Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies

Rui Yuan ruiyuan@fb.com
FAIR, Meta AI & LTCI, Télécom Paris & Institut Polytechnique de Paris
Paris, France

Simon S. Du ssdu@cs.washington.edu
FAIR, Meta AI & University of Washington
Seattle, USA

Robert M. Gower gowerrobert@gmail.com
CCM, Flatiron Institute
New York, USA

Alessandro Lazaric lazaric@fb.com
FAIR, Meta AI
Paris, France

Lin Xiao linx@fb.com
FAIR, Meta AI
Seattle, USA

Abstract

We consider infinite-horizon discounted Markov decision processes and study the convergence rates of the natural policy gradient (NPG) and the Q-NPG methods with the log-linear policy class. Using the compatible function approximation framework, NPG and Q-NPG with log-linear policies can be written as approximate versions of the policy mirror descent (PMD) method. By extending a recent analysis of PMD in the tabular setting, we obtain linear convergence rates and $\mathcal{O}(1/\epsilon^2)$ sample complexities for both NPG and Q-NPG with log-linear policy parametrization using a simple, non-adaptive geometrically increasing step size, without resorting to entropy or other strongly convex regularization. As a byproduct, we obtain sublinear convergence rates for both NPG and Q-NPG with arbitrary large constant step sizes.

Keywords: discounted Markov decision process, natural policy gradient, policy mirror descent, log-linear policy, sample complexity.

1. Introduction

Policy gradient (PG) methods have emerged as a popular class of algorithms for reinforcement learning. Unlike classical methods based on (approximate) dynamic programming (e.g., Puterman, 1994; Sutton and Barto, 2018), PG methods update directly the policy and its parametrization along the gradient direction of the value function (e.g., Williams, 1992; Sutton et al., 2000; Konda and Tsitsiklis, 2000; Baxter and Bartlett, 2001). An important variant of PG is the natural policy gradient (NPG) method (Kakade, 2001). NPG uses the Fisher information matrix of the policy distribution as a preconditioner to improve the policy gradient direction, similar to quasi-Newton methods in classical optimization. Variants of NPG with policy parametrization through deep neural networks were shown to have impressive empirical successes (Schulman et al., 2015; Lillicrap et al., 2016; Mnih et al., 2016; Schulman et al., 2017).

Motivated by the success of NPG in practice, there is now a concerted effort to develop convergence theories for the NPG method. Neu et al. (2017) provide the first interpretation of NPG as a mirror descent (MD) method (Nemirovski and Yudin, 1983; Beck and Teboulle, 2003). By leveraging different techniques for analyzing MD, it has been established that NPG converges to the global optimum in the tabular case (Agarwal et al., 2021; Khodadadian et al., 2021; Xiao, 2022) and some more general settings (Shani et al., 2020; Tomar et al., 2022; Vaswani et al., 2022; Kuba et al., 2022). In order to get fast linear convergence rate for NPG, several recent works consider the regularized NPG methods, such as the entropy-regularized NPG (Cen et al., 2021) and other convex regularized NPG (Lan, 2022; Zhan et al., 2021). By designing appropriate step sizes, Khodadadian et al. (2021) and Xiao (2022) obtain linear convergence of NPG without regularization. However, all these linear convergence results are limited in the tabular setting (direct parametrization). It remains unclear whether same convergence rate can be established in the function approximation regime.

In this paper we provide an affirmative answer to this question for the log-linear policy class. Our approach is based on the framework of *compatible function approximation* (Sutton et al., 2000; Kakade, 2001), which was extensively developed by Agarwal et al. (2021). Using this framework, variants of NPG with log-linear policies can be written as policy mirror descent (PMD) methods with approximate evaluations of the advantage function or Q-function (giving rise to NPG or Q-NPG respectively). Then by extending a recent analysis of PMD (Xiao, 2022), we obtain non-asymptotic linear convergence of both NPG and Q-NPG with log-linear policies. A distinctive feature of this approach is the use of a simple, non-adaptive geometrically increasing step size, without resorting to entropy or other strongly convex regularization. See Appendix A for a thorough review.

2. Preliminaries on Markov Decision Processes

We consider an MDP denoted as $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, c, \gamma\}$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is a Markovian transition model with $\mathcal{P}(s' \mid s, a)$ being the transition probability from state s to s' under action a, c is a cost function with $c(s,a) \in [0,1]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $\gamma \in [0,1)$ is a discounted factor. Here we use cost instead of reward to better align with the convention in the optimization literature.

The agent's behavior is modeled as a stochastic policy $\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}$, where $\pi_s \in \Delta(\mathcal{A})$ is the probability distribution over actions $\mathcal{A}$ in state $s \in \mathcal{S}$. At each time t, the agent takes an action $a_t \in \mathcal{A}$ given the current state $s_t \in \mathcal{S}$, following the policy π, i.e., $a_t \sim \pi_{s_t}$. Then the MDP transitions into the next state s_{t+1} with probability $\mathcal{P}(s_{t+1} \mid s_t, a_t)$ and the agent encounters the cost $c_t = c(s_t, a_t)$. Thus, a policy induces a distribution over trajectories $\{s_t, a_t, c_t\}_{t \geq 0}$. In the infinite-horizon discounted setting, the cost function of π with an initial state s is defined as

$$V_s(\pi) \overset{\text{def}}{=} \mathbb{E}_{\substack{a_t \sim \pi_{s_t} \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \mid s_0 = s \right]. \tag{1}$$

Given an initial state distribution $\rho \in \Delta(\mathcal{S})$, the goal of the agent is to find a policy π that (approximately) minimizes the expected cost function

$$V_\rho(\pi) \overset{\text{def}}{=} \mathbb{E}_{s \sim \rho}\left[V_s(\pi)\right] = \sum_{s \in \mathcal{S}} \rho_s V_s(\pi) = \langle V(\pi), \rho \rangle.$$

A more granular characterization of the performance of a policy is the state-action cost function (Q-function). For any pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, it is defined as

$$Q_{s,a}(\pi) \overset{\text{def}}{=} \mathbb{E}_{\substack{a_t \sim \pi_{s_t} \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \mid s_0 = s, a_0 = a \right]. \tag{2}$$

Let $Q_s \in \mathbb{R}^{|\mathcal{A}|}$ denote the vector $[Q_{s,a}]_{a \in \mathcal{A}}$. Then we have $V_s(\pi) = \mathbb{E}_{a \sim \pi_s}\left[Q_{s,a}(\pi)\right] = \langle \pi_s, Q_s(\pi) \rangle$. The advantage function[1] is a centered version of the Q-function:

$$A_{s,a}(\pi) \overset{\text{def}}{=} Q_{s,a}(\pi) - V_s(\pi), \tag{3}$$

which satisfies $\mathbb{E}_{a \sim \pi_s}\left[A_{s,a}(\pi)\right] = 0$ for all $s \in \mathcal{S}$.

Visitation probabilities. Given a starting state distribution $\rho \in \Delta(\mathcal{S})$, we define the *state visitation distribution* $d^\pi(\rho) \in \Delta(\mathcal{S})$, induced by a policy π, as

$$d_s^\pi(\rho) \overset{\text{def}}{=} (1-\gamma)\, \mathbb{E}_{s_0 \sim \rho} \left[\sum_{t=0}^{\infty} \gamma^t \Pr^\pi(s_t = s \mid s_0) \right],$$

[1] An advantage function should measure how much better is a compared to π, while here A is positive when a is worse than π. We keep calling A advantage function to better align with the convention in the RL literature.

where $\mathrm{Pr}^\pi(s_t = s \mid s_0)$ is the probability that the t-th state is equal to s by following the trajectory generated by π starting from s_0. We define the *state-action visitation distribution* $\bar{d}^\pi(\rho) \in \Delta(\mathcal{S} \times \mathcal{A})$ as

$$\bar{d}^\pi_{s,a}(\rho) \overset{\text{def}}{=} d^\pi_s(\rho)\pi_{s,a} = (1-\gamma)\,\mathbb{E}_{s_0 \sim \rho}\left[\sum_{t=0}^{\infty} \gamma^t \,\mathrm{Pr}^\pi(s_t = s, a_t = a \mid s_0)\right]. \tag{4}$$

In addition, we extend the definition of $\bar{d}^\pi(\rho)$ by specifying the initial state-action distribution $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$, i.e.,

$$\tilde{d}^\pi_{s,a}(\nu) \overset{\text{def}}{=} (1-\gamma)\,\mathbb{E}_{(s_0,a_0)\sim\nu}\left[\sum_{t=0}^{\infty} \gamma^t \,\mathrm{Pr}^\pi(s_t = s, a_t = a \mid s_0, a_0)\right]. \tag{5}$$

The difference in the last two definitions is that for the former, the initial action a_0 is sampled directly from π, whereas for the latter, it is prescribed by the initial state-action distribution ν. We use $\tilde{d}$ compared to $\bar{d}$ to better distinguish the cases with ν and ρ. Without specification, we even omit the argument ν or ρ throughout the paper to simplify the presentation as they are self-evident. From these definitions, we have for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$d^\pi_s \geq (1-\gamma)\rho_s, \qquad \bar{d}^\pi_{s,a} \geq (1-\gamma)\rho_s\pi_{s,a}, \qquad \tilde{d}^\pi_{s,a} \geq (1-\gamma)\nu_{s,a}. \tag{6}$$

Policy parametrization. In general, both the state and action spaces $\mathcal{S}$ and $\mathcal{A}$ can be very large and some form of function approximation is needed to make the computation feasible. In particular, the policy π is often parametrized as $\pi(\theta)$ with $\theta \in \mathbb{R}^m$, where m is much smaller than $|\mathcal{S}|$ and $|\mathcal{A}|$. In this paper, we focus on the log-linear policy class. Specifically, we assume that for each state-action pair (s,a), there is a feature mapping $\phi_{s,a} \in \mathbb{R}^m$ and the policy takes the form

$$\pi_{s,a}(\theta) = \frac{\exp(\phi_{s,a}^\top \theta)}{\sum_{a' \in \mathcal{A}} \exp(\phi_{s,a'}^\top \theta)}. \tag{7}$$

To simplify notation in the rest of this paper, we use the shorthand $V_\rho(\theta)$ for $V_\rho(\pi(\theta))$ and similarly $Q_{s,a}(\theta)$ for $Q_{s,a}(\pi(\theta))$, $A_{s,a}(\theta)$ for $A_{s,a}(\pi(\theta))$, d^θ_s for $d^{\pi(\theta)}_s$, $\bar{d}^\theta_{s,a}$ for $\bar{d}^{\pi(\theta)}_{s,a}$, and $\tilde{d}^\theta_{s,a}$ for $\tilde{d}^{\pi(\theta)}_{s,a}$.

Natural Policy Gradient (NPG) Method. Using the notations defined above, the parametrized policy optimization problem is to minimize the function $V_\rho(\theta)$ over $\theta \in \mathbb{R}^m$. The policy gradient is given by (see, e.g., Williams, 1992; Sutton et al., 2000)

$$\nabla_\theta V_\rho(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\theta,\, a\sim\pi_s(\theta)}\left[Q_{s,a}(\theta)\,\nabla_\theta \log \pi_{s,a}(\theta)\right]. \tag{8}$$

For parametrizations that are differentiable and satisfy $\sum_{a\in\mathcal{A}} \pi_{s,a}(\theta) = 1$, including the log-linear class defined in (7), we can replace $Q_{s,a}(\theta)$ by $A_{s,a}(\theta)$ in the above expression (Agarwal et al., 2021). The NPG method (Kakade, 2001) takes the form

$$\theta^{(k+1)} = \theta^{(k)} - \eta_k F_\rho\big(\theta^{(k)}\big)^\dagger\, \nabla_\theta V_\rho\big(\theta^{(k)}\big), \tag{9}$$

where $\eta_k > 0$ is a scalar step size, $F_\rho(\theta)$ is the Fisher information matrix

$$F_\rho(\theta) \overset{\text{def}}{=} \mathbb{E}_{s\sim d^\theta,\, a\sim\pi_s(\theta)}\left[\nabla_\theta \log \pi_{s,a}(\theta)\big(\nabla_\theta \log \pi_{s,a}(\theta)\big)^\top\right],$$

and $F_\rho(\theta)^\dagger$ denotes the Moore-Penrose pseudoinverse of $F_\rho(\theta)$.

3. NPG with Compatible Function Approximation

The parametrized value function $V_\rho(\theta)$ is non-convex in general (see, e.g., Agarwal et al., 2021). Treating policy optimization as a general non-convex optimization problem thus loses certain problem structure and results in weak convergence results. Following Agarwal et al. (2021), we adopt the framework of the *compatible function approximation* (Sutton et al., 2000; Kakade, 2001), which retains the MDP structure and leads to tight convergence rate analysis.

Kakade (2001) showed that the NPG update (9) is equivalent to (up to a constant scaling of η_k)

$$\theta^{(k+1)} = \theta^{(k)} - \eta_k w_\star^{(k)}, \qquad w_\star^{(k)} \in \mathrm{argmin}_{w\in\mathbb{R}^m}\, L_A\big(w, \theta^{(k)}, \bar{d}^{(k)}\big), \tag{10}$$

where $\bar{d}^{(k)}$ is a shorthand for the state-action visitation distribution $\bar{d}^{\pi(\theta^{(k)})}(\rho)$ defined in (4), and for a state-action distribution ζ, $L_A(w,\theta,\zeta)$ is the *compatible function approximation error* defined as

$$L_A(w,\theta,\zeta) \overset{\text{def}}{=} \mathbb{E}_{(s,a)\sim\zeta}\left[\left(w^\top \nabla_\theta \log \pi_{s,a}(\theta) - A_{s,a}(\theta)\right)^2\right]. \tag{11}$$

A derivation of (10) is provided in Lemma 16 in Appendix B for the completeness. In other words, $w_\star^{(k)}$ is the solution to a regression problem that tries to approximate $A_{s,a}(\theta^{(k)})$ using $\nabla_\theta \log \pi_{s,a}(\theta^{(k)})$ as features. For the log-linear policy class defined in (7), we have

$$\nabla_\theta \log \pi_{s,a}(\theta) \;=\; \bar{\phi}_{s,a}(\theta) \;\overset{\text{def}}{=}\; \phi_{s,a} - \textstyle\sum_{a'\in\mathcal{A}} \pi_{s,a'}(\theta)\phi_{s,a'} \;=\; \phi_{s,a} - \mathbb{E}_{a'\sim\pi_s(\theta)}\left[\phi_{s,a'}\right], \tag{12}$$

where $\bar{\phi}_{s,a}(\theta)$ are called *centered features vectors*.

In practice, we cannot minimize L_A exactly; instead, a sample-based regression problem is solved to obtain an approximate solution $w^{(k)}$. This leads to the following approximate NPG update rule:

$$\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k w^{(k)}, \qquad w^{(k)} \approx \operatorname{argmin}_w L_A\big(w,\theta^{(k)},\bar{d}^{(k)}\big). \tag{13}$$

Alternatively, as proposed by Agarwal et al. (2021), we can define the compatible function approximation error as

$$L_Q(w,\theta,\zeta) \overset{\text{def}}{=} \mathbb{E}_{(s,a)\sim\zeta}\left[\left(w^\top \phi_{s,a} - Q_{s,a}(\theta)\right)^2\right] \tag{14}$$

and use it to derive a variant of the approximate NPG update called Q-NPG:

$$\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k w^{(k)}, \qquad w^{(k)} \approx \operatorname{argmin}_w L_Q\big(w,\theta^{(k)},\bar{d}^{(k)}\big). \tag{15}$$

The approximate NPG and Q-NPG updates require samples of unbiased estimates of $A_{s,a}(\theta)$ and $Q_{s,a}(\theta)$ respectively, and the corresponding sampling procedures are given in the Appendix as Algorithms 4 and 3, respectively.

Following Agarwal et al. (2021), we consider slightly different variants of NPG and Q-NPG, where $\bar{d}^{(k)}$ in (13) and (15) is replaced by a more general state-action visitation distribution $\tilde{d}^{(k)} = \tilde{d}^{\pi(\theta^{(k)})}(\nu)$ defined in (5) with $\nu \in \Delta(\mathcal{S}\times\mathcal{A})$. The advantage of using $\tilde{d}^{(k)}$ is that it allows better exploration than $\bar{d}^{(k)}$ as ν can be chosen to be independent of the policy $\pi(\theta^{(k)})$. For example, it can be seen from (6) that the lower bound of $\tilde{d}^\pi$ is independent of π, which is not the case for $\bar{d}^\pi$. This property is crucial in the forthcoming convergence analysis.

3.1 Formulation as Approximate Policy Mirror Descent

Given an approximate solution $w^{(k)}$ for minimizing $L_Q\big(w,\theta^{(k)},\tilde{d}^{(k)}\big)$, the Q-NPG update rule $\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k w^{(k)}$, when plugged in the log-linear parametrization (7), results in a new policy

$$\pi_{s,a}^{(k+1)} \;=\; \frac{1}{Z_s^{(k)}}\pi_{s,a}^{(k)} \exp\left(-\eta_k\, \phi_{s,a}^T w^{(k)}\right), \qquad \forall\,(s,a)\in\mathcal{S}\times\mathcal{A},$$

where $\pi^{(k)}$ is a shorthand for $\pi_{s,a}(\theta^{(k)})$ and $Z_s^{(k)}$ is a normalization factor to ensure $\sum_{a\in\mathcal{A}}\pi_{s,a}^{(k+1)} = 1$, for each $s\in\mathcal{S}$. Notice that the above $\pi^{(k+1)}$ can also be obtained by a mirror descent update:

$$\pi_s^{(k+1)} = \arg\min_{p\in\Delta(\mathcal{A})}\left\{\eta_k\left\langle \Phi_s w^{(k)}, p\right\rangle + D(p,\pi_s^{(k)})\right\}, \quad \forall s\in\mathcal{S}, \tag{16}$$

where $\Phi_s \in \mathbb{R}^{|\mathcal{A}|\times m}$ is a matrix with rows $(\phi_{s,a})^\top \in \mathbb{R}^m$ for $a\in\mathcal{A}$, and $D(p,q)$ denotes the Kullback-Leibler (KL) divergence between two distributions $p,q\in\Delta(\mathcal{A})$, i.e.,

$$D(p,q) \overset{\text{def}}{=} \sum_{a\in\mathcal{A}} p_a \log\left(\frac{p_a}{q_a}\right).$$

A derivation of (16) is provided in Lemma 17 in Appendix B for the completeness.

If we replace $\Phi_s w^{(k)}$ in (16) by the vector $\left[Q_{s,a}(\pi^{(k)})\right]_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$, then it becomes the *policy mirror descent* (PMD) method in the tabular setting studied by, for example, Shani et al. (2020), Lan (2022) and Xiao (2022). In fact, the update rule (16) can be viewed as an approximate PMD method where $Q_s(\pi^{(k)})$ is linearly approximated by $\Phi_s w^{(k)}$ through compatible function approximation (14). Similarly, we can write the approximate NPG update rule as

$$\pi_s^{(k+1)} = \arg\min_{p\in\Delta(\mathcal{A})} \left\{\eta_k \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, p\right\rangle + D(p, \pi_s^{(k)})\right\}, \quad \forall s \in \mathcal{S}, \tag{17}$$

where $w^{(k)}$ is an approximate solution for minimizing $L_A\big(w, \theta^{(k)}, \tilde{d}^{(k)}\big)$ defined in (11), and $\bar{\Phi}_s^{(k)} \in \mathbb{R}^{|\mathcal{A}|\times m}$ is a matrix whose rows consist of the centered feature maps $\bar{\phi}_{s,a}(\theta^{(k)})$, as defined in (12).

Reformulating Q-NPG and NPG into the mirror descent forms (16) and (17), respectively, allows us to adapt the analysis of PMD method developed in Xiao (2022) to obtain sharper convergence rates. In particular, we show that with an increasing step size $\eta_k \propto \gamma^k$, both NPG and Q-NPG with log-linear policy parametrization has linear convergence up to an error floor determined by the quality of the compatible function approximation.

