LINEAR CONVERGENCE OF NATURAL POLICY GRADIENT METHODS WITH LOG-LINEAR POLICIES

Anonymous authors
Paper under double-blind review

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, both methods with log-linear policies can be written as inexact versions of the policy mirror descent (PMD) method. We show that both methods attain linear convergence rates and $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexities using a simple, non-adaptive geometrically increasing step size, without resorting to entropy or other strongly convex regularization. Lastly, as a byproduct, we obtain sublinear convergence rates for both methods with arbitrary constant step size.

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 & 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; Baxter & 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 & Yudin, 1983; Beck & 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., 2021b; Xiao, 2022) and some more general settings (Shani et al., 2020; Vaswani et al., 2022; Grudzien et al., 2022). In order to get a 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 methods (Lan, 2022; Zhan et al., 2021). By designing appropriate step sizes, Khodadadian et al. (2021b) and Xiao (2022) obtain linear convergence of NPG without regularization. However, all these linear convergence results are limited in the tabular setting with a direct parametrization. It remains unclear whether this same linear convergence rate can be established in function approximation settings.

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 inexact 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 a 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. The extensions are highly nontrivial and require quite different techniques. This linear convergence leads to the $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexities for both methods. In particular, our sample complexity analysis also fixes errors of previous work. Lastly, as a byproduct, we obtain sublinear convergence rates for both methods with arbitrary constant step size. 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 minimization 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

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

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

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

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

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

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

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

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.,

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

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 $(s, a) \in \mathcal{S} \times \mathcal{A}$,

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

[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.

Policy parametrization. In practice, both the state and action spaces $\mathcal{S}$ and $\mathcal{A}$ can be very large and some form of function approximation is needed to reduce the dimensions and 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

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

This setting is important since it is the simplest instantiation of the widely-used neural policy parametrization. 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_s^θ for $d_s^{\pi(\theta)}$, $\bar{d}_{s,a}^\theta$ for $\bar{d}_{s,a}^{\pi(\theta)}$, and $\tilde{d}_{s,a}^\theta$ for $\tilde{d}_{s,a}^{\pi(\theta)}$.

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

$$F_\rho(\theta) \;\stackrel{\text{def}}{=}\; \mathbb{E}_{s \sim d^\theta,\, a \sim \pi_s(\theta)} \left[\nabla_\theta \log \pi_{s,a}(\theta) \big(\nabla_\theta \log \pi_{s,a}(\theta) \big)^\top \right], \tag{10}$$

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). Despite being a non-convex optimization problem, there is still additional structure we can leverage to ensure convergence. Following Agarwal et al. (2021), we adopt the framework of *compatible function approximation* (Sutton et al., 2000; Kakade, 2001), which exploits the MDP structure and leads to tight convergence rate analysis.

For any $w, \theta \in \mathbb{R}^m$ and state-action distribution $\zeta \in \Delta(\mathcal{S} \times \mathcal{A})$, we define the *compatible function approximation error* as

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

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

$$\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k w_\star^{(k)}, \qquad w_\star^{(k)} \in \operatorname{argmin}_{w \in \mathbb{R}^m} L_A\big(w, \theta^{(k)}, \bar{d}^{(k)}\big), \tag{12}$$

where $\bar{d}^{(k)}$ is a shorthand for the state-action visitation distribution $\bar{d}^{\pi(\theta^{(k)})}(\rho)$ defined in (4). A derivation of (12) is provided in Appendix B (Lemma 1) for 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. This is where the term "compatible function approximation error" comes from. For the log-linear policy class defined in (7), we have

$$\nabla_\theta \log \pi_{s,a}(\theta) \;=\; \bar{\phi}_{s,a}(\theta) \;\stackrel{\text{def}}{=}\; \phi_{s,a} - \textstyle\sum_{a' \in \mathcal{A}} \pi_{s,a'}(\theta) \phi_{s,a'} \;=\; \phi_{s,a} - \mathbb{E}_{a' \sim \pi_s(\theta)} \left[\phi_{s,a'} \right], \tag{13}$$

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 inexact NPG update rule:

$$\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k w^{(k)}, \qquad w^{(k)} \approx \operatorname{argmin}_w L_A\big(w, \theta^{(k)}, \bar{d}^{(k)}\big). \tag{14}$$

The inexact NPG updates require samples of unbiased estimates of $A_{s,a}(\theta)$, the corresponding sampling procedure is given in Algorithm 4, and a sample-based regression solver to minimize L_A is given in Algorithm 5 in the Appendix.

