Noname manuscript No.
(will be inserted by the editor)

Policy Mirror Descent for Reinforcement Learning: Linear Convergence, New Sampling Complexity, and Generalized Problem Classes

Guanghui Lan

Submitted: Feb 5, 2021

Abstract We present new policy mirror descent (PMD) methods for solving reinforcement learning (RL) problems with either strongly convex or general convex regularizers. By exploring the structural properties of these overall highly nonconvex problems we show that the PMD methods exhibit fast linear rate of convergence to the global optimality. We develop stochastic counterparts of these methods, and establish an $\mathcal{O}(1/\epsilon)$ (resp., $\mathcal{O}(1/\epsilon^2)$) sampling complexity for solving these RL problems with strongly (resp., general) convex regularizers using different sampling schemes, where ϵ denote the target accuracy. We further show that the complexity for computing the gradients of these regularizers, if necessary, can be bounded by $\mathcal{O}\{(\log_\gamma \epsilon)[(1-\gamma)L/\mu]^{1/2}\log(1/\epsilon)\}$ (resp., $\mathcal{O}\{(\log_\gamma \epsilon)(L/\epsilon)^{1/2}\}$) for problems with strongly (resp., general) convex regularizers. Here γ denotes the discounting factor. To the best of our knowledge, these complexity bounds, along with our algorithmic developments, appear to be new in both optimization and RL literature. The introduction of these convex regularizers also greatly enhances the flexibility and thus expands the applicability of RL models.

1 Introduction

In this paper, we study a general class of reinforcement learning (RL) problems involving either covex or strongly convex regularizers in their cost functions. Consider the finite Markov decision process $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, c, \gamma)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $P : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 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 $\pi : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ determines the probability of selecting a particular action at a given state.

For a given policy π, we measure its performance by the action-value function (Q-function) $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defined as

$$Q^\pi(s,a) := \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t [c(s_t,a_t) + h^\pi(s_t)] \right.$$
$$\left. \mid s_0 = s, a_0 = a, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t)\right]. \tag{1.1}$$

Here h^π is a closed convex function w.r.t. the policy π, i.e., there exist some $\mu \geq 0$ s.t.

$$h^\pi(s) - [h^{\pi'}(s) + \langle (h')^{\pi'}(s,\cdot), \pi(\cdot|s) - \pi'(\cdot|s)\rangle] \geq \mu D^\pi_{\pi'}(s), \tag{1.2}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product over the action space $\mathcal{A}$, $(h')^{\pi'}(s,\cdot)$ denotes a subgradient of $h(s)$ at π', and $D^\pi_{\pi'}(s)$ is the Bregman's distance or Kullback–Leibler (KL) divergence between π and π' (see Subsection 1.1 for more discussion).

Clearly, if $h^\pi = 0$, then Q^π becomes the classic action-value function. If $h^\pi(s) = \mu D^\pi_{\pi_0}(s)$ for some $\mu > 0$, then Q^π reduces to the so-called entropy regularized action-value function. The incorporation of a more general convex regularizer h^π allows us to not only unify these two cases, but also to greatly enhance the expression power and thus the applicability of RL. For example, by using either the indicator function, quadratic penalty or barrier functions, h^π can model the set of constraints that an optimal policy should satisfy. It can describe the correlation among different actions for different states. h^π can also model some risk or utility function associated

This research was partially supported by the NSF grant 1953199 and NIFA grant 2020-67021-31526. The paper was first released at https://arxiv.org/abs/2102.00135 on 01/30/2021.

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. (email: george.lan@isye.gatech.edu).

Address(es) of author(s) should be given

with the policy π. Throughout this paper, we say that h^π is a strongly convex regularizer if $\mu > 0$. Otherwise, we call h^π a general convex regularizer. Clearly the latter class of problems covers the regular case with $h^\pi = 0$.

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

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t [c(s_t, a_t) + h^\pi(s_t)] \right. \\ \left. | \; s_0 = s, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)\right]. \tag{1.3}$$

It can be easily seen from the definitions of Q^π and V^π that

$$V^\pi(s) = \textstyle\sum_{a\in\mathcal{A}} \pi(a|s) Q^\pi(s,a) = \langle Q^\pi(s,\cdot), \pi(\cdot|s)\rangle, \tag{1.4}$$

$$Q^\pi(s,a) = c(s,a) + h^\pi(s) + \gamma \textstyle\sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a) V^\pi(s'). \tag{1.5}$$

The main objective in RL is to find an optimal policy $\pi^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ s.t.

$$V^{\pi^*}(s) \le V^\pi(s), \forall \pi(\cdot|s) \in \Delta_{|\mathcal{A}|}, \forall s \in \mathcal{S}. \tag{1.6}$$

for any $s \in \mathcal{S}$. Here $\Delta_{|\mathcal{A}|}$ denotes the simplex constraint given by

$$\Delta_{|\mathcal{A}|} := \{p \in \mathbb{R}^{|\mathcal{A}|} | \textstyle\sum_{i=1}^{|\mathcal{A}|} p_i = 1, p_i \ge 0\}, \forall s \in \mathcal{S}. \tag{1.7}$$

By examining Bellman's optimality condition for dynamic programming ([3] and Chapter 6 of [19]), we can show the existence of a policy π^* which satisfies (1.6) simultaneously for all $s \in \mathcal{S}$. Hence, we can formulate (1.6) as an optimization problem with a single objective by taking the weighted sum of V^π over s (with weights $\rho_s > 0$ and $\sum_{s\in S} \rho_s = 1$):

$$\begin{array}{ll} \min_\pi & \mathbb{E}_{s\sim\rho}[V^\pi(s)] \\ \text{s.t.} & \pi(\cdot|s) \in \Delta_{|\mathcal{A}|}, \forall s \in \mathcal{S}. \end{array} \tag{1.8}$$

While the weights ρ can be arbitrarily chosen, a reasonable selection of ρ would be the stationary state distribution induced by the optimal policy π^*, denoted by $\nu^* \equiv \nu(\pi^*)$. As such, problem (1.8) reduces to

$$\begin{array}{ll} \min_\pi & \{f(\pi) := \mathbb{E}_{s\sim\nu^*}[V^\pi(s)]\} \\ \text{s.t.} & \pi(\cdot|s) \in \Delta_{|\mathcal{A}|}, \forall s \in \mathcal{S}. \end{array} \tag{1.9}$$

It has been observed recently (eg., [14]) that one can simplify the analysis of various algorithms by setting ρ to ν^*. As we will also see later, even though the definition of the objective f in (1.9) depends on ν^* and hence the unknown optimal policy π^*, the algorithms for solving (1.6) and (1.9) do not really require the input of π^*.

Recently, there has been considerable interest in the development of first-order methods for solving RL problems in (1.8) -(1.9). While these methods have been derived under various names (e.g., policy gradient, natural policy gradient, trust region policy optimization), they all utilize the gradient information of f (i.e., Q function) in some form to guide the search of optimal policy (e.g., [21,9,7,1,20,5,23,15]). As pointed out by a few authors recently, many of these algorithms are intrinsically connected to the classic mirror descent method originally presented by Nemirovski and Yudin [17,2,16], and some analysis techniques in mirror descent method have thus been adapted to reinforcement learning [20,23,22]. In spite of the popularity of these methods in practice, a few significant issues remain on their theoretical studies. Firstly, most policy gradient methods converge only sublinearly, while many other classic algorithms (e.g., policy iteration) can converge at a linear rate due to the contraction properties of the Bellman operator. Recently, there are some interesting works relating first-order methods with the Bellman operator to establish their linear convergence [4,5]. However, in a nutshell these developments rely on the contraction of the Bellman operator, and as a consequence, they either require unrealistic algorithmic assumptions (e.g., exact line search [4]) or apply only for some restricted problem classes (e.g., entropy regularized problems [5]). Secondly, the convergence of stochastic policy gradient methods has not been well-understood in spite of intensive research effort. Due to unavoidable bias, stochastic policy gradient methods exhibit much slower rate of convergence than related methods, e.g., stochastic Q-learning.

Our contributions in this paper mainly exist in the following several aspects. Firstly, we present a policy mirror descent (PMD) method and show that it can achieve a linear rate of convergence for solving RL problems with strongly convex regularizers. We then develop a variant of PMD, namely approximate policy mirror descent (APMD) method, obtained by applying an adaptive perturbation term into PMD, and show that it can achieve a linear rate of convergence for solving RL problems with general convex regularizers. Even though the overall problem is highly nonconvex, we exploit the generalized monotonicity [6,13,11] associated with the variational inequality (VI) reformulation of (1.8)-(1.9) (see [8] for a comprehensive introduction to VI). As a consequence, our convergence analysis does not rely on the contraction properties of the Bellman operator. This fact not only enables us to define h^π as a general (strongly) convex function of π and thus expand the problem classes considered in RL, but also facilitates the study of PMD methods under the stochastic settings.

Secondly, we develop the stochastic policy mirror descent (SPMD) and stochastic approximate policy mirror descent (SAPMD) method to handle stochastic first-order information. One key idea of SPMD and SAPMD is

to handle separately the bias and expected error of the stochastic estimation of the action-value functions in our convergence analysis, since we can usually reduce the bias term much faster than the total expected error. We establish general convergence results for both SPMD and SAPMD applied to solve RL problems with strongly convex and general convex regularizers, under different conditions about the bias and expected error associated with the estimation of value functions.

Thirdly, we establish the overall sampling complexity of these algorithms by employing different schemes to estimate the action-value function. More specifically, we present an $\mathcal{O}(|\mathcal{S}||\mathcal{A}|/\mu\epsilon)$ and $\mathcal{O}(|\mathcal{S}||\mathcal{A}|/\epsilon^2)$ sampling complexity for solving RL problems with strongly convex and general convex regularizers, when one has access to multiple independent sampling trajectories. To the best of our knowledge, the former sampling complexity is new in the RL literature, while the latter one has not been reported before for policy gradient type methods. We further enhance a recently developed conditional temporal difference (CTD) method [12] so that it can reduce the bias term faster. We show that with CTD, the aforementioned $\mathcal{O}(1/\mu\epsilon)$ and $\mathcal{O}(1/\epsilon^2)$ sampling complexity bounds can be achieved in the single trajectory setting with Markovian noise under certain regularity assumptions.

Fourthly, observe that unless h^π is relatively simple (e.g., h^π does not exist or it is given as the KL divergence), the subproblems in the SPMD and SAPMD methods do not have an explicit solution in general and require an efficient solution procedure to find some approximate solutions. We establish the general conditions on the accuracy for solving these subproblems, so that the aforementioned linear rate of convergence and new sampling complexity bounds can still be maintained. We further show that if h^π is a smooth convex function, by employing an accelerated gradient descent method for solving these subproblems, the overall gradient computations for h^π can be bounded by $\mathcal{O}\{(\log_\gamma \epsilon)\sqrt{(1-\gamma)L/\mu}\log(1/\epsilon)\}$ and $\mathcal{O}\{(\log_\gamma \epsilon)\sqrt{L/\epsilon}\}$, respectively, for the case when h^π is a strongly convex and general convex function. To the best of our knowledge, such gradient complexity has not been considered before in the RL and optimization literature.

This paper is organized as follows. In Section 2, we discuss the optimality conditions and generalized monotonicity about RL with convex regularizers. Sections 3 and 4 are dedicated to the deterministic and stochastic policy mirror descent methods, respectively. In Section 5 we establish the sampling complexity bounds under different sampling schemes, while the gradient complexity of computing ∇h^π is shown in Section 6. Some concluding remarks are made in Section 7.

1.1 Notation and terminology

For any two points $\pi(\cdot|s), \pi'(\cdot|s) \in \Delta_{|\mathcal{A}|}$, we measure their Kullback–Leibler (KL) divergence by

$$\mathrm{KL}(\pi(\cdot|s) \parallel \pi'(\cdot|s)) = \textstyle\sum_{a\in\mathcal{A}} \pi(a|s)\log\frac{\pi(a|s)}{\pi'(a|s)}.$$

Observe that the KL divergence can be viewed as is a special instance of the Bregman's distance (or prox-function) widely used in the optimization literature. Let the distance generating function $\omega(\pi(\cdot|s)) := \sum_{a\in\mathcal{A}} \pi(a|s)\log\pi(a|s)$ [1]. The Bregman's distance associated with ω is given by

$$\begin{aligned} D^\pi_{\pi'}(s) &:= \omega(\pi(\cdot|s)) - [\omega(\pi'(\cdot|s)) + \langle\nabla\omega(\pi'(\cdot|s)), \pi(\cdot|s) - \pi'(\cdot|s)\rangle] \\ &= \textstyle\sum_{a\in\mathcal{A}} \big[\pi(a|s)\log\pi(a|s) - \pi'(a|s)\log\pi'(a|s) \\ &\qquad -(1+\log\pi'(a|s))(\pi(a|s)-\pi'(a|s))\big] \\ &= \textstyle\sum_{a\in\mathcal{A}} \pi(a|s)\log\frac{\pi(a|s)}{\pi'(a|s)}, \end{aligned} \tag{1.10}$$

where the last equation follows from the fact that $\sum_{a\in\mathcal{A}}(\pi(a|s)-\pi'(a|s)) = 0$. Therefore, we will use the KL divergence $\mathrm{KL}(\pi(\cdot|s) \parallel \pi'(\cdot|s))$ and Bregman's distance $D^\pi_{\pi'}(s)$ interchangeably throughout this paper. It should be noted that our algorithmic framework allows us to use other distance generating functions, such as $\|\cdot\|_p^2$ for some $p > 1$, which, different from the KL divergence, has a bounded prox-function over $\Delta_{|\mathcal{A}|}$.

2 Optimality Conditions and Generalized Monotonicity

It is well-known that the value function $V^\pi(s)$ in (1.3) is highly nonconvex w.r.t. π, because the components of $\pi(\cdot|s)$ are multiplied by each other in their definitions (see also Lemma 3 of [1] for an instructive counterexample). However, we will show in this subsection that problem (1.9) can be formulated as a variational inequality (VI) which satisfies certain generalized monotonicity properties (see [6], Section 3.8.2 of [13] and [11]).

[1] It is worth noting that we do not enforce $\pi(a|s) > 0$ when defining $\omega(\pi(\cdot|s))$ as all the search points generated by our algorithms will satisfy this assumption.

Let us first compute the gradient of the value function $V^\pi(s)$ in (1.3). For simplicity, we assume for now that h^π is differentiable and will relax this assumption later. For a given policy π, we define the discounted state visitation distribution by

$$d^\pi_{s_0}(s) := (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathrm{Pr}^\pi(s_t = s|s_0), \tag{2.1}$$

where $\mathrm{Pr}^\pi(s_t = s|s_0)$ denotes the state visitation probability of $s_t = s$ after we follow the policy π starting at state s_0. Let $\mathcal{P}^\pi$ denote the transition probability matrix associated with policy π, i.e., $\mathcal{P}^\pi(i,j) = \sum_{a\in\mathcal{A}} \pi(a|i)\mathcal{P}(j|i,a)$, and e_i be the i-th unit vector. Then $\mathrm{Pr}^\pi(s_t = s|s_0) = e_{s_0}^T(\mathcal{P}^\pi)^t e_s$ and

$$d^\pi_{s_0}(s) = (1-\gamma)\sum_{t=0}^\infty \gamma^t e_{s_0}^T(\mathcal{P}^\pi)^t e_s. \tag{2.2}$$

Lemma 1 *For any $(s_0, s, a) \in \mathcal{S}\times\mathcal{S}\times\mathcal{A}$, we have*

$$\frac{\partial V^\pi(s_0)}{\partial \pi(a|s)} = \frac{1}{1-\gamma} d^\pi_{s_0}(s)\left[Q^\pi(s,a) + \nabla h^\pi(s,a)\right],$$

where $\nabla h^\pi(s,\cdot)$ denotes the gradient of $h^\pi(s)$ w.r.t. π.

Proof. It follows from (1.4) that

$$\begin{aligned}\frac{\partial V^\pi(s_0)}{\partial \pi(a|s)} &= \frac{\partial}{\partial \pi(a'|s)}\sum_{a'\in\mathcal{A}} \pi(a'|s_0)Q^\pi(s_0,a') \\ &= \sum_{a'\in\mathcal{A}}\left[\frac{\partial \pi(a'|s_0)}{\partial \pi(a|s)}Q^\pi(s_0,a') + \pi(a'|s_0)\frac{\partial Q^\pi(s_0,a')}{\partial \pi(a|s)}\right].\end{aligned}$$

Also the relation in (1.5) implies that

$$\frac{\partial Q^\pi(s_0,a')}{\partial \pi(a|s)} = \frac{\partial h^\pi(s_0)}{\partial \pi(a|s)} + \gamma\sum_{s'\in\mathcal{S}}\mathcal{P}(s'|s_0,a')\frac{\partial V^\pi(s')}{\partial \pi(a|s)}.$$

Combining the above two relations, we obtain

$$\begin{aligned}\frac{\partial V^\pi(s_0)}{\partial \pi(a|s)} &= \sum_{a'\in\mathcal{A}}\left[\frac{\partial \pi(a'|s_0)}{\partial \pi(a|s)}Q^\pi(s_0,a') + \pi(a'|s_0)\frac{\partial h^\pi(s_0)}{\partial \pi(a|s)}\right] \\ &\quad + \gamma\sum_{a'\in\mathcal{A}}\pi(a'|s_0)\sum_{s'\in\mathcal{S}}\mathcal{P}(s'|s_0,a')\frac{\partial V^\pi(s')}{\partial \pi(a|s)} \\ &= \sum_{x\in\mathcal{S}}\sum_{t=0}^\infty \gamma^t \mathrm{Pr}^\pi(s_t = x|s_0) \\ &\quad \sum_{a'\in\mathcal{A}}\left[\frac{\partial \pi(a'|x)}{\partial \pi(a|s)}Q^\pi(x,a') + \pi(a'|x)\frac{\partial h^\pi(x)}{\partial \pi(a|s)}\right] \\ &= \frac{1}{1-\gamma}\sum_{x\in\mathcal{S}} d^\pi_{s_0}(x)\left\{\sum_{a'\in\mathcal{A}}\left[\frac{\partial \pi(a'|x)}{\partial \pi(a|s)}Q^\pi(x,a')\right] + \frac{\partial h^\pi(x)}{\partial \pi(a|s)}\right\} \\ &= \frac{1}{1-\gamma} d^\pi_{s_0}(s)\left[Q^\pi(s,a) + \frac{\partial h^\pi(s)}{\partial \pi(a|s)}\right],\end{aligned}$$

where the second equality follows by expanding $\frac{\partial V^\pi(s')}{\partial \pi(a|s)}$ recursively, and the third equality follows from the definition of $d^\pi_{s_0}(s)$ in (2.1), and the last identity follows from $\frac{\partial \pi(a'|x)}{\partial \pi(a|s)} = 0$ for $x \neq s$ or $a' \neq a$, and $\frac{\partial h^\pi(x)}{\partial \pi(a|s)} = 0$ for $x \neq s$. ∎

In view of Lemma 1, the gradient of the objective $f(\pi)$ in (1.9) at the optimal policy π^* is given by

$$\begin{aligned}\frac{\partial f(\pi^*)}{\partial \pi(a|s)} &= \mathbb{E}_{s_0\sim\nu^*}\left[\frac{\partial V^{\pi^*}(s_0)}{\partial \pi(a|s)}\right] = \frac{1}{1-\gamma}\mathbb{E}_{s_0\sim\nu^*}\left[d^{\pi^*}_{s_0}(s)[Q^{\pi^*}(s,a) + \nabla h^{\pi^*}(s,a)]\right] \\ &= \sum_{t=0}^\infty \gamma^t(\nu^*)^T(\mathcal{P}^{\pi^*})^t e_s\left[Q^{\pi^*}(s,a) + \nabla h^{\pi^*}(s,a)\right] \\ &= \frac{1}{1-\gamma}(\nu^*)^T e_s\left[Q^{\pi^*}(s,a) + \nabla h^{\pi^*}(s,a)\right] \\ &= \frac{1}{1-\gamma}\nu^*(s)\left[Q^{\pi^*}(s,a) + \nabla h^{\pi^*}(s,a)\right],\end{aligned} \tag{2.3}$$

where the third identity follows from (2.2) and the last one follows from the fact that $(\nu^*)^T(\mathcal{P}^{\pi^*})^t = (\nu^*)^T$ for any $t \geq 0$ since ν^* is the steady state distribution of π^*. Therefore, the optimality condition of (1.9) suggests us to solve the following variational inequality

$$\mathbb{E}_{s\sim\nu^*}\left[\langle Q^{\pi^*}(s,\cdot) + \nabla h^{\pi^*}(s,\cdot), \pi(\cdot|s) - \pi^*(\cdot|s)\rangle\right] \geq 0. \tag{2.4}$$

However, the above VI requires h^π to be differentiable. In order to handle the possible non-smoothness of h^π, we instead solve the following problem

$$\mathbb{E}_{s\sim\nu^*}\left[\langle Q^{\pi^*}(s,\cdot), \pi(\cdot|s) - \pi^*(\cdot|s)\rangle + h^\pi(s) - h^{\pi^*}(s)\right] \geq 0. \tag{2.5}$$

It turns out this variational inequality satisfies certain generalized monotonicity properties thanks to the following performance difference lemma obtained by generalizing some previous results (e.g., Lemma 6.1 of [9]).