4. Analysis of Q-NPG with Log-Linear Policies

In this section, we provide the convergence analysis of the following approximate Q-NPG method:

$$\theta^{(k+1)} \; = \; \theta^{(k)} - \eta_k w^{(k)}, \qquad w^{(k)} \approx \operatorname{argmin}_w L_Q\big(w, \theta^{(k)}, \tilde{d}^{\pi(\theta^{(k)})}(\nu)\big), \tag{18}$$

where $\nu \in \Delta(\mathcal{S}\times\mathcal{A})$ is an arbitrary state-action distribution and does not depend on ρ. The exact minimizer is denoted as $w_\star^{(k)} \in \operatorname{argmin}_w L_Q\big(w, \theta^{(k)}, \tilde{d}^{(k)}\big)$, where $\tilde{d}^{(k)}$ is a shorthand for $\tilde{d}^{\pi(\theta^{(k)})}(\nu)$. The corresponding update from $\pi^{(k)}$ to $\pi^{(k+1)}$ can be described by the PMD method (16). We note that the variant of Q-NPG analyzed in Agarwal et al. (2021) requires that $w^{(k)}$ has a bounded norm. It is not needed here because we rely on quite different techniques for convergence analysis.

The compatible function approximation error can be decomposed as

$$L_Q\big(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big) \; = \; \underbrace{L_Q\big(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big) - L_Q\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big)}_{\text{Statistical error (excess risk)}} + \underbrace{L_Q\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big)}_{\text{Approximation error}}.$$

The statistical error measures how accurate we solve the regression problem, i.e., how good $w^{(k)}$ is compared with $w_\star^{(k)}$. The approximation error measures the best possible quality of approximating $Q_{s,a}(\theta^{(k)})$ using $\phi_{s,a}$ as features in the regression problem (modeling error). One way to proceed with the analysis is to assume that both the statistical error and the approximation error are bounded for all iterations, which is the approach we take in Section 4.2. In Section 4.1, we first take an alternative approach proposed by Agarwal et al. (2021), where the assumption of bounded approximation error is replaced by a bounded *transfer error*. The transfer error refers to $L_Q\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^*\big)$, where the iteration-dependent visitation distribution $\tilde{d}^{(k)}$ is shifted to a fixed one $\tilde{d}^*$ (defined in Section 4.1). These two approaches require different additional assumptions and result in slightly different convergence rates. Here we first state the common assumption on the bounded statistical error.

Assumption 1 (Bounded statistical error, Assumption 6.1.1 in Agarwal et al. (2021)) *There exists $\epsilon_{\text{stat}} > 0$ such that for all iterations $k \geq 0$ of the Q-NPG method (18), we have*

$$\mathbb{E}\left[L_Q\big(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big) - L_Q\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big)\right] \; \leq \; \epsilon_{\text{stat}}. \tag{19}$$

By solving the regression problem with sampling based approaches, we can expect $\epsilon_{\text{stat}} = \mathcal{O}(1/\sqrt{T})$ (Agarwal et al., 2021) or $\epsilon_{\text{stat}} = \mathcal{O}(1/T)$ (see Corollary 24, (also see Liu et al., 2020)) where T is the number of iterations used to find the approximate solution $w^{(k)}$.

4.1 Analysis with Bounded Transfer Error

Here we introduce some additional notation. For any state distributions $p, q \in \Delta(\mathcal{S})$, we define the *distribution mismatch coefficient* of p relative to q as

$$\left\|\frac{p}{q}\right\|_\infty \overset{\text{def}}{=} \max_{s\in\mathcal{S}} \frac{p_s}{q_s}.$$

Let π^* be an arbitrary *comparator policy*, which is not necessarily an optimal policy and does not need to belong to the log-linear policy class. Fix a state distribution $\rho \in \Delta(\mathcal{S})$. We denote d^* as $d^{\pi^*}(\rho)$ and $d^{(k)}$ as $d^{\pi(\theta^{(k)})}(\rho)$, and define the following distribution mismatch coefficients:

$$\vartheta_k \overset{\text{def}}{=} \left\|\frac{d^*}{d^{(k)}}\right\|_\infty \overset{(6)}{\leq} \frac{1}{1-\gamma}\left\|\frac{d^*}{\rho}\right\|_\infty \qquad \text{and} \qquad \vartheta_\rho \overset{\text{def}}{=} \frac{1}{1-\gamma}\left\|\frac{d^*}{\rho}\right\|_\infty \geq \frac{1}{1-\gamma}. \tag{20}$$

Thus, for all $k \geq 0$, we have $\vartheta_k \leq \vartheta_\rho$. We assume that $\vartheta_\rho < \infty$, which is the case, for example, if $\rho_s > 0$ for all $s \in \mathcal{S}$.

We also introduce a weighted KL divergence given by

$$D_k^* \overset{\text{def}}{=} \mathbb{E}_{s\sim d^*}\left[D(\pi_s^*, \pi_s^{(k)})\right].$$

If we choose the uniform initial policy, i.e., $\pi_{s,a}^{(0)} = 1/|\mathcal{A}|$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ (or $\theta^{(0)} = 0$), then $D_0^* \leq \log|\mathcal{A}|$ for all $\rho \in \Delta(\mathcal{S})$ and for any $\pi^* \in \Delta(\mathcal{A})^{\mathcal{S}}$. The choice of the step size will directly depend on D_0^* in all our forthcoming convergence results.

Given a state distribution ρ and a comparator policy π^*, we define a state-action measure $\tilde{d}^*$ as

$$\tilde{d}_{s,a}^* \overset{\text{def}}{=} d_s^* \cdot \text{Unif}_{\mathcal{A}}(a) \overset{\text{def}}{=} \frac{d_s^*}{|\mathcal{A}|}, \tag{21}$$

and use it to express the transfer error as $L_Q\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^*\big)$.

Assumption 2 (Bounded transfer error, Assumption 6.1.2 in Agarwal et al. (2021)) *There exists $\epsilon_{\text{bias}} > 0$ such that for all iterations $k \geq 0$ of the Q-NPG method (18), we have*

$$\mathbb{E}\left[L_Q\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^*\big)\right] \leq \epsilon_{\text{bias}}. \tag{22}$$

The transfer error bound ϵ_{bias} characterizes how well the Q-values can be linearly approximated by the feature maps $\phi_{s,a}$. It can be shown that $\epsilon_{\text{bias}} = 0$ when $\pi^{(k)}$ is the softmax tabular policy (Agarwal et al., 2021) or the MDP has certain low-rank structure (Jiang et al., 2017; Yang and Wang, 2019; Jin et al., 2020). For rich neural parametrizations, ϵ_{bias} can be made small (Wang et al., 2020).

The next assumption concerns the relative condition number between two covariance matrices of $\phi_{s,a}$ defined under different state-action distributions.

Assumption 3 (Bounded relative condition number, Assumption 6.2 in Agarwal et al. (2021)) *Fix a state distribution ρ, a state-action distribution ν and a comparator policy π^*. Let*

$$\Sigma_{\tilde{d}^*} \overset{\textit{def}}{=} \mathbb{E}_{(s,a)\sim\tilde{d}^*}\left[\phi_{s,a}\phi_{s,a}^\top\right], \qquad \textit{and} \qquad \Sigma_\nu \overset{\textit{def}}{=} \mathbb{E}_{(s,a)\sim\nu}\left[\phi_{s,a}\phi_{s,a}^\top\right], \tag{23}$$

where $\tilde{d}^*$ is specified in (21). We define the relative condition number between $\Sigma_{\tilde{d}^*}$ and Σ_ν as

$$\kappa_\nu \overset{\textit{def}}{=} \max_{w\in\mathbb{R}^m} \frac{w^\top \Sigma_{\tilde{d}^*} w}{w^\top \Sigma_\nu w}, \tag{24}$$

and assume that κ_ν is finite.

Notice that Assumption 3 benefits from the us e of ν. In fact, it is shown in Agarwal et al. (2021, Remark 22 and Lemma 23) that κ_ν can be reasonably small (e.g., $\kappa_\nu \leq m$ is always possible) and independent of the size of the state space by controlling ν.

Our analysis also needs the following assumption, which does not appear in Agarwal et al. (2021).

Assumption 4 (Concentrability coefficient for state visitation) *There exists a finite $C_\rho > 0$ such that for all iterations $k \geq 0$ of the Q-NPG method* (18)*, it holds that*

$$\mathbb{E}_{s\sim d^*}\left[\left(\frac{d_s^{(k)}}{d_s^*}\right)^2\right] \leq C_\rho. \tag{25}$$

Let $\rho_{\min} = \min_{s\in\mathcal{S}} \rho_s$. A sufficient condition for Assumption 4 to hold is that $\rho_{\min} > 0$. Indeed,

$$\sqrt{\mathbb{E}_{s\sim d^*}\left[\left(\frac{d_s^{(k)}}{d_s^*}\right)^2\right]} \leq \left\|\frac{d^{(k)}}{d^*}\right\|_\infty \overset{(6)}{\leq} \frac{1}{1-\gamma}\left\|\frac{d^{(k)}}{\rho}\right\|_\infty \leq \frac{1}{(1-\gamma)\rho_{\min}}. \tag{26}$$

In reality, $\sqrt{C_\rho}$ can be much smaller than the pessimistic bound shown above. This is especially the case if we choose π^* to be the optimal policy and $d^{(k)} \to d^*$. We further replace C_ρ by a weaker one independent of ρ in Section 4.2.

Now we present our first main result.

Theorem 5 *Fix a state distribution ρ, an state-action distribution ν and a comparator policy π^*. We consider the Q-NPG method* (18) *with the step sizes satisfying $\eta_0 \geq \frac{1-\gamma}{\gamma} D_0^*$ and $\eta_{k+1} \geq \frac{1}{\gamma}\eta_k$. Suppose that Assumptions 1, 2, 3 and 4 all hold. Then we have for all $k \geq 0$,*

$$\mathbb{E}\left[V_\rho(\theta^{(k)})\right] - V_\rho(\pi^*) \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^k \frac{2}{1-\gamma} + \frac{2\sqrt{|\mathcal{A}|}\left(\vartheta_\rho\sqrt{C_\rho}+1\right)}{1-\gamma}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{bias}}}\right). \tag{27}$$

The main differences between our Theorem 5 and Theorem 20 of Agarwal et al. (2021), which is their corresponding result on the approximate Q-NPG method, are summarized as follows.

- The convergence rate of Agarwal et al. (2021, Theorem 20) is $\mathcal{O}(1/\sqrt{k})$ up to an error floor determined by ϵ_{stat} and ϵ_{bias}. We have linear convergence up to an error floor that also depends on ϵ_{stat} and ϵ_{bias}. However, the magnitude of our error floor is worse (larger) by a factor of $\vartheta_\rho\sqrt{C_\rho}$, due to the concentrability and the distribution mismatch coefficients used in our proof. A very pessimistic bound on this factor is as large as $|\mathcal{S}|^2/(1-\gamma)^2$.
- In terms of required conditions, both results use Assumptions 1, 2 and 3. Agarwal et al. (2021, Theorem 20) further assume that the norms of the feature maps $\phi_{s,a}$ are uniformly bounded and $w^{(k)}$ has a bounded norm (e.g., obtained by a projected stochastic gradient descent). Due to different analysis techniques referred next, we instead rely on a concentrability coefficient C_ρ defined in Assumption 4.
- Agarwal et al. (2021, Theorem 20) uses a diminishing step size $\eta \propto 1/\sqrt{k}$ where k is the total number of iterations, but we uses a geometrically increasing step size $\eta_k \propto \gamma^k$ for all $k \geq 0$. This discrepancy reflects the quite different analysis techniques adopted. The key analysis tool in Agarwal et al. (2021) is a *NPG Regret Lemma* (their Lemma 34) which relies on the smoothness of the functions $\log \pi_{s,a}(\theta)$ (thus the boundedness of $\|\phi_{s,a}\|$) and the boundedness of $\|w^{(k)}\|$, and thus the classical $\mathcal{O}(1/\sqrt{k})$ diminishing step size in the optimization literature. Our analysis exploits the three-point descent lemma (Chen and Teboulle, 1993) and the performance difference lemma (Kakade and Langford, 2002), without reliance on smoothness parameters. As a consequence, we take advantage of exponentially growing step sizes and avoid assuming the boundedness of $\|\phi_{s,a}\|$ or $\|w^{(k)}\|$.

As a by product, we also obtain a sublinear convergence result while using a constant step size.

Theorem 6 *Fix a state distribution ρ, an state-action distribution ν and an* optimal policy π^*. *We consider the Q-NPG method* (18) *with a constant step size $\eta_k = \eta$ satisfying $\eta \geq \frac{1}{2\vartheta_\rho} D_0^*$. Suppose that Assumptions 1, 2, 3 and 4 all hold. Then we have for all $k \geq 0$,*

$$\frac{1}{k}\sum_{t=0}^{k-1} \mathbb{E}\left[V_\rho(\theta^{(t)})\right] - V_\rho(\pi^*) \leq \frac{4\vartheta_\rho}{(1-\gamma)k} + \frac{2\sqrt{|\mathcal{A}|}\left(\vartheta_\rho\sqrt{C_\rho}+1\right)}{1-\gamma}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{bias}}}\right). \tag{28}$$

A deviation from the setting of Theorem 5 is that here we require π^* to be an optimal policy [2]. Compared to Theorem 20 in Agarwal et al. (2021), our convergence rate is also sublinear, but improves from $\mathcal{O}(1/\sqrt{k})$ to $\mathcal{O}(1/k)$. Moreover, they use a diminishing step size of order $\mathcal{O}(1/\sqrt{k})$ while our constant step size is unbounded, independent of k and can be arbitrary large.

4.2 Analysis with Bounded Approximation Error

In this section, instead of assuming bounded transfer error, we provide a convergence analysis based on the usual notion of approximation error and a different concentrability coefficient.

Assumption 7 (Bounded approximation error) *There exists $\epsilon_{\text{approx}} > 0$ such that for all iterations $k \geq 0$ of the Q-NPG method* (18)*, it holds that*

$$\mathbb{E}\left[L_Q\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big)\right] \;\leq\; \epsilon_{\text{approx}}. \tag{29}$$

As mentioned in Agarwal et al. (2021), Assumption 7 is stronger than Assumption 2 (bounded transfer error). Indeed,

$$L_Q\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^*\big) \;\leq\; \left\|\frac{\tilde{d}^*}{\tilde{d}^{(k)}}\right\|_\infty L_Q\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big) \;\overset{(6)}{\leq}\; \frac{1}{1-\gamma}\left\|\frac{\tilde{d}^*}{\nu}\right\|_\infty L_Q\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big).$$

Assumption 8 (Concentrability coefficient for state-action visitation) *There exists $C_\nu < \infty$ such that for all iterations of the Q-NPG method* (18)*, we have*

$$\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{h_{s,a}^{(k)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right] \leq C_\nu, \tag{30}$$

where $h_{s,a}^{(k)}$ represents any of the following quantities:

$$d_s^{(k+1)}\pi_{s,a}^{(k+1)}\,, \qquad d_s^{(k+1)}\pi_{s,a}^{(k)}\,, \qquad d_s^*\pi_{s,a}^{(k)}\,, \qquad \textit{and} \quad d_s^*\pi_{s,a}^*\,. \tag{31}$$

Since ν is completely at our disposal and independent of ρ, it suffices to choose $\nu_{s,a} > 0$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ for Assumption 8 to hold. Indeed, with $\nu_{\min}$ denoting $\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \nu_{s,a}$, we have

$$\sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{h_{s,a}^{(k)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right]} \;\leq\; \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{h_{s,a}^{(k)}}{\tilde{d}_{s,a}^{(k)}} \;\overset{(6)}{\leq}\; \frac{1}{(1-\gamma)\nu_{\min}}, \tag{32}$$

where the upper bound can be smaller than that in (26) if $\rho_{\min}$ is smaller than $\nu_{\min}$.

[2] In our analysis, we need to drop the positive term $\mathbb{E}\left[V_\rho(\theta^{(k)}) - V_\rho(\pi^*)\right]$ to obtain a lower bound, thus require π^* to be an optimal policy.

Theorem 9 *Fix a state distribution ρ, an state-action distribution ν and a comparator policy π^*. We consider the Q-NPG method (18) with the step sizes satisfying $\eta_0 \geq \frac{1-\gamma}{\gamma} D_0^*$ and $\eta_{k+1} \geq \frac{1}{\gamma}\eta_k$. Suppose that Assumptions 1, 7 and 8 hold. Then we have for all $k \geq 0$,*

$$\mathbb{E}\left[V_\rho(\theta^{(k)})\right] - V_\rho(\pi^*) \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^k \frac{2}{1-\gamma} + \frac{2\sqrt{C_\nu}\,(\vartheta_\rho + 1)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{approx}}}\right). \tag{33}$$

Compared to Theorem 5, while the approximation error assumption is stronger than the transfer error assumption, we do not require the assumption on relative condition number κ_ν and the error floor does not depends on κ_ν nor explicitly on $|\mathcal{A}|$. Besides, we can always choose ν so that the concentrability coefficient C_ν is finite even if C_ρ is unbounded.

Remark 10 *Note that all Theorem 5, 6 and 9 benefit from using the visitation distribution $\tilde{d}^{(k)}$ instead of $\bar{d}^{(k)}$ (i.e., benefit from using ν instead of ρ). In particular, from (6), $\tilde{d}^{(k)}$ has a lower bound that is independent of the policy $\pi^{(k)}$ or ρ. This property allows us to define a weak notion of relative condition number (Assumption 3) that is independent of the iterates, and also get a finite upper bound of C_ν (Assumption 8 and (32)) that is independent of ρ.*

By further assuming that the feature maps are bounded and has non-singular covariance matrix, we obtain an $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity for Q-NPG with log-linear policies. We postpone the formal statement of this result to Appendix E for the sake of space.

5. Analysis of NPG with Log-Linear Policies

We now return to the convergence analysis of the approximate NPG method, specifically,

$$\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k w^{(k)}, \qquad w^{(k)} \approx \operatorname{argmin}_w L_A\big(w, \theta^{(k)}, \tilde{d}^{\,\pi(\theta^{(k)})}(\nu)\big), \tag{34}$$

where $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$ is an arbitrary state-action distribution and does not depend on ρ. Again, we use $\tilde{d}^{(k)}$ denote $\tilde{d}^{\,\pi(\theta^{(k)})}(\nu)$ and let $w_\star^{(k)} \in \operatorname{argmin}_w L_A\big(w, \theta^{(k)}, \tilde{d}^{(k)}\big)$ denote the minimizer. Our analysis of NPG is analogous to that of Q-NPG shown in the previous section, by exploiting the approximate PMD formulation (17) using techniques developed in Xiao (2022).

The set of assumptions we use for NPG is analogous to the assumptions used in Section 4.2. In particular, we assume bounded approximation error instead of transfer error (c.f., Assumption 2) in minimizing L_A and do not need the assumption on relative condition number.

Assumption 11 (Bounded statistical error, Assumption 6.5.1 in Agarwal et al. (2021)) *There exists $\epsilon_{\text{stat}} > 0$ such that for all iterations $k \geq 0$ of the NPG method (34), we have*

$$\mathbb{E}\left[L_A\big(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big) - L_A\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big)\right] \;\leq\; \epsilon_{\text{stat}}. \tag{35}$$

Assumption 12 (Bounded approximation error) *There exists $\epsilon_{\text{approx}} > 0$ such that for all terations $k \geq 0$ of the NPG method (34), we have*

$$\mathbb{E}\left[L_A\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big)\right] \;\leq\; \epsilon_{\text{approx}}. \tag{36}$$

Assumption 13 (Concentrability coefficient for state-action visitation) *There exists $C_\nu < \infty$ such that for all iterations $k \geq 0$ of the NPG method (34), we have*

$$\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{\bar{d}_{s,a}^{(k+1)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right] \;\leq\; C_\nu \qquad \textit{and} \qquad \mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{\bar{d}_{s,a}^{\pi^*}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right] \;\leq\; C_\nu. \tag{37}$$

Under the above assumptions, we have the following result.