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

$$L_Q(w, \theta, \zeta) \stackrel{\text{def}}{=} \mathbb{E}_{(s,a)\sim\zeta}\left[\left(w^\top \phi_{s,a} - Q_{s,a}(\theta)\right)^2\right] \tag{15}$$

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

$$\theta^{(k+1)} = \theta^{(k)} - \eta_k w^{(k)}, \qquad w^{(k)} \approx \operatorname{argmin}_w L_Q\big(w, \theta^{(k)}, \bar{d}^{(k)}\big). \tag{16}$$

For Q-NPG, the sampling procedure for estimating $Q_{s,a}(\theta)$ is given in Algorithm 3 and a sample-based regression solver for $w^{(k)}$ is proposed in Algorithm 6 in the Appendix.

The sampling procedure and the regression solver of NPG are less efficient than those of Q-NPG. Indeed, the sampling procedure for $A_{s,a}(\theta)$ in Algorithm 4 not only estimates $Q_{s,a}(\theta)$, but also requires an additional estimation of $V_s(\theta)$, and thus doubles the amount of samples as compared to Algorithm 3. Furthermore, the stochastic gradient estimator of L_Q in Algorithm 6 only computes on a single action of the feature map $\phi_{s,a}$. Whereas the one of L_A in Algorithm 5 computes on the centered feature map $\bar{\phi}_{s,a}(\theta)$ defined in (13), which needs to go through the entire action space, thus is $|\mathcal{A}|$ times more expensive to run. See Appendix C for more details.

Following Agarwal et al. (2021), we consider slightly different variants of NPG and Q-NPG, where $\bar{d}^{(k)}$ in (14) and (16) 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 to the policy $\pi(\theta^{(k)})$. For example, it can be seen from (6) that the lower bound of $\tilde{d}^{\pi}$ is independent to π, which is not the case for $\bar{d}^{\pi}$. This property is crucial in the forthcoming convergence analysis.

3.1 Formulation as Inexact 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 $s \in \mathcal{S}$. We note that the above $\pi^{(k+1)}$ can also be obtained by a mirror descent update:

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

where $\Phi_s \in \mathbb{R}^{|\mathcal{A}|\times m}$ is a matrix with rows $(\phi_{s,a})^\top \in \mathbb{R}^{1\times m}$ 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., $D(p,q) \stackrel{\text{def}}{=} \sum_{a\in\mathcal{A}} p_a \log\big(\frac{p_a}{q_a}\big)$. A derivation of (17) is provided in Appendix B (Lemma 2) for completeness.

If we replace $\Phi_s w^{(k)}$ in (17) by the vector $\big[Q_{s,a}(\pi^{(k)})\big]_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$, 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 (17) can be viewed as an inexact PMD method where $Q_s(\pi^{(k)})$ is linearly approximated by $\Phi_s w^{(k)}$ through compatible function approximation (15). Similarly, we can write the inexact NPG update rule as

$$\pi_s^{(k+1)} = \arg\min_{p\in\Delta(\mathcal{A})} \left\{\eta_k \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, p\right\rangle + D(p, \pi_s^{(k)})\right\}, \quad \forall s \in \mathcal{S}, \tag{18}$$

where $w^{(k)}$ is an inexact 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 $\big(\bar{\phi}_{s,a}(\theta^{(k)})\big)^\top$, as defined in (13).

Reformulating Q-NPG and NPG into the mirror descent forms (17) and (18), respectively, allows us to adapt the analysis of PMD method developed in Xiao (2022) to obtain sharp 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 converge linearly 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 inexact Q-NPG method

$$\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k w^{(k)}, \qquad w^{(k)} \approx \operatorname{argmin}_w L_Q\big(w, \theta^{(k)}, \tilde{d}^{(k)}\big), \tag{19}$$

where $\tilde{d}^{(k)}$ is shorthand for $\tilde{d}^{\pi(\theta^{(k)})}(\nu)$ and $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$ is an arbitrary state-action distribution that does not depend on ρ. The exact minimizer is denoted as $w_\star^{(k)} \in \operatorname{argmin}_w L_Q\big(w, \theta^{(k)}, \tilde{d}^{(k)}\big)$. Following Agarwal et al. (2021), 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 is our solution to the regression problem, i.e., how good $w^{(k)}$ is compared with $w_\star^{(k)}$. The approximation error measures the best possible solution for 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 and is also the approach we take later in Section 5 for the analysis of the NPG method.