Lemma 2 *For any two feasible policies π and π', we have*

$$V^{\pi'}(s) - V^{\pi}(s) = \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi'}}\left[\langle A^{\pi}(s',\cdot), \pi'(\cdot|s')\rangle + h^{\pi'}(s') - h^{\pi}(s')\right],$$

where

$$A^{\pi}(s',a) := Q^{\pi}(s',a) - V^{\pi}(s'). \tag{2.6}$$

Proof. For simplicity, let us denote $\xi^{\pi'}(s_0)$ the random process (s_t, a_t, s_{t+1}), $t \geq 0$, generated by following the policy π' starting with the initial state s_0. It then follows from the definition of $V^{\pi'}$ that

$$\begin{aligned}
&V^{\pi'}(s) - V^{\pi}(s)\\
&= \mathbb{E}_{\xi^{\pi'}(s)}\left[\textstyle\sum_{t=0}^{\infty}\gamma^t[c(s_t,a_t) + h^{\pi'}(s_t)]\right] - V^{\pi}(s)\\
&= \mathbb{E}_{\xi^{\pi'}(s)}\left[\textstyle\sum_{t=0}^{\infty}\gamma^t[c(s_t,a_t) + h^{\pi'}(s_t) + V^{\pi}(s_t) - V^{\pi}(s_t)]\right] - V^{\pi}(s)\\
&\overset{(a)}{=} \mathbb{E}_{\xi^{\pi'}(s)}\left[\textstyle\sum_{t=0}^{\infty}\gamma^t[c(s_t,a_t) + h^{\pi'}(s_t) + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)]\right]\\
&\qquad + \mathbb{E}_{\xi^{\pi'}(s)}[V^{\pi}(s_0)] - V^{\pi}(s)\\
&\overset{(b)}{=} \mathbb{E}_{\xi^{\pi'}(s)}\left[\textstyle\sum_{t=0}^{\infty}\gamma^t[c(s_t,a_t) + h^{\pi'}(s_t) + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)]\right]\\
&= \mathbb{E}_{\xi^{\pi'}(s)}\left[\textstyle\sum_{t=0}^{\infty}\gamma^t[c(s_t,a_t) + h^{\pi}(s_t) + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)\right.\\
&\qquad \left. + h^{\pi'}(s_t) - h^{\pi}(s_t)]\right]\\
&\overset{(c)}{=} \mathbb{E}_{\xi^{\pi'}(s)}\left[\textstyle\sum_{t=0}^{\infty}\gamma^t\left[Q^{\pi}(s_t,a_t) - V^{\pi}(s_t) + h^{\pi'}(s_t) - h^{\pi}(s_t)\right]\right],
\end{aligned}$$

where (a) follows by taking the term $V^{\pi}(s_0)$ outside the summation, (b) follows from the fact that $\mathbb{E}_{\xi^{\pi'}(s)}[V^{\pi}(s_0)] = V^{\pi}(s)$ since the random process starts with $s_0 = s$, and (c) follows from (1.5). The previous conclusion, together with (2.6) and the definition $d_s^{\pi'}$ in (2.1), then imply that

$$\begin{aligned}
&V^{\pi'}(s) - V^{\pi}(s)\\
&= \tfrac{1}{1-\gamma}\textstyle\sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} d_s^{\pi'}(s')\pi'(a'|s')\left[A^{\pi}(s',a') + h^{\pi'}(s') - h^{\pi}(s')\right]\\
&= \tfrac{1}{1-\gamma}\textstyle\sum_{s'\in\mathcal{S}} d_s^{\pi'}(s')\left[\langle A^{\pi}(s',\cdot), \pi'(\cdot|s')\rangle + h^{\pi'}(s') - h^{\pi}(s')\right],
\end{aligned}$$

which immediately implies the result. ∎

We are now ready to prove the generalized monotonicity for the variational inequality in (2.5).

Lemma 3 *The VI problem in (2.5) satisfies*

$$\begin{aligned}
&\mathbb{E}_{s\sim\nu^*}\left[\langle Q^{\pi}(s,\cdot), \pi(\cdot|s) - \pi^*(\cdot|s)\rangle + h^{\pi}(s) - h^{\pi^*}(s)\right]\\
&\qquad = \mathbb{E}_{s\sim\nu^*}[(1-\gamma)(V^{\pi}(s) - V^{\pi^*}(s))].
\end{aligned}\tag{2.7}$$

Proof. It follows from Lemma 2 (with $\pi' = \pi^*$) that

$$(1-\gamma)[V^{\pi^*}(s) - V^{\pi}(s)] = \mathbb{E}_{s'\sim d_s^{\pi^*}}\left[\langle A^{\pi}(s',\cdot), \pi^*(\cdot|s')\rangle + h^{\pi^*}(s') - h^{\pi}(s')\right].$$

Let e denote the vector of all 1's. Then, we have

$$\begin{aligned}
\langle A^{\pi}(s',\cdot), \pi^*(\cdot|s')\rangle &= \langle Q^{\pi}(s',\cdot) - V^{\pi}(s')e, \pi^*(\cdot|s')\rangle\\
&= \langle Q^{\pi}(s',\cdot), \pi^*(\cdot|s')\rangle - V^{\pi}(s')\\
&= \langle Q^{\pi}(s',\cdot), \pi^*(\cdot|s')\rangle - \langle Q^{\pi}(s',\cdot), \pi(\cdot|s')\\
&= \langle Q^{\pi}(s',\cdot), \pi^*(\cdot|s') - \pi(\cdot|s')\rangle,
\end{aligned}\tag{2.8}$$

where the first identity follows from the definition of $A^{\pi}(s',\cdot)$ in (2.6), the second equality follows from the fact that $\langle e, \pi^*(\cdot|s')\rangle = 1$, and the third equality follows from the definition of V^{π} in (1.3). Combining the above two relations and taking expectation w.r.t. ν^*, we obtain

$$\begin{aligned}
&(1-\gamma)\mathbb{E}_{s\sim\nu^*}[V^{\pi^*}(s) - V^{\pi}(s)]\\
&= \mathbb{E}_{s\sim\nu^*, s'\sim d_s^{\pi^*}}\left[\langle Q^{\pi}(s',\cdot), \pi^*(\cdot|s') - \pi(\cdot|s')\rangle + h^{\pi^*}(s') - h^{\pi}(s')\right]\\
&= \mathbb{E}_{s\sim\nu^*}\left[\langle Q^{\pi}(s,\cdot), \pi^*(\cdot|s) - \pi(\cdot|s)\rangle + h^{\pi^*}(s') - h^{\pi}(s')\right],
\end{aligned}$$

where the second identity follows similarly to (2.3) since ν^* is the steady state distribution induced by π^*. The result then follows by rearranging the terms. ∎

Since $V^{\pi}(s) - V^{\pi^*}(s) \geq 0$ for any feasible policity π, we conclude from Lemma 3 that

$$\mathbb{E}_{s\sim\nu^*}\left[\langle Q^{\pi}(s,\cdot), \pi(\cdot|s) - \pi^*(\cdot|s)\rangle + h^{\pi}(s) - h^{\pi^*}(s)\right] \geq 0.$$

Therefore, the VI in (2.5) satisfies the generalized monotonicity. In the next few sections, we will exploit the generalized monotonicity and some other structural properties to design efficient algorithms for solving the RL problem.

3 Deterministic Policy Mirror Descent

In this section, we present the basic schemes of policy mirror descent (PMD) and establish their convergence properties.

3.1 Prox-mapping

In the proposed PMD methods, we will update a given policy π to π^+ through the following proximal mapping:

$$\pi^+(\cdot|s) = \underset{p(\cdot|s)\in\Delta_{|\mathcal{A}|}}{\arg\min} \ \eta[\langle G^{\pi}(s,\cdot), p(\cdot|s)\rangle + h^p(s)] + D^p_{\pi}(s). \tag{3.1}$$

Here $\eta > 0$ denotes a certain stepsize (or learning rate), and G^{π} can be the operator for the VI formulation, e.g., $G^{\pi}(s,\cdot) = Q^{\pi}(s,\cdot)$ or its approximation.

It is well-known that one can solve (3.1) explicitly for some interesting special cases, e.g., when $h^p(s) = 0$ or $h^p(s) = \tau D^p_{\pi_0}(s)$ for some $\tau > 0$ and given π_0. For both these cases, the solution of (3.1) boils down to solving a problem of the form

$$p^* := \underset{p(\cdot|s)\in\Delta_{|\mathcal{A}|}}{\arg\min} \ \sum_{i=1}^{|\mathcal{A}|} (g_i p_i + p_i \log p_i)$$

for some $g \in \mathbb{R}^{|\mathcal{A}|}$. It can be easily checked from the Karush-Kuhn-Tucker conditions that its optimal solution is given by

$$p^*_i = \exp(-g_i)/[\textstyle\sum_{i=1}^{|\mathcal{A}|} \exp(-g_i)]. \tag{3.2}$$

For more general convex functions h^p, problem (3.1) usually does not have an explicit solution, and one can only solve it approximately. In fact, we will show in Section 6 that by applying the accelerated gradient descent method, we only need to compute a small number of updates in the form of (3.2) in order to approximately solve (3.1) without slowing down the efficiency of the overall PMD algorithms.

3.2 Basic PMD method

As shown in Algorithm 1, each iteration of the PMD method applies the prox-mapping step discussed in Subsection 3.1 to update the policy π_k. It involves the stepsize parameter η_k and requires the selection of an initial point π_0. For the sake of simplicity, we will assume throughout the paper that

$$\pi_0(a|s) = 1/|\mathcal{A}|, \ \forall a \in \mathcal{A}, \forall s \in \mathcal{S}. \tag{3.3}$$

In this case, we have

$$D^{\pi}_{\pi_0}(s) = \textstyle\sum_{a\in\mathcal{A}} \pi(a|s)\log\pi(a|s) + \log|\mathcal{A}| \leq \log|\mathcal{A}|, \ \forall \pi(\cdot|s) \in \Delta_{|\mathcal{A}|}. \tag{3.4}$$

Observe also that we can replace $Q^{\pi_k}(s,\cdot)$ in (3.5) with $A^{\pi_k}(s,a)$ defined in (2.6) without impacting the updating of $\pi_{k+1}(s,\cdot)$, since this only introduces an extra constant into the objective function of (3.5).

Algorithm 1 The policy mirror descent (PMD) method

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

$$\pi_{k+1}(\cdot|s) = \underset{p(\cdot|s)\in\Delta_{|\mathcal{A}|}}{\arg\min} \ \left\{\eta_k[\langle Q^{\pi_k}(s,\cdot), p(\cdot|s)\rangle + h^p(s)] + D^p_{\pi_k}(s)\right\}, \forall s \in \mathcal{S}. \tag{3.5}$$

end for

Below we establish some general convergence properties about the PMD method.

The following result characterizes the optimality condition of problem (3.5) (see Lemma 3.5 of [13]). We add a proof for the sake of completeness.

Lemma 4 *For any $p(\cdot|s) \in \Delta_{|\mathcal{A}|}$, we have*

$$\begin{aligned}&\eta_k[\langle Q^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - p(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^p(s)] + D^{\pi_{k+1}}_{\pi_k}(s)\\ &\le D^p_{\pi_k}(s) - (1+\eta_k\mu)D^p_{\pi_{k+1}}(s).\end{aligned}$$

Proof. By the optimality condition of (3.5),

$$\langle \eta_k[Q^{\pi_k}(s,\cdot) + (h')^{\pi_{k+1}}(s,\cdot)] + \nabla D^{\pi_{k+1}}_{\pi_k}(s,\cdot), p(\cdot|s) - \pi_{k+1}(\cdot|s)\rangle \ge 0, \quad \forall p(\cdot|s) \in \Delta_{|\mathcal{A}|},$$

where $(h')^{\pi_{k+1}}$ denotes the subgradient of h at π_{k+1} and $\nabla D^{\pi_{k+1}}_{\pi_k}(s,\cdot)$ denotes the gradient of $D^{\pi_{k+1}}_{\pi_k}(s)$ at π_{k+1}. Using the definition of Bregman's distance, it is easy to verify that

$$D^p_{\pi_k}(s) = D^{\pi_{k+1}}_{\pi_k}(s) + \langle \nabla D^{\pi_{k+1}}_{\pi_k}(s,\cdot), p(\cdot|s) - \pi_{k+1}(\cdot|s)\rangle + D^p_{\pi_{k+1}}(s). \tag{3.6}$$

The result then immediately follows by combining the above two relations together with (1.2). ∎

Lemma 5 *For any $s \in \mathcal{S}$, we have*

$$V^{\pi_{k+1}}(s) \le V^{\pi_k}(s), \tag{3.7}$$

$$\langle Q^{\pi_k}(s,\cdot),\, \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s) \ge V^{\pi_{k+1}}(s) - V^{\pi_k}(s). \tag{3.8}$$

Proof. It follows from Lemma 2 (with $\pi' = \pi_{k+1}$, $\pi = \pi_k$ and $\tau = \tau_k$) that

$$\begin{aligned}&V^{\pi_{k+1}}(s) - V^{\pi_k}(s)\\ &= \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\left[\langle A^{\pi_k}(s',\cdot), \pi_{k+1}(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s')\right].\end{aligned} \tag{3.9}$$

Similarly to (2.8), we can show that

$$\begin{aligned}\langle A^{\pi_k}(s',\cdot), \pi_{k+1}(\cdot|s')\rangle &= \langle Q^{\pi_k}(s',\cdot) - V^{\pi_k}_{\tau_k}(s')e, \pi_{k+1}(\cdot|s')\rangle\\ &= \langle Q^{\pi_k}(s',\cdot), \pi_{k+1}(\cdot|s')\rangle - V^{\pi_k}_{\tau_k}(s')\\ &= \langle Q^{\pi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle.\end{aligned}$$

Combining the above two identities, we then obtain

$$\begin{aligned}V^{\pi_{k+1}}(s) - V^{\pi_k}(s) = \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}&\left[\langle Q^{\pi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle\right.\\ &\left.+h^{\pi_{k+1}}(s') - h^{\pi_k}(s')\right].\end{aligned} \tag{3.10}$$

Now we conclude from Lemma 4 applied to (3.5) with $p(\cdot|s') = \pi_k(\cdot|s')$ that

$$\begin{aligned}&\langle Q^{\pi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s')\\ &\le -\tfrac{1}{\eta_k}[(1+\eta_k\mu)D^{\pi_k}_{\pi_{k+1}}(s') + D^{\pi_{k+1}}_{\pi_k}(s')].\end{aligned} \tag{3.11}$$

The previous two conclusions then clearly imply the result in (3.7). It also follows from (3.11) that

$$\begin{aligned}&\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\left[\langle Q^{\pi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s')\right]\\ &\le d_s^{\pi_{k+1}}(s)\left[\langle Q^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s)\right]\\ &\le (1-\gamma)\left[\langle Q^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s)\right],\end{aligned} \tag{3.12}$$

where the last inequality follows from the fact that $d_s^{\pi_{k+1}}(s) \ge (1-\gamma)$ due to the definition of $d_s^{\pi_{k+1}}$ in (2.1). The result in (3.7) then follows immediately from (3.10) and the above inequality. ∎

Now we show that with a constant stepsize rule, the PMD method can achieve a linear rate of convergence for solving RL problems with strongly convex regularizers (i.e., $\mu > 0$).

Theorem 1 *Suppose that $\eta_k = \eta$ for any $k \ge 0$ in the PMD method with*

$$1 + \eta\mu \ge \tfrac{1}{\gamma}. \tag{3.13}$$

Then we have

$$f(\pi_k) - f(\pi^*) + \tfrac{\mu}{1-\gamma}\mathcal{D}(\pi_k, \pi^*) \le \gamma^k[f(\pi_0) - f(\pi^*_\tau) + \tfrac{\mu}{1-\gamma}\log|\mathcal{A}|]$$

for any $k \ge 0$, where

$$\mathcal{D}(\pi_k, \pi^*) := \mathbb{E}_{s\sim\nu^*}[D^{\pi^*}_{\pi_k}(s)]. \tag{3.14}$$

Proof. By Lemma 4 applied to (3.5) (with $\eta_k = \eta$ and $p = \pi^*$), we have

$$\eta[\langle Q^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi^*(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi^*}(s)] + D_{\pi_k}^{\pi_{k+1}}(s)$$
$$\leq D_{\pi_k}^{\pi^*}(s) - (1+\eta\mu)D_{\pi_{k+1}}^{\pi^*}(s),$$

which, in view of (3.8), then implies that

$$\eta[\langle Q^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle + h^{\pi_k}(s) - h^{\pi^*}(s)]$$
$$+ \eta[V^{\pi_{k+1}}(s) - V^{\pi_k}(s)] + D_{\pi_k}^{\pi_{k+1}}(s) \leq D_{\pi_k}^{\pi^*}(s) - (1+\eta\mu)D_{\pi_{k+1}}^{\pi^*}(s).$$

Taking expectation w.r.t. ν^* on both sides of the above inequality and using Lemma 3, we arrive at

$$\mathbb{E}_{s\sim\nu^*}[\eta(1-\gamma)(V^{\pi_k}(s) - V^{\pi_\tau^*}(s))] + \eta\mathbb{E}_{s\sim\nu^*}[V^{\pi_{k+1}}(s) - V^{\pi_k}(s)] + \mathbb{E}_{s\sim\nu^*}[D_{\pi_k}^{\pi_{k+1}}(s)]$$
$$\leq \mathbb{E}_{s\sim\nu^*}[D_{\pi_k}^{\pi^*}(s) - (1+\eta\mu)D_{\pi_{k+1}}^{\pi^*}(s)].$$

Noting $V^{\pi_{k+1}}(s) - V^{\pi_k}(s) = V^{\pi_{k+1}}(s) - V^{\pi^*}(s) - [V^{\pi_k}(s) - V^{\pi^*}(s)]$ and rearranging the terms in the above inequality, we have

$$\mathbb{E}_{s\sim\nu^*}[\eta(V^{\pi_{k+1}}(s) - V^{\pi^*}(s)) + (1+\eta\mu)D_{\pi_{k+1}}^{\pi^*}(s)] + \mathbb{E}_{s\sim\nu^*}[D_{\pi_k}^{\pi_{k+1}}(s)]$$
$$\leq \gamma\mathbb{E}_{s\sim\nu^*}[\eta(V^{\pi_k}(s) - V^{\pi^*}(s)) + \tfrac{1}{\gamma}D_{\pi_k}^{\pi^*}(s)], \tag{3.15}$$

which, in view of the assumption (3.13) and the definition of f in (1.9)

$$f(\pi_{k+1}) - f(\pi^*) + \tfrac{\mu}{1-\gamma}\mathbb{E}_{s\sim\nu^*}[D_{\pi_{k+1}}^{\pi^*}(s)]$$
$$\leq \gamma\left[(f(\pi_k) - f(\pi^*)) + \tfrac{\mu}{1-\gamma}\mathbb{E}_{s\sim\nu^*}[D_{\pi_k}^{\pi^*}(s)]\right].$$

Applying this relation recursively and using the bound in (3.4) we then conclude the result. ∎

According to Theorem 1, the PMD method converges linearly in terms of both function value and the distance to the optimal solution for solving RL problems with strongly convex regularizers. Now we show that a direct application of the PMD method only achieves a sublinear rate of convergence for the case when $\mu = 0$.

Theorem 2 *Suppose that $\eta_k = \eta$ in the PMD method. Then we have*

$$f(\pi_{k+1}) - f(\pi^*) \leq \tfrac{\eta\gamma[f(\pi_0) - f(\pi^*)] + \log|\mathcal{A}|}{\eta(1-\gamma)(k+1)}$$

for any $k \geq 0$.

Proof. It follows from (3.15) with $\mu = 0$ that

$$\mathbb{E}_{s\sim\nu^*}[\eta(V^{\pi_{k+1}}(s) - V^{\pi^*}(s)) + D_{\pi_{k+1}}^{\pi^*}(s)] + \mathbb{E}_{s\sim\nu^*}[D_{\pi_k}^{\pi_{k+1}}(s)]$$
$$\leq \eta\gamma\mathbb{E}_{s\sim\nu^*}[V^{\pi_k}(s) - V^{\pi^*}(s)] + \mathbb{E}_{s\sim\nu^*}[D_{\pi_k}^{\pi^*}(s)].$$

Taking the telescopic sum of the above inequalities and using the fact that $V^{\pi_{k+1}}(s) \leq V^{\pi_k}(s)$ due to (3.7) , we obtain

$$(k+1)\eta(1-\gamma)\mathbb{E}_{s\sim\nu^*}[V^{\pi_{k+1}}(s) - V^{\pi^*}(s)] \leq \mathbb{E}_{s\sim\nu^*}[\eta\gamma(V^{\pi_0}(s) - V^{\pi^*}(s)) + D_{\pi_0}^{\pi^*}(s)],$$

which clearly implies the result in view of the definition of f in (1.9) and the bound on $D_{\pi_0}^{\pi^*}$ in (3.4). ∎

The result in Theorem 2 shows that the PMD method requires $\mathcal{O}(1/(1-\gamma)\epsilon)$ iterations to find an ϵ-solution for general RL problems. This bound already matches, in terms of its dependence on $(1-\gamma)$ and ϵ, the previously best-known complexity for natural policy gradient methods [1]. We will develop a variant of the PMD method that can achieve a linear rate of convergence for the case when $\mu = 0$ in next subsection.

3.3 Approximate policy mirror descent method

In this subsection, we propose a novel variant of the PMD method by adding adaptively a perturbation term into the definition of the value functions.