Theorem 14 *Fix a state distribution ρ, a state-action distribution ν, and a comparator policy π^*. We consider the NPG method (34) with the step sizes satisfying $\eta_0 \geq \frac{1-\gamma}{\gamma} D_0^*$ and $\eta_{k+1} \geq \frac{1}{\gamma}\eta_k$. Suppose that Assumptions 11, 12 and 13 hold. Then we have for all $k \geq 0$,*

$$\mathbb{E}\left[V_\rho(\theta^{(k)})\right] - V_\rho(\pi^*) \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^k \frac{2}{1-\gamma} + \frac{\sqrt{C_\nu}\,(\vartheta_\rho + 1)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{approx}}}\right). \tag{38}$$

Now we compare Theorem 14 with Theorem 29 in Agarwal et al. (2021), which is their corresponding results on the approximate NPG method. The main differences are similar to those for Q-NPG as summarized right after Theorem 5: Their convergence rate is sublinear while ours is linear; they assume uniformly bounded $\phi_{s,a}$ and $w^{(k)}$ while we required bounded concentrability coefficient C_ν due to different proof techniques; they use diminishing step sizes and we use geometrically increasing ones. Moreover, Theorem 14 requires bounded approximation error, which is a stronger assumption than the bounded transfer error used by their Theorem 29, but we do not need the assumption on bounded relative condition number.

We note that the bounded relative condition number required by Agarwal et al. (2021, Theorem 29) needs to be held for the covariance matrix of $\bar{\phi}_{s,a}^{(k)}$ for all $k \geq 0$ because the centered feature maps $\bar{\phi}_{s,a}^{(k)}$ depends on the iterates $\theta^{(k)}$. This is in contrast to the use of a single fixed covariance matrix for Q-NPG as defined in (23).

In addition, the inequalities in (37) only involve half of the state-action visitation distributions listed in (31), i.e., the first and the fourth terms. From (32), the upper bound of C_ν is obtained only through (6), which is the property of $\tilde{d}^\pi$ itself for all policy $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$. Thus, C_ν in (37) can share the same upper bound in (32) independent to the use of the algorithm Q-NPG or NPG. Consequently, our concentrability coefficient assumption is weaker than Assumption 2 in Cayci et al. (2021) which studies the linear convergence of NPG with entropy regularization for the log-linear policy class. The reason is that the bound on C_ν in (32) does not depend on the policies throughout the iterations thanks to the use of $\tilde{d}^{(k)}$ instead of $\bar{d}^{(k)}$ (see Remark 10 as well).

Similar to Theorem 6, we also obtain a sublinear rate for NPG while using a constant step size.

Theorem 15 *Fix a state distribution ρ, an state-action distribution ν and an* optimal *policy π^*. We consider the NPG method (34) with a constant step size $\eta_k = \eta$ satisfying $\eta \geq \frac{1}{2\vartheta_\rho} D_0^*$. Suppose that Assumptions 11, 12 and 13 hold. Then we have for all $k \geq 0$,*

$$\frac{1}{k}\sum_{t=0}^{k-1} \mathbb{E}\left[V_\rho(\theta^{(t)})\right] - V_\rho(\pi^*) \leq \frac{4\vartheta_\rho}{(1-\gamma)k} + \frac{\sqrt{C_\nu}\,(\vartheta_\rho + 1)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{approx}}}\right). \tag{39}$$

Despite the difference of using ν instead of ρ, note that same convergence rate $\mathcal{O}(1/k)$ is established by Liu et al. (2020) for NPG with constant step size, while they require that the feature maps are bounded and the Fisher information matrix is strictly lower bounded for all parameters $\theta \in \mathbb{R}^m$. With such additional conditions, we are able to provide an $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity result of NPG in Appendix G.

References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In *Advances in Neural Information Processing Systems*, volume 26, 2013.

J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15: 319–350, Nov 2001. ISSN 1076-9757. doi: 10.1613/jair.806.

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

Amir Beck. *First-Order Methods in Optimization*. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2017. ISBN 1611974984.

Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite mdps. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2386–2394. PMLR, 13–15 Apr 2021.

Semih Cayci, Niao He, and R. Srikant. Linear convergence of entropy-regularized natural policy gradient with linear function approximation, 2021.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. In *Operations Research*, 2021.

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

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

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR, 06–11 Aug 2017.

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

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

Sham M Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

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

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

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

Jakub Grudzien Kuba, Christian Schroeder de Witt, and Jakob Foerster. Mirror learning: A unifying framework of policy optimisation, 2022.

Guanghui Lan. Policy mirror descent for reinforcement learning: linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, Apr 2022. ISSN 1436-4646.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *Advances in Neural Information Processing Systems*, volume 33, pages 7624–7636. Curran Associates, Inc., 2020.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6820–6829. PMLR, 13–18 Jul 2020.

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

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR.

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

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

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

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, Inc., USA, 1994. ISBN 0471619779.

R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

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

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

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014. ISBN 978-1-10-705713-5.

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

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* The MIT Press, second edition, 2018.

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

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. In *International Conference on Learning Representations*, 2022.

Sharan Vaswani, Olivier Bachem, Simone Totaro, Robert Müller, Shivam Garg, Matthieu Geist, Marlos C. Machado, Pablo Samuel Castro, and Nicolas Le Roux. A general class of surrogate functions for stable and efficient reinforcement learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 8619–8649. PMLR, 28–30 Mar 2022.

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

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

Lin Xiao. On the convergence rates of policy gradient methods, 2022.

Lin Yang and Mengdi Wang. Sample-Optimal Parametric Q-Learning Using Linearly Additive Features. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

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

Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D. Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence, 2021.

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

Appendix A. Related work

NPG for the softmax tabular policies. For the softmax tabular policies, Shani et al. (2020) show that the unregularized NPG has a $\mathcal{O}(1/\sqrt{k})$ convergence rate and the regularized NPG has a faster $\mathcal{O}(1/k)$ convergence rate by using a decaying step size. Agarwal et al. (2021) improve the convergence rate to $\mathcal{O}(1/k)$ with constant step sizes. By using the entropy regularization, Cen et al. (2021) achieve linear convergence rate of NPG. Similar linear convergence result is obtained by rewriting the NPG update under the PMD framework with the Kullback–Leibler (KL) divergence (Lan, 2022) or with a more general convex regularizer (Zhan et al., 2021). However, adding regularization might induce bias for the solution. Recently, Bhandari and Russo (2021), Khodadadian et al. (2021, 2022) and Xiao (2022) show that regularization is unnecessary for obtaining linear convergence, and it suffices to use appropriate step sizes for NPG. In particular, Bhandari and Russo (2021) use an exact line search for the step size and Khodadadian et al. (2021, 2022) choose an adaptive step size. In this paper, with non-adaptive geometrically increasing step size, we extend the results of Xiao (2022) from the tabular setting to the log-linear policies, obtaining the linear convergence rate of NPG without regularization.

NPG with function approximation. In the function approximation regime, Wang et al. (2020) establish the $\mathcal{O}(1/\sqrt{k})$ convergence rate for two-layer neural-network parametrization with a projection step. Same convergence rate is obtained for the smooth policy with projections by Agarwal et al. (2021), which is later improved to $\mathcal{O}(1/k)$ by Liu et al. (2020) by replacing the projection step with a strong regularity condition on the Fisher information matrix. With entropy regularization and a projection step, Cayci et al. (2021) obtains linear convergence for log-linear policies. In contrast, we show that by using a simple geometrically increasing step size, fast linear convergence can be achieved for log-linear policies without any regularization nor a projection step.

Fast linear convergence of other policy gradient methods. Different to the PMD analysis approach, by leveraging a gradient dominance property (Polyak, 1963), fast linear convergence results are also established in the PG methods under different settings, such as the linear quadratic control problems (Fazel et al., 2018) and the exact PG method with softmax tabular policy and entropy regularization (Mei et al., 2020; Yuan et al., 2022). Linear convergence of PG can also be obtained by exploiting non-uniform smoothness (Mei et al., 2021).

Appendix B. Standard Reinforcement Learning Results

In this section, we prove the standard reinforcement learning results used in our main paper, including the NPG updates written through the compatible function approximation (10) and the NPG updates formalized as policy mirror descent ((16) and (17)). Then, we also prove the performance difference lemma (Kakade and Langford, 2002) used in our proofs, which is the first key ingredient for our PMD analysis. See later the three-point descent lemma 28, the second key ingredient for our PMD analysis.

Lemma 16 (NPG updates via compatible function approximation, Theorem 1 in Kakade (2001)) *Consider the NPG updates* (9)

$$\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k F_\rho\big(\theta^{(k)}\big)^\dagger \, \nabla_\theta V_\rho\big(\theta^{(k)}\big),$$

and the updates using the compatible function approximation (10)

$$\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k w_\star^{(k)},$$

where $w_\star^{(k)} \in \mathrm{argmin}_{w\in\mathbb{R}^m} L_A\big(w, \theta^{(k)}, \bar{d}^{\,(k)}\big)$*. If the parametrized policy is differentiable for all* $\theta \in \mathbb{R}^m$*, then the two updates are equivalent up to a constant scaling* $(1-\gamma)$ *of* η_k*.*

Proof Indeed, by the policy gradient (8) and by using the fact that $\sum_{a\in\mathcal{A}}\nabla\pi_{s,a}(\theta)=0$ for all $s\in\mathcal{S}$, as $\pi(\theta)$ is differentiable on θ and $\sum_{a\in\mathcal{A}}\pi_{s,a}=1$, we have the policy gradient theorem (Sutton et al., 2000)

$$\nabla_\theta V_\rho(\theta)=\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\theta,\,a\sim\pi_s(\theta)}\left[A_{s,a}(\theta)\,\nabla_\theta\log\pi_{s,a}(\theta)\right]. \tag{40}$$

Furthermore, consider the optima $w_\star^{(k)}$. By the first-order optimality condition, we have

$$\begin{aligned}
&\nabla_w L_A(w_\star^{(k)},\theta^{(k)},\bar{d}^{(k)})=0\\
\iff\quad&\mathbb{E}_{(s,a)\sim\bar{d}^{(k)}}\left[\left((w_\star^{(k)})^\top\nabla_\theta\log\pi_{s,a}^{(k)}-A_{s,a}(\theta^{(k)})\right)\nabla_\theta\log\pi_{s,a}^{(k)}\right]=0\\
\iff\quad&\mathbb{E}_{(s,a)\sim\bar{d}^{(k)}}\left[\nabla_\theta\log\pi_{s,a}^{(k)}\left(\nabla_\theta\log\pi_{s,a}^{(k)}\right)^\top\right]w_\star^{(k)}=\mathbb{E}_{(s,a)\sim\bar{d}^{(k)}}\left[A_{s,a}(\theta^{(k)})\nabla_\theta\log\pi_{s,a}^{(k)}\right]\\
\overset{(9)+(40)}{\iff}\quad&F_\rho(\theta^{(k)})w_\star^{(k)}=(1-\gamma)\nabla_\theta V_\rho(\theta^{(k)}).
\end{aligned}$$

Thus, we have

$$w_\star^{(k)}=(1-\gamma)F_\rho(\theta)^\dagger\nabla_\theta V_\rho(\theta^{(k)})$$

which yields the update (9) up to a constant scaling $(1-\gamma)$ of η_k. ■

Lemma 17 (NPG updates as policy mirror descent) *The closed form solution to (16) is given by*

$$\begin{aligned}
\pi_s^{(k+1)}&=\pi_s^{(k)}\odot\frac{\exp\left(-\eta_k\Phi_s w^{(k)}\right)}{\sum_{a\in\mathcal{A}}\pi_{s,a}^{(k)}\exp\left(-\eta_k\phi_{s,a}^\top w^{(k)}\right)} &(41)\\
&=\pi_s^{(k)}\odot\frac{\exp\left(-\eta_k\bar{\Phi}_s^{(k)}w^{(k)}\right)}{\sum_{a\in\mathcal{A}}\pi_{s,a}^{(k)}\exp\left(-\eta_k\left(\bar{\phi}_{s,a}(\theta^{(k)})\right)^\top w^{(k)}\right)} &(42)\\
&=\arg\min_{p\in\Delta(\mathcal{A})}\left\{\eta_k\left\langle\bar{\Phi}_s^{(k)}w^{(k)},p\right\rangle+D(p,\pi_s^{(k)})\right\},\quad\forall s\in\mathcal{S}, &(43)
\end{aligned}$$

where $\odot$ is the element-wise product between vectors, and $\bar{\Phi}_s^{(k)}\in\mathbb{R}^{|\mathcal{A}|\times m}$ is defined in (17), i.e.

$$\left(\bar{\Phi}_{s,a}^{(k)}\right)^\top\overset{def}{=}\bar{\phi}_{s,a}(\theta^{(k)})\overset{(12)}{=}\phi_{s,a}-\mathbb{E}_{a'\sim\pi_s^{(k)}}\left[\phi_{s,a'}\right].$$

Such policy update coincides the approximate NPG updates (34) of the log-linear policy, if $\theta^{(k+1)}=\theta^{(k)}-\eta_k w^{(k)}$ with $w^{(k)}\approx\mathrm{argmin}_w\,L_A(w,\theta^{(k)},\tilde{d}^{(k)})$; and coincides the approximate Q-NPG updates (18) of the log-linear policy, if $\theta^{(k+1)}=\theta^{(k)}-\eta_k w^{(k)}$ with $w^{(k)}\approx\mathrm{argmin}_w\,L_Q(w,\theta^{(k)},\tilde{d}^{(k)})$.

Proof For shorthand, let $g=\Phi_s w^{(k)}$. Thus, (16) fits the format of Lemma 27 where $q=\pi_s^{(k)}$. Consequently, the closed form solution is given by (104), that is

$$\begin{aligned}
\pi_s^{(k+1)}&=\frac{\pi_s^{(k)}\odot e^{-\eta_k g}}{\sum_{a\in\mathcal{A}}\pi_{s,a}^{(k)}e^{-\eta_k g_a}}=\frac{\pi_s^{(k)}\odot e^{-\eta_k\Phi_s w^{(k)}}}{\sum_{a\in\mathcal{A}}\pi_{s,a}^{(k)}e^{-\eta_k\phi_{s,a}^\top w^{(k)}}}\\
&=\pi_s^{(k)}\odot\frac{\exp\left(-\eta_k\bar{\Phi}_s(\theta^{(k)})w^{(k)}\right)}{\sum_{a\in\mathcal{A}}\pi_{s,a}^{(k)}\exp\left(-\eta_k\left(\bar{\phi}_{s,a}(\theta^{(k)})\right)^\top w^{(k)}\right)}, &(44)
\end{aligned}$$

where the last equality is obtained as

$$\bar{\phi}_{s,a}(\theta^{(k)})=\phi_{s,a}-\mathbb{E}_{a'\sim\pi_s^{(k)}}\left[\phi_{s,a'}\right]=\phi_{s,a}-c_s,$$

with $c_s \in \mathbb{R}$ some constant independent to a.

Similarly, by applying Lemma 27 with $g = \bar{\Phi}_s^{(k)} w^{(k)}$, the closed form solution to (43) is (44).

As for the closed form updates of the policy for NPG (34) and Q-NPG (18) with the parameter updates $\theta^{(k+1)} = \theta^{(k)} - \eta_k w^{(k)}$, it is straightforward to verify that it coincides (41) and (42) given the specific structure of the log-linear policy (7), which concludes the proof. ■

Lemma 18 (Performance difference lemma (Kakade and Langford, 2002)) *For any policy $\pi, \pi' \in \Delta(\mathcal{A})^{\mathcal{S}}$ and $\rho \in \Delta(\mathcal{S})$,*

$$V_\rho(\pi) - V_\rho(\pi') = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim \bar{d}^\pi}\left[A_{s,a}(\pi')\right] \tag{45}$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi}\left[\langle Q_s(\pi'), \pi_s - \pi'_s\rangle\right], \tag{46}$$

where $Q_s(\pi)$ is the shorthand for $[Q_{s,a}(\pi)]_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ for any policy π.

Proof From Lemma 2 in Agarwal et al. (2021), we have

$$V_\rho(\pi) - V_\rho(\pi') = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim \bar{d}^\pi}\left[A_{s,a}(\pi')\right] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi}\left[\langle A_s(\pi'), \pi_s\rangle\right],$$

where $A_s(\pi)$ is the shorthand for $[A_{s,a}(\pi)]_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ for any policy π. To show (46), it suffices to show

$$\langle A_s(\pi'), \pi_s\rangle = \langle Q_s(\pi'), \pi_s - \pi'_s\rangle, \qquad \text{for all } s \in \mathcal{S} \text{ and } \pi, \pi' \in \Delta(\mathcal{A})^{\mathcal{S}}.$$

Denote $\mathbf{1}_n$ as a vector in $\mathbb{R}^n$ with coordinates equal to 1 element-wisely. Indeed, we have

$$\begin{aligned}
\langle A_s(\pi'), \pi_s\rangle &\overset{(3)}{=} \left\langle Q_s(\pi') - V_s(\pi')\cdot \mathbf{1}_{|\mathcal{A}|}, \pi_s\right\rangle \\
&= \langle Q_s(\pi'), \pi_s\rangle - \left\langle V_s(\pi')\cdot \mathbf{1}_{|\mathcal{A}|}, \pi_s\right\rangle \\
&= \langle Q_s(\pi'), \pi_s\rangle - V_s(\pi') \\
&\overset{(1)}{=} \langle Q_s(\pi'), \pi_s - \pi'_s\rangle,
\end{aligned}$$

from which we conclude the proof. ■

Appendix C. Algorithms

C.1 NPG and Q-NPG Algorithm

Algorithm 1 combined with the sampling procedure (Algorithm 4) and the averaged SGD procedure, called `NPG-SGD` (Algorithm 5), provide the sample-based NPG methods.

Idem, Algorithm 2 combined with the sampling procedure (Algorithm 3) and the averaged SGD procedure, called `Q-NPG-SGD` (Algorithm 6), provide the sample-based Q-NPG methods.

C.2 Sampling Procedures

In practice, we cannot compute the true minimizer $w_\star^{(k)}$ of the regression problem in either (34) or (18), since computing the expectation L_A or L_Q requires averaging over all state-action pairs $(s,a) \sim \tilde{d}^{(k)}$ and averaging over all trajectories

```
Algorithm 1: Natural policy gradient
Input: Initial state-action distribution $\nu$, policy $\pi^{(0)}$, discounted factor $\gamma \in [0, 1)$, step size $\eta_0 > 0$ for NPG update,
       step size $\alpha > 0$ for NPG-SGD update, number of iterations $T$ for NPG-SGD
for $k = 0$ to $K - 1$ do
    Compute $w^{(k)}$ of (34) by NPG-SGD, i.e., Algorithm 5 with inputs $(T, \nu, \pi^{(k)}, \gamma, \alpha)$
    Update $\theta^{(k+1)} = \theta^{(k)} - \eta_k w^{(k)}$ and $\eta_k$
Output: $\pi^{(K)}$
```

```
Algorithm 2: Q-Natural policy gradient
Input: Initial state-action distribution $\nu$, policy $\pi^{(0)}$, discounted factor $\gamma \in [0, 1)$, step size $\eta_0 > 0$ for Q-NPG
       update, step size $\alpha > 0$ for Q-NPG-SGD update, number of iterations $T$ for Q-NPG-SGD
for $k = 0$ to $K - 1$ do
    Compute $w^{(k)}$ of (18) by Q-NPG-SGD, i.e., Algorithm 6 with inputs $(T, \nu, \pi^{(k)}, \gamma, \alpha)$
    Update $\theta^{(k+1)} = \theta^{(k)} - \eta_k w^{(k)}$ and $\eta_k$
Output: $\pi_{\theta^{(K)}}$
```

$(s_0, a_0, c_0, s_1, \cdots)$ to compute the values of $Q^{(k)}_{s,a}$ and $A^{(k)}_{s,a}$. So instead, we provide a sampler which is able to obtain unbiased estimates of $Q_{s,a}(\theta)$ (or $A_{s,a}(\theta)$) with $(s, a) \sim \tilde{d}^{\theta}(\nu)$ for any $\pi(\theta)$.