However, 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 (19), we have*

$$\mathbb{E}\left[L_Q\big(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big) - L_Q\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big)\right] \;\leq\; \epsilon_{\text{stat}}. \tag{20}$$

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 1) where T is the number of iterations used to find the inexact solution $w^{(k)}$.

4.1 Analysis with Bounded Transfer Error

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

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

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 $d^{\pi^*}(\rho)$ as d^* and $d^{\pi(\theta^{(k)})}(\rho)$ as $d^{(k)}$, and define the following distribution mismatch coefficients:

$$\vartheta_k \stackrel{\text{def}}{=} \left\|\frac{d^*}{d^{(k)}}\right\|_\infty \stackrel{(6)}{\leq} \frac{1}{1-\gamma}\left\|\frac{d^*}{\rho}\right\|_\infty \qquad \text{and} \qquad \vartheta_\rho \stackrel{\text{def}}{=} \frac{1}{1-\gamma}\left\|\frac{d^*}{\rho}\right\|_\infty \geq \frac{1}{1-\gamma}. \tag{21}$$

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 $s \in \mathcal{S}$. This is commonly used in the literature on policy gradient methods (e.g., Zhang et al., 2020; Wang et al., 2020) and the NPG convergence analysis (e.g., Cayci et al., 2021; Xiao, 2022). We further relax this condition in Appendix H.

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

$$\tilde{d}^*_{s,a} \stackrel{\text{def}}{=} d^*_s \cdot \text{Unif}_{\mathcal{A}}(a) \stackrel{\text{def}}{=} \frac{d^*_s}{|\mathcal{A}|}, \tag{22}$$

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 (19), we have*

$$\mathbb{E}\left[L_Q\left(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^*\right)\right] \;\leq\; \epsilon_{\text{bias}}. \tag{23}$$

The ϵ_{bias} is often referred to as the transfer error, since it is the error due to replacing the relevant distribution $\tilde{d}^{(k)}$ by $\tilde{d}^*$. This transfer error bound 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 a certain low-rank structure (Jiang et al., 2017; Yang & 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*

$$\Sigma_{\tilde{d}^*} \overset{\text{def}}{=} \mathbb{E}_{(s,a)\sim\tilde{d}^*}\left[\phi_{s,a}\phi_{s,a}^\top\right], \qquad \text{and} \qquad \Sigma_\nu \overset{\text{def}}{=} \mathbb{E}_{(s,a)\sim\nu}\left[\phi_{s,a}\phi_{s,a}^\top\right], \tag{24}$$

where $\tilde{d}^*$ is specified in (22). We define the relative condition number between $\Sigma_{\tilde{d}^*}$ and Σ_ν as

$$\kappa_\nu \overset{\text{def}}{=} \max_{w\in\mathbb{R}^m} \frac{w^\top \Sigma_{\tilde{d}^*} w}{w^\top \Sigma_\nu w}, \tag{25}$$

and assume that κ_ν is finite.

The κ_ν is referred to as the relative condition number, since the ratio is between two different matrix induced norm. Assumption 3 benefits from the use 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 to 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 (19), it holds that*

$$\mathbb{E}_{s\sim d^*}\left[\left(\frac{d_s^{(k)}}{d_s^*}\right)^2\right] \leq C_\rho. \tag{26}$$

The concentrability coefficient is studied in the analysis of approximate dynamic programming algorithms (Munos, 2003; 2005; Munos & Szepesvári, 2008). It measures how much ρ can get amplified in k steps as compared to the reference distribution d_s^*. Let $\rho_{\min} = \min_{s\in\mathcal{S}} \rho_s$. A sufficient condition for Assumption 4 to hold is that $\rho_{\min} > 0$. Indeed,

$$\sqrt{\mathbb{E}_{s\sim d^*}\left[\left(\frac{d_s^{(k)}}{d_s^*}\right)^2\right]} \leq \left\|\frac{d^{(k)}}{d^*}\right\|_\infty \overset{(6)}{\leq} \frac{1}{1-\gamma}\left\|\frac{d^{(k)}}{\rho}\right\|_\infty \leq \frac{1}{(1-\gamma)\rho_{\min}}. \tag{27}$$

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 C_ν defined in Section 4.2 that is independent to ρ and thus is more easily satisfied.

Now we present our first main result.

Theorem 1. *Fix a state distribution ρ, an state-action distribution ν and a comparator policy π^*. We consider the Q-NPG method (19) with the uniform initial policy [2] and the step sizes satisfying $\eta_0 \geq \frac{1-\gamma}{\gamma}\log|\mathcal{A}|$ 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 $k \geq 0$,*

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

[2] For simplicity, we present all our results with the uniform initial policy in the main paper. See Theorem 5, 7 and 8 in the Appendix for more general results with any initial policy.