For some $\tau \geq 0$ and a given initial policy $\pi_0(a|s) > 0$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$, we define the perturbed action-value and state-value functions, respectively, by

$$\begin{aligned} Q_\tau^\pi(s,a) &:= \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \left[c(s_t,a_t) + h^\pi(s_t) + \tau D_{\pi_0}^\pi(s_t)\right]\right. \\ &\qquad \left. \mid s_0 = s, a_0 = a, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t)\right], \qquad (3.16) \\ V_\tau^\pi(s) &:= \langle Q_\tau^\pi(s,\cdot), \pi(\cdot|s)\rangle. \qquad (3.17) \end{aligned}$$

Clearly, if $\tau = 0$, then the perturbed value functions reduce to the usual value functions, i.e.,

$$Q_0^\pi(s,a) = Q^\pi(s,a) \ \text{ and } \ V_0^\pi(s) = V^\pi(s).$$

The following result relates the value functions with different τ.

Lemma 6 *For any given $\tau, \tau' \geq 0$, we have*

$$V_\tau^\pi(s) - V_{\tau'}^\pi(s) = \tfrac{\tau-\tau'}{1-\gamma}\mathbb{E}_{s'\sim d_s^\pi}[D_{\pi_0}^\pi(s')]. \qquad (3.18)$$

As a consequence, if $\tau \geq \tau' \geq 0$ then

$$V_{\tau'}^\pi(s) \leq V_\tau^\pi(s) \leq V_{\tau'}^\pi(s) + \tfrac{\tau-\tau'}{1-\gamma}\log|\mathcal{A}|. \qquad (3.19)$$

Proof. By the definitions of V_τ^π and d_s^π, we have

$$\begin{aligned} &V_\tau^\pi(s) \\ &= \mathbb{E}\left[\textstyle\sum_{t=0}^\infty \gamma^t [c(s_t,a_t) + h^\pi(s_t) + \tau D_{\pi_0}^\pi(s)] \mid s_0 = s, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t)\right] \\ &= \mathbb{E}\left[\textstyle\sum_{t=0}^\infty \gamma^t [c(s_t,a_t) + h^\pi(s_t) + \tau' D_{\pi_0}^\pi(s)] \mid s_0 = s, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t)\right] \\ &\quad + \mathbb{E}\left[\textstyle\sum_{t=0}^\infty \gamma^t (\tau-\tau') D_{\pi_0}^\pi(s)] \mid s_0 = s, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t)\right] \\ &= V_{\tau'}^\pi(s) + \tfrac{\tau-\tau'}{1-\gamma}\mathbb{E}_{s'\sim d_s^\pi}[D_{\pi_0}^\pi(s')], \end{aligned}$$

which together with the bound on $D_{\pi_0}^\pi$ in (3.4) then imply (3.19). ∎

As shown in Algorithm 2, the approximate policy mirror descent (APMD) method is obtained by replacing $Q^{\pi_k}(s,\cdot)$ with its approximation $Q_{\tau_k}^{\pi_k}(s,\cdot)$ and adding the perturbation $\tau_k D_{\pi_0}^\pi(s_t)$ for the updating of π_{k+1} in the basic PMD method. As discussed in Subsection 3.1, the incorporation of the perturbation term does not impact the difficulty of solving the subproblem in (3.20).

Algorithm 2 The approximate policy mirror descent (APMD) method

Input: initial points π_0, stepsizes $\eta_k \geq 0$ and perturbation $\tau_k \geq 0$.
for $k = 0, 1, \ldots,$ **do**

$$\pi_{k+1}(\cdot|s) = \operatorname*{arg\,min}_{p(\cdot|s)\in\Delta_{|\mathcal{A}|}} \left\{\eta_k[\langle Q_{\tau_k}^{\pi_k}(s,\cdot), p(\cdot|s)\rangle + h^p(s) + \tau_k D_{\pi_0}^p(s_t)] + D_{\pi_k}^p(s)\right\}, \forall s \in \mathcal{S}. \qquad (3.20)$$

end for

Our goal in the remaining part of this subsection is to show that the APMD method, when employed with proper selection of τ_k, can achieve a linear rate of convergence for solving general RL problems. First we observe that Lemma 3 can still be applied to the perturbed value functions. The difference between the following result and Lemma 3 exists in that the RHS of (3.21) is no longer nonnegative, i.e., $V_\tau^\pi(s) - V_\tau^{\pi^*}(s) \not\geq 0$. However, this relation will be approximately satisfied if τ is small enough.

Lemma 7 *The VI problem in (2.5) satisfies*

$$\begin{aligned} &\mathbb{E}_{s\sim\nu^*}\left[\langle Q_\tau^\pi(s,\cdot), \pi(\cdot|s) - \pi^*(\cdot|s)\rangle + h^\pi(s) - h^{\pi^*}(s) + \tau[D_{\pi_0}^\pi(s) - D_{\pi_0}^{\pi^*}(s)]\right] \\ &\quad = \mathbb{E}_{s\sim\nu^*}[(1-\gamma)(V_\tau^\pi(s) - V_\tau^{\pi^*}(s))]. \qquad (3.21) \end{aligned}$$

Proof. The proof is the same as that for Lemma 3 except that we will apply the performance difference lemma (i.e., Lemma 2) to the perturbed value function V_τ^π. ∎

Next we establish some general convergence properties about the APMD method. Lemma 8 below characterizes the optimal solution of (3.20) (see, e.g., Lemma 3.5 of [13]).

Lemma 8 *Let $\pi_{k+1}(\cdot|s)$ be defined in (3.20). For any $p(\cdot|s) \in \Delta_{|\mathcal{A}|}$, we have*

$$\begin{aligned}&\eta_k[\langle Q_{\tau_k}^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - p(\cdot|s)\rangle + h^{\pi^+}(s) - h^p(s)]\\ &+ \eta_k\tau_k[D_{\pi_0}^{\pi_{k+1}}(s_t) - D_{\pi_0}^p(s_t)] + D_{\pi_k}^{\pi_{k+1}}(s) \le D_{\pi_k}^p(s) - (1+\eta\tau_k)D_{\pi_{k+1}}^p(s).\end{aligned}$$

Lemma 9 below is similar to Lemma 5 for the PMD method.

Lemma 9 *For any $s \in \mathcal{S}$, we have*

$$\begin{aligned}&\langle Q_{\tau_k}^{\pi_k}(s,\cdot),\, \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s)\\ &+ \tau_k[D_{\pi_0}^{\pi_{k+1}}(s) - D_{\pi_0}^{\pi_k}(s)] \ge V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi_k}(s).\end{aligned} \tag{3.22}$$

Proof. By applying Lemma 2 to the perturbed value function V_τ^π and using an argument similar to (3.10), we can show that

$$\begin{aligned}V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi_k}(s) = \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}&\left[\langle Q_{\tau_k}^{\pi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle\right.\\ &\left.+h^{\pi_{k+1}}(s') - h^{\pi_k}(s') + \tau_k[D_{\pi_0}^{\pi_{k+1}}(s) - D_{\pi_0}^{\pi_k}(s)]\right].\end{aligned} \tag{3.23}$$

Now we conclude from Lemma 8 with $p(\cdot|s') = \pi_k(\cdot|s')$ that

$$\begin{aligned}&\langle Q_{\tau_k}^{\pi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s')\\ &+ \tau_k[D_{\pi_0}^{\pi_{k+1}}(s') - D_{\pi_0}^{\pi_k}(s')] \le -\tfrac{1}{\eta_k}[(1+\eta_k\tau_k)D_{\pi_{k+1}}^{\pi_k}(s') + D_{\pi_k}^{\pi_{k+1}}(s')],\end{aligned} \tag{3.24}$$

which implies that

$$\begin{aligned}&\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\left[\langle Q_{\tau_k}^{\pi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s')\right.\\ &\quad\left.+\tau_k[D_{\pi_0}^{\pi_{k+1}}(s') - D_{\pi_0}^{\pi_k}(s')]\right]\\ &\le d_s^{\pi_{k+1}}(s)\left[\langle Q_{\tau_k}^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s)\right.\\ &\quad\left.+\tau_k[D_{\pi_0}^{\pi_{k+1}}(s) - D_{\pi_0}^{\pi_k}(s)]\right]\\ &\le (1-\gamma)\left[\langle Q_{\tau_k}^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s)\right.\\ &\quad\left.+\tau_k[D_{\pi_0}^{\pi_{k+1}}(s) - D_{\pi_0}^{\pi_k}(s)]\right],\end{aligned} \tag{3.25}$$

where the last inequality follows from the fact that $d_s^{\pi_{k+1}}(s) \ge (1-\gamma)$ due to the definition of $d_s^{\pi_{k+1}}$ in (2.1). The result in (3.22) then follows immediately from (3.23) and the above inequality. ∎

The following general result holds for different stepsize rules for APMD.

Lemma 10 *Suppose $1 + \eta_k\tau_k = 1/\gamma$ and $\tau_k \ge \tau_{k+1}$ in the APMD method. Then for any $k \ge 0$, we have*

$$\begin{aligned}&\mathbb{E}_{s\sim\nu^*}[V_{\tau_{k+1}}^{\pi_{k+1}}(s) - V_{\tau_{k+1}}^{\pi^*}(s) + \tfrac{\tau_{k+1}}{1-\gamma}D_{\pi_{k+1}}^{\pi^*}(s)]\\ &\le \mathbb{E}_{s\sim\nu^*}[\gamma[V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s) + \tfrac{\tau_k}{1-\gamma}D_{\pi_k}^{\pi^*}(s)] + \tfrac{\tau_k - \tau_{k+1}}{1-\gamma}\log|\mathcal{A}|.\end{aligned} \tag{3.26}$$

Proof. By Lemma 8 with $p = \pi^*$, we have

$$\begin{aligned}&\eta_k\left[\langle Q_{\tau_k}^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi^*(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi^*}(s)\right]\\ &+ \eta_k\tau_k[D_{\pi_0}^{\pi_{k+1}}(s_t) - D_{\pi_0}^{\pi^*}(s_t)] + D_{\pi_k}^{\pi_{k+1}}(s)\\ &\le D_{\pi_k}^{\pi^*}(s) - (1+\eta_k\tau_k)D_{\pi_{k+1}}^{\pi^*}(s).\end{aligned}$$

Moreover, by Lemma 9,

$$\begin{aligned}&\langle Q_{\tau_k}^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s) + \tau_k[D_{\pi_0}^{\pi_{k+1}}(s_t) - D_{\pi_0}^{\pi_k}(s_t)]\\ &\ge V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi_k}(s).\end{aligned}$$

Combining the above two relations, we obtain

$$\begin{aligned}&\eta_k\left[\langle Q_{\tau_k}^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle + h^{\pi_k}(s) - h^{\pi^*}(s)\right] + \eta_k\tau_k[D_{\pi_0}^{\pi_k}(s_t) - D_{\pi_0}^{\pi^*}(s_t)]\\ &+ \eta_k[V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi_k}(s)] + D_{\pi_k}^{\pi_{k+1}}(s) \le D_{\pi_k}^{\pi^*}(s) - (1+\eta_k\tau_k)D_{\pi_{k+1}}^{\pi^*}(s).\end{aligned}$$

Taking expectation w.r.t. ν^* on both sides of the above inequality and using Lemma 7, we arrive at

$$\mathbb{E}_{s\sim\nu^*}[\eta_k(1-\gamma)(V_{\tau_k}^{\pi_k}(s)-V_{\tau_k}^{\pi^*}(s))]+\eta_k\mathbb{E}_{s\sim\nu^*}[V_{\tau_k}^{\pi_{k+1}}(s)-V_{\tau_k}^{\pi_k}(s)]$$
$$+\mathbb{E}_{s\sim\nu^*}[D_{\pi_k}^{\pi_{k+1}}(s)]\le\mathbb{E}_{s\sim\nu^*}[D_{\pi_k}^{\pi^*}(s)-(1+\eta_k\tau_k)D_{\pi_{k+1}}^{\pi^*}(s)].$$

Noting $V_{\tau_k}^{\pi_{k+1}}(s)-V_{\tau_k}^{\pi_k}(s)=V_{\tau_k}^{\pi_{k+1}}(s)-V_{\tau_k}^{\pi^*}(s)-[V_{\tau_k}^{\pi_k}(s)-V_{\tau_k}^{\pi^*}(s)]$ and rearranging the terms in the above inequality, we have

$$\mathbb{E}_{s\sim\nu^*}[\eta_k(V_{\tau_k}^{\pi_{k+1}}(s)-V_{\tau_k}^{\pi^*}(s))+(1+\eta_k\tau_k)D_{\pi_{k+1}}^{\pi^*}(s)+D_{\pi_k}^{\pi_{k+1}}(s)]$$
$$\le\eta_k\gamma\mathbb{E}_{s\sim\nu^*}[V_{\tau_k}^{\pi_k}(s)-V_{\tau_k}^{\pi^*}(s)]+\mathbb{E}_{s\sim\nu^*}[D_{\pi_k}^{\pi^*}(s)]. \tag{3.27}$$

Using the above inequality, the assumption $\tau_k\ge\tau_{k+1}$ and (3.19), we have

$$\mathbb{E}_{s\sim\nu^*}[\eta_k(V_{\tau_{k+1}}^{\pi_{k+1}}(s)-V_{\tau_{k+1}}^{\pi^*}(s))+(1+\eta_k\tau_k)D_{\pi_{k+1}}^{\pi^*}(s)+D_{\pi_k}^{\pi_{k+1}}(s)]$$
$$\le\mathbb{E}_{s\sim\nu^*}[\eta_k\gamma(V_{\tau_k}^{\pi_k}(s)-V_{\tau_k}^{\pi^*}(s))+D_{\pi_k}^{\pi^*}(s)]+\tfrac{\eta_k(\tau_k-\tau_{k+1})}{1-\gamma}\log|\mathcal{A}|, \tag{3.28}$$

which implies the result by the assumption $1+\eta_k\tau_k=1/\gamma$. ∎

We are now ready to establish the rate of convergence of the APMD method with dynamic stepsize rules to select η_k and τ_k for solving general RL problems.

Theorem 3 *Suppose that $\tau_k=\tau_0\gamma^k$ for some $\tau_0\ge0$ and that $1+\eta_k\tau_k=1/\gamma$ for any $k\ge0$ in the APMD method. Then for any $k\ge0$, we have*

$$f(\pi_k)-f(\pi^*)\le\gamma^k\left[f(\pi_0)-f(\pi^*)+\tau_0\left(\tfrac{2}{1-\gamma}+\tfrac{k}{\gamma}\right)\log|\mathcal{A}|\right]. \tag{3.29}$$

Proof. Applying the result in Lemma 10 recursively, we have

$$\mathbb{E}_{s\sim\nu^*}[V_{\tau_k}^{\pi_k}(s)-V_{\tau_k}^{\pi^*}(s)]\le\gamma^k\mathbb{E}_{s\sim\nu^*}[V_{\tau_0}^{\pi_0}(s)-V_{\tau_0}^{\pi^*}(s)+\tfrac{\tau_0}{1-\gamma}D_{\pi_0}^{\pi^*}(s)]$$
$$+\textstyle\sum_{i=1}^k\frac{(\tau_{i-1}-\tau_i)\gamma^{k-i}}{1-\gamma}\log|\mathcal{A}|.$$

Noting that $V_{\tau_k}^{\pi_k}(s)\ge V^{\pi_k}(s)$, $V_{\tau_k}^{\pi^*}(s)\le V^{\pi^*}(s)+\frac{\tau_k}{1-\gamma}\log|\mathcal{A}|$, and $V_{\tau_0}^{\pi^*}(s)\ge V^{\pi^*}(s)$ due to (3.18), and that $V_{\tau_0}^{\pi_0}(s)=V^{\pi_0}(s)$ due to $D_{\pi_0}^{\pi_0}(s)=0$, we conclude from the previous inequality that

$$\mathbb{E}_{s\sim\nu^*}[V^{\pi_k}(s)-V^{\pi^*}(s)]\le\gamma^k\mathbb{E}_{s\sim\nu^*}[V^{\pi_0}(s)-V^{\pi^*}(s)+\tfrac{\tau_0}{1-\gamma}D_{\pi_0}^{\pi^*}(s)]$$
$$+\left[\tfrac{\tau_k}{1-\gamma}+\textstyle\sum_{i=1}^k\frac{(\tau_{i-1}-\tau_i)\gamma^{k-i}}{1-\gamma}\right]\log|\mathcal{A}|. \tag{3.30}$$

The result in (3.29) immediately follows from the above relation, the definition of f in (1.9), and the selection of τ_k. ∎

According to (3.29), if τ_0 is a constant, then the rate of convergence of the APMD method is $\mathcal{O}(k\gamma^k)$. If the total number of iterations k is given a priori, we can improve the rate of convergence to $\mathcal{O}(\gamma^k)$ by setting $\tau_0=1/k$. Below we propose a different way to specify τ_k for the APMD method so that it can achieve this $\mathcal{O}(\gamma^k)$ rate of convergence without fixing k a priori.

We first establish a technical result that will also be used later for the analysis of stochastic PMD methods.

Lemma 11 *Assume that the nonnegative sequences $\{X_k\}_{k\ge0},\{Y_k\}_{k\ge0}$ and $\{Z_k\}_{k\ge0}$ satisfy*

$$X_{k+1}\le\gamma X_k+(Y_k-Y_{k+1})+Z_k. \tag{3.31}$$

Let us denote $l=\lceil\log_\gamma\frac{1}{4}\rceil$. If $Y_k=Y\cdot2^{-(\lfloor k/l\rfloor+1)}$ and $Z_k=Z\cdot2^{-(\lfloor k/l\rfloor+2)}$ for some $Y\ge0$ and $Z\ge0$, then

$$X_k\le2^{-\lfloor k/l\rfloor}(X_0+Y+\tfrac{5Z}{4(1-\gamma)}). \tag{3.32}$$

Proof. Let us group the indices $\{0,\dots,k\}$ into $\bar p\equiv\lfloor k/l\rfloor+1$ epochs with each of the first $\bar p-1$ epochs consisting of l iterations. Let $p=0,\dots,\bar p$ be the epoch indices. We first show that for any $p=0,\dots,\bar p-1$,

$$X_{pl}\le2^{-p}(X_0+Y+\tfrac{Z}{1-\gamma}). \tag{3.33}$$

This relation holds obviously for $p = 0$. Let us assume that (3.33) holds at the beginning of epoch p ad examine the progress made in epoch p. Note that for any indices $k = pl, \dots, (p+1)l-1$ in epoch p, we have $Y_k = Y \cdot 2^{-(p+1)}$ and $Z = Z \cdot 2^{-(p+2)}$. By applying (3.31) recursively, we have

$$\begin{aligned} X_{(p+1)l} &\le \gamma^l X_{pl} + Y_{pl} - Y_{(p+1)l} + Z_{pl} \textstyle\sum_{i=0}^{l-1} \gamma^i \\ &= \gamma^l X_{pl} + Y_{(p+1)l} + Z_{pl} \tfrac{1-\gamma^l}{1-\gamma} \\ &\le \gamma^l X_{pl} + Y \cdot 2^{-(p+2)} + \tfrac{Z \cdot 2^{-(p+2)}}{1-\gamma} \\ &\le \tfrac{1}{4} X_{pl} + Y \cdot 2^{-(p+2)} + \tfrac{Z \cdot 2^{-(p+2)}}{1-\gamma} \\ &\le \tfrac{1}{4} 2^{-p} (X_0 + Y + \tfrac{Z}{1-\gamma}) + Y \cdot 2^{-(p+2)} + \tfrac{Z \cdot 2^{-(p+2)}}{1-\gamma} \\ &\le 2^{-(p+1)} (X_0 + Y + \tfrac{Z}{1-\gamma}), \end{aligned}$$

where the second inequality follows from the definition of Z_{pl} and $\gamma^l \ge 0$, the third one follows from $\gamma^l \le 1/4$, the fourth one follows by induction hypothesis, and the last one follows by regrouping the terms. Since $k = (\bar{p}-1)l + k \pmod l$, we have

$$\begin{aligned} X_k &\le \gamma^{k \pmod l} A_{(\bar{p}-1)l} + Z_{(\bar{p}-1)} \textstyle\sum_{i=0}^{k \pmod l - 1} \\ &\le 2^{-(\bar{p}-1)} (X_0 + Y + \tfrac{Z}{1-\gamma}) + \tfrac{Z \cdot 2^{-(\bar{p}+1)}}{1-\gamma} \\ &= 2^{-(\bar{p}-1)} (X_0 + Y + \tfrac{5Z}{4(1-\gamma)}), \end{aligned}$$

which implies the result. ∎

We are now ready to present a more convenient selection of τ_k and η_k for the APMD method.