To solve (18), we sample $(s, a) \sim \tilde{d}^{(k)}$ and $\widehat{Q}^{(k)}_{s,a}$ by a standard rollout, formalized in Algorithm 3. This sampling procedure is commonly used, for example in Agarwal et al. (2021, Algorithm 1).

```
Algorithm 3: Sampler for: $(s, a) \sim \tilde{d}^{\theta}(\nu)$ and unbiased estimate $\widehat{Q}_{s,a}(\theta)$ of $Q_{s,a}(\theta)$
Input: Initial state-action distribution $\nu$, policy $\pi(\theta)$, discounted factor $\gamma \in [0, 1)$
Initialize $(s_0, a_0) \sim \nu$, the time step $h, t = 0$, the variable $X = 1$
while $X = 1$ do
    With probability $\gamma$:
        Sample $s_{h+1} \sim \mathcal{P}(\cdot \mid s_h, a_h)$
        Sample $a_{h+1} \sim \pi_{s_{h+1}}(\theta)$
        $h \leftarrow h + 1$
    Otherwise with probability $(1 - \gamma)$:
        $X = 0$                                                  ▷ Accept $(s_h, a_h)$
$X = 1$
Set the estimate $\widehat{Q}_{s_h,a_h}(\theta) = c(s_h, a_h)$                ▷ Start to estimate $\widehat{Q}_{s_h,a_h}(\theta)$
$t = h$
while $X = 1$ do
    With probability $\gamma$:
        Sample $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)$
        Sample $a_{t+1} \sim \pi_{s_{t+1}}(\theta)$
        $\widehat{Q}_{s_h,a_h}(\theta) \leftarrow \widehat{Q}_{s_h,a_h}(\theta) + c(s_{t+1}, a_{t+1})$
        $t \leftarrow t + 1$
    Otherwise with probability $(1 - \gamma)$:
        $X = 0$                                                  ▷ Accept $\widehat{Q}_{s_h,a_h}(\theta)$
Output: $(s_h, a_h)$ and $\widehat{Q}_{s_h,a_h}(\theta)$
```

It is straightforward to verify that (s_h, a_h) and $\widehat{Q}_{s_h,a_h}(\theta)$ obtained in Algorithm 3 are unbiased for any $\pi(\theta)$. The expected length of the trajectory is $\frac{1}{1-\gamma}$. We provide its proof here for the completeness.

Lemma 19 *Consider the output (s_h, a_h) and $\widehat{Q}_{s_h,a_h}(\theta)$ of Algorithm 3. It follows that*

$$\begin{aligned}
&\mathbb{E}\left[h+1\right] = \frac{1}{1-\gamma},\\
&\Pr(s_h = s, a_h = a) = \tilde{d}^{\theta}_{s,a}(\nu),\\
&\mathbb{E}\left[\widehat{Q}_{s_h,a_h}(\theta) \ \mid \ s_h, a_h\right] = Q_{s_h,a_h}(\theta).
\end{aligned}$$

Proof The expected length $(h+1)$ of sampling (s, a) is

$$\mathbb{E}\left[h+1\right] = \sum_{k=0}^{\infty} \Pr(h=k)(k+1) = (1-\gamma)\sum_{k=0}^{\infty}\gamma^k(k+1) = \frac{1}{1-\gamma}.$$

The probability of the state-action pair (s, a) being sampled by Algorithm 3 is

$$\begin{aligned}
\Pr(s_h = s, a_h = a) &= \sum_{(s_0,a_0)\in\mathcal{S}\times\mathcal{A}} \nu_{s_0,a_0}\sum_{k=0}^{\infty}\Pr(h=k)\Pr^{\pi(\theta)}(s_h = s, a_h = a \mid h = k, s_0, a_0)\\
&= \sum_{(s_0,a_0)\in\mathcal{S}\times\mathcal{A}} \nu_{s_0,a_0}(1-\gamma)\sum_{k=0}^{\infty}\gamma^k\Pr^{\pi(\theta)}(s_k = s, a_k = a \mid s_0, a_0) \overset{(5)}{=} \tilde{d}^{\theta}_{s,a}(\nu).
\end{aligned}$$

Now we verify that $\widehat{Q}_{s_h,a_h}(\theta)$ obtained from Algorithm 3 is an unbiased estimate of $Q_{s_h,a_h}(\theta)$. Indeed, from Algorithm 3, we have

$$\widehat{Q}_{s_h,a_h}(\theta) = \sum_{t=0}^{H} c(s_{t+h}, a_{t+h}), \tag{47}$$

where $(H+1)$ is the length of the horizon executed between lines 13 and 19 in Algorithm 3 for calculating $\widehat{Q}_{s_h,a_h}(\theta)$. Taking expectation, we have

$$\begin{aligned}
\mathbb{E}\left[\widehat{Q}_{s_h,a_h}(\theta) \ \mid \ s_h, a_h\right] &= \mathbb{E}\left[\sum_{t=0}^{H} c(s_t, a_t) \mid s_0 = s_h, a_0 = a_h\right]\\
&= \sum_{k=0}^{\infty}\Pr(H=k)\mathbb{E}\left[\sum_{t=0}^{H} c(s_t, a_t) \mid s_0 = s_h, a_0 = a_h, H = k\right]\\
&= \sum_{k=0}^{\infty}(1-\gamma)\gamma^k\mathbb{E}\left[\sum_{t=0}^{k} c(s_t, a_t) \mid s_0 = s_h, a_0 = a_h\right]\\
&= (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} c(s_t, a_t)\sum_{k=t}^{\infty}\gamma^k \mid s_0 = s_h, a_0 = a_h\right]\\
&= \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^k c(s_t, a_t) \mid s_0 = s_h, a_0 = a_h\right] \overset{(2)}{=} Q_{s_h,a_h}(\theta).
\end{aligned}$$

■

Similar to Algorithm 3, to solve (34), we sample $(s, a) \sim \tilde{d}^{(k)}$ by the same procedure and estimate $\widehat{A}^{(k)}_{s,a}$ with a slight modification, namely Algorithm 4 (also see Agarwal et al., 2021, Algorithm 3).

Algorithm 4: Sampler for: $(s,a) \sim \tilde{d}^{\theta}(\nu)$ and unbiased estimate $\widehat{A}_{s,a}(\theta)$ of $A_{s,a}(\theta)$

```
Input: Initial state-action distribution $\nu$, policy $\pi(\theta)$, discounted factor $\gamma \in [0,1)$
Initialize $(s_0, a_0) \sim \nu$, the time step $h, t = 0$, the variable $X = 1$
while $X = 1$ do
    With probability $\gamma$:
        Sample $s_{h+1} \sim \mathcal{P}(\cdot \mid s_h, a_h)$
        Sample $a_{h+1} \sim \pi_{s_{h+1}}(\theta)$
        $h \leftarrow h + 1$
    Otherwise with probability $(1-\gamma)$:
        $X = 0$                                              ▷ Accept $(s_h, a_h)$
$X = 1$
Set the estimate $\widehat{Q}_{s_h,a_h}(\theta) = c(s_h, a_h)$            ▷ Start to estimate $\widehat{Q}_{s_h,a_h}(\theta)$
$t = h$
while $X = 1$ do
    With probability $\gamma$:
        Sample $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)$
        Sample $a_{t+1} \sim \pi_{s_{t+1}}(\theta)$
        $\widehat{Q}_{s_h,a_h}(\theta) \leftarrow \widehat{Q}_{s_h,a_h}(\theta) + c(s_{t+1}, a_{t+1})$
        $t \leftarrow t + 1$
    Otherwise with probability $(1-\gamma)$:
        $X = 0$                                              ▷ Accept $\widehat{Q}_{s_h,a_h}(\theta)$
$X = 1$
Set the estimate $\widehat{V}_{s_h}(\theta) = 0$                          ▷ Start to estimate $\widehat{V}_{s_h}(\theta)$
$t = h$
while $X = 1$ do
    Sample $a_t \sim \pi_{s_t}(\theta)$
    $\widehat{V}_{s_h}(\theta) \leftarrow \widehat{V}_{s_h}(\theta) + c(s_t, a_t)$
    With probability $\gamma$:
        Sample $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)$
        $t \leftarrow t + 1$
    Otherwise with probability $(1-\gamma)$:
        $X = 0$                                              ▷ Accept $\widehat{V}_{s_h}(\theta)$
Output: $(s_h, a_h)$ and $\widehat{A}_{s_h,a_h}(\theta) = \widehat{Q}_{s_h,a_h}(\theta) - \widehat{V}_{s_h}(\theta)$
```

Notice that the sampling procedure for estimating $Q_{s,a}(\theta)$ in Algorithm 3 is simpler than that for estimating $A_{s,a}(\theta)$ in Algorithm 4, as Algorithm 4 requires an additional estimation of $V_s(\theta)$. As in Lemma 19, we verify in the following lemma that the output $(s_h, a_h) \sim \tilde{d}^\theta$ and $\widehat{A}_{s_h,a_h}(\theta)$ in Algorithm 4 is an unbiased estimator of $A_{s_h,a_h}(\theta)$ for all policy $\pi(\theta)$.

Lemma 20 *Consider the output (s_h, a_h) and $\widehat{A}_{s_h,a_h}(\theta)$ of Algorithm 4. It follows that*

$$\begin{aligned}
&\mathbb{E}\left[h+1\right] = \frac{1}{1-\gamma},\\
&\Pr(s_h = s, a_h = a) = \tilde{d}^\theta_{s,a}(\nu),\\
&\mathbb{E}\left[\widehat{A}_{s_h,a_h}(\theta) \ \mid \ s_h, a_h\right] = A_{s_h,a_h}(\theta).
\end{aligned}$$

Proof Since the procedure of sampling (s_h, a_h) in Algorithm 4 is identical to the one in Algorithm 3, from Lemma 19, the first two results are verified. It remains to show that $\widehat{A}_{s_h,a_h}(\theta)$ is unbiased.

The estimation of $\widehat{A}_{s_h,a_h}(\theta)$ is decomposed into the estimations of $\widehat{Q}_{s_h,a_h}(\theta)$ and $\widehat{V}_{s_h}(\theta)$. The procedure of estimating $\widehat{Q}_{s_h,a_h}(\theta)$ is also identical to the one in Algorithm 3. Thus, from Lemma 19, we have

$$\mathbb{E}\left[\widehat{Q}_{s_h,a_h}(\theta) \ \mid \ s_h, a_h\right] = Q_{s_h,a_h}(\theta).$$

By following the similar arguments of Lemma 19, one can verify that

$$\mathbb{E}\left[\widehat{V}_{s_h}(\theta) \ \mid \ s_h, a_h\right] = V_{s_h}(\theta).$$

Combine the above two equalities and obtain that

$$\mathbb{E}\left[\widehat{A}_{s_h,a_h}(\theta) \ \mid \ s_h, a_h\right] = \mathbb{E}\left[\widehat{Q}_{s_h,a_h}(\theta) - \widehat{V}_{s_h}(\theta) \ \mid \ s_h, a_h\right] = Q_{s_h,a_h}(\theta) - V_{s_h}(\theta) \overset{(3)}{=} A_{s_h,a_h}(\theta).$$

■

C.3 SGD Procedures for Solving the Regression Problems of NPG and Q-NPG

Once we obtain the sampled (s, a) and $\widehat{A}_{s,a}(\theta^{(k)})$ from Algorithm 4, we can apply the averaged SGD algorithm as in Bach and Moulines (2013) to solve the regression problem (34) of NPG for every iteration k.

Here we suppress the superscript (k). For any parameter $\theta \in \mathbb{R}^m$, recall the compatible function approximation L_A in (34)

$$L_A(w, \theta, \tilde{d}^\theta) = \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\left(w^\top \bar{\phi}_{s,a}(\theta) - A_{s,a}(\theta)\right)^2\right].$$

With the output $(s, a) \sim \tilde{d}^\theta$ and $\widehat{A}_{s,a}(\theta)$ from Algorithm 4 (here we suppress the subscript h), we compute the stochastic gradient estimator of the function L_A in (34) by

$$\widehat{\nabla}_w L_A(w, \theta, \tilde{d}^\theta) \overset{\text{def}}{=} 2\left(w^\top \bar{\phi}_{s,a}(\theta) - \widehat{A}_{s,a}(\theta)\right)\bar{\phi}_{s,a}(\theta). \tag{48}$$

Next, we show that (48) is an unbiased gradient estimator of the loss function L_A

Lemma 21 *Consider the output (s, a) and $\widehat{A}_{s,a}(\theta)$ of Algorithm 4 and the stochastic gradient (48). It follows that*

$$\mathbb{E}\left[\widehat{\nabla}_w L_A(w, \theta, \tilde{d}^\theta)\right] = \nabla_w L_A(w, \theta, \tilde{d}^\theta),$$

where the expectation is with respect to the randomness in the sequence of the sampled $s_0, a_0, \cdots, s_t, a_t$ from Algorithm 4.

Proof The total expectation of the stochastic gradient is given by

$$\begin{aligned}\mathbb{E}\left[\widehat{\nabla}_w L_A(w,\theta,\tilde{d}^\theta)\right] &\overset{(48)}{=} \mathbb{E}_{s,\,a,\,\widehat{A}_{s,a}(\theta)}\left[2\left(w^\top\bar{\phi}_{s,a}(\theta)-\widehat{A}_{s,a}(\theta)\right)\bar{\phi}_{s,a}(\theta)\right] \\ &= \mathbb{E}_{(s,a)\sim\tilde{d}^\theta,\,\widehat{A}_{s,a}(\theta)}\left[2\left(w^\top\bar{\phi}_{s,a}(\theta)-\widehat{A}_{s,a}(\theta)\right)\bar{\phi}_{s,a}(\theta)\mid s,a\right],\end{aligned} \tag{49}$$

where the second line is obtained by $(s,a)\sim\tilde{d}^\theta$ from Lemma 20.

From Lemma 20, we have

$$\mathbb{E}_{s_0,a_0,\cdots,s_t,a_t}\left[\widehat{A}_{s,a}(\theta)\mid s_0=s,a_0=a\right]=A_{s,a}(\theta). \tag{50}$$

Combining the above two equalities yield

$$\begin{aligned}\mathbb{E}\left[\widehat{\nabla}_w L_A(w,\theta,\tilde{d}^\theta)\right] &\overset{(49)}{=} \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[2\left(w^\top\bar{\phi}_{s,a}(\theta)-\mathbb{E}\left[\widehat{A}_{s,a}(\theta)\mid s,a\right]\right)\bar{\phi}_{s,a}(\theta)\right] \\ &\overset{(50)}{=} \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[2\left(w^\top\bar{\phi}_{s,a}(\theta)-A_{s,a}(\theta)\right)\bar{\phi}_{s,a}(\theta)\right] \\ &= \nabla_w L_A(w,\theta,\tilde{d}^\theta).\end{aligned}$$

■

Since (48) is unbiased shown in Lemma 21, we can use it for the averaged SGD algorithm to minimize L_A, called `NPG-SGD` in Algorithm 5 (also see Agarwal et al., 2021, Algorithm 4).

Algorithm 5: NPG-SGD

```
Input: Number of iterations $T$, step size $\alpha > 0$, initialization $w_0 \in \mathbb{R}^m$, initial state-action measure $\nu$, policy $\pi(\theta)$,
       discounted factor $\gamma \in [0,1)$
for $t = 0$ to $T-1$ do
    Call Algorithm 4 with the inputs $(\nu, \pi(\theta), \gamma)$ to sample $(s,a) \sim \tilde{d}^\theta$ and $\widehat{A}_{s,a}(\theta)$
    Update $w_{t+1} = w_t - \alpha\widehat{\nabla}_w L_A(w,\theta,\tilde{d}^\theta)$ by using (48)
Output: $w_{\text{out}} = \frac{1}{T}\sum_{t=1}^{T} w_t$
```

Similar to Algorithm 5, once we obtain the sampled (s,a) and $\widehat{Q}_{s,a}(\theta)$ from Algorithm 3, we can apply the averaged SGD algorithm to solve (18) of Q-NPG.

Recall the compatible function approximation L_Q in (18)

$$L_Q(w,\theta,\tilde{d}^\theta)=\mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\left(w^\top\phi_{s,a}(\theta)-Q_{s,a}(\theta)\right)^2\right].$$

With the output $(s,a)\sim\tilde{d}^\theta$ and $\widehat{Q}_{s,a}(\theta)$ from Algorithm 3, we compute the stochastic gradient estimator of the function L_Q in (18) by

$$\widehat{\nabla}_w L_Q(w,\theta,\tilde{d}^\theta)\overset{\text{def}}{=}2\left(w^\top\phi_{s,a}(\theta)-\widehat{Q}_{s,a}(\theta)\right)\phi_{s,a}(\theta), \tag{51}$$

and use it for the averaged SGD algorithm to minimize L_Q, called `Q-NPG-SGD` in Algorithm 6 (also see Agarwal et al., 2021, Algorithm 2).

The estimator $\widehat{\nabla}_w L_Q(w,\theta,\tilde{d}^\theta)$ is also unbiased following the similar argument of the proof of Lemma 21. We formalize this in the following and omit the proof.

Lemma 22 *Consider the output (s,a) and $\widehat{Q}_{s,a}(\theta)$ of Algorithm 3 and the stochastic gradient (51). It follows that*

$$\mathbb{E}\left[\widehat{\nabla}_w L_Q(w,\theta,\tilde{d}^\theta)\right]=\nabla_w L_Q(w,\theta,\tilde{d}^\theta),$$

where the expectation is with respect to the randomness in the sequence of the sampled $s_0,a_0,\cdots,s_t,a_t$ from Algorithm 3.

Algorithm 6: Q-NPG-SGD

```
Input: Number of iterations $T$, step size $\alpha > 0$, initialization $w_0 \in \mathbb{R}^m$, initial state-action measure $\nu$, policy $\pi(\theta)$,
       discounted factor $\gamma \in [0, 1)$
for $t = 0$ to $T - 1$ do
    Call Algorithm 3 with the inputs $(\nu, \pi(\theta), \gamma)$ to sample $(s, a) \sim \tilde{d}^\theta$ and $\widehat{Q}_{s,a}(\theta)$
    Update $w_{t+1} = w_t - \alpha \widehat{\nabla}_w L_Q(w, \theta, \tilde{d}^\theta)$ by using (51)
Output: $w_{\text{out}} = \frac{1}{T}\sum_{t=1}^{T} w_t$
```

Appendix D. Proof of Section 4

Throughout this section and the next, we use the shorthand $V_\rho^{(k)}$ for $V_\rho(\theta^{(k)})$ and similarly, $Q_{s,a}^{(k)}$ for $Q_{s,a}(\theta^{(k)})$ and $A_{s,a}^{(k)}$ for $A_{s,a}(\theta^{(k)})$. We also use the shorthand $Q_s^{(k)}$ for the vector $\left[Q_{s,a}^{(k)}\right]_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ and $A_s^{(k)}$ for the vector $\left[A_{s,a}^{(k)}\right]_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$.

We first provide the one step analysis of the Q-NPG update, which will be helpful for proving Theorem 5, 6 and 9.

D.1 The One Step Q-NPG Lemma

The following one step analysis of Q-NPG is based on the mirror descent approach developed in Xiao (2022).