The main differences between our Theorem 1 and Theorem 20 of Agarwal et al. (2021), which is their corresponding result on the inexact 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 avoid such boundedness assumptions but rely on the 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 use a geometrically increasing step size $\eta_k \propto \gamma^k$ for all $k \geq 0$. This discrepancy reflects the 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 & Teboulle, 1993) and the performance difference lemma (Kakade & Langford, 2002), without reliance on smoothness parameters. As a consequence, we can take advantage of exponentially growing step sizes and avoid assuming the boundedness of $\|\phi_{s,a}\|$ or $\|w^{(k)}\|$.

The reason of using increasing step size can be interpreted as approximate policy iteration. See Appendix H the connection with policy iteration for more details.

As a by product, we also obtain a sublinear $\mathcal{O}(1/k)$ convergence result while using arbitrary constant step size.

Theorem 2. *Fix a state distribution ρ, an state-action distribution ν and an* optimal policy π^*. *We consider the Q-NPG method* (19) *with the uniform initial policy and any constant step size $\eta_k = \eta > 0$. Suppose that Assumptions 1, 2, 3 and 4 all hold. Then we have for all $k \geq 0$,*

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

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 weaker concentrability coefficient.

Assumption 5 (Bounded approximation error)**.** *There exists $\epsilon_{\text{approx}} > 0$ such that for all iterations $k \geq 0$ of the Q-NPG method* (19)*, it holds that*

$$\mathbb{E}\left[L_Q\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big)\right] \leq \epsilon_{\text{approx}}. \tag{28}$$

As mentioned in Agarwal et al. (2021), Assumption 5 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 6 (Concentrability coefficient for state-action visitation)**.** *There exists $C_\nu < \infty$ such that for all iterations of the Q-NPG method* (19)*, we have*

$$\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{h_{s,a}^{(k)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right] \leq C_\nu, \tag{29}$$

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

$$d_s^{(k+1)}\pi_{s,a}^{(k+1)}, \qquad d_s^{(k+1)}\pi_{s,a}^{(k)}, \qquad d_s^*\pi_{s,a}^{(k)}, \qquad \text{and} \quad d_s^*\pi_{s,a}^*. \tag{30}$$

Since we are free to choose ν independently to ρ, we can choose $\nu_{s,a} > 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ for Assumption 6 to hold. Indeed, with $\nu_{\min}$ denoting $\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \nu_{s,a}$, we have

$$\sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{h_{s,a}^{(k)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right]} \leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{h_{s,a}^{(k)}}{\tilde{d}_{s,a}^{(k)}} \overset{(6)}{\leq} \frac{1}{(1-\gamma)\nu_{\min}}, \tag{31}$$

where the upper bound can be smaller than that in (27) if $\rho_{\min}$ is smaller than $\nu_{\min}$.

Theorem 3. *Fix a state distribution ρ, an state-action distribution ν and a comparator policy π^*. We consider Q-NPG method (19) with uniform initial policy and step sizes satisfying $\eta_0 \geq \frac{1-\gamma}{\gamma}\log|\mathcal{A}|$ and $\eta_{k+1} \geq \frac{1}{\gamma}\eta_k$. Suppose that Assumptions 1, 5 and 6 hold. Then we have for all $k \geq 0$,*

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

Compared to Theorem 1, 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. However, it is not clear if Theorem 3 is better than Theorem 1.

Remark 1. *Note that Theorems 1, 2 and 3 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 to the policy $\pi^{(k)}$ or ρ. This property allows us to define a weak notion of relative condition number (Assumption 3) that is independent to the iterates, and also allows us to get a finite upper bound of C_ν (Assumption 6 and (31)) that is independent to ρ.*

4.3 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 method (Algorithm 2 in Appendix C). Combined with a simple stochastic gradient descent (SGD) 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 (32).