Theorem 4 *Let us denote* $l := \left\lceil \log_\gamma \frac{1}{4} \right\rceil$. *If* $\tau_k = 2^{-(\lfloor k/l \rfloor + 1)}$ *and* $1 + \eta_k \tau_k = 1/\gamma$, *then*

$$f(\pi_k) - f(\pi^*) \le 2^{-\lfloor k/l \rfloor} [f(\pi_0) - f(\pi^*) + \tfrac{2 \log |\mathcal{A}|}{1-\gamma}].$$

Proof. By using Lemma 10 and Lemma 11 (with $X_k = \mathbb{E}_{s \sim \nu^*}[V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s) + \frac{\tau_k}{1-\gamma} D_{\pi_k}^{\pi^*}(s)]$ and $Y_k = \frac{\tau_k}{1-\gamma} \log |\mathcal{A}|$), we have

$$\begin{aligned} &\mathbb{E}_{s \sim \nu^*}[V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s) + \tfrac{\tau_k}{1-\gamma} D_{\pi_k}^{\pi^*}(s)] \\ &\le 2^{-\lfloor k/l \rfloor} \left\{ \mathbb{E}_{s \sim \nu^*}[V_{\tau_0}^{\pi_k}(s) - V_{\tau_0}^{\pi^*}(s) + \tfrac{\tau_0}{1-\gamma} D_{\pi_0}^{\pi^*}(s)] + \tfrac{\log |\mathcal{A}|}{1-\gamma} \right\}. \end{aligned}$$

Noting that $V_{\tau_k}^{\pi_k}(s) \ge V^{\pi_k}(s)$, $V_{\tau_k}^{\pi^*}(s) \le V^{\pi^*}(s) + \frac{\tau_k}{1-\gamma} \log |\mathcal{A}|$, $V_{\tau_0}^{\pi^*}(s) \ge V^{\pi^*}(s)$ due to (3.18), and that $V_{\tau_0}^{\pi_0}(s) = V^{\pi_0}(s)$ due to $D_{\pi_0}^{\pi_0}(s) = 0$, we conclude from the previous inequality and the definition of τ_k that

$$\begin{aligned} &\mathbb{E}_{s \sim \nu^*}[V^{\pi_k}(s) - V^{\pi^*}(s) + \tfrac{\tau_k}{1-\gamma} D_{\pi_k}^{\pi^*}(s)] \\ &\le 2^{-\lfloor k/l \rfloor} \left\{ \mathbb{E}_{s \sim \nu^*}[V^{\pi_k}(s) - V_{\tau_0}^{\pi^*}(s) + \tfrac{\tau_0}{1-\gamma} D_{\pi_0}^{\pi^*}(s)] + \tfrac{\log |\mathcal{A}|}{1-\gamma} \right\} + \tfrac{\tau_k \log |\mathcal{A}|}{1-\gamma} \\ &\le 2^{-\lfloor k/l \rfloor} \left\{ \mathbb{E}_{s \sim \nu^*}[V^{\pi_0}(s) - V_{\tau_0}^{\pi^*}(s) + \tfrac{2 \log |\mathcal{A}|}{1-\gamma} \right\}. \end{aligned}$$

∎

In view of Theorem 4, a policy $\bar{\pi}$ s.t. $f(\bar{\pi}) - f(\pi^*) \le \epsilon$ will be found in at most $\mathcal{O}(\log(1/\epsilon))$ epochs and hence at most $\mathcal{O}(l \log(1/\epsilon)) = \mathcal{O}(\log_\gamma(\epsilon))$ iterations, which matches the one for solving RL problems with strongly convex regularizers. The difference exists in that for general RL problems, we cannot guarantee the linear convergence of $D_{\pi_{k+1}}^{\pi^*}(s)$ since its coefficient τ_k will become very small eventually.

4 Stochastic Policy Mirror Descent

The policy mirror descent methods described in the previous section require the input of the exact action-value functions Q^{π_k}. This requirement can hardly be satisfied in practice even for the case when $\mathcal{P}$ is given explicitly, since Q^{π_k} is defined as an infinite sum. In addition, in RL one does not know the transition dynamics $\mathcal{P}$ and thus only stochastic estimators of action-value functions are available. In this section, we propose stochastic versions for the PMD and APMD methods to address these issues.

4.1 Basic stochastic policy mirror descent

In this subsection, we assume that for a given policy π_k, there exists a stochastic estimator $\mathcal{Q}^{\pi_k,\xi_k}$ s.t.

$$\mathbb{E}_{\xi_k}[\mathcal{Q}^{\pi_k,\xi_k}] = \bar{\mathcal{Q}}^{\pi_k}, \tag{4.1}$$

$$\mathbb{E}_{\xi_k}[\|\mathcal{Q}^{\pi_k,\xi_k} - Q^{\pi_k}\|_\infty^2] \le \sigma_k^2, \tag{4.2}$$

$$\|\bar{\mathcal{Q}}^{\pi_k} - Q^{\pi_k}\|_\infty \le \varsigma_k, \tag{4.3}$$

for some $\sigma_k \ge$ and $\varsigma_k \ge 0$, where ξ_k denotes the random vector used to generate the stochastic estimator $\mathcal{Q}^{\pi_k,\xi_k}$. Clearly, if $\sigma_k = 0$, then we have exact information about Q^{π_k}. One key insight we have for the stochastic PMD methods is to handle separately the bias term ς_k from the overall expected error term σ_k, because one can reduce the bias term much faster than the total error. While in this section we focus on the convergence analysis of the algorithms, we will show in next section that such separate treatment of bias and total error enables us to substantially improve the sampling complexity for solving RL problems by using policy gradient type methods.

The stochastic policy mirror descent (SPMD) is obtained by replacing Q^{π_k} in (3.5) with its stochastic estimator $\mathcal{Q}^{\pi_k,\xi_k}$, i.e.,

$$\pi_{k+1}(\cdot|s) = \mathop{\arg\min}_{p(\cdot|s)\in\Delta_{|\mathcal{A}|}} \left\{\Phi_k(p) := \eta_k[\langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot), p(\cdot|s)\rangle + h^p(s)] + D^p_{\pi_k}(s)\right\}. \tag{4.4}$$

In the sequel, we denote $\xi_{\lceil k\rceil}$ the sequence of random vectors $\xi_0,\ldots,\xi_k$ and define

$$\delta_k := \mathcal{Q}^{\pi_k,\xi_k} - Q^{\pi_k}. \tag{4.5}$$

By using the assumptions in (4.1) and (4.3) and the decomposition

$$\begin{aligned}\langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle &= \langle Q^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle \\ &\quad + \langle \bar{\mathcal{Q}}^{\pi_k}(s,\cdot) - Q^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle \\ &\quad + \langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot) - \bar{\mathcal{Q}}^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle,\end{aligned}$$

we can see that

$$\begin{aligned}&\mathbb{E}_{\xi_k}[\langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle \mid \xi_{\lceil \xi_{k-1}\rceil}] \\ &\ge \langle Q^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle - 2\varsigma_k.\end{aligned} \tag{4.6}$$

Similar to Lemma 5, below we show some general convergence properties about the SPMD method. Unlike PMD, SPMD does not guarantee the non-increasing property of $V^{\pi_k}(s)$ anymore.

Lemma 12 *For any $s \in \mathcal{S}$, we have*

$$\begin{aligned}V^{\pi_{k+1}}(s) - V^{\pi_k}(s) &\le \langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s) \\ &\quad + \tfrac{1}{\eta_k} D^{\pi_{k+1}}_{\pi_k}(s) + \tfrac{\eta_k\|\delta_k\|_\infty^2}{2(1-\gamma)}.\end{aligned} \tag{4.7}$$

Proof. Observe that (3.10) still holds, and hence that

$$\begin{aligned}&V^{\pi_{k+1}}(s) - V^{\pi_k}(s) \\ &= \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\left[\langle Q^{\pi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s')\right] \\ &= \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\Big[\langle \mathcal{Q}^{\pi_k,\xi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s') \\ &\quad -\langle \delta_k, \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle\Big] \\ &\le \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\Big[\langle \mathcal{Q}^{\pi_k,\xi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s') \\ &\quad +\tfrac{1}{2\eta_k}\|\pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\|_1^2 + \tfrac{\eta_k\|\delta_k\|_\infty^2}{2}\Big] \\ &\le \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\Big[\langle \mathcal{Q}^{\pi_k,\xi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s') \\ &\quad +\tfrac{1}{\eta_k} D^{\pi_{k+1}}_{\pi_k}(s') + \tfrac{\eta_k\|\delta_k\|_\infty^2}{2}\Big],\end{aligned} \tag{4.8}$$

where the first inequality follows from Young's inequality and the second one follows from the strong convexity of $D^{\pi_k}_{\pi_{k+1}}$ w.r.t. to $\|\cdot\|_1$. Moreover, we conclude from Lemma 4 applied to (4.4) with Q^{π_k} replaced by $\mathcal{Q}^{\pi_k,\xi_k}$ and $p(\cdot|s') = \pi_k(\cdot|s')$ that

$$\begin{aligned}&\langle \mathcal{Q}^{\pi_k,\xi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s') + \tfrac{1}{\eta_k} D^{\pi_{k+1}}_{\pi_k}(s') \\ &\le -\tfrac{1}{\eta_k}[(1+\eta_k\mu) D^{\pi_k}_{\pi_{k+1}}(s')] \le 0,\end{aligned}$$

which implies that

$$\begin{aligned}
&\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\left[\langle \mathcal{Q}^{\pi_k,\xi_k}(s',\cdot),\pi_{k+1}(\cdot|s')-\pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s')-h^{\pi_k}(s')+\tfrac{1}{\eta_k}D_{\pi_{k+1}}^{\pi_k}(s')\right]\\
&\le d_s^{\pi_{k+1}}(s)\left[\langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot),\pi_{k+1}(\cdot|s)-\pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s)-h^{\pi_k}(s)+\tfrac{1}{\eta_k}D_{\pi_k}^{\pi_{k+1}}(s)\right]\\
&\le (1-\gamma)\left[\langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot),\pi_{k+1}(\cdot|s)-\pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s)-h^{\pi_k}(s)+\tfrac{1}{\eta_k}D_{\pi_k}^{\pi_{k+1}}(s)\right],
\end{aligned}$$

where the last inequality follows from the fact that $d_s^{\pi_{k+1}}(s)\ge 1-\gamma$ due to the definition of $d_s^{\pi_{k+1}}$ in (2.1). The result in (4.7) then follows immediately from (4.8) and the above inequality. ∎

We now establish an important recursion about the SPMD method.

Lemma 13 *For any $k\ge 0$, we have*

$$\begin{aligned}
&\mathbb{E}_{\xi_{\lceil k\rceil}}[f(\pi_{k+1})-f(\pi^*)+(\tfrac{1}{\eta_k}+\mu)\mathcal{D}(\pi_{k+1},\pi^*)]\\
&\le \mathbb{E}_{\xi_{\lceil k-1\rceil}}[\gamma(f(\pi_k)-f(\pi^*))+\tfrac{1}{\eta_k}\mathcal{D}(\pi_k,\pi^*)]+2\varsigma_k+\tfrac{\eta_k\sigma_k^2}{2(1-\gamma)}.
\end{aligned}$$

Proof. By applying Lemma 4 to (3.5) (with Q^{π_k} replaced by $\mathcal{Q}^{\pi_k,\xi_k}$ and $p=\pi^*$), we have

$$\begin{aligned}
&\eta_k[\langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot),\pi_{k+1}(\cdot|s)-\pi^*(\cdot|s)\rangle + h^{\pi_{k+1}}(s)-h^{\pi^*}(s)]+D_{\pi_k}^{\pi_{k+1}}(s)\\
&\le D_{\pi_k}^{\pi^*}(s)-(1+\eta_k\mu)D_{\pi_{k+1}}^{\pi^*}(s),
\end{aligned}$$

which, in view of (4.7), then implies that

$$\begin{aligned}
&\langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot),\pi_k(\cdot|s)-\pi^*(\cdot|s)\rangle + h^{\pi_k}(s)-h^{\pi^*}(s)+V^{\pi_{k+1}}(s)-V^{\pi_k}(s)\\
&\le \tfrac{1}{\eta_k}D_{\pi_k}^{\pi^*}(s)-(\tfrac{1}{\eta_k}+\mu)D_{\pi_{k+1}}^{\pi^*}(s)+\tfrac{\eta_k\|\delta_k\|_\infty^2}{2(1-\gamma)}.
\end{aligned}$$

Taking expectation w.r.t. $\xi_{\lceil k\rceil}$ and ν^* on both sides of the above inequality, and using Lemma 3 and the relation in (4.6), we arrive at

$$\begin{aligned}
&\mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}\left[(1-\gamma)(V^{\pi_k}(s)-V^{\pi_\tau^*}(s))+V^{\pi_{k+1}}(s)-V^{\pi_k}(s)\right]\\
&\le \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}[\tfrac{1}{\eta_k}D_{\pi_k}^{\pi^*}(s)-(\tfrac{1}{\eta_k}+\mu)D_{\pi_{k+1}}^{\pi^*}(s)]+2\varsigma_k+\tfrac{\eta_k\sigma_k^2}{2(1-\gamma)}.
\end{aligned}$$

Noting $V^{\pi_{k+1}}(s)-V^{\pi_k}(s)=V^{\pi_{k+1}}(s)-V^{\pi^*}(s)-[V^{\pi_k}(s)-V^{\pi^*}(s)]$, rearranging the terms in the above inequality, and using the definition of f in (1.9), we arrive at the result. ∎

We are now ready to establish the convergence rate of the SPMD method. We start with the case when $\mu>0$ and state a constant stepsize rule which requires both ς_k and σ_k, $k\ge 0$, to be small enough to guarantee the convergence of the SPMD method.

Theorem 5 *Suppose that $\eta_k=\eta=\frac{1-\gamma}{\gamma\mu}$ in the SPMD method. If $\varsigma_k=2^{-(\lfloor k/l\rfloor+2)}$ and $\sigma_k^2=2^{-(\lfloor k/l\rfloor+2)}$ for any $k\ge 0$ with $l:=\lceil\log_\gamma(1/4)\rceil$, then*

$$\begin{aligned}
&\mathbb{E}_{\xi_{\lceil k-1\rceil}}[f(\pi_k)-f(\pi^*)+\tfrac{\mu}{1-\gamma}\mathcal{D}(\pi_k,\pi^*)]\\
&\le 2^{-\lfloor k/l\rfloor}\left[f(\pi_0)-f(\pi^*)+\tfrac{1}{1-\gamma}(\mu\log|\mathcal{A}|+\tfrac{5}{2}+\tfrac{5}{8\gamma\mu})\right].
\end{aligned}\tag{4.9}$$

Proof. By Lemma 13 and the selection of η, we have

$$\begin{aligned}
&\mathbb{E}_{\xi_{\lceil k\rceil}}[f(\pi_{k+1})-f(\pi^*)+\tfrac{\mu}{1-\gamma}\mathcal{D}(\pi_{k+1},\pi^*)]\\
&\le \gamma[\mathbb{E}_{\xi_{\lceil k-1\rceil}}[f(\pi_k)-f(\pi^*)+\tfrac{\mu}{1-\gamma}\mathcal{D}(\pi_k,\pi^*)]+2\varsigma_k+\tfrac{\sigma_k^2}{2\gamma\mu},
\end{aligned}$$

which, in view of Lemma 11 with $X_k=\mathbb{E}_{\xi_{\lceil k-1\rceil}}[f(\pi_k)-f(\pi^*)+\frac{\mu}{1-\gamma}\mathcal{D}(\pi_k,\pi^*)$ and $Z_k=2\varsigma_k+\frac{\sigma_k^2}{2\gamma\mu}$, then implies that

$$\begin{aligned}
&\mathbb{E}_{\xi_{\lceil k-1\rceil}}[f(\pi_k)-f(\pi^*)+\tfrac{\mu}{1-\gamma}\mathcal{D}(\pi_k,\pi^*)\\
&\le \gamma^{\lfloor k/l\rfloor}\left[f(\pi_0)-f(\pi^*)+\tfrac{\mu\mathcal{D}(\pi_0,\pi^*)}{1-\gamma}+\tfrac{5}{4}(\tfrac{2}{1-\gamma}+\tfrac{1}{2\gamma(1-\gamma)\mu})\right]\\
&\le \gamma^{\lfloor k/l\rfloor}\left[f(\pi_0)-f(\pi^*)+\tfrac{1}{1-\gamma}(\mu\log|\mathcal{A}|+\tfrac{5}{2}+\tfrac{5}{8\gamma\mu})\right].
\end{aligned}$$

∎

We now turn our attention to the convergence properties of the SPMD method for the case when $\mu=0$.

Theorem 6 *Suppose that $\eta_k = \eta$ for any $k \geq 0$ in the SPMD method. If $\varsigma_k \leq \varsigma$ and $\sigma_k \leq \sigma$ for any $k \geq 0$, then we have*

$$\mathbb{E}_{\xi_{\lceil k \rceil}, R}[f(\pi_R) - f(\pi^*)] \leq \frac{\gamma[f(\pi_0) - f(\pi^*)]}{(1-\gamma)k} + \frac{\log|\mathcal{A}|}{\eta(1-\gamma)k} + \frac{2\varsigma}{1-\gamma} + \frac{\eta\sigma^2}{2(1-\gamma)^2}, \tag{4.10}$$

where R denotes a random number uniformly distributed between 1 and k. In particular, if the number of iterations k is given a priori and $\eta = (\frac{2(1-\gamma)\log|\mathcal{A}|}{k\sigma^2})^{1/2}$, then

$$\mathbb{E}_{\xi_{\lceil k \rceil}, R}[f(\pi_R) - f(\pi^*)] \leq \frac{\gamma[f(\pi_0) - f(\pi^*)]}{(1-\gamma)k} + \frac{2\varsigma}{1-\gamma} + \frac{\sigma\sqrt{2\log|\mathcal{A}|}}{(1-\gamma)^{3/2}\sqrt{k}}. \tag{4.11}$$

Proof. By Lemma 13 and the fact that $\mu = 0$, we have

$$\begin{aligned}
&\mathbb{E}_{\xi_{\lceil k \rceil}}[f(\pi_{k+1}) - f(\pi^*) + \tfrac{1}{\eta}\mathcal{D}(\pi_{k+1}, \pi^*)] \\
&\leq \mathbb{E}_{\xi_{\lceil k-1 \rceil}}[\gamma(f(\pi_k) - f(\pi^*)) + \tfrac{1}{\eta}\mathcal{D}(\pi_k, \pi^*)] + 2\varsigma_k + \tfrac{\eta\sigma_k^2}{2(1-\gamma)}.
\end{aligned}$$

Taking the telescopic sum of the above relations, we have

$$(1-\gamma)\textstyle\sum_{i=1}^k \mathbb{E}_{\xi_{\lceil k \rceil}}[f(\pi_i) - f(\pi^*)] \leq [\gamma(f(\pi_0) - f(\pi^*)) + \tfrac{1}{\eta}\mathcal{D}(\pi_0, \pi^*)] + 2k\varsigma + \tfrac{k\eta\sigma^2}{2(1-\gamma)}.$$

Dividing both sides by $(1-\gamma)k$ and using the definition of R, we obtain the result in (4.10). ∎

We add some remarks about the results in Theorem 6. In comparison with the convergence results of SPMD for the case $\mu > 0$, there exist some possible shortcomings for the case when $\mu = 0$. Firstly, one needs to output a randomly selected π_R from the trajectory. Secondly, since the first term in (4.11) converges sublinearly, one has to update π_{k+1} at least $\mathcal{O}(1/\epsilon)$ times, which may also impact the gradient complexity of computing ∇h^π if π_{k+1} cannot be computed explicitly. We will address these issues by developing the stochastic APMD method in next subsection.

4.2 Stochastic approximate policy mirror descent

The stochastic approximate policy mirror descent (SAPMD) method is obtained by replacing $Q_{\tau_k}^{\pi_k}$ in (3.20) with its stochastic estimator $\mathcal{Q}_{\tau_k}^{\pi_k,\xi_k}$. As such, its updating formula is given by

$$\pi_{k+1}(\cdot|s) = \operatorname*{arg\,min}_{p(\cdot|s)\in\Delta_{|\mathcal{A}|}} \left\{\eta_k[\langle \mathcal{Q}_{\tau_k}^{\pi_k,\xi_k}(s,\cdot), p(\cdot|s)\rangle + h^p(s) + \tau_k D_{\pi_0}^{\pi}(s_t)] + D_{\pi_k}^p(s)\right\}. \tag{4.12}$$

With a little abuse of notation, we still denote $\delta_k := \mathcal{Q}_{\tau_k}^{\pi_k,\xi_k} - Q_{\tau_k}^{\pi_k}$ and assume that

$$\mathbb{E}_{\xi_k}[\mathcal{Q}_{\tau_k}^{\pi_k,\xi_k}] = \bar{\mathcal{Q}}_{\tau_k}^{\pi_k}, \tag{4.13}$$

$$\mathbb{E}_{\xi_k}[\|\mathcal{Q}_{\tau_k}^{\pi_k,\xi_k} - Q_{\tau_k}^{\pi_k}\|_\infty^2] \leq \sigma_k^2, \tag{4.14}$$

$$\|\bar{\mathcal{Q}}_{\tau_k}^{\pi_k} - Q_{\tau_k}^{\pi_k}\|_\infty \leq \varsigma_k, \tag{4.15}$$

for some $\sigma_k \geq$ and $\varsigma_k \geq 0$. Similarly to (4.6) we have

$$\begin{aligned}
&\mathbb{E}_{\xi_k}[\langle \mathcal{Q}_{\tau_k}^{\pi_k,\xi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle \mid \xi_{\lceil \xi_{k-1} \rceil}] \\
&\geq \langle Q_{\tau_k}^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle - 2\varsigma_k.
\end{aligned} \tag{4.16}$$

Lemma 14 and Lemma 15 below show the improvement for each SAPMD iteration.