Lemma 23 (One step Q-NPG lemma) *Fix a state distribution ρ; an initial state-action distribution ν; an arbitrary comparator policy π^*. Denote $w_\star^{(k)} \in \operatorname{argmin}_w L_Q(w, \theta^{(k)}, \tilde{d}^{(k)})$ as the exact minimizer. Consider the $w^{(k)}$ and $\pi^{(k)}$ given in (18) and (16) respectively. We have that*

$$
\begin{aligned}
&\vartheta_\rho(1-\gamma)\left(V_\rho^{(k+1)} - V_\rho^{(k)}\right) + (1-\gamma)\left(V_\rho^{(k)} - V_\rho(\pi^*)\right)\\
&+\vartheta_\rho\Bigg(\underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k+1)}\phi_{s,a}^\top\left(w^{(k)} - w_\star^{(k)}\right)}_{\textcircled{1}} + \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k+1)}\left(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right)}_{\textcircled{2}}\\
&+\underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k)}\phi_{s,a}^\top\left(w_\star^{(k)} - w^{(k)}\right)}_{\textcircled{3}} + \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k)}\left(Q_{s,a}^{(k)} - \phi_{s,a}^\top w_\star^{(k)}\right)}_{\textcircled{4}}\Bigg)\\
&+\underbrace{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^*\pi_{s,a}^{(k)}\phi_{s,a}^\top\left(w^{(k)} - w_\star^{(k)}\right)}_{\textcircled{a}} + \underbrace{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^*\pi_{s,a}^{(k)}\left(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right)}_{\textcircled{b}}\\
&+\underbrace{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^*\pi_{s,a}^*\phi_{s,a}^\top\left(w_\star^{(k)} - w^{(k)}\right)}_{\textcircled{c}} + \underbrace{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^*\pi_{s,a}^*\left(Q_{s,a}^{(k)} - \phi_{s,a}^\top w_\star^{(k)}\right)}_{\textcircled{d}}\\
\leq\ &\frac{1}{\eta_k}D_k^* - \frac{1}{\eta_k}D_{k+1}^*.
\end{aligned}
\tag{52}
$$

Proof In the context of the PMD method (16), we apply the three-point descent lemma 28 with $\mathcal{C} = \Delta(\mathcal{A})$, f is the linear function $\eta_k\left\langle\Phi_s w^{(k)}, \cdot\right\rangle$ and $h : \Delta(\mathcal{A}) \to \mathbb{R}$ is the negative entropy with $h(p) = \sum_{a\in\mathcal{A}} p_a \log p_a$. Thus, h is of

Legendre type with $\operatorname{rint}\operatorname{dom} h \cap \mathcal{C} = \operatorname{rint}\Delta(\mathcal{A}) \neq \emptyset$ and $D_h(\cdot,\cdot)$ is the KL divergence $D(\cdot,\cdot)$. From Lemma 28, we obtain that for any $p \in \Delta(\mathcal{A})$, we have

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

Rearranging terms and dividing both sides by η_k, we get

$$\left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - p \right\rangle + \frac{1}{\eta_k} D(\pi_s^{(k+1)}, \pi_s^{(k)}) \leq \frac{1}{\eta_k} D(p, \pi_s^{(k)}) - \frac{1}{\eta_k} D(p, \pi_s^{(k+1)}). \tag{53}$$

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

$$\left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \leq -\frac{1}{\eta_k} D(\pi_s^{(k+1)}, \pi_s^{(k)}) - \frac{1}{\eta_k} D(\pi_s^{(k)}, \pi_s^{(k+1)}) \leq 0. \tag{54}$$

Letting $p = \pi_s^*$ and subtract and add $\pi_s^{(k)}$ within the inner product term in (53) yields

$$\left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle + \left\langle \Phi_s w^{(k)}, \pi_s^{(k)} - \pi_s^* \right\rangle \leq \frac{1}{\eta_k} D(\pi_s^*, \pi_s^{(k)}) - \frac{1}{\eta_k} D(\pi_s^*, \pi_s^{(k+1)}).$$

Note that we dropped the nonnegative term $\frac{1}{\eta_k} D(\pi_s^{(k+1)}, \pi_s^{(k)})$ on the left hand side to the inequality.

Taking expectation with respect to the distribution d^*, we have

$$\mathbb{E}_{s\sim d^*}\left[\left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle\right] + \mathbb{E}_{s\sim d^*}\left[\left\langle \Phi_s w^{(k)}, \pi_s^{(k)} - \pi_s^* \right\rangle\right] \leq \frac{1}{\eta_k} D_k^* - \frac{1}{\eta_k} D_{k+1}^*. \tag{55}$$

For the first expectation in (55), we have

$$\begin{aligned}
& \mathbb{E}_{s\sim d^*}\left[\left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle\right] \\
= \;& \sum_{s\in\mathcal{S}} d_s^* \left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
= \;& \sum_{s\in\mathcal{S}} \frac{d_s^*}{d_s^{(k+1)}} d_s^{(k+1)} \left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
\geq \;& \vartheta_{k+1} \sum_{s\in\mathcal{S}} d_s^{(k+1)} \left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
\geq \;& \vartheta_\rho \sum_{s\in\mathcal{S}} d_s^{(k+1)} \left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
= \;& \vartheta_\rho \sum_{s\in\mathcal{S}} d_s^{(k+1)} \left\langle Q_s^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle + \vartheta_\rho \sum_{s\in\mathcal{S}} d_s^{(k+1)} \left\langle \Phi_s w^{(k)} - Q_s^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
= \;& \vartheta_\rho (1-\gamma) \left(V_\rho^{(k+1)} - V_\rho^{(k)} \right) + \vartheta_\rho \sum_{s\in\mathcal{S}} d_s^{(k+1)} \left\langle \Phi_s w^{(k)} - Q_s^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle,
\end{aligned} \tag{56}$$

where the last equality is due to the performance difference lemma (46) in Lemma 18 and the two inequalities above are obtained by the negative sign of $\left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle$ shown in (54) and by using the following inequality

$$\frac{d_s^*}{d_s^{(k+1)}} \overset{(20)}{\leq} \vartheta_{k+1} \overset{(20)}{\leq} \vartheta_\rho.$$

The second term of (56) can be decomposed into four terms. That is,

$$\begin{aligned}
&\sum_{s\in\mathcal{S}} d_s^{(k+1)} \left\langle \Phi_s w^{(k)} - Q_s^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k+1)} \left(\phi_{s,a}^\top w^{(k)} - Q_{s,a}^{(k)}\right) + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k)} \left(Q_{s,a}^{(k)} - \phi_{s,a}^\top w^{(k)}\right) \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k+1)} \phi_{s,a}^\top \left(w^{(k)} - w_\star^{(k)}\right) + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k+1)} \left(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right) \\
&\quad + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k)} \phi_{s,a}^\top \left(w_\star^{(k)} - w^{(k)}\right) + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k)} \left(Q_{s,a}^{(k)} - \phi_{s,a}^\top w_\star^{(k)}\right) \\
&= \textcircled{1} + \textcircled{2} + \textcircled{3} + \textcircled{4}, \qquad (57)
\end{aligned}$$

where ①, ②, ③ and ④ are defined in (52).

For the second expectation in (55), by applying again the performance difference lemma (46), we have

$$\begin{aligned}
&\mathbb{E}_{s\sim d^*}\left[\left\langle \Phi_s w^{(k)}, \pi_s^{(k)} - \pi_s^* \right\rangle\right] \\
= \;& \mathbb{E}_{s\sim d^*}\left[\left\langle Q_s^{(k)}, \pi_s^{(k)} - \pi_s^* \right\rangle\right] + \mathbb{E}_{s\sim d^*}\left[\left\langle \Phi_s w^{(k)} - Q_s^{(k)}, \pi_s^{(k)} - \pi_s^* \right\rangle\right] \\
\overset{(46)}{=} \;& (1-\gamma)\left(V_\rho^{(k)} - V_\rho(\pi^*)\right) + \mathbb{E}_{s\sim d^*}\left[\left\langle \Phi_s w^{(k)} - Q_s^{(k)}, \pi_s^{(k)} - \pi_s^* \right\rangle\right]. \qquad (58)
\end{aligned}$$

Similarly, we decompose the second term of (58) into four terms. That is,

$$\begin{aligned}
&\mathbb{E}_{s\sim d^*}\left[\left\langle \Phi_s w^{(k)} - Q_s^{(k)}, \pi_s^{(k)} - \pi_s^* \right\rangle\right] \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^* \pi_{s,a}^{(k)} \left(\phi_{s,a}^\top w^{(k)} - Q_{s,a}^{(k)}\right) + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^* \pi_{s,a}^* \left(Q_{s,a}^{(k)} - \phi_{s,a}^\top w^{(k)}\right) \\
&= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \pi_{s,a}^{(k)} \phi_{s,a}^\top \left(w^{(k)} - w_\star^{(k)}\right) + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \pi_{s,a}^{(k)} \left(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right) \\
&\quad + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \pi_{s,a}^* \phi_{s,a}^\top \left(w_\star^{(k)} - w^{(k)}\right) + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \pi_{s,a}^* \left(Q_{s,a}^{(k)} - \phi_{s,a}^\top w_\star^{(k)}\right) \\
&= \textcircled{a} + \textcircled{b} + \textcircled{c} + \textcircled{d}, \qquad (59)
\end{aligned}$$

where ⓐ, ⓑ, ⓒ and ⓓ are defined in (52).

Plugging (56) with the decomposition (57) and (58) with the decomposition (59) into (55) concludes the proof. ■

Consequently, the convergence analysis of Q-NPG (Theorem 5, 6 and 9) will be obtained by upper bounding the absolute values of ①, ②, ③, ④, ⓐ, ⓑ, ⓒ, ⓓ in (52) with different set of assumptions (assumptions in Theorem 5 or assumptions in Theorem 9) and with different step size scheme (geometrically increasing step size for Theorem 5 and 9 or constant step size for Theorem 6).

D.2 Proof of Theorem 5

Proof From (52) in Lemma 23, we will upper bound |①| and |③| by the statistical error assumption (19) and upper bound |②| and |④| by using the transfer error assumption (22).

Indeed, to upper bound $|①|$, by Cauchy-Schwartz's inequality, we have

$$\begin{aligned}
|①| &\leq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k+1)}\left|\phi_{s,a}^\top\left(w^{(k)}-w_\star^{(k)}\right)\right| \\
&\leq \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\left(d_s^{(k+1)}\right)^2\left(\pi_{s,a}^{(k+1)}\right)^2}{d_s^*\cdot\mathrm{Unif}_{\mathcal{A}}(a)}\cdot\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^*\cdot\mathrm{Unif}_{\mathcal{A}}(a)\left(\phi_{s,a}^\top\left(w^{(k)}-w_\star^{(k)}\right)\right)^2} \\
&\overset{(23)}{=} \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\left(d_s^{(k+1)}\right)^2\left(\pi_{s,a}^{(k+1)}\right)^2}{d_s^*\cdot\mathrm{Unif}_{\mathcal{A}}(a)}\left\|w^{(k)}-w_\star^{(k)}\right\|^2_{\Sigma_{\tilde{d}^*}}} \\
&\leq \sqrt{\mathbb{E}_{s\sim d^*}\left[\left(\frac{d_s^{(k+1)}}{d_s^*}\right)^2\right]|\mathcal{A}|\left\|w^{(k)}-w_\star^{(k)}\right\|^2_{\Sigma_{\tilde{d}^*}}} \\
&\overset{(25)}{\leq} \sqrt{C_\rho|\mathcal{A}|\left\|w^{(k)}-w_\star^{(k)}\right\|^2_{\Sigma_{\tilde{d}^*}}},
\end{aligned} \tag{60}$$

where the second inequality is obtained by Cauchy-Schwartz's inequality, and the third inequality is obtained by the following inequality

$$\sum_{a\in\mathcal{A}}\left(\pi_{s,a}^{(k+1)}\right)^2 \leq \sum_{a\in\mathcal{A}}\pi_{s,a}^{(k+1)} = 1. \tag{61}$$

Then, by using Assumption 3 with the definition of κ_ν, (60) is upper bounded by

$$\begin{aligned}
|①| &\overset{(24)}{\leq} \sqrt{C_\rho|\mathcal{A}|\kappa_\nu\left\|w^{(k)}-w_\star^{(k)}\right\|^2_{\Sigma_\nu}} \\
&\overset{(6)}{\leq} \sqrt{\frac{C_\rho|\mathcal{A}|\kappa_\nu}{1-\gamma}\left\|w^{(k)}-w_\star^{(k)}\right\|^2_{\Sigma_{\tilde{d}^{(k)}}}},
\end{aligned} \tag{62}$$

where we use the shorthand

$$\Sigma_{\tilde{d}^{(k)}} \overset{\text{def}}{=} \mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\phi_{s,a}\phi_{s,a}^\top\right]. \tag{63}$$

Besides, by the first-order optimality conditions for the optima $w_\star^{(k)} \in \mathrm{argmin}_w L_Q(w,\theta^{(k)},\tilde{d}^{(k)})$, we have

$$(w-w_\star^{(k)})^\top\nabla_w L_Q(w_\star^{(k)},\theta^{(k)},\tilde{d}^{(k)}) \geq 0, \qquad \text{for all } w\in\mathbb{R}^m. \tag{64}$$

Therefore, for all $w\in\mathbb{R}^m$,

$$\begin{aligned}
& L_Q(w,\theta^{(k)},\tilde{d}^{(k)}) - L_Q(w_\star^{(k)},\theta^{(k)},\tilde{d}^{(k)}) \\
=\ & \mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\phi_{s,a}^\top w-\phi_{s,a}^\top w_\star^{(k)}+\phi_{s,a}^\top w_\star^{(k)}-Q_{s,a}^{(k)}\right)^2\right] - L_Q(w_\star^{(k)},\theta^{(k)},\tilde{d}^{(k)}) \\
=\ & \mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[(\phi_{s,a}^\top w-\phi_{s,a}^\top w_\star^{(k)})^2\right] + 2(w-w_\star^{(k)})^\top\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[(\phi_{s,a}^\top w_\star^{(k)}-Q_{s,a}^{(k)})\phi_{s,a}\right] \\
=\ & \left\|w-w_\star^{(k)}\right\|^2_{\Sigma_{\tilde{d}^{(k)}}} + (w-w_\star^{(k)})^\top\nabla_w L_Q(w_\star^{(k)},\theta^{(k)},\tilde{d}^{(k)}) \\
\overset{(64)}{\geq}\ & \left\|w-w_\star^{(k)}\right\|^2_{\Sigma_{\tilde{d}^{(k)}}}.
\end{aligned} \tag{65}$$

Define

$$\epsilon_{\text{stat}}^{(k)} \overset{\text{def}}{=} L_Q(w^{(k)},\theta^{(k)},\tilde{d}^{(k)}) - L_Q(w_\star^{(k)},\theta^{(k)},\tilde{d}^{(k)}).$$

Note that from (19), we have

$$\mathbb{E}\left[\epsilon_{\text{stat}}^{(k)}\right] \leq \epsilon_{\text{stat}}. \tag{66}$$

Plugging (65) into (62), we have

$$|①| \quad \leq \quad \sqrt{\frac{C_\rho |\mathcal{A}| \kappa_\nu}{1-\gamma} \epsilon_{\text{stat}}^{(k)}}. \tag{67}$$

Similar to (60), we get the same upper bound for $|③|$ by just replacing $\pi_{s,a}^{(k+1)}$ into $\pi_{s,a}^{(k)}$. That is,

$$|③| \leq \sqrt{\frac{C_\rho |\mathcal{A}| \kappa_\nu}{1-\gamma} \epsilon_{\text{stat}}^{(k)}}. \tag{68}$$

To upper bound $|②|$ and $|④|$, we introduce the following term

$$\epsilon_{\text{bias}}^{(k)} \overset{\text{def}}{=} L_Q(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^*).$$

Note that from (22), we have

$$\mathbb{E}\left[\epsilon_{\text{bias}}^{(k)}\right] \leq \epsilon_{\text{bias}}. \tag{69}$$

By Cauchy-Schwartz's inequality, we have

$$\begin{aligned}
|②| \quad &\leq \quad \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k+1)} \left| \phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)} \right| \\
&\leq \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(d_s^{(k+1)}\right)^2 \left(\pi_{s,a}^{(k+1)}\right)^2}{d_s^* \cdot \text{Unif}_{\mathcal{A}}(a)} \cdot \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \cdot \text{Unif}_{\mathcal{A}}(a) \left(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right)^2} \\
&= \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(d_s^{(k+1)}\right)^2 \left(\pi_{s,a}^{(k+1)}\right)^2}{d_s^* \cdot \text{Unif}_{\mathcal{A}}(a)} \cdot \epsilon_{\text{bias}}^{(k)}} \\
&\overset{(61)}{\leq} \sqrt{\mathbb{E}_{s\sim d^*}\left[\left(\frac{d_s^{(k+1)}}{d_s^*}\right)^2\right] |\mathcal{A}| \epsilon_{\text{bias}}^{(k)}} \quad \overset{(25)}{\leq} \quad \sqrt{C_\rho |\mathcal{A}| \epsilon_{\text{bias}}^{(k)}}.
\end{aligned} \tag{70}$$

Similar to (70), we get the same upper bound for $|④|$ by just replacing $\pi_{s,a}^{(k+1)}$ into $\pi_{s,a}^{(k)}$. That is,

$$|④| \leq \sqrt{C_\rho |\mathcal{A}| \epsilon_{\text{bias}}^{(k)}}. \tag{71}$$

Next, we will upper bound the absolute values of ⓐ, ⓑ, ⓒ and ⓓ of (52) separately by using again the statistical error (19) and by using the transfer error assumption (22).

Indeed, to upper bound $|\textcircled{a}|$, by Cauchy-Schwartz's inequality, we have

$$\begin{aligned}
|\textcircled{a}| &\le \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^*\pi_{s,a}^{(k)}\left|\phi_{s,a}^\top\left(w^{(k)}-w_\star^{(k)}\right)\right| \\
&\le \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(d_s^*\right)^2\left(\pi_{s,a}^{(k)}\right)^2}{d_s^*\cdot\text{Unif}_{\mathcal{A}}(a)} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^*\cdot\text{Unif}_{\mathcal{A}}(a)\left(\phi_{s,a}^\top\left(w^{(k)}-w_\star^{(k)}\right)\right)^2} \\
&\overset{(23)}{=} \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(d_s^*\right)^2\left(\pi_{s,a}^{(k)}\right)^2}{d_s^*\cdot\text{Unif}_{\mathcal{A}}(a)} \left\|w^{(k)}-w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^*}}^2} \\
&\overset{(61)}{\le} \sqrt{|\mathcal{A}|\left\|w^{(k)}-w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^*}}^2}.
\end{aligned}$$

From the definition of κ_ν, we further obtain

$$\begin{aligned}
|\textcircled{a}| &\overset{(24)}{\le} \sqrt{|\mathcal{A}|\kappa_\nu\left\|w^{(k)}-w_\star^{(k)}\right\|_{\Sigma_\nu}^2} \\
&\overset{(6)}{\le} \sqrt{\frac{|\mathcal{A}|\kappa_\nu}{1-\gamma}\left\|w^{(k)}-w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^{(k)}}}^2} \\
&\overset{(65)}{\le} \sqrt{\frac{|\mathcal{A}|\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}^{(k)}}.
\end{aligned} \tag{72}$$

Similar to (72), we get the same upper bound for $|\textcircled{c}|$ by just replacing $\pi_{s,a}^{(k)}$ into $\pi_{s,a}^*$. That is,

$$|\textcircled{c}| \le \sqrt{\frac{|\mathcal{A}|\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}^{(k)}}. \tag{73}$$

To upper bound $|\textcircled{b}|$, by Cauchy-Schwartz's inequality, we have

$$\begin{aligned}
|\textcircled{b}| &\le \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^*\pi_{s,a}^{(k)}\left|\left(\phi_{s,a}^\top w_\star^{(k)}-Q_{s,a}^{(k)}\right)\right| \\
&\le \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(d_s^*\right)^2\left(\pi_{s,a}^{(k)}\right)^2}{d_s^*\cdot\text{Unif}_{\mathcal{A}}(a)} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^*\cdot\text{Unif}_{\mathcal{A}}(a)\left(\phi_{s,a}^\top w_\star^{(k)}-Q_{s,a}^{(k)}\right)^2} \\
&= \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(d_s^*\right)^2\left(\pi_{s,a}^{(k)}\right)^2}{d_s^*\cdot\text{Unif}_{\mathcal{A}}(a)}\epsilon_{\text{bias}}^{(k)}} \\
&\overset{(61)}{\le} \sqrt{|\mathcal{A}|\epsilon_{\text{bias}}^{(k)}}.
\end{aligned} \tag{74}$$

Similar to (74), we get the same upper bound for $|\textcircled{d}|$ by just replacing $\pi_{s,a}^{(k)}$ into $\pi_{s,a}^*$. That is,

$$|\textcircled{d}| \le \sqrt{|\mathcal{A}|\epsilon_{\text{bias}}^{(k)}}. \tag{75}$$