Corollary 1. *Consider the setting of Theorem 3. 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 followed by the initial state-action distribution ν satisfies*

$$\mathbb{E}_{(s,a)\sim\nu}\left[\phi_{s,a}\phi_{s,a}^\top\right] \overset{(24)}{=} \Sigma_\nu \;\geq\; \mu\mathbf{I}_m, \tag{32}$$

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

$$\mathbb{E}\left[V_\rho(\pi^{(K)})\right] - V_\rho(\pi^*) \leq \left(1 - \tfrac{1}{\vartheta_\rho}\right)^K \tfrac{2}{1-\gamma} + \tfrac{2(\vartheta_\rho+1)\sqrt{C_\nu\epsilon_{\text{approx}}}}{1-\gamma} + \tfrac{4\sqrt{C_\nu}(\vartheta_\rho+1)}{(1-\gamma)^3\sqrt{T}}\left(\tfrac{B^2}{\mu}(\sqrt{2m}+1) + (1-\gamma)\sqrt{2m}\right).$$

In `Q-NPG-SGD`, each trajectory has the expected length $1/(1-\gamma)$ (see Lemma 4). Consequently, with $K = \mathcal{O}(\log(1/\epsilon)\log(1/(1-\gamma)))$ and $T = \mathcal{O}\big(\frac{1}{(1-\gamma)^6\epsilon^2}\big)$, Q-NPG requires $K * T/(1-\gamma) = \tilde{\mathcal{O}}\big(\frac{1}{(1-\gamma)^7\epsilon^2}\big)$ samples such that $\mathbb{E}\left[V_\rho(\pi^{(K)})\right] - V_\rho(\pi^*) \leq \mathcal{O}(\epsilon) + \mathcal{O}\big(\frac{\sqrt{\epsilon_{\text{approx}}}}{1-\gamma}\big)$. The $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity matches with the one of value-based algorithms such as Q-learning (Li et al., 2020) .

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 improvement on the convergence rate for K, they use the optimization results of Shalev-Shwartz & Ben-David (2014, Theorem 14.8) to obtain $\epsilon_{\text{stat}} = \mathcal{O}(1/\sqrt{T})$, while we use the one of Bach & Moulines (2013, Theorem 1) (see Theorem 12 as well) to establish faster $\epsilon_{\text{stat}} = \mathcal{O}(1/T)$. Besides, they use the projected SGD method and require that the stochastic gradient is bounded which is incorrectly verified in their proof [3]. In contrast, to apply Theorem 12, the boundedness of the stochastic gradient is not necessary. Alternatively, we require a different condition (32). A proof sketch of our corollary and a discussion of the condition (32) are provided in Appendix D.5 for more details.

[3] Indeed, the stochastic gradient of L_Q is unbounded, since the estimate $\widehat{Q}_{s,a}(\theta)$ of $Q_{s,a}(\theta)$ is unbounded. This is because each single sampled trajectory has unbounded length. See Appendix D.5 for more explanations.

5 Analysis of NPG with Log-Linear Policies

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

$$\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k w^{(k)}, \qquad w^{(k)} \approx \operatorname{argmin}_w L_A\big(w, \theta^{(k)}, \tilde{d}^{(k)}\big). \tag{33}$$

Again, 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. That is, we again exploit the inexact PMD formulation (18) and use 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 a 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 7 (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 (33), we have*

$$\mathbb{E}\left[L_A\big(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big) - L_A\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big)\right] \;\leq\; \epsilon_{\text{stat}}. \tag{34}$$

Assumption 8 (Bounded approximation error)**.** *There exists $\epsilon_{\text{approx}} > 0$ such that for all iterations $k \geq 0$ of the NPG method (33), we have*

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

Assumption 9 (Concentrability coefficient for state-action visitation)**.** *There exists $C_\nu < \infty$ such that for all terations $k \geq 0$ of the NPG method (33), we have*

$$\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{\bar{d}_{s,a}^{(k+1)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right] \;\leq\; C_\nu \qquad \textit{and} \qquad \mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{\bar{d}_{s,a}^{\pi^*}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right] \;\leq\; C_\nu. \tag{36}$$

Under the above assumptions, we have the following result.

Theorem 4. *Fix a state distribution ρ, a state-action distribution ν, and a comparator policy π^*. We consider the NPG method (33) with uniform initial policy and step sizes satisfying $\eta_0 \geq \frac{1-\gamma}{\gamma}\log|\mathcal{A}|$ and $\eta_{k+1} \geq \frac{1}{\gamma}\eta_k$. Suppose that Assumptions 7, 8 and 9 hold. Then we have for all $k \geq 0$,*

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

Now we compare Theorem 4 with Theorem 29 in Agarwal et al. (2021) for the NPG analysis. The main differences are similar to those for Q-NPG as summarized right after Theorem 1: Their convergence rate is sublinear while ours is linear; they assume uniformly bounded $\phi_{s,a}$ and $w^{(k)}$ while we require bounded concentrability coefficient C_ν due to different proof techniques; they use diminishing step sizes and we use geometrically increasing ones. Moreover, Theorem 4 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. In particular, such bounded relative condition number must hold for the covariance matrix of $\bar{\phi}_{s,a}(\theta^{(k)})$ for all $k \geq 0$ which depends on the iterates $\theta^{(k)}$. This is in contrast to our Assumption 3, where we use a single fixed covariance matrix for Q-NPG that is independent to the iterates, as defined in (24).