Lemma 14 *For any $k \geq 0$, we have*

$$\begin{aligned}
V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi_k}(s) \leq\ & \langle \mathcal{Q}_{\tau_k}^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s) \\
&+ \tau_k[D_{\pi_0}^{\pi_{k+1}}(s) - D_{\pi_0}^{\pi^*}(s)] + \tfrac{1}{\eta_k} D_{\pi_k}^{\pi_{k+1}}(s) + \tfrac{\eta_k\|\delta_k\|_\infty^2}{2(1-\gamma)}.
\end{aligned} \tag{4.17}$$

Proof. The proof is similar to the one for Lemma 12 except that we will apply Lemma 2 to the perturbed value functions $V_{\tau_k}^{\pi}$ instead of V^π. ∎

Lemma 15 *If $1 + \eta_k\tau_k = 1/\gamma$ and $\tau_k \geq \tau_{k+1}$ in the SAPMD method, then for any $k \geq 0$,*

$$\begin{aligned}
&\mathbb{E}_{s\sim\nu^*, \xi_{\lceil k \rceil}}[V_{\tau_{k+1}}^{\pi_{k+1}}(s) - V_{\tau_{k+1}}^{\pi^*}(s) + \tfrac{\tau_k}{1-\gamma} D_{\pi_{k+1}}^{\pi^*}(s)] \\
&\leq \mathbb{E}_{s\sim\nu^*, \xi_{\lceil k-1 \rceil}}[\gamma[V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s) + \tfrac{\tau_k}{1-\gamma} D_{\pi_k}^{\pi^*}(s)] \\
&\quad + \tfrac{\tau_k - \tau_{k+1}}{1-\gamma}\log|\mathcal{A}| + 2\varsigma_k + \tfrac{\sigma_k^2}{2\gamma\tau_k}.
\end{aligned} \tag{4.18}$$

Proof. By Lemma 8 with $p = \pi^*$ and $Q_{\tau_k}^{\pi_k}$ replaced by $\mathcal{Q}_{\tau_k}^{\pi_k,\xi_k}$, we have

$$\begin{aligned}&\langle \mathcal{Q}_{\tau_k}^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi^*(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi^*}(s)\\ &+ \tau_k[D_{\pi_0}^{\pi_{k+1}}(s) - D_{\pi_0}^{\pi^*}(s)] + \tfrac{1}{\eta_k} D_{\pi_k}^{\pi_{k+1}}(s)\\ &\le \tfrac{1}{\eta_k} D_{\pi_k}^{\pi^*}(s) - (\tfrac{1}{\eta_k} + \tau_k) D_{\pi_{k+1}}^{\pi^*}(s),\end{aligned}$$

which, in view of (4.17), implies that

$$\begin{aligned}&\langle \mathcal{Q}_{\tau_k}^{\pi_k,\xi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle + h^{\pi_k}(s) - h^{\pi^*}(s) + \tau_k[D_{\pi_0}^{\pi_k}(s) - D_{\pi_0}^{\pi^*}(s)]\\ &+ V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi_k}(s) \le \tfrac{1}{\eta_k} D_{\pi_k}^{\pi^*}(s) - (\tfrac{1}{\eta_k} + \tau_k) D_{\pi_{k+1}}^{\pi^*}(s) + \tfrac{\eta_k \|\delta_k\|_\infty^2}{2(1-\gamma)}.\end{aligned}$$

Taking expectation w.r.t. $\xi_{\lceil k\rceil}$ and ν^* on both sides of the above inequality, and using Lemma 7 and the relation in (4.16), we arrive at

$$\begin{aligned}&\mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}[(1-\gamma)(V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s))] + \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}[V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi_k}(s)]\\ &\le \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}[\tfrac{1}{\eta_k} D_{\pi_k}^{\pi^*}(s) - (\tfrac{1}{\eta_k} + \tau_k) D_{\pi_{k+1}}^{\pi^*}(s)] + 2\varsigma_k + \tfrac{\eta_k \sigma_k^2}{2(1-\gamma)}.\end{aligned}$$

Noting $V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi_k}(s) = V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi^*}(s) - [V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s)]$ and rearranging the terms in the above inequality, we have

$$\begin{aligned}&\mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}[V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi^*}(s) + (\tfrac{1}{\eta_k} + \tau_k) D_{\pi_{k+1}}^{\pi^*}(s)]\\ &\le \gamma \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k-1\rceil}}[V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s)] + \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k-1\rceil}}[\tfrac{1}{\eta_k} D_{\pi_k}^{\pi^*}(s)] + 2\zeta_k + \tfrac{\eta_k \sigma_k^2}{2(1-\gamma)},\end{aligned}$$

which, in view of the assumption $\tau_k \ge \tau_{k+1}$ and (3.19), then implies that

$$\begin{aligned}V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi_k}(s) \le\ &\langle \mathcal{Q}_{\tau_k}^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s)\\ &+ \tau_k[D_{\pi_0}^{\pi_{k+1}}(s) - D_{\pi_0}^{\pi^*}(s)] + \tfrac{1}{\eta_k} D_{\pi_k}^{\pi_{k+1}}(s) + \tfrac{\eta_k \|\delta_k\|_\infty^2}{2(1-\gamma)}.\end{aligned} \tag{4.19}$$

The result then immediately follows from the assumption that $1 + \eta_k \tau_k = 1/\gamma$. ∎

We are now ready to establish the convergence of the SAPMD method.

Theorem 7 *Suppose that $\eta_k = \frac{1-\gamma}{\gamma\tau_k}$ in the SAPMD method. If $\tau_k = \frac{1}{\sqrt{\gamma \log|\mathcal{A}|}} 2^{-(\lfloor k/l\rfloor+1)}$, $\varsigma_k = 2^{-(\lfloor k/l\rfloor+2)}$, and $\sigma_k^2 = 4^{-(\lfloor k/l\rfloor+2)}$ with $l := \lceil \log_\gamma(1/4)\rceil$, then*

$$\mathbb{E}_{\xi_{\lceil k-1\rceil}}[f(\pi_k) - f(\pi^*)] \le 2^{-\lfloor k/l\rfloor}[f(\pi_0) - f(\pi^*) + \tfrac{3\sqrt{\log|\mathcal{A}|}}{(1-\gamma)\sqrt{\gamma}} + \tfrac{5}{2(1-\gamma)}]. \tag{4.20}$$

Proof. By Lemma 15 and the selection of τ_k, ς_k and σ_k, we have

$$\begin{aligned}&\mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}[V_{\tau_{k+1}}^{\pi_{k+1}}(s) - V_{\tau_{k+1}}^{\pi^*}(s) + \tfrac{\tau_k}{1-\gamma} D_{\pi_{k+1}}^{\pi^*}(s)]\\ &\le \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k-1\rceil}}[\gamma[V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s) + \tfrac{\tau_k}{1-\gamma} D_{\pi_k}^{\pi^*}(s)]]\\ &\quad + \tfrac{\tau_k - \tau_{k+1}}{1-\gamma}\log|\mathcal{A}| + (2 + \tfrac{\sqrt{\log|\mathcal{A}|}}{2\sqrt{\gamma}}) 2^{-(\lfloor k/l\rfloor+2)}.\end{aligned} \tag{4.21}$$

Using the above inequality and Lemma 11 (with $X_k = \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k-1\rceil}}[\gamma[V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s) + \frac{\tau_k}{1-\gamma} D_{\pi_k}^{\pi^*}(s)]]$, $Y_k = \frac{\tau_k}{1-\gamma}\log|\mathcal{A}|$ and $Z_k = (2 + \frac{\sqrt{\log|\mathcal{A}|}}{2\sqrt{\gamma}}) 2^{-(\lfloor k/l\rfloor+2)}$), we conclude

$$\begin{aligned}&\mathbb{E}_{s\sim\nu^*,\xi_{\lceil k-1\rceil}}[V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s) + \tfrac{\tau_k}{1-\gamma} D_{\pi_k}^{\pi^*}(s)]\\ &\le 2^{-\lfloor k/l\rfloor}\{\mathbb{E}_{s\sim\nu^*}[V_{\tau_0}^{\pi_k}(s) - V_{\tau_0}^{\pi^*}(s) + \tfrac{\sqrt{\log|\mathcal{A}|}}{2(1-\gamma)\sqrt{\gamma}}] + \tfrac{\sqrt{\log|\mathcal{A}|}}{(1-\gamma)\sqrt{\gamma}} + \tfrac{5}{2(1-\gamma)} + \tfrac{5\sqrt{\log|\mathcal{A}|}}{8(1-\gamma)\sqrt{\gamma}}\}\\ &= 2^{-\lfloor k/l\rfloor}\{\mathbb{E}_{s\sim\nu^*}[V_{\tau_0}^{\pi_k}(s) - V_{\tau_0}^{\pi^*}(s)] + \tfrac{17\sqrt{\log|\mathcal{A}|}}{8(1-\gamma)\sqrt{\gamma}} + \tfrac{5}{2(1-\gamma)}\}.\end{aligned}$$

Noting that $V_{\tau_k}^{\pi_k}(s) \ge V^{\pi_k}(s)$, $V_{\tau_k}^{\pi^*}(s) \le V^{\pi^*}(s) + \frac{\tau_k}{1-\gamma}\log|\mathcal{A}|$, $V_{\tau_0}^{\pi^*}(s) \ge V^{\pi^*}(s)$ due to (3.18), and that $V_{\tau_0}^{\pi_0}(s) = V^{\pi_0}(s)$ due to $D_{\pi_0}^{\pi_0}(s) = 0$, we conclude from the previous inequality and the definition of τ_k that

$$\mathbb{E}_{s\sim\nu^*,\xi_{\lceil k-1\rceil}}[V^{\pi_k}(s) - V^{\pi^*}(s)] \le 2^{-\lfloor k/l\rfloor}\{\mathbb{E}_{s\sim\nu^*}[V^{\pi_0}(s) - V^{\pi^*}(s) + \tfrac{3\sqrt{\log|\mathcal{A}|}}{(1-\gamma)\sqrt{\gamma}} + \tfrac{5}{2(1-\gamma)}\},$$

from which the result immediately follows. ∎

In view of Theorem 7, the SAPMD method does not need to randomly output a solution as most existing nonconvex stochastic gradient descent methods did. Instead, the linear rate of convergence in (4.20) has been established for the last iterate π_k generated by this algorithm.

5 Stochastic Estimation for Action-value Functions

In this section, we discuss the estimation of the action-value functions Q^π or Q^π_τ through two different approaches. In Subsection 5.1, we assume the existence of a generative model for the Markov Chains so that we can estimate value functions by generating multiple independent trajectories starting from an arbitrary pair of state and action. In Subsection 5.2, we consider a more challenging setting where we only have access to a single trajectory observed when the dynamic system runs online. In this case, we employ and enhance the conditional temporal difference (CTD) method recently developed in [12] to estimate value functions. Throughout the section we assume that

$$c(s,a) \le \bar{c}, \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{5.1}$$

$$h^\pi(s) \le \bar{h}, \forall s \in \mathcal{S}, \pi \in \Delta_{|\mathcal{A}|}. \tag{5.2}$$

5.1 Multiple independent trajectories

In the multiple trajectory setting, starting from state-action pair (s,a) and following policy π_k, we can generate M_k independent trajectories of length T_k, denoted by

$$\zeta_k^i \equiv \zeta_k^i(s,a) := \{(s_0^i = s, a_0^i = a); (s_1^i, a_1^i), \ldots, (s_{T_k-1}^i, a_{T_k-1}^i)\}, i = 1, \ldots, M_k.$$

Let $\xi_k := \{\zeta_k^i(s,a), i = 1, \ldots, M_k, s \in \mathcal{S}, a \in \mathcal{A}\}$ denote all these random variables. We can estimate Q^{π_k} in the SPMD method by

$$\mathcal{Q}^{\pi_k,\xi_k}(s,a) = \tfrac{1}{M_k} \textstyle\sum_{i=1}^{M_k} \sum_{t=0}^{T_k-1} \gamma^t [c(s_t^i, a_t^i) + h^{\pi_k}(s_t^i)].$$

It is easy to see that $\mathcal{Q}^{\pi_k,\xi_k}$ satisfy (4.1)-(4.3) with

$$\varsigma_k = \tfrac{(\bar{c}+\bar{h})\gamma^{T_k}}{1-\gamma} \quad \text{and} \quad \sigma_k^2 = 2\varsigma_k^2 + \tfrac{2(\bar{c}+\bar{h})^2}{(1-\gamma)^2 M_k} = \tfrac{2(\bar{c}+\bar{h})^2}{(1-\gamma)^2}(\gamma^{2T_k} + \tfrac{1}{M_k}). \tag{5.3}$$

By choosing T_k and M_k properly, we can show the convergence of the SPMD method employed with different stepsize rules as stated in Theorems 5 and 6.

Proposition 1 *Suppose that $\eta_k = \frac{1-\gamma}{\gamma\mu}$ in the SPMD method. If T_k and M_k are chosen such that*

$$T_k \ge \tfrac{l}{2}(\lfloor k/l \rfloor + \log_2 \tfrac{\bar{c}+\bar{h}}{1-\gamma} + 2) \quad \textit{and} \quad M_k \ge \tfrac{(\bar{c}+\bar{h})^2}{(1-\gamma)^2} 2^{\lfloor k/l \rfloor + 4}$$

with $l := \lceil \log_\gamma(1/4) \rceil$, then the relation in (4.9) holds. As a consequence, an ϵ-solution of (1.9), i.e., a solution $\bar{\pi}$ s.t. $\mathbb{E}[f(\bar{\pi}) - f(\pi^*) + \frac{\mu}{1-\gamma}\mathcal{D}(\bar{\pi}, \pi^*)] \le \epsilon$, can be found in at most $\mathcal{O}(\log_\gamma \epsilon)$ SPMD iterations. In addition, the total number of samples for (s_t, a_t) pairs can be bounded by

$$\mathcal{O}(\tfrac{|\mathcal{S}||\mathcal{A}| \log |\mathcal{A}| \log_\gamma \epsilon}{\mu(1-\gamma)^3 \epsilon}). \tag{5.4}$$

Proof. Using the fact that $\gamma^l \le 1/4$, we can easily check from (5.3) and the selection of T_k and M_k that (4.1)-(4.3) hold with $\varsigma_k = 2^{-(\lfloor k/l \rfloor + 2)}$ and $\sigma_k^2 = 2^{-(\lfloor k/l \rfloor + 2)}$. Suppose that an ϵ-solution $\bar{\pi}$ will be found at the $\bar{k}$ iteration. By (4.9), we have

$$\lfloor \bar{k}/l \rfloor \le \log_2 \{[f(\pi_0) - f(\pi^*) + \tfrac{1}{1-\gamma}(\mu \log |\mathcal{A}| + \tfrac{5}{2} + \tfrac{5}{8\gamma\mu})]\epsilon^{-1}\},$$

which implies that the number of iterations is bounded by $\mathcal{O}(l\lfloor \bar{k}/l \rfloor) = \mathcal{O}(\log_\gamma \epsilon)$. Moreover by the definition of T_k and M_k, the total number of samples is bounded by

$$\begin{aligned} |\mathcal{S}||\mathcal{A}| \textstyle\sum_{p=0}^{\lfloor \bar{k}/l \rfloor} T_k M_k &= |\mathcal{S}||\mathcal{A}| \textstyle\sum_{p=0}^{\lfloor \bar{k}/l \rfloor} [\tfrac{l}{2}(p + \log_2 \tfrac{\bar{c}+\bar{h}}{1-\gamma} + 2) \tfrac{(\bar{c}+\bar{h})^2}{(1-\gamma)^2} 2^{p+4}] \\ &= \mathcal{O}\{|\mathcal{S}||\mathcal{A}| l (\lfloor \bar{k}/l \rfloor + \log_2 \tfrac{\bar{c}+\bar{h}}{1-\gamma}) \tfrac{(\bar{c}+\bar{h})^2}{(1-\gamma)^2} 2^{\lfloor \bar{k}/l \rfloor}\} = \mathcal{O}(\tfrac{|\mathcal{S}||\mathcal{A}| \log |\mathcal{A}| \log_\gamma \epsilon}{\mu(1-\gamma)^3 \epsilon}). \end{aligned}$$

■

To the best of our knowledge, this is the first time in the literature that an $\mathcal{O}(\log(1/\epsilon)/\epsilon)$ sampling complexity, after disregarding all constant factors, has been obtained for solving RL problems with strongly convex regularizers, even though problem (1.9) is still nonconvex. The previously best-known sampling complexity for RL problems with entropy regularizer was $\tilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}|^2/\epsilon^3)$ [20], and the author was not aware of an $\tilde{\mathcal{O}}(1/\epsilon)$ sampling complexity results for any RL problems.

Below we discuss the sampling complexities of SPMD and SAPMD for solving RL problems with general convex regularizers.

Proposition 2 *Consider the general RL problems with $\mu = 0$. Suppose that the number of iterations k is given a priori and $\eta_k = (\frac{2(1-\gamma)\log|\mathcal{A}|}{k\sigma^2})^{1/2}$. If $T_k \geq T \equiv \log_\gamma \frac{(1-\gamma)\epsilon}{3(\bar{c}+\bar{h})}$ and $M_k = 1$, then an ϵ-solution of problem of (1.9), i.e., a solution $\bar{\pi}$ s.t. $\mathbb{E}[f(\bar{\pi}) - f(\pi^*)] \leq \epsilon$, can be found in at most $\mathcal{O}(\log|\mathcal{A}|/[(1-\gamma)^5\epsilon^2])$ SPMD iterations. In addition, the total number of state-action samples can be bounded by*

$$\mathcal{O}(\frac{|\mathcal{S}||\mathcal{A}|\log|\mathcal{A}|\log_\gamma \epsilon}{(1-\gamma)^5\epsilon^2}). \tag{5.5}$$

Proof. We can easily check from (5.3) and the selection of T_k and M_k that (4.1)-(4.3) holds with $\varsigma_k = \epsilon/3$ and $\sigma_k^2 = 2(\frac{\epsilon^2}{3^2} + \frac{2(\bar{c}+\bar{h})^2}{(1-\gamma)^2})$. Using these bounds in (4.10), we conclude that an ϵ-solution will be found in at most

$$\bar{k} = \frac{4[(\epsilon/3)^2 + (\bar{c}+\bar{h})^2/(1-\gamma)^2)]\log|\mathcal{A}|}{(1-\gamma)^3(\epsilon/3)^2} + \frac{\gamma[f(\pi_0) - f(\pi^*)]}{(1-\gamma)(\epsilon/3)}.$$

Moreover, the total number of samples is bounded by $|\mathcal{S}||\mathcal{A}|T\bar{k}$ and hence by (5.5). ∎

We can further improve the above iteration and sampling complexities by using the SAPMD method, in which we estimate $Q_{\tau_k}^{\pi_k}$ by

$$\mathcal{Q}_{\pi_k}^{\pi_k,\xi_k}(s,a) = \frac{1}{M_k}\sum_{i=1}^{M_k}\sum_{t=0}^{T_k-1}\gamma^t[c(s_t^i,a_t^i) + h^{\pi_k}(s_t^i) + \tau_k D_{\pi_0}^{\pi_k}(s_t^i)].$$

Since $\tau_0 \geq \tau_k$, we can easily see that $\mathcal{Q}^{\pi_k,\xi_k}$ satisfy (4.13)-(4.15) with

$$\varsigma_k = \frac{(\bar{c}+\bar{h}+\tau_0\log|\mathcal{A}|)\gamma^{T_k}}{1-\gamma} \quad \text{and} \quad \sigma_k^2 = \frac{2(\bar{c}+\bar{h}+\tau_0\log|\mathcal{A}|)^2}{(1-\gamma)^2}(\gamma^{2T_k} + \frac{1}{M_k}). \tag{5.6}$$

Proposition 3 *Suppose that $\eta_k = \frac{1-\gamma}{\gamma\tau_k}$ and $\tau_k = \frac{1}{\sqrt{\gamma\log|\mathcal{A}|}}2^{-(\lfloor k/l\rfloor+1)}$ in the SAPMD method. If T_k and M_k are chosen such that*

$$T_k \geq \frac{l}{2}(\lfloor k/l\rfloor + \log_2\frac{\bar{c}+\bar{h}+\tau_0\log|\mathcal{A}|}{1-\gamma} + 4) \quad \text{and} \quad M_k \geq \frac{(\bar{c}+\bar{h}+\tau_0\log|\mathcal{A}|)^2}{(1-\gamma)^2}4^{\lfloor k/l\rfloor+3}$$

with $l := \lceil\log_\gamma(1/4)\rceil$, then the relation in (4.20) holds. As a consequence, an ϵ-solution of (1.9), i.e., a solution $\bar{\pi}$ s.t. $\mathbb{E}[f(\bar{\pi}) - f(\pi^*)] \leq \epsilon$, can be found in at most $\mathcal{O}(\log_\gamma\epsilon)$ SAPMD iterations. In addition, the total number of samples for (s_t,a_t) pairs can be bounded by

$$\mathcal{O}(\frac{|\mathcal{S}||\mathcal{A}|\log^2|\mathcal{A}|\log_\gamma\epsilon}{(1-\gamma)^4\epsilon^2}). \tag{5.7}$$

Proof. Using the fact that $\gamma^l \leq 1/4$, we can easily check from (5.6) and the selection of T_k and M_k that (4.13)-(4.15) hold with $\varsigma_k = 2^{-(\lfloor k/l\rfloor+2)}$ and $\sigma_k^2 = 4^{-(\lfloor k/l\rfloor+2)}$. Suppose that an ϵ-solution $\bar{\pi}$ will be found at the $\bar{k}$ iteration. By (4.20), we have

$$\lfloor\bar{k}/l\rfloor \leq \log_2\{[f(\pi_0) - f(\pi^*) + \frac{3\sqrt{\log|\mathcal{A}|}}{(1-\gamma)\sqrt{\gamma}} + \frac{5}{2(1-\gamma)}]\epsilon^{-1}\},$$

which implies that the number of iterations is bounded by $\mathcal{O}(l\lfloor\bar{k}/l\rfloor) = \mathcal{O}(\log_\gamma\epsilon)$. Moreover by the definition of T_k and M_k, the number of samples is bounded by

$$\begin{aligned}
&|\mathcal{S}||\mathcal{A}|\textstyle\sum_{p=1}^{\lfloor\bar{k}/l\rfloor+1}[\frac{l}{2}(p + \log_2\frac{\bar{c}+\bar{h}+\tau_0\log|\mathcal{A}|}{1-\gamma} + 4)\frac{(\bar{c}+\bar{h}+\tau_0\log|\mathcal{A}|)^2}{(1-\gamma)^2}4^{p+3}] \\
&= \mathcal{O}\{|\mathcal{S}||\mathcal{A}|l(\lfloor\bar{k}/l\rfloor + \log_2\frac{\bar{c}+\bar{h}+\tau_0\log|\mathcal{A}|}{1-\gamma})\frac{(\bar{c}+\bar{h}+\tau_0\log|\mathcal{A}|)^2}{(1-\gamma)^2}4^{\lfloor\bar{k}/l\rfloor}\} = \mathcal{O}(\frac{|\mathcal{S}||\mathcal{A}|\log^2|\mathcal{A}|\log_\gamma\epsilon}{(1-\gamma)^4\epsilon}).
\end{aligned}$$

∎

To the best of our knowledge, the results in Propositions 2 and 3 appear to be new for policy gradient type methods. The previously best-known sampling complexity for policy gradient methods for RL problems was $\tilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}|^2/\epsilon^4)$ (e.g., [20]) although some improvements have been made under certain specific settings (e.g., [24]).