Plugging all the upper bounds (67) of $|\textcircled{1}|$, (70) of $|\textcircled{2}|$, (68) of $|\textcircled{3}|$, (71) of $|\textcircled{4}|$, (72) of $|\textcircled{a}|$, (74) of $|\textcircled{b}|$, (73) of $|\textcircled{c}|$ and (75) of $|\textcircled{d}|$ into (52) yields

$$\vartheta_\rho\left(\delta_{k+1}-\delta_k\right)+\delta_k \le \frac{D_k^*}{(1-\gamma)\eta_k} - \frac{D_{k+1}^*}{(1-\gamma)\eta_k} + \frac{2\sqrt{|\mathcal{A}|}\left(\vartheta_\rho\sqrt{C_\rho}+1\right)}{1-\gamma}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}^{(k)}}+\sqrt{\epsilon_{\text{bias}}^{(k)}}\right), \tag{76}$$

where $\delta_k \overset{\text{def}}{=} V_\rho^{(k)} - V_\rho(\pi^*)$. Dividing both sides by ϑ_ρ and rearranging terms, we get

$$\delta_{k+1} + \frac{D^*_{k+1}}{(1-\gamma)\eta_k\vartheta_\rho} \le \left(1 - \frac{1}{\vartheta_\rho}\right)\left(\delta_k + \frac{D^*_k}{(1-\gamma)\eta_k(\vartheta_\rho - 1)}\right) + \frac{2\sqrt{|\mathcal{A}|}\left(\sqrt{C_\rho} + \frac{1}{\vartheta_\rho}\right)}{1-\gamma}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}^{(k)}} + \sqrt{\epsilon_{\text{bias}}^{(k)}}\right).$$

If the step sizes satisfy $\eta_{k+1}(\vartheta_\rho - 1) \ge \eta_k\vartheta_\rho$, which is implied by $\eta_{k+1} \ge \eta_k/\gamma$ and (20), then

$$\begin{aligned}\delta_{k+1} + \frac{D^*_{k+1}}{(1-\gamma)\eta_{k+1}(\vartheta_\rho - 1)} &\le \left(1 - \frac{1}{\vartheta_\rho}\right)\left(\delta_k + \frac{D^*_k}{(1-\gamma)\eta_k(\vartheta_\rho - 1)}\right) \\ &\quad + \frac{2\sqrt{|\mathcal{A}|}\left(\sqrt{C_\rho} + \frac{1}{\vartheta_\rho}\right)}{1-\gamma}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}^{(k)}} + \sqrt{\epsilon_{\text{bias}}^{(k)}}\right) \\ &\le \left(1 - \frac{1}{\vartheta_\rho}\right)^{k+1}\left(\delta_0 + \frac{D^*_0}{(1-\gamma)\eta_0(\vartheta_\rho - 1)}\right) \\ &\quad + \sum_{t=0}^{k}\left(1 - \frac{1}{\vartheta_\rho}\right)^{k-t}\frac{2\sqrt{|\mathcal{A}|}\left(\sqrt{C_\rho} + \frac{1}{\vartheta_\rho}\right)}{1-\gamma}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}^{(t)}} + \sqrt{\epsilon_{\text{bias}}^{(t)}}\right).\end{aligned}$$

Finally, by choosing $\eta_0 \ge \frac{1-\gamma}{\gamma}D^*_0$ and using the fact that

$$(1-\gamma)(\vartheta_\rho - 1) \overset{(20)}{\ge} (1-\gamma)\left(\frac{1}{1-\gamma} - 1\right) = \gamma,$$

we obtain

$$\begin{aligned}\delta_k \le \delta_k + \frac{D^*_k}{(1-\gamma)\eta_k\vartheta_\rho} &\le \left(1 - \frac{1}{\vartheta_\rho}\right)^k\frac{2}{1-\gamma} \\ &\quad + \frac{2\sqrt{|\mathcal{A}|}\left(\sqrt{C_\rho} + \frac{1}{\vartheta_\rho}\right)}{1-\gamma}\sum_{t=0}^{k-1}\left(1 - \frac{1}{\vartheta_\rho}\right)^{k-1-t}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}^{(t)}} + \sqrt{\epsilon_{\text{bias}}^{(t)}}\right).\end{aligned}$$

Taking the total expectation with respect to the randomness in the sequence of the iterates $w^{(0)}, \cdots, w^{(k-1)}$, we have

$$\begin{aligned}&\mathbb{E}\left[V_\rho(\theta^{(k)})\right] - V_\rho(\pi^*) \\ \le\quad &\left(1 - \frac{1}{\vartheta_\rho}\right)^k\frac{2}{1-\gamma} \\ &+ \frac{2\sqrt{|\mathcal{A}|}\left(\sqrt{C_\rho} + \frac{1}{\vartheta_\rho}\right)}{1-\gamma}\sum_{t=0}^{k-1}\left(1 - \frac{1}{\vartheta_\rho}\right)^{k-1-t}\left(\mathbb{E}\left[\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}^{(t)}}\right] + \mathbb{E}\left[\sqrt{\epsilon_{\text{bias}}^{(t)}}\right]\right) \\ \le\quad &\left(1 - \frac{1}{\vartheta_\rho}\right)^k\frac{2}{1-\gamma} \\ &+ \frac{2\sqrt{|\mathcal{A}|}\left(\sqrt{C_\rho} + \frac{1}{\vartheta_\rho}\right)}{1-\gamma}\sum_{t=0}^{k-1}\left(1 - \frac{1}{\vartheta_\rho}\right)^{k-1-t}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\mathbb{E}\left[\epsilon_{\text{stat}}^{(t)}\right]} + \sqrt{\mathbb{E}\left[\epsilon_{\text{bias}}^{(t)}\right]}\right) \\ \overset{(66)+(69)}{\le}\quad &\left(1 - \frac{1}{\vartheta_\rho}\right)^k\frac{2}{1-\gamma} \\ &+ \frac{2\sqrt{|\mathcal{A}|}\left(\sqrt{C_\rho} + \frac{1}{\vartheta_\rho}\right)}{1-\gamma}\sum_{t=0}^{k-1}\left(1 - \frac{1}{\vartheta_\rho}\right)^{k-1-t}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{bias}}}\right) \\ \le\quad &\left(1 - \frac{1}{\vartheta_\rho}\right)^k\frac{2}{1-\gamma} + \frac{2\sqrt{|\mathcal{A}|}\left(\vartheta_\rho\sqrt{C_\rho} + 1\right)}{1-\gamma}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{bias}}}\right),\end{aligned}$$

where the second inequality is obtained by Jensen's inequality, which concludes the proof. ■

D.3 Proof of Theorem 6

Proof By (76) and using a constant step size η, we have

$$\vartheta_\rho\left(\delta_{k+1}-\delta_k\right)+\delta_k \le \frac{D_k^*}{(1-\gamma)\eta} - \frac{D_{k+1}^*}{(1-\gamma)\eta} + \frac{2\sqrt{|\mathcal{A}|}\left(\vartheta_\rho\sqrt{C_\rho}+1\right)}{1-\gamma}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}^{(k)}}+\sqrt{\epsilon_{\text{bias}}^{(k)}}\right).$$

Taking the total expectation with respect to the randomness in the sequence of the iterates $w^{(0)},\cdots,w^{(k-1)}$, summing up from 0 to $k-1$ and rearranging terms, we have

$$\vartheta_\rho\mathbb{E}\left[\delta_k\right]+\sum_{t=0}^{k-1}\mathbb{E}\left[\delta_t\right] \le \frac{D_0^*}{(1-\gamma)\eta}+\vartheta_\rho\delta_0+k\cdot\frac{2\sqrt{|\mathcal{A}|}\left(\vartheta_\rho\sqrt{C_\rho}+1\right)}{1-\gamma}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}}+\sqrt{\epsilon_{\text{bias}}}\right),$$

where we use the following inequalities

$$\mathbb{E}\left[\sqrt{\epsilon_{\text{stat}}^{(t)}}\right] \le \sqrt{\mathbb{E}\left[\epsilon_{\text{stat}}^{(t)}\right]} \overset{(66)}{\le} \sqrt{\epsilon_{\text{stat}}},$$

$$\mathbb{E}\left[\sqrt{\epsilon_{\text{bias}}^{(t)}}\right] \le \sqrt{\mathbb{E}\left[\epsilon_{\text{bias}}^{(t)}\right]} \overset{(69)}{\le} \sqrt{\epsilon_{\text{bias}}}.$$

Finally, dropping the positive term $\mathbb{E}\left[\delta_k\right]$ on the left hand side as π^* is the optimal policy and dividing both side by k yields

$$\begin{aligned}\frac{1}{k}\sum_{t=0}^{k-1}\mathbb{E}\left[V_\rho(\theta^{(t)})\right]-V_\rho(\pi^*) \le{}& \frac{D_0^*}{(1-\gamma)\eta k}+\frac{2\vartheta_\rho}{(1-\gamma)k}\\ &+\frac{2\sqrt{|\mathcal{A}|}\left(\vartheta_\rho\sqrt{C_\rho}+1\right)}{1-\gamma}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}}+\sqrt{\epsilon_{\text{bias}}}\right).\end{aligned}$$

With the constant step size $\eta \ge \frac{D_0^*}{2\vartheta_\rho}$, we have

$$\frac{1}{k}\sum_{t=0}^{k-1}\mathbb{E}\left[V_\rho(\theta^{(t)})\right]-V_\rho(\pi^*) \le \frac{4\vartheta_\rho}{(1-\gamma)k}+\frac{2\sqrt{|\mathcal{A}|}\left(\vartheta_\rho\sqrt{C_\rho}+1\right)}{1-\gamma}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}}+\sqrt{\epsilon_{\text{bias}}}\right).$$

■

D.4 Proof of Theorem 9

Proof Similar to the proof of Theorem 5, by Lemma 23, we upper bound the absolute values of ①, ②, ③, ④, ⓐ, ⓑ, ⓒ, ⓓ introduced in (52), separately, with the set of assumptions in Theorem 9.

In comparison with the proof of Theorem 5, we will also upper bound |①|, |③|, |ⓐ| and |ⓒ| by the statistical error assumption (19) as in the proof of Theorem 5. However, we will upper bound |②|, |④|, |ⓑ| and |ⓓ| by using the approximation error assumption (29) instead of the transfer error assumption (22).

To upper bound |①|, by Cauchy-Schwartz's inequality, we get

$$\begin{aligned}
|\textcircled{1}| &\leq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k+1)}\left|\phi_{s,a}^\top\left(w^{(k)}-w_\star^{(k)}\right)\right| \\
&\leq \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\left(d_s^{(k+1)}\right)^2\left(\pi_{s,a}^{(k+1)}\right)^2}{\tilde{d}_{s,a}^{(k)}}\cdot\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\tilde{d}_{s,a}^{(k)}\left(\phi_{s,a}^\top\left(w^{(k)}-w_\star^{(k)}\right)\right)^2} \\
&\overset{(63)}{=} \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{d_s^{(k+1)}\pi_{s,a}^{(k+1)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right]\left\|w^{(k)}-w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^{(k)}}}^2} \\
&\overset{(30)}{\leq} \sqrt{C_\nu\left\|w^{(k)}-w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^{(k)}}}^2} \\
&\overset{(65)}{\leq} \sqrt{C_\nu\epsilon_{\text{stat}}^{(k)}}.
\end{aligned}$$

Similar to |①|, by using Assumption 8 and Cauchy-Schwartz's inequality, and by simply replacing $\pi^{(k+1)}$ into $\pi^{(k)}$ or π^* and replacing $d^{(k+1)}$ into d^*, we obtain the same upper bound of |③|, |ⓐ| and |ⓒ|, that is

$$|\textcircled{3}|, |\textcircled{a}|, |\textcircled{c}| \quad\leq\quad \sqrt{C_\nu\epsilon_{\text{stat}}^{(k)}}.$$

Next, we define

$$\epsilon_{\text{approx}}^{(k)} \overset{\text{def}}{=} L_Q(w_\star^{(k)},\theta^{(k)},\tilde{d}^{(k)})$$

By Assumption 7, we know that

$$\mathbb{E}\left[\epsilon_{\text{approx}}^{(k)}\right] \leq \epsilon_{\text{approx}}.$$

To upper bound |②|, by Cauchy-Schwartz's inequality, we have

$$\begin{aligned}
|\textcircled{2}| &\leq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k+1)}\left|\phi_{s,a}^\top w_\star^{(k)}-Q_{s,a}^{(k)}\right| \\
&\leq \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\left(d_s^{(k+1)}\right)^2\left(\pi_{s,a}^{(k+1)}\right)^2}{\tilde{d}_{s,a}^{(k)}}\cdot\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\tilde{d}_{s,a}^{(k)}\left(\phi_{s,a}^\top w_\star^{(k)}-Q_{s,a}^{(k)}\right)^2} \\
&= \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{d_s^{(k+1)}\pi_{s,a}^{(k+1)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right]\cdot\epsilon_{\text{approx}}^{(k)}} \\
&\overset{(30)}{\leq} \sqrt{C_\nu\epsilon_{\text{approx}}^{(k)}}.
\end{aligned}$$

Similar to |②|, by using Assumption 7 and Cauchy-Schwartz's inequality, and by simply replacing $\pi^{(k+1)}$ into $\pi^{(k)}$ or π^* and replacing $d^{(k+1)}$ into d^*, we obtain the same upper bound for |④|, |ⓑ| and |ⓓ|, that is

$$|\textcircled{4}|, |\textcircled{b}|, |\textcircled{d}| \quad\leq\quad \sqrt{C_\nu\epsilon_{\text{approx}}^{(k)}}.$$

Consequently, plugging all these upper bounds into (52) leads to the following recurrent inequality

$$\vartheta_\rho\left(\delta_{k+1}-\delta_k\right)+\delta_k \leq \frac{D_k^*}{(1-\gamma)\eta_k}-\frac{D_{k+1}^*}{(1-\gamma)\eta_k}+\frac{2\sqrt{C_\nu}\left(\vartheta_\rho+1\right)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}^{(k)}}+\sqrt{\epsilon_{\text{approx}}^{(k)}}\right).$$

By using the same increasing step size as in Theorem 5 and following the same arguments in the proof of Theorem 5 after (76), we obtain the final performance bound with the linear convergence rate

$$\mathbb{E}\left[V_\rho(\theta^{(k)})\right] - V_\rho(\pi^*) \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^k \frac{2}{1-\gamma} + \frac{2\sqrt{C_\nu}\,(\vartheta_\rho + 1)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{approx}}}\right).$$

■

Appendix E. Sample complexity of Q-NPG

Here we establish the sample complexity results (i.e., total number of samples of single-step interaction with the environment) of a sample-based Q-NPG Algorithm 2 in Appendix C. Combined with a regression solver, `Q-NPG-SGD` in Algorithm 6, the following corollary shows that Algorithm 2 converges globally by further assuming that the feature map is bounded and has non-singular covariance matrix.

Corollary 24 *Consider the setting of Theorem 9. Suppose that the sample-based Q-NPG Algorithm 2 is run for K iterations, with T gradient steps of `Q-NPG-SGD` (Algorithm 6) per iteration. Furthermore, suppose that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $\|\phi_{s,a}\| \leq B$ with $B > 0$, and we choose the step size $\alpha = \frac{1}{2B^2}$ and the initialization $w_0 = 0$ for `Q-NPG-SGD`. If for all $\theta \in \mathbb{R}^m$, the covariance matrix of the feature map induced by the policy $\pi(\theta)$ and the initial state-action distribution ν satisfies*

$$\mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\phi_{s,a}\phi_{s,a}^\top\right] \;\geq\; \mu \mathbf{I}_m, \tag{77}$$

where $\mathbf{I}_m \in \mathbb{R}^{m\times m}$ is the identity matrix and $\mu > 0$, then

$$\begin{aligned}\mathbb{E}\left[V_\rho(\theta^{(K)})\right] - V_\rho(\pi^*) \leq &\left(1 - \frac{1}{\vartheta_\rho}\right)^K \frac{2}{1-\gamma} + \frac{2\,(\vartheta_\rho + 1)\sqrt{C_\nu \epsilon_{\text{approx}}}}{1-\gamma} \\ &+ \frac{4\sqrt{C_\nu}\,(\vartheta_\rho+1)}{(1-\gamma)^2\sqrt{T}}\left(\frac{B^2}{\mu}\left(\sqrt{2m}+1\right) + \sqrt{2m}\right).\end{aligned} \tag{78}$$

In `Q-NPG-SGD`, each trajectory has the expected length $1/(1-\gamma)$. Consequently, with $K = \mathcal{O}(\log(1/\epsilon)\log(1/(1-\gamma)))$ and $T = \mathcal{O}\big(\frac{1}{(1-\gamma)^4\epsilon^2}\big)$, Q-NPG requires $K * T/(1-\gamma) = \tilde{\mathcal{O}}\big(\frac{1}{(1-\gamma)^5\epsilon^2}\big)$ samples such that $\mathbb{E}\left[V_\rho(\theta^{(K)})\right] - V_\rho(\pi^*) \leq \mathcal{O}(\epsilon) + \mathcal{O}\left(\frac{\sqrt{\epsilon_{\text{approx}}}}{1-\gamma}\right)$.

Compared to Agarwal et al. (2021, Corollary 26) for the sampled based Q-NPG Algorithm 2, their sample complexity is $\mathcal{O}\big(\frac{1}{(1-\gamma)^{11}\epsilon^6}\big)$ with $K = \frac{1}{(1-\gamma)^2\epsilon^2}$ and $T = \frac{1}{(1-\gamma)^8\epsilon^4}$. Despite the difference on the convergence rate for K, they use the optimization results of Shalev-Shwartz and Ben-David (2014, Theorem 14.8) to obtain $\epsilon_{\text{stat}} = \mathcal{O}(1/\sqrt{T})$, while we use the one of Bach and Moulines (2013, Theorem 1) to establish $\epsilon_{\text{stat}} = \mathcal{O}(1/T)$. Thus, they consider the projected SGD and require that the feature map is bounded and the stochastic gradient is bounded[3]. To apply Bach and Moulines (2013, Theorem 1), we do not require the projection step nor the stochastic gradient bounded. Instead, we verify conditions on the covariance matrix of the stochastic gradient at the optimum (see (vi) in Theorem 29). Thus, we require that the feature map has non-singular covariance matrix (77).

Proof From Theorem 9, it remains to upper bound the statistical error $\sqrt{\epsilon_{\text{stat}}}$ produced from the `Q-NPG-SGD` procedure (Algorithm 6) for each iteration k. We suppress the superscript (k). Let w_{out} be the output of T steps `Q-NPG-SGD` with the constant step size $\frac{1}{2B^2}$ and the initialization $w_0 = 0$, and let $w_\star \in \operatorname{argmin}_w L_Q(w, \theta, \tilde{d}^\theta)$ be the exact minimizer. To upper bound ϵ_{stat} from (19), we aim to apply the standard analysis for the averaged SGD, i.e., Theorem 29. Now we verify all the assumptions in order for `Q-NPG-SGD`.

First, (i) is verified by considering the Euclidean space $\mathcal{H} = \mathbb{R}^m$.

[3] which is not correctly verified in their proof, since each single sampled trajectory has unbounded length.

The observations $\left(\phi_{s,a}\,,\ \widehat{Q}_{s,a}(\theta)\phi_{s,a}\right) \in \mathbb{R}^m \times \mathbb{R}^m$ are independent and identically distributed, sampled from Algorithm 3. Thus, (ii) is verified with $x_n = \phi_{s,a} \in \mathbb{R}^m$ and $z_n = \widehat{Q}_{s,a}(\theta)\phi_{s,a} \in \mathbb{R}^m$.