In addition, the inequalities in (36) only involve half of the state-action visitation distributions listed in (30), i.e., the first and the fourth terms. Thus, C_ν in (36) can share the same upper bound in (31) independent to the use of the algorithm Q-NPG or NPG. Consequently, our concentrability coefficient C_ν in (36) 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 (31) does not depend on the policies throughout the iterations thanks to the use of $\tilde{d}^{(k)}$ instead of $\bar{d}^{(k)}$ (see Remark 1 as well). See Appendix H for a thorough discussion on the concentrability coefficient C_ν.

As a by product, we also obtain a sublinear rate for NPG while using an unconstrained constant step size in Theorem 9 in Appendix F. By further assuming that the feature map is bounded and the Fisher information matrix (10) is non-singular, we also provide a $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity result of NPG as for Q-NPG and fix the errors occurred in the NPG sample complexity proofs of Agarwal et al. (2021) and Liu et al. (2020) 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. (Cited on pages 1, 3, 4, 5, 6, 7, 8, 9, 16, 17, 18, 20, 22, 25, 33, 35, 43, 44, 46, and 47.)

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

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. Curran Associates, Inc., 2013. (Cited on pages 8, 24, 35, 36, 48, and 50.)

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. (Cited on page 1.)

A Beck and M Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. (Cited on page 1.)

Amir Beck. *First-Order Methods in Optimization*. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2017. ISBN 1611974984. (Cited on page 48.)

D. Bertsekas. *Dynamic Programming and Optimal Control: Volume II; Approximate Dynamic Programming*. Athena Scientific optimization and computation series. Athena Scientific, 2012. (Cited on page 48.)

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

Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite mdps. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2386–2394. PMLR, 13–15 Apr 2021. (Cited on pages 16 and 17.)

Shalabh Bhatnagar, Richard S. Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009. ISSN 0005-1098. (Cited on page 44.)

L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967. ISSN 0041-5553. (Cited on page 49.)

R. H. Byrd, S. L. Hansen, Jorge Nocedal, and Y. Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016. doi: 10.1137/140954362. (Cited on page 44.)

Semih Cayci, Niao He, and R. Srikant. Linear convergence of entropy-regularized natural policy gradient with linear function approximation, 2021. (Cited on pages 5, 9, 17, 18, 36, and 47.)

Semih Cayci, Niao He, and R. Srikant. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm, 2022. (Cited on page 17.)

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. (Cited on pages 1 and 16.)

Y. Censor and S.A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, USA, 1997. (Cited on page 49.)

Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993. (Cited on pages 7 and 48.)

Zaiwei Chen and Siva Theja Maguluri. Sample complexity of policy-based methods under off-policy sampling and linear function approximation. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 11195–11214. PMLR, 28–30 Mar 2022. (Cited on pages 17, 18, and 47.)

Zaiwei Chen, Sajad Khodadadian, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor–critic with linear function approximation. *IEEE Control Systems Letters*, 6:2611–2616, 2022. (Cited on page 17.)

Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global optimum convergence of momentum-based policy gradient. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 1910–1934. PMLR, 28–30 Mar 2022. (Cited on page 44.)

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 1329–1338. JMLR.org, 2016. (Cited on page 44.)

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*, pp. 1467–1476. PMLR, 10–15 Jul 2018. (Cited on page 17.)

Robert Gower, Donald Goldfarb, and Peter Richtarik. Stochastic block BFGS: Squeezing more curvature out of data. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1869–1878, New York, New York, USA, 20–22 Jun 2016. PMLR. (Cited on page 44.)

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

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. In Shie Mannor, Nathan Srebro, and Robert C. Williamson (eds.), *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 9.1–9.24, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. (Cited on page 35.)

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

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4422–4433. PMLR, 13–18 Jul 2020. (Cited on page 44.)

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*, pp. 1704–1713. PMLR, 06–11 Aug 2017. (Cited on page 6.)

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*, pp. 2137–2143. PMLR, 09–12 Jul 2020. (Cited on pages 6 and 17.)

