arXiv preprint arXiv:2201.09457*, 2022.

[17] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the Global Convergence Rates of Softmax Policy Gradient Methods. *arXiv e-prints*, page arXiv:2005.06392, May 2020.

[18] S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability.* 1993.

[19] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2009. Initial version: March, 2002.

[20] Georg Ch. Pflug. Derivatives of probability measures-concepts and applications to the optimization of stochastic systems. In Pravin Varaiya and Alexander B. Kurzhanski, editors, *Discrete Event Systems: Models and Applications*, pages 252–274, Berlin, Heidelberg, 1988. Springer Berlin Heidelberg.

[21] Georg Ch. Pflug. On-line optimization of simulated markovian processes. *Mathematics of Operations Research*, 15(3):381–395, 1990.

[22] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., USA, 1st edition, 1994.

[23] Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 5668–5675. AAAI Press, 2020.

[24] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[25] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[26] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2018.

[27] R.S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS'99: Proceedings of the 12th International Conference on Neural Information Processing Systems*, pages 1057–1063, 1999.

[28] L. Wang, Q. Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *ArXiv*, abs/1909.01150, 2020.

[29] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, pages 2543–2596, 2010.

[30] Lin Xiao. On the convergence rates of policy gradient methods. *arXiv preprint arXiv:2201.07443*, 2022.

[31] T. Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for actor-critic algorithms. *ArXiv*, abs/2004.12956, 2020.

[32] 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. *arXiv preprint arXiv:2105.11066*, 2021.