5.2 Conditional temporal difference

In this subsection, we enhance a recently developed temporal different (TD) type method, i.e., conditional temporal difference (CTD) method, and use it to estimate the action-value functions in an online manner. We focus on estimating Q^π in SPMD since the estimation of Q_τ^π in SAPMD is similar.

For a given policy π, we denote the Bellman operator

$$T^\pi Q(s,a) := c(s,a) + h^\pi(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) \sum_{a' \in \mathcal{A}} \pi(a'|s) Q(s',a'). \tag{5.8}$$

The action value function Q^π corresponding to policy π satisfies the Bellman equation

$$Q^\pi(s,a) = T^\pi Q^\pi(s,a). \tag{5.9}$$

We also need to define a positive-definite weighting matrix $M^\pi \in \mathbb{R}^{n\times n}$ to define the sampling scheme to evaluate policies using TD-type methods. A natural weighting matrix is the diagonal matrix $M^\pi = \mathrm{Diag}(\nu(\pi)) \otimes \mathrm{Diag}(\pi)$, where $\nu(\pi)$ is the steady state distribution induced by π and $\otimes$ denotes the Kronecker product. A widely used assumption in the RL literature is that $M^\pi \succ 0$. This assumption requires: (a) the policy π is sufficiently random, i.e., $\pi(s,a) \geq \underline{\pi}$ for some $\underline{\pi} > 0$, which can be enforced, for example, by adding some corresponding constraints through h^π; and (b) $\nu(\pi)(s) \geq \underline{\nu}$ for some $\underline{\nu} > 0$, which holds when the Markov chain employed with policy π has a single ergodic class with unique stationary distribution, i.e., $\nu(\pi) = \nu(\pi)\mathcal{P}^\pi$. With the weighting matrix M^π, we define the operator F^π as

$$F^\pi(\theta) := M^\pi\big(\theta - T^\pi\theta\big),$$

where T^π is the Bellman operator defined in (5.8). Our goal is to find the root $\theta^* \equiv Q^\pi$ of $F(\theta)$, i.e., $F(\theta^*) = 0$. We can show that F is strongly monotone with strong monotonicity modulus bounded from below by $\Lambda_{\min} := (1-\gamma)\lambda_{\min}(M^\pi)$. Here $\lambda_{\min}(A)$ denotes the smallest eigenvalue of A. It can also be easily seen that F^π is Lipschitz continuous with Lipschitz constant bounded by $\Lambda_{\max} := (1-\gamma)(\lambda_{\max}(M^\pi)$, where $\lambda_{\max}(A)$ denotes the largest eigenvalue of A.

Remark 1 If $\nu(\pi)(s) \cdot \pi(s,a) = 0$ for some $(s,a) \in \mathcal{S}\times\mathcal{A}$, one may define the weighting matrix $M^\pi = (1-\lambda)\mathrm{Diag}(\nu(\pi)) \otimes \mathrm{Diag}(\pi) + \frac{\lambda}{n} I$ for some sufficiently small $\lambda \in (0,1)$ which depends on the target accuracy for solving the RL problem, where $n = |\mathcal{S}| \times |\mathcal{A}|$. Obviously, the selection of λ will impact the efficiency estimte for policy evaluation. However, the algorithmic frameworks of CTD and SPMD, and their convergence analysis are still applicable to this more general setting. In this paper we focus on the more restrictive assumption $M^\pi \succ 0$ in order to compare our results with the existing ones in the literature. ∎

Remark 2 For problems of high dimension (i.e., $n \equiv |\mathcal{S}| \times |\mathcal{A}|$ is large), one often resorts to a parametric approximation of the value function. It is possible to define a more general operator $F^\pi(\theta) := \Phi^T M^\pi\big(\Phi\theta - T^\pi\Phi\theta\big)$ for some feature matrix Φ (see Section 4 of [12] for a discussion about CTD with function approximation). ∎

At time instant $t \in \mathbb{Z}_+$, we define the stochastic operator of F^π as

$$\tilde{F}^\pi(\theta_t, \zeta_t) = \big(\langle e(s_t,a_t), \theta_t\rangle - c(s_t,a_t) - h^\pi(s_t) - \gamma\langle e(s_{t+1},a_{t+1}), \theta_t\rangle\big)\ e(s_t,a_t),$$

where $\zeta_t = (s_t, a_t, s_{t+1}, a_{t+1})$ denotes the state transition steps following policy π and $e(s_t,a_t)$ denotes the unit vector. The CTD method uses the stochastic operator $\tilde{F}^\pi(\theta_t,\zeta_t)$ to update the parameters θ_t iteratively as shown in Algorithm 3. It involves two algorithmic parameters: $\alpha \geq 0$ determines how often θ_t is updated and $\beta_t \geq 0$ defines the learning rate. Observe that if $\alpha = 0$, then CTD reduces to the classic TD learning method.

Algorithm 3 Conditional Temporal Difference (CTD) for evaluating policy π

Let θ_1, the nonnegative parameters α and $\{\beta_t\}$ be given.
for $t = 1, \ldots, T$ **do**
Collect α state transition steps without updating $\{\theta_t\}$, denoted as $\{\zeta_t^1, \zeta_t^2, \ldots, \zeta_t^\alpha\}$.
Set

$$\theta_{t+1} = \theta_t - \beta_t \tilde{F}^\pi(\theta_t, \zeta_t^\alpha). \tag{5.10}$$

end for

When applying the general convergence results of CTD to our setting, we need to handle the following possible pitfalls. Firstly, current analysis of TD-type methods only provides bounds on $\mathbb{E}[\|\theta_t - \theta_*\|_2^2]$, which gives an upper bound on $\mathbb{E}[\|\theta_t - Q^\pi\|_\infty^2]$ and thus the bound on the total expected error (c.f., (4.2)). One needs to develop a tight enough bound on the bias $\|\mathbb{E}[\theta_t] - \theta_*\|_\infty$ (c.f., (4.3)) to derive the overall best rate of convergence for the SPMD method. Secondly, the selection of α and $\{\beta_t\}$ that gives the best rate of convergence in terms of $\mathbb{E}[\|\theta_t - \theta_*\|_2^2]$

does not necessarily result in the best rate of convergence for SPMD, since we need to deal with the bias term explicitly.

The following result can be shown similarly to Lemma 4.1 of [12].

Lemma 16 *Given the single ergodic class Markov chain $\zeta_1^1,\dots,\zeta_1^\alpha,\zeta_2^2,\dots,\zeta_2^\alpha,\dots$, there exists a constant $C>0$ and $\rho\in[0,1)$ such that for every $t,\alpha\in\mathbb{Z}_+$ with probability 1,*

$$\|F^\pi(\theta_t)-\mathbb{E}[\tilde F^\pi(\theta_t,\zeta_t^\alpha)|\zeta_{\lceil t-1\rceil}]\|_2\le C\rho^\alpha\|\theta_t-\theta^*\|_2.$$

We can also show that the variance of $\tilde F^\pi$ is bounded as follows.

$$\begin{aligned}&\mathbb{E}[\|\tilde F^\pi(\theta_t,\zeta_t^\alpha)-\mathbb{E}[\tilde F^\pi(\theta_t,\zeta_t^\alpha)|\zeta_{\lceil t-1\rceil}]\|_2^2]\\&\le 2(1+\gamma)^2\mathbb{E}[\|\theta_t\|_2^2]+2(\bar c+\bar h)^2\\&\le 4(1+\gamma)^2\mathbb{E}[\|\theta_t-\theta^*\|_2^2]+\|\theta^*\|_2^2+2(\bar c+\bar h)^2.\end{aligned}\tag{5.11}$$

The following result has been shown in Proposition 6.2 of [12].

Lemma 17 *If the algorithmic parameters in CTD are chosen such that*

$$\alpha\ge\tfrac{\log(1/\Lambda_{\min})+\log(9C)}{\log(1/\rho)}\quad\text{and}\quad\beta_t=\tfrac{2}{\Lambda_{\min}(t+t_0-1)}\tag{5.12}$$

with $t_0=8\max\{\Lambda_{\max}^2,8(1+\gamma)^2\}/\Lambda_{\min}^2$, then

$$\mathbb{E}[\|\theta_{t+1}-\theta^*\|_2^2]\le\tfrac{2(t_0+1)(t_0+2)\|\theta_1-\theta^*\|^2}{(t+t_0)(t+t_0+1)}+\tfrac{12t\sigma_F^2}{\Lambda_{\min}^2(t+t_0)(t+t_0+1)},$$

where $\sigma_F^2:=4(1+\gamma)^2R^2+\|\theta^*\|_2^2+2(\bar c+\bar h)^2$ and $R^2:=8\|\theta_1-\theta_*\|_2^2+\frac{3[\|\theta^*\|_2^2+2(\bar c+\bar h)^2]}{4(1+\gamma)^2}$. Moreover, we have $\mathbb{E}[\|\theta_t-\theta^*\|_2^2]\le R^2$ for any $t\ge1$.

We now enhance the above result with a bound on the bias term given by $\|\mathbb{E}[\theta_{t+1}]-\theta^*\|_2$. The proof of this result is put in the appendix since it is more technical.

Lemma 18 *Suppose that the algorithmic parameters in CTD are set according to Lemma 17. Then we have*

$$\|\mathbb{E}[\theta_{t+1}]-\theta^*\|_2^2\le\tfrac{(t_0-1)(t_0-2)(t_0-3)\|\theta_1-\theta^*\|_2^2}{(t+t_0-1)(t+t_0-2)(t+t_0-3)}+\tfrac{8CR^2\rho^\alpha}{3\Lambda_{\min}}+\tfrac{C^2R^2\rho^{2\alpha}}{\Lambda_{\min}^2}.$$

We are now ready to establish the convergence of the SMPD method by using the CTD method to estimate the action-value functions. We focus on the case when $\mu>0$, and the case for $\mu=0$ can be shown similarly.

Proposition 4 *Suppose that $\eta_k=\frac{1-\gamma}{\gamma\mu}$ in the SPMD method. If the initial point of CTD is set to $\theta_1=0$ and the number of iterations T and the parameter α in CTD are set to*

$$T_k=t_0(3\bar\theta2^{\lfloor k/l\rfloor+2})^{2/3}+(4t_0^2\bar\theta^2 2^{\lfloor k/l\rfloor+2})^{1/2}+24\sigma_F^2\Lambda_{\min}^{-2}2^{\lfloor k/l\rfloor+2},\tag{5.13}$$

$$\alpha_k=\max\{2(\lfloor\tfrac{k}{l}\rfloor+2)\log_\rho\tfrac12+\log_\rho\tfrac{\Lambda_{\min}}{24CR^2},(\lfloor\tfrac{k}{l}\rfloor+2)\log_\rho\tfrac12+\log_\rho\tfrac{\Lambda_{\min}}{3CR^2}\},\tag{5.14}$$

where $l:=\lceil\log_\gamma(1/4)\rceil$ and $\bar\theta:=\sqrt{n}\frac{\bar c+\bar h}{1-\gamma}$, then the relation in (4.9) holds. As a consequence, an ϵ-solution of (1.9), i.e., a solution $\bar\pi$ s.t. $\mathbb{E}[f(\bar\pi)-f(\pi^*)+\frac{\mu}{1-\gamma}\mathcal{D}(\bar\pi,\pi^*)]\le\epsilon$, can be found in at most $\mathcal{O}(\log_\gamma\epsilon)$ SPMD iterations. In addition, the total number of samples for (s_t,a_t) pairs can be bounded by

$$\mathcal{O}\{(\log_\gamma\tfrac12)(\log_2\tfrac1\epsilon)(\log_\rho\tfrac{\Lambda_{\min}}{CR^2})(\tfrac{t_0\bar\theta^{2/3}}{(\mu(1-\gamma)\epsilon)^{2/3}}+\tfrac{t_0\bar\theta}{\sqrt{\mu(1-\gamma)\epsilon}}+\tfrac{\sigma_F^2}{\mu(1-\gamma)\Lambda_{\min}^2\epsilon})\}.\tag{5.15}$$

Proof. Using the fact that $\gamma^l\le1/4$, we can easily check from Lemma 17, Lemma 18, and the selection of T and α that (4.1)-(4.3) hold with $\varsigma_k=2^{-(\lfloor k/l\rfloor+2)}$ and $\sigma_k^2=2^{-(\lfloor k/l\rfloor+2)}$. Suppose that an ϵ-solution $\bar\pi$ will be found at the $\bar k$ iteration. By (4.9), we have

$$\lfloor\bar k/l\rfloor\le\log_2\{[f(\pi_0)-f(\pi^*)+\tfrac{1}{1-\gamma}(\mu\log|\mathcal{A}|+\tfrac52+\tfrac{5}{8\gamma\mu})]\epsilon^{-1}\},$$

which implies that the number of iterations is bounded by $\mathcal{O}(l\lfloor\bar k/l\rfloor)=\mathcal{O}(\log_\gamma\epsilon)$. Moreover by the definition of T_k and α_k, the number of samples is bounded by

$$\begin{aligned}&\textstyle\sum_{p=0}^{\lfloor\bar k/l\rfloor}l\alpha_kT_k=\mathcal{O}\{\log_\gamma\frac12\sum_{p=0}^{\lfloor\bar k/l\rfloor}(p\log_\rho\frac12+\log_\rho\frac{\Lambda_{\min}}{CR^2})(t_0\bar\theta^{2/3}2^{2p/3}+t_0\bar\theta2^{p/2}+\sigma_F^2\Lambda_{\min}^{-2}2^p)\}\\&=\mathcal{O}\{\log_\gamma\tfrac12(\lfloor\bar k/l\rfloor\log_\rho\tfrac12+\log_\rho\tfrac{\Lambda_{\min}}{CR^2})(t_0\bar\theta^{2/3}2^{2\lfloor\bar k/l\rfloor/3}+t_0\bar\theta2^{\lfloor\bar k/l\rfloor/2}+\sigma_F^2\Lambda_{\min}^{-2}2^{\lfloor\bar k/l\rfloor})\}\\&=\mathcal{O}\{\log_\gamma\tfrac12(\log_2\tfrac1\epsilon\log_\rho\tfrac12+\log_\rho\tfrac{\Lambda_{\min}}{CR^2})(\tfrac{t_0\bar\theta^{2/3}}{(\mu(1-\gamma)\epsilon)^{2/3}}+\tfrac{t_0\bar\theta}{\sqrt{\mu(1-\gamma)\epsilon}}+\tfrac{\sigma_F^2}{\mu(1-\gamma)\Lambda_{\min}^2\epsilon})\}.\end{aligned}$$

■

The following result shows the convergence properties of the SAMPD method when the action-value function is estimated by using the CTD method.

Proposition 5 *Suppose that* $\eta_k = \frac{1-\gamma}{\gamma\tau_k}$ *and* $\tau_k = \frac{1}{\sqrt{\gamma \log|\mathcal{A}|}} 2^{-(\lfloor k/l \rfloor+1)}$ *in the SAPMD method. If the initial point of CTD is set to* $\theta_1 = 0$*, the number of iterations* T *is set to*

$$T_k = t_0(3\bar{\theta}2^{\lfloor k/l\rfloor+2})^{2/3} + (4t_0^2\bar{\theta}^2 4^{\lfloor k/l\rfloor+2})^{1/2} + 24\sigma_F^2 \Lambda_{\min}^{-2} 4^{\lfloor k/l\rfloor+2}, \tag{5.16}$$

and the parameter α *in CTD is set to (5.14), where* $l := \lceil \log_\gamma(1/4) \rceil$ *and* $\bar{\theta} := \frac{\bar{c}+\bar{h}+\tau_0 \log|\mathcal{A}|}{1-\gamma}$*, then the relation in (4.20) holds. As a consequence, an* ϵ*-solution of (1.9), i.e., a solution* $\bar{\pi}$ *s.t.* $\mathbb{E}[f(\bar{\pi}) - f(\pi^*)] \le \epsilon$*, can be found in at most* $\mathcal{O}(\log_\gamma \epsilon)$ *SPMD iterations. In addition, the total number of samples for* (s_t, a_t) *pairs can be bounded by*

$$\mathcal{O}\{(\log_\gamma \tfrac{1}{2})(\log_2 \tfrac{1}{\epsilon})(\log_\rho \tfrac{\Lambda_{\min}}{CR^2})(\tfrac{t_0\bar{\theta}}{(1-\gamma)\epsilon} + \tfrac{\sigma_F^2}{(1-\gamma)^2\Lambda_{\min}^2\epsilon^2})\}. \tag{5.17}$$

Proof. The proof is similar to that of Proposition 4 except that we will show that (4.1)-(4.3) hold with $\varsigma_k = 2^{-(\lfloor k/l\rfloor+2)}$ and $\sigma_k^2 = 4^{-(\lfloor k/l\rfloor+2)}$. Moreover, we will use (4.20) instead of (4.9) to bound the number of iterations. ∎

To the best of our knowledge, the complexity result in (5.15) is new in the RL literature, while the one in (5.17) is new for policy gradient type methods. It seems that this bound significantly improves the previously best-known $\mathcal{O}(1/\epsilon^3)$ sampling complexity result for stochastic policy gradient methods (see [24] and Appendix C of [10] for more explanation).

6 Efficient Solution for General Subproblems

In this section, we study the convergence properties of the PMD methods for the situation where we do not have exact solutions for prox-mapping subprobems. Throughout this section, we assume that h^π is differentiable and its gradients are Lipschitz continuous with Lipschitz constant L. We will first review Nesterov's accelerated gradient descent (AGD) method [18], and then discuss the overall gradient complexity of using this method for solving prox-mapping in the PMD methods. We will focus on the stochastic PMD methods since they cover deterministic methods as certain special cases.

6.1 Review of accelerated gradient descent

Let us denote $X \equiv \Delta_{|\mathcal{A}|}$ and consider the problem of

$$\min_{x\in X}\{\Phi(x) := \phi(x) + \chi(x)\}, \tag{6.1}$$

where $\phi : X \to \mathbb{R}$ is a smooth convex function such that

$$\mu_\phi D_{x'}^x \le \phi(x) - [\phi(x') + \langle \nabla\phi(x'), x - x'\rangle] \le \tfrac{L_\phi}{2}\|x - x'\|_1.$$

Moreover, we assume that $\chi : X \to \mathbb{R}$ satisfies

$$\chi(x) - [\chi(x') + \langle \chi'(x'), x - x'\rangle] \ge \mu_\chi D_{x'}^x$$

for some $\mu_\chi \ge 0$. Given $(x_{t-1}, y_{t-1}) \in X \times X$, the accelerated gradient method performs the following updates:

$$\underline{x}_t = (1 - q_t)y_{t-1} + q_t x_{t-1}, \tag{6.2}$$

$$x_t = \arg\min_{x\in X}\{r_t[\langle \nabla\phi(\underline{x}_t), x\rangle + \mu_\phi D_{\underline{x}_t}^x + \chi(x)] + D_{x_{t-1}}^x\}, \tag{6.3}$$

$$y_t = (1 - \rho_t)y_{t-1} + \rho_t x_t, \tag{6.4}$$

for some $q_t \in [0, 1]$, $r_t \ge 0$, and $\rho_t \in [0, 1]$.

Below we slightly generalize the convergence results for the AGD method so that they depend on the distance $D_{x_0}^x$ rather than $\Phi(y_0) - \Phi(x)$ for any $x \in X$. This result better fits our need to analyze the convergence of inexact SPMD and SAPMD methods in the next two subsections.