As the feature map $\|\phi_{s,a}\| \leq B$, we have $\mathbb{E}\left[\|\phi_{s,a}\|^2\right]$ finite. From (77), we know that the covariance $\mathbb{E}\left[\phi_{s,a}\phi_{s,a}^\top\right]$ is invertible. To verify (iii), it remains to verify that $\mathbb{E}\left[\left\|\widehat{Q}_{s,a}(\theta)\phi_{s,a}\right\|^2\right]$ is finite. Indeed, by using $\|\phi_{s,a}\| \leq B$, we have

$$\mathbb{E}\left[\left\|\widehat{Q}_{s,a}(\theta)\phi_{s,a}\right\|^2\right] \leq B^2\mathbb{E}\left[\widehat{Q}_{s,a}(\theta)^2\right].$$

Thus, it remains to show $\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta)\right)^2\right]$ finite for (iii). From (47), we rewrite $\widehat{Q}_{s,a}(\theta)$ as

$$\widehat{Q}_{s,a}(\theta) = \sum_{t=0}^{H} c(s_t, a_t)$$

with $(s_0, a_0) = (s, a) \sim \tilde{d}^\theta$ and H is the length of the trajectory for estimating $Q_{s,a}(\theta)$. Thus, (iii) is verified as the variance of $\widehat{Q}_{s,a}(\theta)$ is upper bounded by

$$\begin{aligned}
\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta)\right)^2\right] &= \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\sum_{k=0}^{\infty}\Pr(H=k)\mathbb{E}\left[\left(\sum_{t=0}^{k}c(s_t,a_t)\right)^2 \mid H=k, s_0=s, a_0=a\right]\right] \\
&= \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[(1-\gamma)\sum_{k=0}^{\infty}\gamma^k\mathbb{E}\left[\left(\sum_{t=0}^{k}c(s_t,a_t)\right)^2 \mid H=k, s_0=s, a_0=a\right]\right] \\
&\leq \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[(1-\gamma)\sum_{k=0}^{\infty}\gamma^k(k+1)^2\right] \leq \frac{2}{(1-\gamma)^2},
\end{aligned} \tag{79}$$

where the first inequality is obtained as $|c(s_t, a_t)| \in [0, 1]$ for all $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$.

Next, we introduce the residual

$$\xi \overset{\text{def}}{=} \left(\widehat{Q}_{s,a}(\theta) - w_\star^\top\phi_{s,a}\right)\phi_{s,a} \overset{(51)}{=} \frac{1}{2}\widehat{\nabla}_w L_Q(w_\star, \theta, \tilde{d}^\theta). \tag{80}$$

From Lemma 22, we know that

$$\mathbb{E}\left[\widehat{\nabla}_w L_Q(w_\star, \theta, \tilde{d}^\theta)\right] = \nabla_w L_Q(w_\star, \theta, \tilde{d}^\theta).$$

So, we have that

$$\mathbb{E}\left[\xi\right] = \frac{1}{2}\nabla_w L_Q(w_\star, \theta, \tilde{d}^\theta) = 0,$$

where the last equality is obtained as $w_\star$ is the exact minimizer of the loss function L_Q. Thus, (iv) is verified with that f is $\frac{1}{2}L_Q$, ξ_n is ξ and θ is w in our context.

From Q-NPG-SGD update 51, we have (v) verified with step size $\alpha/2$ in our context.

Finally, for (vi), from the boundedness of the feature map $\|\phi_{s,a}\| \leq B$, we take $R = B$ such that $\mathbb{E}\left[\|\phi_{s,a}\|^2\phi_{s,a}\phi_{s,a}^\top\right] \leq B^2\mathbb{E}\left[\phi_{s,a}\phi_{s,a}^\top\right]$. It remains to find $\sigma > 0$ such that

$$\mathbb{E}\left[\xi\xi^\top\right] \leq \sigma^2\mathbb{E}\left[\phi_{s,a}\phi_{s,a}^\top\right].$$

We rewrite the covariance of ξ as

$$\begin{aligned}
\mathbb{E}\left[\xi\xi^\top\right] &\overset{(80)}{=} \mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta) - w_\star^\top\phi_{s,a}\right)^2\phi_{s,a}\phi_{s,a}^\top\right] \\
&= \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\left(\widehat{Q}_{s,a}(\theta) - w_\star^\top\phi_{s,a}\right)^2\phi_{s,a}\phi_{s,a}^\top \mid s, a\right] \\
&= \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta) - w_\star^\top\phi_{s,a}\right)^2 \mid s, a\right]\phi_{s,a}\phi_{s,a}^\top\right].
\end{aligned}$$

Thus, it suffices to find $\sigma > 0$ such that

$$\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta) - w_\star^\top \phi_{s,a}\right)^2 \mid s,a\right] = \mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta)\right)^2 \mid s,a\right] - 2Q_{s,a}(\theta) w_\star^\top \phi_{s,a} + \left(w_\star^\top \phi_{s,a}\right)^2 \leq \sigma^2 \tag{81}$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ to verify (vi). Besides, we know that

$$\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta)\right)^2 \mid s,a\right] \overset{(79)}{\leq} \frac{2}{(1-\gamma)^2}.$$

We also know that $|Q_{s,a}(\theta)| \leq \frac{1}{1-\gamma}$ and $\|\phi_{s,a}\| \leq B$. Now we need to bound $\|w_\star\|$. Again, since $w_\star$ is the exact minimizer, we have $\nabla_w L_Q(w_\star, \theta, \tilde{d}^\theta) = 0$. That is

$$\mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\left(w_\star^\top \phi_{s,a} - Q_{s,a}(\theta)\right)\phi_{s,a}\right] = 0,$$

which implies

$$w_\star = \left(\mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\phi_{s,a}\phi_{s,a}^\top\right]\right)^\dagger \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[Q_{s,a}(\theta)\phi_{s,a}\right].$$

By the boundness of the feature map $\|\phi_{s,a}\| \leq B$ and the Q-function $|Q_{s,a}(\theta)| \leq \frac{1}{1-\gamma}$, and the condition (77), we have the minimizer $w_\star$ bounded by

$$\|w_\star\| \overset{(77)}{\leq} \frac{B}{\mu(1-\gamma)}.$$

By using the upper bounds of $\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta)\right)^2 \mid s,a\right]$, $|Q_{s,a}(\theta)|$, $\|w_\star\|$ and $\|\phi_{s,a}\|$, the left hand side of (81) can be upper bounded by

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta) - w_\star^\top \phi_{s,a}\right)^2 \mid s,a\right] &\leq \frac{2}{(1-\gamma)^2} + \frac{2B^2}{\mu(1-\gamma)^2} + \frac{B^4}{\mu^2(1-\gamma)^2} \\ &= \frac{1}{(1-\gamma)^2}\left(\left(\frac{B^2}{\mu}+1\right)^2 + 1\right) \\ &\leq \frac{2}{(1-\gamma)^2}\left(\frac{B^2}{\mu}+1\right)^2.\end{aligned}$$

Thus, we choose

$$\sigma = \frac{\sqrt{2}}{1-\gamma}\left(\frac{B^2}{\mu}+1\right).$$

Now all the conditions (i) - (vi) in Theorem 29 are verified. With step size $\alpha = \frac{1}{2B^2}$, the initialization $w_0 = 0$ and T steps of `Q-NPG-SGD` updates (51), we have

$$\begin{aligned}\mathbb{E}\left[L_Q(w_{\mathrm{out}}, \theta, \tilde{d}^\theta)\right] - L_Q(w_\star, \theta, \tilde{d}^\theta) &\leq \frac{4}{T}\left(\sigma\sqrt{m} + B\|w_\star\|\right)^2 \\ &\leq \frac{4}{T}\left(\frac{\sqrt{2m}}{1-\gamma}\left(\frac{B^2}{\mu}+1\right) + \frac{B^2}{\mu(1-\gamma)}\right)^2\end{aligned}$$

Consequently, Assumption 1 is verified by

$$\sqrt{\epsilon_{\mathrm{stat}}} \leq \frac{2}{(1-\gamma)\sqrt{T}}\left(\frac{B^2}{\mu}\left(\sqrt{2m}+1\right) + \sqrt{2m}\right).$$

The proof is completed by replacing the above upper bound of $\sqrt{\epsilon_{\mathrm{stat}}}$ in the results of Theorem 9. ■

Appendix F. Proof of Section 5

F.1 The One Step NPG Lemma

To prove Theorem 14 and 15, we start from providing the one step analysis of the NPG update.

Lemma 25 (One step NPG lemma) *Fix a state distribution ρ; an initial state-action distribution ν; an arbitrary comparator policy π^*. At the k-th iteration, denote $w_\star^{(k)} \in \operatorname{argmin}_w L_A(w, \theta^{(k)}, \tilde{d}^{(k)})$ as the exact minimizer. Consider the $w^{(k)}$ and $\pi^{(k)}$ NPG iterates given in (34) and (17) respectively. Note*

$$\epsilon_{\text{stat}}^{(k)} \overset{\text{def}}{=} L_A(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}) - L_A(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}), \tag{82}$$

$$\epsilon_{\text{approx}}^{(k)} \overset{\text{def}}{=} L_A(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}), \tag{83}$$

$$\delta_k \overset{\text{def}}{=} V_\rho^{(k)} - V_\rho(\pi^*).$$

If Assumptions 11, 12 and 13 hold for all $k \geq 0$, then we have that

$$\vartheta_\rho \left(\delta_{k+1} - \delta_k\right) + \delta_k \leq \frac{D_k^*}{(1-\gamma)\eta_k} - \frac{D_{k+1}^*}{(1-\gamma)\eta_k} + \frac{\sqrt{C_\nu}\left(\vartheta_\rho + 1\right)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}^{(k)}} + \sqrt{\epsilon_{\text{approx}}^{(k)}}\right). \tag{84}$$

Proof From the three-point descent lemma 28 and (17), we obtain that for any $p \in \Delta(\mathcal{A})$, we have

$$\eta_k \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)} \right\rangle + D(\pi_s^{(k+1)}, \pi_s^{(k)}) \leq \eta_k \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, p \right\rangle + D(p, \pi_s^{(k)}) - D(p, \pi_s^{(k+1)}).$$

Rearranging terms and dividing both sides by η_k, we get

$$\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)} - p \right\rangle + \frac{1}{\eta_k} D(\pi_s^{(k+1)}, \pi_s^{(k)}) \leq \frac{1}{\eta_k} D(p, \pi_s^{(k)}) - \frac{1}{\eta_k} D(p, \pi_s^{(k+1)}).$$

Letting $p = \pi_s^{(k)}$ and knowing that

$$\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k)} \right\rangle = 0 \qquad \text{for all } k \geq 0,$$

yields

$$\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)} \right\rangle \leq -\frac{1}{\eta_k} D(\pi_s^{(k+1)}, \pi_s^{(k)}) - \frac{1}{\eta_k} D(\pi_s^{(k)}, \pi_s^{(k+1)}) \leq 0. \tag{85}$$

Letting $p = \pi_s^*$ yields

$$\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)} - \pi_s^* \right\rangle \leq \frac{1}{\eta_k} D(\pi_s^*, \pi_s^{(k)}) - \frac{1}{\eta_k} D(\pi_s^*, \pi_s^{(k+1)}).$$

Note that we dropped the nonnegative term $\frac{1}{\eta_k} D(\pi_s^{(k+1)}, \pi_s^{(k)})$ on the left hand side to the inequality.

Taking expectation with respect to the distribution d^*, we have

$$\mathbb{E}_{s \sim d^*}\left[\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)} \right\rangle\right] - \mathbb{E}_{s \sim d^*}\left[\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^* \right\rangle\right] \leq \frac{1}{\eta_k} D_k^* - \frac{1}{\eta_k} D_{k+1}^*. \tag{86}$$

For the first expectation in (86), we have

$$\begin{aligned}
& \mathbb{E}_{s\sim d^*}\left[\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)}\right\rangle\right] \\
= \quad & \sum_{s\in\mathcal{S}} d_s^* \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)}\right\rangle \\
= \quad & \sum_{s\in\mathcal{S}} \frac{d_s^*}{d_s^{(k+1)}} d_s^{(k+1)} \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)}\right\rangle \\
\overset{(20)+(85)}{\geq} \quad & \vartheta_{k+1} \sum_{s\in\mathcal{S}} d_s^{(k+1)} \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)}\right\rangle \\
\overset{(20)+(85)}{\geq} \quad & \vartheta_{\rho} \sum_{s\in\mathcal{S}} d_s^{(k+1)} \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)}\right\rangle \\
= \quad & \vartheta_\rho \mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[(\bar{\phi}_{s,a}^{(k)})^\top w^{(k)}\right] \\
= \quad & \vartheta_\rho \mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[A_{s,a}^{(k)}\right] + \vartheta_\rho \mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[(\bar{\phi}_{s,a}^{(k)})^\top w^{(k)} - A_{s,a}^{(k)}\right] \\
= \quad & \vartheta_\rho (1-\gamma)\left(V_\rho^{(k+1)} - V_\rho^{(k)}\right) + \vartheta_\rho \mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[(\bar{\phi}_{s,a}^{(k)})^\top w^{(k)} - A_{s,a}^{(k)}\right], \qquad (87)
\end{aligned}$$

where the last line is obtained by the performance difference lemma (45), and we use the shorthand $\bar{\phi}_{s,a}^{(k)}$ as $\bar{\phi}_{s,a}(\theta^{(k)})$.

The second term of (87) can be lower bounded. To do it, we first decompose it into two terms. That is,

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[(\bar{\phi}_{s,a}^{(k)})^\top w^{(k)} - A_{s,a}^{(k)}\right] = & \underbrace{\mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[(\bar{\phi}_{s,a}^{(k)})^\top (w^{(k)} - w_\star^{(k)})\right]}_{\textcircled{1}} \\
& + \underbrace{\mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[(\bar{\phi}_{s,a}^{(k)})^\top w_\star^{(k)} - A_{s,a}^{(k)}\right]}_{\textcircled{2}}. \qquad (88)
\end{aligned}$$

We will upper bound the absolute values of the above two terms $|\textcircled{1}|$ and $|\textcircled{2}|$ separately. More precisely, similar to the proof of Theorem 9, we will upper bound the first term $|\textcircled{1}|$ by the statistical error assumption (35) and upper bound the second term $|\textcircled{2}|$ by using the approximation error assumption (36).

To upper bound $\textcircled{1}$, we first define the following covariance matrix of the centered feature map

$$\Sigma_{\tilde{d}^{(k)}}^{(k)} \overset{\text{def}}{=} \mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\bar{\phi}_{s,a}^{(k)}(\bar{\phi}_{s,a}^{(k)})^\top\right]. \qquad (89)$$

Here we use the superscript (k) for $\Sigma_{\tilde{d}^{(k)}}^{(k)}$ to distinguish the covariance matrix of the feature map $\Sigma_{\tilde{d}^{(k)}}$ defined in (63) in the proof of Theorem 5, as the centered feature map $\bar{\phi}_{s,a}^{(k)}$ depends on the iterates $\theta^{(k)}$.

By Cauchy-Schwartz's inequality, we have

$$\begin{aligned}
|\textcircled{1}| \quad \leq \quad & \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bar{d}_{s,a}^{(k+1)} \left|(\bar{\phi}_{s,a}^{(k)})^\top (w^{(k)} - w_\star^{(k)})\right| \\
\leq \quad & \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(\bar{d}_{s,a}^{(k+1)}\right)^2}{\tilde{d}_{s,a}^{(k)}} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \tilde{d}_{s,a}^{(k)} \left((\bar{\phi}_{s,a}^{(k)})^\top (w^{(k)} - w_\star^{(k)})\right)^2} \\
\overset{(89)}{=} \quad & \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{\bar{d}_{s,a}^{(k+1)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right] \left\|w^{(k)} - w_\star^{(k)}\right\|^2_{\Sigma_{\tilde{d}^{(k)}}^{(k)}}}.
\end{aligned}$$

By further using the concentrability assumption 13, we have

$$
\begin{aligned}
|①| &\overset{(37)}{\leq} \sqrt{C_\nu \left\|w^{(k)} - w_\star^{(k)}\right\|^2_{\Sigma^{(k)}_{\tilde{d}^{(k)}}}} \\
&\leq \sqrt{C_\nu \left(L_A(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}) - L_A(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)})\right)} \qquad (90)\\
&\overset{(82)}{=} \sqrt{C_\nu \epsilon_{\text{stat}}^{(k)}}, \qquad (91)
\end{aligned}
$$

where (90) uses that $w_\star^{(k)}$ is a minimizer of L_A and $w_\star^{(k)}$ is feasible (see the same arguments of (65) in the proof of Theorem 5).

For the second term $|②|$ in (88), by Cauchy-Schwartz's inequality, we have

$$
\begin{aligned}
|②| &\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bar{d}_{s,a}^{(k+1)} \left|(\bar{\phi}_{s,a}^{(k)})^\top w_\star^{(k)} - A_{s,a}^{(k)}\right| \\
&\leq \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(\bar{d}_{s,a}^{(k+1)}\right)^2}{\tilde{d}_{s,a}^{(k)}} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \tilde{d}_{s,a}^{(k)} \left((\bar{\phi}_{s,a}^{(k)})^\top w_\star^{(k)} - A_{s,a}^{(k)}\right)^2} \\
&= \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{\bar{d}_{s,a}^{(k+1)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right] L_A(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)})} \\
&\overset{(37)+(83)}{\leq} \sqrt{C_\nu \epsilon_{\text{approx}}^{(k)}}. \qquad (92)
\end{aligned}
$$

Plugging (91) and (92) into (87) yields

$$
\mathbb{E}_{s\sim d^*}\left[\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)}\right\rangle\right] \geq \vartheta_\rho(1-\gamma)\left(V_\rho^{(k+1)} - V_\rho^{(k)}\right) - \vartheta_\rho\sqrt{C_\nu}\left(\sqrt{\epsilon_{\text{stat}}^{(k)}} + \sqrt{\epsilon_{\text{approx}}^{(k)}}\right). \qquad (93)
$$

Now for the second expectation in (86), by using the performance difference lemma (45) in Lemma 18, we have

$$
\begin{aligned}
-\mathbb{E}_{s\sim d^*}\left[\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^*\right\rangle\right] &= -\mathbb{E}_{(s,a)\sim\bar{d}^{\pi^*}}\left[A_{s,a}^{(k)}\right] + \mathbb{E}_{(s,a)\sim\bar{d}^{\pi^*}}\left[A_{s,a}^{(k)} - (\bar{\phi}_{s,a}^{(k)})^\top w^{(k)}\right] \\
&= (1-\gamma)\left(V_\rho^{(k)} - V_\rho(\pi^*)\right) + \mathbb{E}_{(s,a)\sim\bar{d}^{\pi^*}}\left[A_{s,a}^{(k)} - (\bar{\phi}_{s,a}^{(k)})^\top w^{(k)}\right]. \qquad (94)
\end{aligned}
$$

The second term of (94) can be lower bounded. We first decompose it into two terms. That is,

$$
\begin{aligned}
\mathbb{E}_{(s,a)\sim\bar{d}^{\pi^*}}\left[A_{s,a}^{(k)} - (\bar{\phi}_{s,a}^{(k)})^\top w^{(k)}\right] &= \underbrace{\mathbb{E}_{(s,a)\sim\bar{d}^{\pi^*}}\left[A_{s,a}^{(k)} - (\bar{\phi}_{s,a}^{(k)})^\top w_\star^{(k)}\right]}_{ⓐ} \\
&\quad + \underbrace{\mathbb{E}_{(s,a)\sim\bar{d}^{\pi^*}}\left[(\bar{\phi}_{s,a}^{(k)})^\top (w_\star^{(k)} - w^{(k)})\right]}_{ⓑ}. \qquad (95)
\end{aligned}
$$

Now we will upper bound the absolute values of the above two terms $|ⓐ|$ and $|ⓑ|$ separately.