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of 19th International Conference on Machine Learning*, pp. 267–274, 2002. (Cited on pages 7, 18, and 20.)

Sham M Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. (Cited on pages 1, 3, 18, and 44.)

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

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)*, pp. 3794–3799. IEEE Press, 2021b. (Cited on pages 1 and 16.)

Sajad Khodadadian, Thinh T. Doan, Justin Romberg, and Siva Theja Maguluri. Finite sample analysis of two-time-scale natural actor-critic algorithm. *IEEE Transactions on Automatic Control*, pp. 1–16, 2022a. (Cited on page 16.)

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, 2022b. ISSN 0167-6911. (Cited on page 16.)

Guanghui Lan. Policy mirror descent for reinforcement learning: linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, Apr 2022. ISSN 1436-4646. (Cited on pages 1, 4, 16, 18, and 47.)

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7031–7043, 2020. (Cited on page 8.)

Tianjiao Li, Feiyang Wu, and Guanghui Lan. Stochastic first-order methods for average-reward markov decision processes, 2022a. (Cited on pages 16 and 18.)

Yan Li, Tuo Zhao, and Guanghui Lan. Homotopic policy mirror descent: Policy convergence, implicit regularization, and improved sample complexity, 2022b. (Cited on page 16.)

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. (Cited on page 1.)

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, pp. 7624–7636. Curran Associates, Inc., 2020. (Cited on pages 9, 16, 17, 43, 44, and 45.)

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

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*, pp. 6820–6829. PMLR, 13–18 Jul 2020. (Cited on page 17.)

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

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*, pp. 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR. (Cited on page 1.)

Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pp. 560–567. AAAI Press, 2003. ISBN 1577351894. (Cited on page 6.)

Rémi Munos. Error bounds for approximate value iteration. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*, AAAI'05, pp. 1006–1011. AAAI Press, 2005. ISBN 157735236x. (Cited on page 6.)

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008. (Cited on page 6.)

A Nemirovski and D Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983. (Cited on page 1.)

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes, 2017. (Cited on page 1.)

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4026–4035. PMLR, 2018. (Cited on page 44.)

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7–9):1180–1190, mar 2008. ISSN 0925-2312. (Cited on page 44.)

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

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, Inc., USA, 1994. ISBN 0471619779. (Cited on pages 1 and 48.)

R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970. (Cited on page 49.)

Bruno Scherrer. Approximate policy iteration schemes: A comparison. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. II–1314–II–1322. JMLR.org, 2014. (Cited on page 47.)

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*, pp. 1889–1897, Lille, France, 07–09 Jul 2015. PMLR. (Cited on page 1.)

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. (Cited on page 1.)

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014. ISBN 978-1-10-705713-5. (Cited on pages 8 and 35.)

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*, pp. 5668–5675, 2020. (Cited on pages 1, 4, and 16.)

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. (Cited on page 1.)

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*, pp. 1057–1063. MIT Press, 2000. (Cited on pages 1, 3, and 19.)

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*, pp. 8619–8649. PMLR, 28–30 Mar 2022. (Cited on page 1.)

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2020. (Cited on pages 5, 6, and 17.)

Xiao Wang, Shiqian Ma, Donald Goldfarb, and Wei Liu. Stochastic quasi-newton methods for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 27(2):927–956, 2017. doi: 10.1137/15M1053141. (Cited on page 44.)

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. (Cited on pages 1 and 3.)

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022. (Cited on pages 1, 4, 5, 9, 15, 16, 17, 18, 26, 46, 47, 48, and 49.)

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

Lin Yang and Mengdi Wang. Sample-Optimal Parametric Q-Learning Using Linearly Additive Features. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6995–7004. PMLR, 2019. (Cited on page 6.)

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*, pp. 3332–3380. PMLR, 28–30 Mar 2022. (Cited on pages 17 and 44.)

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

Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D. Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence, 2021. (Cited on pages 1 and 16.)

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4572–4583. Curran Associates, Inc., 2020. (Cited on pages 5 and 17.)

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

Appendix

Table of Contents

A Related Work 15

B Standard Reinforcement Learning Results 18

C Algorithms 20

D Proof of Section 4 26

E Proof of Section 5 38

F Sublinear Convergence of NPG with Constant step size 42

G Sample Complexity of NPG 44

H Discussion on the Distribution Mismatch Coefficients, the Concentrability Coefficients and the Increasing Step Size 46

I Standard Optimization Results 48

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

A Related Work

A.1 Technical Contribution and Novelty Compared to Xiao (2022)

Our technical novelty compared to Xiao (2022) is summarized as follows.