Lemma 19 *Let us denote* $\mu_\Phi := \mu_\phi + \mu_\chi$ *and* $t_0 := \lfloor 2\sqrt{L_\phi/\mu_\Phi} - 1\rfloor$*. If*

$$\rho_t = \begin{cases} \frac{2}{t+1} & t \le t_0 \\ \sqrt{\mu_\Phi/L_\phi} & o.w. \end{cases}, q_t = \begin{cases} \frac{2}{t+1} & t \le t_0 \\ \frac{\sqrt{\mu_\Phi/L_\phi} - \mu_\Phi/L_\phi}{1-\mu_\Phi/L_\phi} & o.w. \end{cases}, r_t = \begin{cases} \frac{t}{2L_\phi} & t \le t_0 \\ \frac{1}{\sqrt{L_\phi\mu_\Phi} - \mu_\Phi} & o.w. \end{cases},$$

then for any $x \in X$*,*

$$\Phi(y_t) - \Phi(x) + \mu_\Phi D_{x_t}^x \le \varepsilon(t) D_{x_0}^x, \tag{6.5}$$

where

$$\varepsilon(t) := 2L_\phi \min\left\{\left(1 - \sqrt{\mu_\Phi/L_\phi}\right)^{t-1}, \tfrac{2}{t(t+1)}\right\}. \tag{6.6}$$

Proof. Using the discussions in Corollary 3.5 of [13] (and the possible strong convexity of χ), we can check that the conclusions in Theorems 3.6 and 3.7 of [13] hold for the AGD method applied to problem (6.1). It then follows from Theorem 3.6 of [13] that

$$\Phi(y_t) - \Phi(x) + \tfrac{\rho_t}{r_t} D_{x_k}^{x^*} \le \tfrac{4L_\phi}{t(t+1)} D_{x_0}^x, \forall t = 1, \ldots, t_0. \tag{6.7}$$

Moreover, it follows from Theorem 3.7 of [13] that for any $t \ge t_0$,

$$\begin{aligned}\Phi(y_t) - \Phi(x) + \mu_\Phi D_{x_k}^x &\le \left(1 - \sqrt{\mu_\Phi / L_\phi}\right)^{t-t_0} [\Phi(y_{t_0}) - \Phi(x) + \mu_\Phi D_{x_{t_0}}^x] \\ &\le 2\left(1 - \sqrt{\mu_\Phi / L_\phi}\right)^{t-1} L_\phi D_{x_0}^x,\end{aligned}$$

where the last inequality follows from (6.7) (with $t = t_0$) and the facts that

$$\tfrac{\rho_t}{r_t} \ge \tfrac{\rho_{t_0}}{r_{t_0}} \ge \mu_\Phi \text{ and } \tfrac{2}{t(t+1)} = \textstyle\prod_{i=2}^t (1 - \frac{2}{i+1}) \le (1 - \sqrt{\mu_\Phi / L_\phi})^{t-1}$$

for any $2 \le t \le t_0$. The result then follows by combining these observations. ∎

6.2 Convergence of inexact SPMD

In this subsection, we study the convergence properties of the SPMD method when its subproblems are solved inexactly by using the AGD method (see Algorithm 4). Observe that we use the same initial point π_0 whenever calling the AGD method. To use a dynamic initial point (e.g., v_k) will make the analysis more complicated since we do not have a uniform bound on the KL divergence $D_{v_k}^\pi$ for an arbitrary v_k. To do so probably will require us to use other distance generating functions than the entropy function.

Algorithm 4 The SPMD method with inexact subproblem solutions

Input: initial points $\pi_0 = v_0$ and stepsizes $\eta_k \ge 0$.
for $k = 0, 1, \ldots,$ **do**
Apply T_k AGD iterations (with initial points $x_0 = y_0 = \pi_0$) to

$$\pi_{k+1}(\cdot|s) = \operatorname*{arg\,min}_{p(\cdot|s) \in \Delta_{|\mathcal{A}|}} \left\{ \Phi_k(p) := \eta_k [\langle \mathcal{Q}^{\pi_k, \xi_k}(s, \cdot), p(\cdot|s) \rangle + h^p(s)] + D_{v_k}^p(s) \right\}. \tag{6.8}$$

Set $(\pi_{k+1}, v_{k+1}) = (y_{T_k+1}, x_{T_k+1})$.
end for

In the sequel, we will denote $\varepsilon_k \equiv \varepsilon(T_k)$ to simplify notations. The following result will take place of Lemma 4 in our convergence analysis.

Lemma 20 *For any $\pi(\cdot|s) \in X$, we have*

$$\begin{aligned}&\eta_k [\langle \mathcal{Q}^{\pi_k, \xi_k}(s, \cdot), \pi_{k+1}(\cdot|s) - \pi(\cdot|s) \rangle + h^{\pi_{k+1}}(s) - h^\pi(s)] \\ &+ D_{v_k}^{\pi_{k+1}}(s) + (1 + \mu\eta_k) D_{v_{k+1}}^\pi(s) \le D_{v_k}^\pi(s) + \varepsilon_k \log |\mathcal{A}|.\end{aligned} \tag{6.9}$$

Moreover, we have

$$\begin{aligned}&\eta_k [\langle \mathcal{Q}^{\pi_k, \xi_k}(s, \cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s) \rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s)] \\ &+ D_{v_k}^{\pi_{k+1}}(s) + (1 + \mu\eta_k) D_{v_{k+1}}^{\pi_{k+1}}(s) \le (\varepsilon_k + \tfrac{\varepsilon_{k-1}}{1 + \mu\eta_{k-1}}) \log |\mathcal{A}|.\end{aligned} \tag{6.10}$$

Proof. It follows from Lemma 19 (with $\mu_\Phi = 1 + \mu\eta_k$ and $L_\phi = L$) that

$$\Phi_k(\pi_{k+1}) - \Phi_k(\pi) + (1 + \mu\eta_k) D_{v_{k+1}}^\pi(s) \le \epsilon_k D_{\pi_0}^\pi(s) \le \varepsilon_k \log |\mathcal{A}|.$$

Using the definition of Φ_k, we have

$$\begin{aligned}&\eta_k [\langle \mathcal{Q}^{\pi_k, \xi_k}(s, \cdot), \pi_{k+1}(\cdot|s) - \pi(\cdot|s) \rangle + h^{\pi_{k+1}}(s) - h^\pi(s)] \\ &+ D_{v_k}^{\pi_{k+1}}(s) - D_{v_k}^\pi(s) + (1 + \mu\eta_k) D_{v_{k+1}}^\pi(s) \le \varepsilon_k \log |\mathcal{A}|,\end{aligned}$$

which proves (6.9). Setting $\pi = \pi_k$ and $\pi = \pi_{k+1}$ respectively, in the above conclusion, we obtain

$$\begin{aligned}&\eta_k [\langle \mathcal{Q}^{\pi_k, \xi_k}(s, \cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s) \rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s)] \\ &+ D_{v_k}^{\pi_{k+1}}(s) + (1 + \mu\eta_k) D_{v_{k+1}}^{\pi_k}(s) \le D_{v_k}^{\pi_k}(s) + \varepsilon_k \log |\mathcal{A}|, \\ &(1 + \mu\eta_k) D_{v_{k+1}}^{\pi_{k+1}}(s) \le \varepsilon_k \log |\mathcal{A}|.\end{aligned}$$

Then (6.10) follows by combining these two inequalities. ∎

Proposition 6 *For any $s \in \mathcal{S}$, we have*

$$\begin{aligned} V^{\pi_{k+1}}(s) - V^{\pi_k}(s) \le & \langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s) \\ & + \tfrac{1}{\eta_k} D_{\pi_k}^{\pi_{k+1}}(s) + \tfrac{\eta_k \|\delta_k\|_\infty^2}{2(1-\gamma)} + \tfrac{\gamma}{(1-\gamma)\eta_k}(\varepsilon_k + \tfrac{\varepsilon_{k-1}}{1+\mu\eta_{k-1}}) \log|\mathcal{A}| \\ & - \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}[\langle \delta_k, v_k(\cdot|s') - \pi_k(\cdot|s')\rangle]. \end{aligned} \tag{6.11}$$

Proof. Similar to (4.8), we have

$$\begin{aligned} & V^{\pi_{k+1}}(s) - V^{\pi_k}(s) \\ & = \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\Big[\langle \mathcal{Q}^{\pi_k,\xi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s') \\ & \quad - \langle \delta_k, \pi_{k+1}(\cdot|s') - v_k(\cdot|s')\rangle - \langle \delta_k, v_k(\cdot|s') - \pi_k(\cdot|s')\rangle\Big] \\ & \le \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\Big[\langle \mathcal{Q}^{\pi_k,\xi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s') \\ & \quad + \tfrac{1}{2\eta_k}\|\pi_{k+1}(\cdot|s') - v_k(\cdot|s')\|_1^2 + \tfrac{\eta_k\|\delta_k\|_\infty^2}{2} - \langle \delta_k, v_k(\cdot|s') - \pi_k(\cdot|s')\rangle\Big] \\ & \le \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\Big[\langle \mathcal{Q}^{\pi_k,\xi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s') \\ & \quad + \tfrac{1}{\eta_k} D_{v_k}^{\pi_{k+1}}(s') + \tfrac{\eta_k\|\delta_k\|_\infty^2}{2} - \langle \delta_k, v_k(\cdot|s') - \pi_k(\cdot|s')\rangle\Big]. \end{aligned} \tag{6.12}$$

It follows from (6.10) that

$$\begin{aligned} & \langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s) \\ & + \tfrac{1}{\eta_k}\Big[D_{v_k}^{\pi_{k+1}}(s) + (1+\mu\eta_k) D_{v_{k+1}}^{\pi_{k+1}}(s) - (\varepsilon_k + \tfrac{\varepsilon_{k-1}}{1+\mu\eta_{k-1}})\log|\mathcal{A}|\Big] \le 0, \end{aligned} \tag{6.13}$$

which implies that

$$\begin{aligned} & \mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\Big[\langle \mathcal{Q}^{\pi_k,\xi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s') \\ & \quad + \tfrac{1}{\eta_k}\Big(D_{\pi_{k+1}}^{\pi_k}(s') - (\varepsilon_k + \tfrac{\varepsilon_{k-1}}{1+\mu\eta_{k-1}})\log|\mathcal{A}|\Big)\Big] \\ & \le d_s^{\pi_{k+1}}(s)\Big[\langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s) \\ & \quad + \tfrac{1}{\eta_k}\Big(D_{\pi_k}^{\pi_{k+1}}(s) - (\varepsilon_k + \tfrac{\varepsilon_{k-1}}{1+\mu\eta_{k-1}})\log|\mathcal{A}|\Big)\Big] \\ & \le (1-\gamma)\Big[\langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s) \\ & \quad + \tfrac{1}{\eta_k}\Big(D_{\pi_k}^{\pi_{k+1}}(s) - (\varepsilon_k + \tfrac{\varepsilon_{k-1}}{1+\mu\eta_{k-1}})\log|\mathcal{A}|\Big)\Big], \end{aligned}$$

where the last inequality follows from the fact that $d_s^{\pi_{k+1}}(s) \ge (1-\gamma)$ due to the definition of $d_s^{\pi_{k+1}}$ in (2.1). The result in (6.11) then follows immediately from (6.12) and the above inequality. ∎

We now establish an important recursion about the inexact SPMD method in Algorithm 4.

Lemma 21 *Suppose that $\eta_k = \eta = \frac{1-\gamma}{\gamma\mu}$ and $\varepsilon_k \le \varepsilon_{k-1}$ for any $k \ge 0$ in the inexact SPMD method, we have*

$$\begin{aligned} & \mathbb{E}_{\xi_{\lceil k\rceil}}[f(\pi_{k+1}) - f(\pi^*) + \tfrac{\mu}{1-\gamma}\mathcal{D}(\pi_{k+1},\pi^*)] \\ & \le \gamma[\mathbb{E}_{\xi_{\lceil k-1\rceil}}[f(\pi_k) - f(\pi^*) + \tfrac{\mu}{1-\gamma}D(\pi_k,\pi^*)] + \tfrac{2(2-\gamma)\varsigma_k}{1-\gamma} + \tfrac{\sigma_k^2}{2\gamma\mu} + \tfrac{\mu\gamma^2(1+\gamma)\log|\mathcal{A}|\varepsilon_{k-1}}{(1-\gamma)^2}. \end{aligned}$$

Proof. By (6.9) (with $p = \pi^*$), we have

$$\begin{aligned} & \langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi^*(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi^*}(s) + \tfrac{1}{\eta_k} D_{v_k}^{\pi_{k+1}}(s) \\ & \le \tfrac{1}{\eta_k} D_{v_k}^{\pi^*}(s) - (\tfrac{1}{\eta_k} + \mu) D_{v_{k+1}}^{\pi^*}(s) + \tfrac{\varepsilon_k}{\eta_k}\log|\mathcal{A}|, \end{aligned}$$

which, in view of (4.7), then implies that

$$\begin{aligned} & \langle \mathcal{Q}^{\pi_k,\xi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle + h^{\pi_k}(s) - h^{\pi^*}(s) + V^{\pi_{k+1}}(s) - V^{\pi_k}(s) \\ & \le \tfrac{1}{\eta_k} D_{v_k}^{\pi^*}(s) - (\tfrac{1}{\eta_k} + \mu) D_{v_{k+1}}^{\pi^*}(s) + \tfrac{\eta_k\|\delta_k\|_\infty^2}{2(1-\gamma)} + \tfrac{\gamma}{(1-\gamma)\eta_k}(\varepsilon_k + \tfrac{\varepsilon_{k-1}}{1+\mu\eta_{k-1}})\log|\mathcal{A}| \\ & \quad - \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}[\langle \delta_k, v_k(\cdot|s') - \pi_k(\cdot|s')\rangle]. \end{aligned}$$

Taking expectation w.r.t. $\xi_{\lceil k\rceil}$ and ν^* on both sides of the above inequality, and using Lemma 3 and the relation in (4.6), we arrive at

$$\begin{aligned}&\mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}\left[(1-\gamma)(V^{\pi_k}(s)-V^{\pi_\tau^*}(s))+V^{\pi_{k+1}}(s)-V^{\pi_k}(s)\right]\\&\le \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}[\tfrac{1}{\eta_k}D_{v_k}^{\pi^*}(s)-(\tfrac{1}{\eta_k}+\mu)D_{v_{k+1}}^{\pi^*}(s)]+2\varsigma_k+\tfrac{\eta_k\sigma_k^2}{2(1-\gamma)}\\&\quad+\tfrac{\gamma}{(1-\gamma)\eta_k}(\varepsilon_k+\tfrac{\varepsilon_{k-1}}{1+\mu\eta_{k-1}})\log|\mathcal{A}|+\tfrac{2}{1-\gamma}\varsigma_k.\end{aligned}$$

Noting $V^{\pi_{k+1}}(s)-V^{\pi_k}(s)=V^{\pi_{k+1}}(s)-V^{\pi^*}(s)-[V^{\pi_k}(s)-V^{\pi^*}(s)]$, rearranging the terms in the above inequality, and using the definition of f in (1.9), we arrive at

$$\begin{aligned}&\mathbb{E}_{\xi_{\lceil k\rceil}}[f(\pi_{k+1})-f(\pi^*)+(\tfrac{1}{\eta_k}+\mu)\mathcal{D}(v_{k+1},\pi^*)]\le\mathbb{E}_{\xi_{\lceil k-1\rceil}}[\gamma(f(\pi_k)-f(\pi^*))+\tfrac{1}{\eta_k}\mathcal{D}(v_k,\pi^*)]\\&\quad+\tfrac{2(2-\gamma)\varsigma_k}{1-\gamma}+\tfrac{\eta_k\sigma_k^2}{2(1-\gamma)}+\tfrac{\gamma}{(1-\gamma)\eta_k}(\varepsilon_k+\tfrac{\varepsilon_{k-1}}{1+\mu\eta_{k-1}})\log|\mathcal{A}|.\end{aligned}$$

The result then follows immediately by the selection of η and the assumption $\varepsilon_k\le\varepsilon_{k-1}$. ∎

We now are now ready to state the convergence rate of the SPMD method with inexact prox-mapping. We focus on the case when $\mu>0$.

Theorem 8 *Suppose that $\eta_k=\eta=\frac{1-\gamma}{\gamma\mu}$ in the inexact SPMD method. If $\varsigma_k=(1-\gamma)2^{-(\lfloor k/l\rfloor+2)}$, $\sigma_k^2=2^{-(\lfloor k/l\rfloor+2)}$ and $\varepsilon_k=(1-\gamma)^2 2^{-(\lfloor (k+1)/l\rfloor+2)}$ for any $k\ge0$ with $l:=\lceil\log_\gamma(1/4)\rceil$, then*

$$\begin{aligned}&\mathbb{E}_{\xi_{\lceil k-1\rceil}}[f(\pi_k)-f(\pi^*)+\tfrac{\mu}{1-\gamma}\mathcal{D}(\pi_k,\pi^*)]\\&\le 2^{-\lfloor k/l\rfloor}\left[f(\pi_0)-f(\pi^*)+\tfrac{1}{1-\gamma}(\mu\log|\mathcal{A}|+\tfrac{5(2-\gamma)}{2}+\tfrac{5}{8\gamma\mu}+\tfrac{5\mu\gamma^2(1+\gamma)\log|\mathcal{A}|}{4})\right].\end{aligned}\tag{6.14}$$

Proof. The result follows as an immediate consequence of Proposition 21 and Lemma 11. ∎

In view of Theorem 8, the inexact solutions of the subproblems barely affect the iteration and sampling complexities of the SPMD method as long as $\varepsilon_k\le(1-\gamma)^2 2^{-(\lfloor(k+1)/l\rfloor+2)}$. Notice that an ϵ-solution of problem (1.9), i.e., a solution $\bar\pi$ s.t. $\mathbb{E}[f(\bar\pi)-f(\pi^*)]\le\epsilon$, can be found at the $\bar k$-th iteration with

$$\lfloor\bar k/l\rfloor\le\log_2\{\epsilon^{-1}[f(\pi_0)-f(\pi^*)+\tfrac{1}{1-\gamma}(4\mu\log|\mathcal{A}|+5+\tfrac{5}{8\gamma\mu})]\}.$$

Also observe that the condition number of the subproblem is given by

$$\tfrac{L\eta_k}{\mu\eta_k+1}=\tfrac{L(1-\gamma)}{\mu}.$$

Combining these observations with Lemma 19, we conclude that the total number of gradient computations of h can be bounded by

$$\begin{aligned}l\textstyle\sum_{p=0}^{\lfloor\bar k/l\rfloor}\sqrt{\frac{L\eta}{\mu\eta+1}}\log(L\eta/\varepsilon_k)&=l\textstyle\sum_{p=0}^{\lfloor\bar k/l\rfloor}\sqrt{\frac{L(1-\gamma)}{\mu}}\log\frac{4L2^{p+1}}{\gamma(1-\gamma)\mu}\\&=\mathcal{O}\{l(\lfloor\bar k/l\rfloor)^2\sqrt{\tfrac{L(1-\gamma)}{\mu}}\log\tfrac{L}{\gamma(1-\gamma)\mu}\}\\&=\mathcal{O}\left\{(\log_\gamma\tfrac14)(\log^2\tfrac1\epsilon)\sqrt{\tfrac{L(1-\gamma)}{\mu}}(\log\tfrac{L}{\gamma(1-\gamma)\mu})\right\}.\end{aligned}$$

6.3 Convergence of inexact SAPMD

In this subsection, we study the convergence properties of the SAPMD method when its subproblems are solved inexactly by using the AGD method (see Algorithm 5).

Algorithm 5 The Inexact SAPMD method

Input: initial points $\pi_0=v_0$, stepsizes $\eta_k\ge0$, and regularization parameters $\tau_k\ge0$.
for $k=0,1,\ldots,$ **do**
Apply T_k AGD iterations (with initial points $x_0=y_0=\pi_0$) to

$$\pi_{k+1}(\cdot|s)=\operatorname*{arg\,min}_{p(\cdot|s)\in\Delta_{|\mathcal{A}|}}\left\{\tilde\Phi_k(p):=\eta_k[\langle\mathcal{Q}_{\tau_k}^{\pi_k,\xi_k}(s,\cdot),p(\cdot|s)\rangle+h^p(s)+\tau_k D_{\pi_0}^p(s)]+D_{v_k}^p(s)\right\}.\tag{6.15}$$

Set $(\pi_{k+1},v_{k+1})=(y_{T_k+1},x_{T_k+1})$.
end for

In the sequel, we will still denote $\varepsilon_k\equiv\varepsilon(T_k)$ to simplify notations. The following result has the same role as Lemma 20 in our convergence analysis.