For the first one $|\textcircled{a}|$, by Cauchy-Schwartz's inequality, we have

$$\begin{aligned}
|\textcircled{a}| \quad &\le \quad \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bar{d}^{\pi^*}_{s,a}\left|A^{(k)}_{s,a} - (\bar{\phi}^{(k)}_{s,a})^\top w^{(k)}_\star\right| \\
&\le \quad \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(\bar{d}^{\pi^*}_{s,a}\right)^2}{\tilde{d}^{(k)}_{s,a}} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \tilde{d}^{(k)}_{s,a}\left((\bar{\phi}^{(k)}_{s,a})^\top w^{(k)}_\star - A^{(k)}_{s,a}\right)^2} \\
&= \quad \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{\bar{d}^{\pi^*}_{s,a}}{\tilde{d}^{(k)}_{s,a}}\right)^2\right] L_A(w^{(k)}_\star, \theta^{(k)}, \tilde{d}^{(k)})} \\
&\overset{(37)+(83)}{\le} \quad \sqrt{C_\nu \epsilon^{(k)}_{\text{approx}}}.
\end{aligned} \tag{96}$$

For the second term $|\textcircled{b}|$ in (95), by Cauchy-Schwartz's inequality, we have

$$\begin{aligned}
|\textcircled{b}| \quad &\le \quad \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bar{d}^{\pi^*}_{s,a}\left|(\bar{\phi}^{(k)}_{s,a})^\top (w^{(k)}_\star - w^{(k)})\right| \\
&\le \quad \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(\bar{d}^{\pi^*}_{s,a}\right)^2}{\tilde{d}^{(k)}_{s,a}} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \tilde{d}^{(k)}_{s,a}\left((\bar{\phi}^{(k)}_{s,a})^\top (w^{(k)} - w^{(k)}_\star)\right)^2} \\
&\overset{(89)}{=} \quad \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{\bar{d}^{\pi^*}_{s,a}}{\tilde{d}^{(k)}_{s,a}}\right)^2\right] \left\|w^{(k)} - w^{(k)}_\star\right\|^2_{\Sigma^{(k)}_{\tilde{d}^{(k)}}}} \\
&\overset{(37)}{\le} \quad \sqrt{C_\nu \left\|w^{(k)} - w^{(k)}_\star\right\|^2_{\Sigma^{(k)}_{\tilde{d}^{(k)}}}} \\
&\overset{(90)}{\le} \quad \sqrt{C_\nu \left(L_A(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}) - L_A(w^{(k)}_\star, \theta^{(k)}, \tilde{d}^{(k)})\right)} \\
&\overset{(82)}{=} \quad \sqrt{C_\nu \epsilon^{(k)}_{\text{stat}}}.
\end{aligned} \tag{97}$$

Thus, we lower bound (95) by

$$-\mathbb{E}_{s\sim d^*}\left[\left\langle \bar{\Phi}^{(k)}_s w^{(k)}, \pi^*_s\right\rangle\right] \overset{(96)+(97)}{\ge} (1-\gamma)\left(V^{(k)}_\rho - V_\rho(\pi^*)\right) - \sqrt{C_\nu}\left(\sqrt{\epsilon^{(k)}_{\text{stat}}} + \sqrt{\epsilon^{(k)}_{\text{approx}}}\right). \tag{98}$$

Substituting (93) and (98) into (86), dividing both side by $1-\gamma$ and rearranging terms, we get

$$\vartheta_\rho\left(\delta_{k+1} - \delta_k\right) + \delta_k \le \frac{D^*_k}{(1-\gamma)\eta_k} - \frac{D^*_{k+1}}{(1-\gamma)\eta_k} + \frac{\sqrt{C_\nu}\left(\vartheta_\rho + 1\right)}{1-\gamma}\left(\sqrt{\epsilon^{(k)}_{\text{stat}}} + \sqrt{\epsilon^{(k)}_{\text{approx}}}\right).$$

■

F.2 Proof of Theorem 14

Proof From (84) in Lemma 25, by using the same increasing step size as in Theorem 5, i.e. $\eta_0 \ge \frac{1-\gamma}{\gamma}D^*_0$ and $\eta_{k+1} \ge \eta_k/\gamma$, and following the same arguments in the proof of Theorem 5 after (76), we obtain the final performance bound with the linear convergence rate

$$\mathbb{E}\left[V_\rho(\theta^{(k)})\right] - V_\rho(\pi^*) \le \left(1 - \frac{1}{\vartheta_\rho}\right)^k \frac{2}{1-\gamma} + \frac{\sqrt{C_\nu}\left(\vartheta_\rho + 1\right)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{approx}}}\right).$$

■

F.3 Proof of Theorem 15

Proof From (84) in Lemma 25 with the constant step size, we have

$$\vartheta_\rho\left(\delta_{k+1}-\delta_k\right)+\delta_k \le \frac{D_k^*}{(1-\gamma)\eta}-\frac{D_{k+1}^*}{(1-\gamma)\eta}+\frac{\sqrt{C_\nu}\left(\vartheta_\rho+1\right)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}^{(k)}}+\sqrt{\epsilon_{\text{approx}}^{(k)}}\right).$$

Taking the total expectation with respect to the randomness in the sequence of the iterates $w^{(0)},\cdots,w^{(k-1)}$ yields

$$\begin{aligned}\vartheta_\rho\left(\mathbb{E}\left[\delta_{k+1}\right]-\mathbb{E}\left[\delta_k\right]\right)+\mathbb{E}\left[\delta_k\right] \quad &\le \quad \frac{\mathbb{E}\left[D_k^*\right]}{(1-\gamma)\eta}-\frac{\mathbb{E}\left[D_{k+1}^*\right]}{(1-\gamma)\eta}\\ &\quad +\frac{\sqrt{C_\nu}\left(\vartheta_\rho+1\right)}{1-\gamma}\left(\mathbb{E}\left[\sqrt{\epsilon_{\text{stat}}^{(k)}}\right]+\mathbb{E}\left[\sqrt{\epsilon_{\text{approx}}^{(k)}}\right]\right)\\ &\le \quad \frac{\mathbb{E}\left[D_k^*\right]}{(1-\gamma)\eta}-\frac{\mathbb{E}\left[D_{k+1}^*\right]}{(1-\gamma)\eta}\\ &\quad +\frac{\sqrt{C_\nu}\left(\vartheta_\rho+1\right)}{1-\gamma}\left(\sqrt{\mathbb{E}\left[\epsilon_{\text{stat}}^{(k)}\right]}+\sqrt{\mathbb{E}\left[\epsilon_{\text{approx}}^{(k)}\right]}\right)\\ &\overset{(35)+(36)}{\le} \quad \frac{\mathbb{E}\left[D_k^*\right]}{(1-\gamma)\eta}-\frac{\mathbb{E}\left[D_{k+1}^*\right]}{(1-\gamma)\eta}+\frac{\sqrt{C_\nu}\left(\vartheta_\rho+1\right)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}}+\sqrt{\epsilon_{\text{approx}}}\right).\end{aligned}$$

By summing up from 0 to $k-1$, we get

$$\vartheta_\rho\mathbb{E}\left[\delta_k\right]+\sum_{t=0}^{k-1}\mathbb{E}\left[\delta_t\right]\le\frac{D_0^*}{(1-\gamma)\eta}+\vartheta_\rho\delta_0+k\cdot\frac{\sqrt{C_\nu}\left(\vartheta_\rho+1\right)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}}+\sqrt{\epsilon_{\text{approx}}}\right).$$

Finally, dropping the positive term $\mathbb{E}\left[\delta_k\right]$ on the left hand side as π^* is the optimal policy and dividing both side by k yields

$$\frac{1}{k}\sum_{t=0}^{k-1}\mathbb{E}\left[V_\rho(\theta^{(t)})\right]-V_\rho(\pi^*)\le\frac{D_0^*}{(1-\gamma)\eta k}+\frac{2\vartheta_\rho}{(1-\gamma)k}+\frac{\sqrt{C_\nu}\left(\vartheta_\rho+1\right)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}}+\sqrt{\epsilon_{\text{approx}}}\right).$$

With the constant step size $\eta\ge\frac{D_0^*}{2\vartheta_\rho}$, we have

$$\frac{1}{k}\sum_{t=0}^{k-1}\mathbb{E}\left[V_\rho(\theta^{(t)})\right]-V_\rho(\pi^*)\le\frac{4\vartheta_\rho}{(1-\gamma)k}+\frac{\sqrt{C_\nu}\left(\vartheta_\rho+1\right)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}}+\sqrt{\epsilon_{\text{approx}}}\right).$$

■

Appendix G. Sample complexity of NPG

Combined with a regression solver, `NPG-SGD` in Algorithm 5, which uses a slight modification of `Q-NPG-SGD` for the unbiased gradient estimates of L_A, we consider a sampled-based NPG Algorithm 1 proposed in Appendix C and show its sample complexity result in the following corollary.

Corollary 26 *Consider the setting of Theorem 14. Suppose that the sample-based NPG Algorithm 1 is run for K iterations, with T gradient steps of* `NPG-SGD` *(Algorithm 5) per iteration. Furthermore, suppose that for all*

$(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $\|\phi_{s,a}\| \leq B$ with $B > 0$, and we choose the step size $\alpha = \frac{1}{8B^2}$ and the initialization $w_0 = 0$ for NPG-SGD. If for all $\theta \in \mathbb{R}^m$, the covariance matrix of the centered feature map induced by the policy $\pi(\theta)$ and the initial state-action distribution ν satisfies

$$\mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\bar{\phi}_{s,a}(\theta)(\bar{\phi}_{s,a}(\theta))^\top\right] \;\geq\; \mu \mathbf{I}_m, \tag{99}$$

where $\mathbf{I}_m \in \mathbb{R}^{m\times m}$ is the identity matrix and $\mu > 0$, then

$$\mathbb{E}\left[V_\rho(\theta^{(K)})\right] - V_\rho(\pi^*) \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^K \frac{2}{1-\gamma} + \frac{(\vartheta_\rho + 1)\sqrt{C_\nu \epsilon_{\text{approx}}}}{1-\gamma} + \frac{4\sqrt{C_\nu}\,(\vartheta_\rho+1)}{(1-\gamma)^2\sqrt{T}}\left(\frac{2B^2}{\mu}\left(\sqrt{2m}+1\right) + \sqrt{2m}\right). \tag{100}$$

Proof Similar to the proof of Corollary 24, we suppress the subscript k. First, the centered feature map is bounded by $\left\|\bar{\phi}_{s,a}(\theta)\right\| \leq 2B$. In order to apply Theorem 1 of Bach and Moulines (2013), it remains to upper bound $\mathbb{E}\left[\left\|\widehat{A}_{s,a}(\theta)\bar{\phi}_{s,a}(\theta)\right\|^2\right]$ and $\|w_\star\|$ with $w_\star \in \operatorname{argmin}_w L_A(w,\theta,\tilde{d}^\theta)$, and find $\sigma > 0$ such that

$$\mathbb{E}\left[\left(\widehat{A}_{s,a}(\theta) - w_\star^\top \bar{\phi}_{s,a}(\theta)\right)^2 \mid s,a\right] = \mathbb{E}\left[\left(\widehat{A}_{s,a}(\theta)\right)^2 \mid s,a\right] - 2A_{s,a}(\theta)w_\star^\top\bar{\phi}_{s,a}(\theta) + \left(w_\star^\top\bar{\phi}_{s,a}(\theta)\right)^2 \leq \sigma^2 \tag{101}$$

holds for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $\theta \in \mathbb{R}^m$.

Similar to the proof of Corollary 24, the closed form solution of $w_\star$ can be written as

$$w_\star = \left(\mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\bar{\phi}_{s,a}(\theta)\bar{\phi}_{s,a}(\theta)^\top\right]\right)^\dagger \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[Q_{s,a}(\theta)\bar{\phi}_{s,a}(\theta)\right].$$

From (99), we have

$$\|w_\star\| \leq \frac{2B}{\mu(1-\gamma)}.$$

Now we need to upper bound $\mathbb{E}\left[\left(\widehat{A}_{s,a}(\theta)\right)^2 \mid s,a\right]$ from (101). Indeed, by using $\widehat{A}_{s,a}(\theta) = \widehat{Q}_{s,a}(\theta) - \widehat{V}_s(\theta)$, we have

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{A}_{s,a}(\theta)\right)^2 \mid s,a\right] &\leq 2\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta)\right)^2 \mid s,a\right] + 2\mathbb{E}\left[\left(\widehat{V}_{s,a}(\theta)\right)^2 \mid s,a\right] \\ &\overset{(79)}{\leq} \frac{8}{(1-\gamma)^2},\end{aligned} \tag{102}$$

where the last line is obtained, as $\mathbb{E}\left[\left(\widehat{V}_{s,a}(\theta)\right)^2 \mid s,a\right]$ shares the same upper bound (79) of $\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta)\right)^2 \mid s,a\right]$ by using the similar argument.

From (102) and $\bar{\phi}_{s,a}(\theta) \leq 2B$, we verify $\mathbb{E}\left[\left\|\widehat{A}_{s,a}(\theta)\bar{\phi}_{s,a}(\theta)\right\|^2\right]$ bounded as well.

By using the upper bounds of $\mathbb{E}\left[\left(\widehat{A}_{s,a}(\theta)\right)^2 \mid s,a\right]$, $\|w_\star\|$, $|A_{s,a}(\theta)| \leq \frac{2}{1-\gamma}$ and $\left\|\bar{\phi}_{s,a}(\theta)\right\| \leq 2B$, the left hand side of (101) is upper bounded by

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{A}_{s,a}(\theta) - w_\star^\top\bar{\phi}_{s,a}(\theta)\right)^2 \mid s,a\right] &\leq \frac{8}{(1-\gamma)^2} + \frac{16B^2}{\mu(1-\gamma)^2} + \frac{16B^4}{\mu^2(1-\gamma)^2} \\ &= \frac{4}{(1-\gamma)^2}\left(\left(\frac{2B^2}{\mu}+1\right)^2 + 1\right) \\ &\leq \frac{8}{(1-\gamma)^2}\left(\frac{2B^2}{\mu}+1\right)^2.\end{aligned}$$

Thus, we choose

$$\sigma = \frac{2\sqrt{2}}{1-\gamma}\left(\frac{2B^2}{\mu}+1\right).$$

Now all the conditions (i) - (vi) in Theorem 29 are verified. The reminder of the proof follows that of Corollary 24. ■

Appendix H. Standard Optimization Results

In this section, we present the standard optimization results from Beck (2017); Xiao (2022); Bach and Moulines (2013) used in our proofs.

First, we present the closed form update of mirror descent with KL divergence on the simplex. We provide its proof for the completeness.

Lemma 27 (Mirror descent on the simplex, Example 9.10 in Beck (2017)) *Let $g \in \mathbb{R}^n$ which will often be a gradient and let $\eta > 0$. For p, q in the unit n-simplex Δ^n, the mirror descent step with respect to the KL divergence*

$$\min_{p \in \Delta^n} \; \eta \langle g, p \rangle + D(p, q) \tag{103}$$

is given by

$$p = \frac{q \odot e^{-\eta g}}{\sum_{i=1}^n q_i e^{-\eta g_i}}, \tag{104}$$

where $\odot$ is the element-wise product between vectors.

Proof The Lagrangian of (103) is given by

$$L(p, \mu, \lambda) = \eta \langle g, p \rangle + D(p, q) + \mu(1 - \sum_{i=1}^n p_i) - \sum_{i=1}^n \lambda_i p_i,$$

where $\mu \in \mathbb{R}$ and $\lambda \in \mathbb{R}^n$ with non-negative coordinates are the Lagrangian multipliers. Thus the Karush–Kuhn–Tucker conditions are given by

$$\begin{aligned} \eta g + \log(p/q) + \mathbf{1}_n &= \mu \mathbf{1}_n + \lambda, \\ \mathbf{1}_n^\top p &= 1, \\ \lambda_i = 0 \text{ or } p_i &= 0, \qquad \text{for all } i = 1, \cdots, n, \end{aligned}$$

where the division p/q is element-wise. Isolating p in the top equation gives

$$p = q \odot e^{(\mu-1)\mathbf{1}_n + \lambda - \eta g} = e^{\mu-1} q \odot e^{\lambda - \eta g}.$$

Using the second constraint $\mathbf{1}_n^\top p = 1$ gives that

$$1 = e^{\mu-1} \sum_{i=1}^n q_i e^{\lambda_i - \eta g_i} \implies e^{\mu-1} = \frac{1}{\sum_{i=1}^n q_i e^{\lambda_i - \eta g_i}}.$$

Consequently, by plugging the above term into p, we have that

$$p = \frac{q \odot e^{\lambda - \eta g}}{\sum_{i=1}^n q_i e^{\lambda_i - \eta g_i}}.$$

It remains to determine λ. If $q_i = 0$ then $p_i = 0$ and thus $\lambda_i > 0$. Conversely, if $q_i > 0$ then $p_i > 0$ and thus $\lambda_i = 0$. In either of these cases, we have that the solution is given by (104). ■

Now we present the three-point descent lemma on proximal optimization with Bregman divergences, which is another key ingredient for our PMD analysis. Following Xiao (2022, Lemma 6), we adopt a slight variation of Lemma 3.2 in Chen and Teboulle (1993). First, we need some technical conditions. We say a function h is of Legendre type (Rockafellar, 1970, Section 26) if it is essentially smooth and strictly convex in the relative interior of $\operatorname{dom} h$, denoted as $\operatorname{rint}\operatorname{dom} h$. Essential smoothness means that h is differentiable and $\|\nabla h(x_n)\| \to \infty$ for every sequence $\{x_n\}$ converging to a boundary point of $\operatorname{dom} h$. The Bregman divergence generated by a function h of Legendre type is a distance-like function defined as

$$D_h(p, p') \overset{\text{def}}{=} h(p) - h(p') - \langle \nabla h(p'), p - p' \rangle \,.$$

Under the above conditions, we have the following result.

Lemma 28 (Three-point decent lemma, Lemma 6 in Xiao (2022)) *Suppose that $\mathcal{C} \subset \mathbb{R}^m$ is a closed convex set, $f : \mathcal{C} \to \mathcal{R}$ is a proper, closed convex function, $D_h(\cdot, \cdot)$ is the Bregman divergence generated by a function h of Lengendre type and* $\operatorname{rint}\operatorname{dom} h \cap \mathcal{C} \neq \emptyset$*. For any $x \in \operatorname{rint}\operatorname{dom} h$, let*

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

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

$$f(x^+) + D_h(x^+, x) \leq f(u) + D_h(u, x) - D_h(u, x^+).$$

Finally, we use the following linear regression analysis for the proof of our sample complexity results, i.e., Corollary 24 and 26.

Theorem 29 (Theorem 1 in Bach and Moulines (2013)) *Consider the following assumptions:*

(i) $\mathcal{H}$ is a m-dimensional Euclidean space.

(ii) The observations $(x_n, z_n) \in \mathcal{H} \times \mathcal{H}$ are independent and identically distributed.

(iii) $\mathbb{E}\left[\|x_n\|^2\right]$ and $\mathbb{E}\left[\|z_n\|^2\right]$ are finite. The covariance $\mathbb{E}\left[x_n x_n^\top\right]$ is assumed invertible.

(iv) The global minimum of $f(\theta) = \frac{1}{2}\mathbb{E}\left[\langle \theta, x_n \rangle^2 - 2\langle \theta, z_n \rangle\right]$ is attained at a certain $\theta_* \in \mathcal{H}$. Denote $\xi_n = z_n - \langle \theta_*, x_n \rangle x_n$ as the residual. We have $\mathbb{E}\left[\xi_n\right] = 0$.

(v) Consider the stochastic gradient recursion defined as

$$\theta_n = \theta_{n-1} - \eta(\langle \theta_{n-1}, x_n \rangle x_n - z_n),$$

started from $\theta_0 \in \mathcal{H}$ and also consider the averaged iterates $\theta_{\text{out}} = \frac{1}{n+1}\sum_{k=0}^{n} \theta_k$.

(vi) There exists $R > 0$ and $\sigma > 0$ such that $\mathbb{E}\left[\xi_n \xi_n^\top\right] \leq \sigma^2 \mathbb{E}\left[x_n x_n^\top\right]$ and $\mathbb{E}\left[\|x_n\|^2 x_n x_n^\top\right] \leq R^2 \mathbb{E}\left[x_n x_n^\top\right]$.

When $\eta = \frac{1}{4R^2}$, we have

$$\mathbb{E}\left[f(\theta_{\text{out}}) - f(\theta_*)\right] \leq \frac{2}{n}\left(\sigma\sqrt{m} + R\|\theta_0 - \theta_*\|\right)^2. \tag{105}$$