Lemma 22 *For any $\pi(\cdot|s) \in X$, we have*

$$\eta_k[\langle \mathcal{Q}_{\tau_k}^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi}(s) + \tau_k(D_{\pi_0}^{\pi_{k+1}}(s) - D_{\pi_0}^{\pi}(s))] + D_{v_k}^{\pi_{k+1}}(s) + (1+\tau_k\eta_k)D_{v_{k+1}}^{\pi}(s) \le D_{v_k}^{\pi}(s) + \varepsilon_k \log|\mathcal{A}|. \tag{6.16}$$

Moreover, we have

$$\eta_k[\langle \mathcal{Q}_{\tau_k}^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s) + \tau_k(D_{\pi_0}^{\pi_{k+1}}(s) - D_{\pi_0}^{\pi_k}(s))] + D_{v_k}^{\pi_{k+1}}(s) + (1+\tau_k\eta_k)D_{v_{k+1}}^{\pi_{k+1}}(s) \le (\varepsilon_k + \tfrac{\varepsilon_{k-1}}{1+\tau_{k-1}\eta_{k-1}})\log|\mathcal{A}|. \tag{6.17}$$

Proof. The proof is the same as that for Lemma 20 except that Lemma 19 will be applied to problem (6.15) (with $\mu_\Phi = 1+\tau_k\eta_k$ and $L_\phi = L$). ∎

Lemma 23 *For any $s \in \mathcal{S}$, we have*

$$\begin{aligned} V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi_k}(s) \le\ & \langle \mathcal{Q}_{\tau_k}^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s) \\ & + \tau_k(D_{\pi_0}^{\pi_{k+1}}(s) - D_{\pi_k}^{\pi}(s)) \\ & + \tfrac{1}{\eta_k}D_{\pi_k}^{\pi_{k+1}}(s) + \tfrac{\eta_k\|\delta_k\|_\infty^2}{2(1-\gamma)} + \tfrac{\gamma}{(1-\gamma)\eta_k}(\varepsilon_k + \tfrac{\varepsilon_{k-1}}{1+\tau_{k-1}\eta_{k-1}})\log|\mathcal{A}| \\ & - \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}[\langle \delta_k, v_k(\cdot|s') - \pi_k(\cdot|s')\rangle]. \end{aligned} \tag{6.18}$$

Proof. The proof is similar to that for Lemma 6 with the following two exceptions: (a) we will apply Lemma 2 (i.e., the performance difference lemma) to the perturbed value functions $V_{\tau_k}^{\pi}$ instead of V^{π} to obtain a result similar to (6.12); and (b) we will use use (6.17) in place of (6.10) to derive a bound similar to (6.13). ∎

Lemma 24 *Suppose that $1+\eta_k\tau_k = 1/\gamma$ and $\varepsilon_k \le \varepsilon_{k-1}$ in the SAPMD method. Then for any $k \ge 0$, we have*

$$\begin{aligned} & \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}[V_{\tau_{k+1}}^{\pi_{k+1}}(s) - V_{\tau_{k+1}}^{\pi^*}(s) + \tfrac{\tau_k}{1-\gamma}D_{\pi_{k+1}}^{\pi^*}(s)] \\ & \le \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k-1\rceil}}[\gamma[V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s) + \tfrac{\tau_k}{1-\gamma}D_{\pi_k}^{\pi^*}(s)] \\ & \quad + \tfrac{(\tau_k-\tau_{k+1})}{1-\gamma}\log|\mathcal{A}| + \tfrac{2(2-\gamma)\varsigma_k}{1-\gamma} + \tfrac{\sigma_k^2}{2\gamma\tau_k} \\ & \quad + \tfrac{\gamma^2(1+\gamma)\varepsilon_{k-1}\tau_k}{(1-\gamma)^2}\log|\mathcal{A}|. \end{aligned} \tag{6.19}$$

Proof. By (6.16) (with $p = \pi^*$), we have

$$\begin{aligned} & \langle Q_{\tau_k}^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi^*(\cdot|s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi^*}(s) \\ & + \tau_k[D_{\pi_0}^{\pi_{k+1}}(s_t) - D_{\pi_0}^{\pi^*}(s_t)] + \tfrac{1}{\eta_k}D_{\pi_k}^{\pi_{k+1}}(s) \\ & \le \tfrac{1}{\eta_k}D_{v_k}^{\pi^*}(s) - (\tfrac{1}{\eta_k}+\tau_k)D_{v_{k+1}}^{\pi^*}(s) + \tfrac{\varepsilon_k}{\eta_k}\log|\mathcal{A}|, \end{aligned}$$

which, in view of (6.18), implies that

$$\begin{aligned} & \langle Q_{\tau_k}^{\pi_k,\xi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle + h^{\pi_k}(s) - h^{\pi^*}(s) + \tau_k[D_{\pi_0}^{\pi_k}(s_t) - D_{\pi_0}^{\pi^*}(s_t)] \\ & + V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi_k}(s) \\ & \le \tfrac{1}{\eta_k}D_{v_k}^{\pi^*}(s) - (\tfrac{1}{\eta_k}+\tau_k)D_{v_{k+1}}^{\pi^*}(s) + \tfrac{\eta_k\|\delta_k\|_\infty^2}{2(1-\gamma)} + \tfrac{\gamma}{(1-\gamma)\eta_k}(\varepsilon_k + \tfrac{\varepsilon_{k-1}}{1+\tau_{k-1}\eta_{k-1}})\log|\mathcal{A}| \\ & \quad - \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}[\langle \delta_k, v_k(\cdot|s') - \pi_k(\cdot|s')\rangle]. \end{aligned}$$

Taking expectation w.r.t. $\xi_{\lceil k\rceil}$ and ν^* on both sides of the above inequality, and using Lemma 3 (with h^π replaced by $h^\pi + \tau_k D_{\pi_0}^\pi(s_t)$ and Q^π replaced by Q_τ^π) and the relation in (4.6), we arrive at

$$\begin{aligned} & \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}[(1-\gamma)(V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s))] + \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}[V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi_k}(s)] \\ & \le \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}[\tfrac{1}{\eta_k}D_{\pi_k}^{\pi^*}(s) - (\tfrac{1}{\eta_k}+\tau_k)D_{\pi_{k+1}}^{\pi^*}(s)] + 2\varsigma_k + \tfrac{\eta_k\sigma_k^2}{2(1-\gamma)} \\ & \quad + \tfrac{\gamma}{(1-\gamma)\eta_k}(\varepsilon_k + \tfrac{\varepsilon_{k-1}}{1+\tau_{k-1}\eta_{k-1}})\log|\mathcal{A}| + \tfrac{2\varsigma_k}{1-\gamma}. \end{aligned}$$

Noting $V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi_k}(s) = V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi^*}(s) - [V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s)]$ and rearranging the terms in the above inequality, we have

$$\begin{aligned} & \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k\rceil}}[V_{\tau_k}^{\pi_{k+1}}(s) - V_{\tau_k}^{\pi^*}(s) + (\tfrac{1}{\eta_k}+\tau_k)D_{v_{k+1}}^{\pi^*}(s)] \\ & \le \gamma\mathbb{E}_{s\sim\nu^*,\xi_{\lceil k-1\rceil}}[V_{\tau_k}^{\pi_k}(s) - V_{\tau_k}^{\pi^*}(s)] + \mathbb{E}_{s\sim\nu^*,\xi_{\lceil k-1\rceil}}[\tfrac{1}{\eta_k}D_{v_k}^{\pi^*}(s)] + \tfrac{2(2-\gamma)\zeta_k}{1-\gamma} + \tfrac{\eta_k\sigma_k^2}{2(1-\gamma)} \\ & \quad + \tfrac{\gamma}{(1-\gamma)\eta_k}(\varepsilon_k + \tfrac{\varepsilon_{k-1}}{1+\tau_{k-1}\eta_{k-1}})\log|\mathcal{A}|. \end{aligned} \tag{6.20}$$

The result then follows from $1+\eta_k\tau_k = 1/\gamma$, the assumptions $\tau_k \ge \tau_{k+1}$, $\varepsilon_k \le \varepsilon_{k-1}$ and (3.19). ∎

Theorem 9 *Suppose that* $\eta_k = \frac{1-\gamma}{\gamma\tau_k}$ *in the SAPMD method. If* $\tau_k = \frac{1}{\sqrt{\gamma \log|\mathcal{A}|}} 2^{-(\lfloor k/l \rfloor+1)}$, $\varsigma_k = 2^{-(\lfloor k/l \rfloor+2)}$, $\sigma_k^2 = 4^{-(\lfloor k/l \rfloor+2)}$, *and* $\varepsilon_k = \frac{(1-\gamma)^2}{2\gamma^2(1+\gamma)}$ *with* $l := \lceil \log_\gamma(1/4) \rceil$, *then*

$$\begin{aligned}&\mathbb{E}_{\xi_{\lceil k-1 \rceil}}[f(\pi_k) - f(\pi^*)] \\ &\le 2^{-\lfloor k/l \rfloor}[f(\pi_0) - f(\pi^*) + \tfrac{1}{1-\gamma}(\tfrac{3\sqrt{\log|\mathcal{A}|}}{\sqrt{\gamma}} + \tfrac{5(2-\gamma)}{2} + \tfrac{5\sqrt{\log|\mathcal{A}|}}{4\sqrt{\gamma}})].\end{aligned} \tag{6.21}$$

Proof. The result follows as an immediate consequence of Lemma 24, Lemma 11, and an argument similar to the one to prove Theorem 7. ∎

In view of Theorem 9, the inexact solution of the subproblem barely affect the iteration and sampling complexities of the SAPMD method as long as $\varepsilon_k \le \frac{(1-\gamma)^2}{2\gamma^2(1+\gamma)}$. Notice that an ϵ-solution of problem (1.9), i.e., a solution $\bar{\pi}$ s.t. $\mathbb{E}[f(\bar{\pi}) - f(\pi^*)] \le \epsilon$, can be found at the $\bar{k}$-th iteration with

$$\lfloor \bar{k}/l \rfloor \le \log_2\{\epsilon^{-1}[f(\pi_0) - f(\pi^*) + \tfrac{5}{1-\gamma}(\tfrac{\sqrt{\log|\mathcal{A}|}}{\sqrt{\gamma}} + 1)]\}.$$

Also observe that the condition number of the subproblem is given by

$$\tfrac{L\eta_k}{\tau_k\eta_k+1} = \tfrac{L(1-\gamma)}{\tau_k} = (1-\gamma)L\sqrt{\gamma \log|\mathcal{A}|}\, 2^{\lfloor k/l \rfloor+1}.$$

Combining these observations with Lemma 19, we conclude that the total number of gradient computations of h can be bounded by

$$\begin{aligned}l\textstyle\sum_{p=0}^{\lfloor \bar{k}/l \rfloor} \sqrt{\tfrac{L\eta_k}{\tau_k\eta_k+1}} \log(L\eta_k/\varepsilon_k) &= l\textstyle\sum_{p=0}^{\lfloor \bar{k}/l \rfloor} \sqrt{\tfrac{L\eta_k}{\tau_k\eta_k+1}} \log \tfrac{L(1-\gamma)^3}{2\gamma^2(1+\gamma)\tau_k} \\ &= \mathcal{O}\{l\lfloor \bar{k}/l \rfloor[(1-\gamma)L]^{1/2} 2^{\lfloor \bar{k}/l \rfloor/2} \log \tfrac{L(1-\gamma)}{\gamma}\} \\ &= \mathcal{O}\left\{(\log_\gamma \epsilon)\sqrt{\tfrac{L}{\epsilon}}(\log \tfrac{L(1-\gamma)}{\gamma})\right\}.\end{aligned}$$

7 Concluding Remarks

In this paper, we present the policy mirror descent (PMD) method and show that it can achieve the linear and sublinear rate of convergence for RL problems with strongly convex or general convex regularizers, respectively. We then present an approximate policy mirror descent (APMD) method obtained by adding adaptive perturbations to the action-value functions and show that it can achieve the linear convergence rate for RL problems with general convex regularizers. We develop the stochastic PMD and APMD methods and derive general conditions on the bias and overall expected error to guarantee the convergence of these methods. Using these conditions, we establish new sampling complexity bounds of RL problems by using two different sampling schemes, i.e., either using a straightforward generative model or a more involved conditional temporal different method. The latter setting requires us to establish a bound on the bias for estimating action-value functions, which might be of independent interest. Finally, we establish the conditions on the accuracy required for the prox-mapping subproblems in these PMD type methods, as well as the overall complexity of computing the gradients of the regularizers.

Reference

1. 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.
2. A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *SIAM Journal on Optimization*, 27:927–956, 2003.
3. R. Bellman and S. Dreyfus. Functional approximations and dynamic programming. *Mathematical Tables and Other Aids to Computation*, 13(68):247–251, 1959.
4. Jalaj Bhandari and Daniel Russo. A Note on the Linear Convergence of Policy Gradient Methods. *arXiv e-prints*, page arXiv:2007.11120, July 2020.
5. 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.
6. C. D. Dang and G. Lan. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and Applications*, 60(2):277–310, 2015.
7. Eyal Even-Dar, Sham. M. Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.
8. F. Facchinei and J. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volumes I and II*. Comprehensive Study in Mathematics. Springer-Verlag, New York, 2003.
9. S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, 2002.

10. S. Khodadadian, Z. Chen, and S. T. Maguluri. Finite-sample analysis of off-policy natural actor-critic algorithm. *arXiv*, pages arXiv–2102.09318, 2021.
11. G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, I: operator extrapolation. *arXiv*, pages arXiv–2011.02987, 2020.
12. G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, II: Markovian noise and policy evaluation in reinforcement learning. *arXiv*, pages arXiv–2011.08434, 2020.
13. G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Nature, Switzerland AG, 2020.
14. B. Liu, Q. Cai, Z. Yang, and Z. Wang. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv*, pages arXiv–1906.10306, 2019.
15. 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.
16. A. S. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.
17. A. S. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, 1983.
18. Y. E. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.
19. Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994.
20. 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.
21. 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.
22. M. Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. *ArXiv*, abs/2005.09814, 2020.
23. L. Wang, Q. Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *ArXiv*, abs/1909.01150, 2020.
24. T. Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for actor-critic algorithms. *ArXiv*, abs/2004.12956, 2020.

Appendix: Bounding Bias for the Conditional Temporal Difference Methods

Proof of Lemma 18.

Proof. For simplicity, let us denote $\bar{\theta}_t \equiv \mathbb{E}[\theta_t]$, $\zeta_t \equiv (\zeta_t^1, \dots, \zeta_t^\alpha)$ and $\zeta_{\lceil t \rceil} = (\zeta_1, \dots, \zeta_t)$. Also let us denote $\delta_t^F := F^\pi(\theta_t) - \mathbb{E}[\tilde{F}^\pi(\theta_t, \zeta_t^\alpha)|\zeta_{\lceil t-1 \rceil}]$ and $\bar{\delta}_t^F = \mathbb{E}_{\zeta_{\lceil t-1 \rceil}}[\delta_t^F]$. It follows from Jensen's ienquality and Lemma 17 that

$$\|\bar{\theta}_t - \theta^*\|_2 = \|\mathbb{E}_{\zeta_{\lceil t-1 \rceil}}[\theta_t] - \theta^*\|_2 \le \mathbb{E}_{\zeta_{\lceil t-1 \rceil}}[\|\theta_t - \theta^*\|_2] \le R. \tag{7.1}$$

Also by Jensen's inequality, Lemma 16 and Lemma 17, we have

$$\begin{aligned} \|\bar{\delta}_t^F\|_2 &= \|\mathbb{E}_{\zeta_{\lceil t-1 \rceil}}[\delta_t^F]\|_2 \le \mathbb{E}_{\zeta_{\lceil t-1 \rceil}}[\|\delta_t^F\|_2] \\ &\le C\rho^\alpha \mathbb{E}_{\zeta_{\lceil t-1 \rceil}}[\|\theta_t - \theta^*\|_2] \le CR\rho^\alpha. \end{aligned} \tag{7.2}$$

Notice that

$$\begin{aligned} \theta_{t+1} &= \theta_t - \beta_t \tilde{F}^\pi(\theta_t, \zeta_t^\alpha) \\ &= \theta_t - \beta_t F^\pi(\theta_t) + \beta_t [F^\pi(\theta_t) - \tilde{F}^\pi(\theta_t, \zeta_t^\alpha)]. \end{aligned}$$

Now conditional on $\zeta_{\lceil t-1 \rceil}$, taking expectation w.r.t. ζ_t on (5.10), we have $\mathbb{E}[\theta_{t+1}|\zeta_{\lceil t-1 \rceil}] = \theta_t - \beta_t F^\pi(\theta_t) + \beta_t \delta_t^F$. Taking further expectation w.r.t. $\zeta_{\lceil t-1 \rceil}$ and using the linearity of F, we have $\bar{\theta}_{t+1} = \bar{\theta}_t - \beta_t F^\pi(\bar{\theta}_t) + \beta_t \bar{\delta}_t^F$, which implies

$$\begin{aligned} \|\bar{\theta}_{t+1} - \theta^*\|_2^2 &= \|\bar{\theta}_t - \theta^* - \beta_t F^\pi(\bar{\theta}_t) + \beta_t \bar{\delta}_t^F\|_2^2 \\ &= \|\bar{\theta}_t - \theta^*\|_2^2 - 2\beta_t \langle F^\pi(\bar{\theta}_t) - \bar{\delta}_t^F, \bar{\theta}_t - \theta^* \rangle + \beta_t^2 \|F^\pi(\bar{\theta}_t) - \bar{\delta}_t^F\|_2^2 \\ &\le \|\bar{\theta}_t - \theta^*\|_2^2 - 2\beta_t \langle F^\pi(\bar{\theta}_t) - \bar{\delta}_t^F, \bar{\theta}_t - \theta^* \rangle + 2\beta_t^2 [\|F^\pi(\bar{\theta}_t)\|_2^2 + \|\bar{\delta}_t^F\|_2^2]. \end{aligned}$$

The above inequality, together with (7.1), (7.2) and the facts that

$$\begin{aligned} \langle F^\pi(\bar{\theta}_t), \bar{\theta}_t - \theta^* \rangle = \langle F^\pi(\bar{\theta}_t) - F^\pi(\theta^*), \bar{\theta}_t - \theta^* \rangle &\ge \Lambda_{\min} \|\bar{\theta}_t - \theta^*\|_2^2 \\ \|F^\pi(\bar{\theta}_t)\|_2 = \|F^\pi(\bar{\theta}_t) - F^\pi(\theta^*)\|_2 &\le \Lambda_{\max} \|\bar{\theta}_t - \theta^*\|_2, \end{aligned}$$

then imply that

$$\begin{aligned} \|\bar{\theta}_{t+1} - \theta^*\|_2^2 &\le (1 - 2\beta_t \Lambda_{\min} + 2\beta_t^2 \Lambda_{\max}^2) \|\bar{\theta}_t - \theta^*\|_2^2 + 2\beta_t CR^2 \rho^\alpha + 2\beta_t^2 C^2 R^2 \rho^{2\alpha} \\ &\le (1 - \tfrac{3}{t + t_0 - 1}) \|\bar{\theta}_t - \theta^*\|_2^2 + 2\beta_t CR^2 \rho^\alpha + 2\beta_t^2 C^2 R^2 \rho^{2\alpha}, \end{aligned} \tag{7.3}$$

where the last inequality follows from

$$\begin{aligned}2(\beta_t\Lambda_{\min}-\beta_t^2\Lambda_{\max}^2)&=2\beta_t(\Lambda_{\min}-\beta_t\Lambda_{\max}^2)=2\beta_t(\Lambda_{\min}-\tfrac{2\Lambda_{\max}^2}{\Lambda_{\min}(t+t_0-1)})\\&\geq 2\beta_t(\Lambda_{\min}-\tfrac{2\Lambda_{\max}^2}{\Lambda_{\min}t_0})\geq\tfrac{3}{2}\beta_t\Lambda_{\min}=\tfrac{3}{t+t_0-1}\end{aligned}$$

due to the selection of β_t in (5.12). Now let us denote $\Gamma_t:=\begin{cases}1 & t=0,\\ (1-\frac{3}{t+t_0-1})\Gamma_{t-1} & t\geq 1,\end{cases}$ or equivalently, $\Gamma_t:=\frac{(t_0-1)(t_0-2)(t_0-3)}{(t+t_0-1)(t+t_0-2)(t+t_0-3))}$. Dividing both sides of (7.3) by Γ_t and taking the telescopic sum, we have

$$\tfrac{1}{\Gamma_t}\|\bar\theta_{t+1}-\theta^*\|_2^2\leq\|\bar\theta_1-\theta^*\|_2^2+2CR^2\rho^\alpha\textstyle\sum_{i=1}^t\frac{\beta_i}{\Gamma_i}+2C^2R^2\rho^{2\alpha}\sum_{i=1}^t\frac{\beta_i^2}{\Gamma_i}.$$

Noting that

$$\begin{aligned}\textstyle\sum_{i=1}^t\frac{\beta_i}{\Gamma_i}&=\frac{2}{\Lambda_{\min}}\sum_{i=1}^t\frac{(i+t_0-2)(i+t_0-3)}{(t_0-1)(t_0-2)(t_0-3)}\leq\frac{2\sum_{i=1}^t(i+t_0-2)^2}{\Lambda_{\min}(t_0-1)(t_0-2)(t_0-3)}\\&\leq\frac{2(t+t_0-1)^3}{3\Lambda_{\min}(t_0-1)(t_0-2)(t_0-3)},\\\textstyle\sum_{i=1}^t\frac{\beta_i^2}{\Gamma_i}&\leq\frac{4\sum_{i=1}^t(i+t_0-3)}{\Lambda_{\min}^2(t_0-1)(t_0-2)(t_0-3)}\leq\frac{2(t+t_0-2)^2}{\Lambda_{\min}^2(t_0-1)(t_0-2)(t_0-3)},\end{aligned}$$

we conclude

$$\begin{aligned}\|\bar\theta_{t+1}-\theta^*\|_2^2&\leq\tfrac{(t_0-1)(t_0-2)(t_0-3)}{(t+t_0-1)(t+t_0-2)(t+t_0-3)}\|\bar\theta_1-\theta^*\|_2^2+2CR^2\rho^\alpha\tfrac{2(t+t_0-1)^2}{3\Lambda_{\min}(t+t_0-2)(t+t_0-3)}\\&\quad+2C^2R^2\rho^{2\alpha}\tfrac{2(t+t_0-2)}{\Lambda_{\min}^2(t+t_0-1)(t+t_0-3)}\\&\leq\tfrac{(t_0-1)(t_0-2)(t_0-3)}{(t+t_0-1)(t+t_0-2)(t+t_0-3)}\|\bar\theta_1-\theta^*\|_2^2+\tfrac{8CR^2\rho^\alpha}{3\Lambda_{\min}}+\tfrac{C^2R^2\rho^{2\alpha}}{\Lambda_{\min}^2},\end{aligned}$$

from which the result holds since $\bar\theta_1=\theta_1$. ∎