- Our linear convergence results (i.e., Theorem 1, 3 and 4) are not direct applications of Theorem 10 in Xiao (2022). Indeed, Xiao (2022) establishes the connection between NPG and a specific form of policy mirror descent (PMD) with the use of the weighted Bregman divergence for the tabular setting, while we show that this connection can also be established for the function approximation setting via the compatible function approximation framework (11). We also modify the PMD framework of Xiao (2022) with the linear approximation of the advantage function in (18), inspired from the compatible function approximation framework. Thus, the approaches of deriving the PMD form update are different. Without this work of using the compatible function approximation framework to bridge NPG and PMD, it was not clear at all that the analysis of Xiao (2022) could be extended in the log-linear policy setting. So our work is the first step of showing that the proof techniques used in Xiao (2022) can be extended in function approximation regime. In fact, the extension is highly nontrivial and requires significant innovation (see details below). As for future work, one can extend our work to other function approximation setting through a similar compatible function approximation framework. See Appendix A.3 for more details about the future work.
- Besides, our linear convergence results only consider the inexact NPG update. Compared to Theorem 14 in Xiao (2022), which is their corresponding result on the inexact PMD method, we improve their analysis by making much weaker assumptions on the accuracy of the estimation $Q(\pi)$. Xiao (2022) requires an L_∞ supremum norm bound on the estimation error of Q, i.e., $\|\widehat{Q}(\pi) - Q(\pi)\|_\infty \leq \epsilon_{\text{stat}}$, whereas our convergence guarantee depends on the expected L_2 error of the estimate, i.e., Assumption 1 and 7. For instance, Assumption 1 from equation (63) can be written as $\mathbb{E}\left[(\phi_{s,a}^\top w^{(k)} - \phi_{s,a}^\top w_\star^{(k)})^2\right] \leq \epsilon_{\text{stat}}$, which can

be interpreted as $\mathbb{E}\left[(\widehat{Q}(\pi) - Q(\pi))^2\right] \leq \epsilon_{\text{stat}}$ under the linear approximation setting. The techniques for handling L_∞ and L_2 errors are very different. Not only our assumption is weaker, it also benefits from the sample complexity analysis that we explain next.

- Consequently, when considering the sample complexity results we derived for sample-based (Q)-NPG in Corollary 1 and 2, the difference between our work and Theorem 16 in Xiao (2022), which corresponds to their sample complexity results, is even more significant. Corollary 1 with Algorithm `Q-NPG-SGD` (Algorithm 6) satisfies Assumption 1 with a number of samples that depends only on the feature dimension m of ϕ and does not depend on the cardinality of state space $|\mathcal{S}|$ or action space $|\mathcal{A}|$. In contrast, the assumption $\|\widehat{Q}(\pi) - Q(\pi)\|_\infty \leq \epsilon_{\text{stat}}$ with the L_∞ norm in Xiao (2022, Theorem 16) causes the sample complexity to depend on $|\mathcal{S}||\mathcal{A}|$.

  Furthermore, Xiao (2022) uses a Monte-Carlo approach with multiple independent rollouts per iteration, while our sample-based (Q)-NPG uses one single rollout (Algorithm 3 and 4) combined with regression solvers; Xiao (2022) derives a high probability sample complexity result, while we derive the convergence of the optimality gap $\mathbb{E}\left[V_\rho(\pi^{(K)})\right] - V_\rho(\pi^*)$ which can guarantee that the variance of $V_\rho(\pi^{(K)})$ converges to zero. Thus, our sample-based algorithms had not been considered in Xiao (2022) and our proofs of Corollary 1 and 2 require a different approach.

  In particular, our sample complexity analysis regarding to the policy evaluation is novel. Although our sample-based algorithms had been considered previously in Agarwal et al. (2021) and Liu et al. (2020), none of their analysis on the sample complexity was correct. Indeed, Agarwal et al. (2021) required the boundedness of the stochastic gradient estimator, which might not hold as we extensively discussed in Appendix D.5. We fixed this by showing that $\mathbb{E}\left[\widehat{Q}_{s,a}(\theta)^2\right]$ is bounded. See Appendix D.5 for all the subtleties, including a proof sketch of Corollary 1. Liu et al. (2020) also incorrectly used an inequality where the random variables are correlated. See the detailed explanation (Footnote 7) in Appendix G.1. We fixed this error with a careful conditional expectation argument. Please refer to Appendix G.1 for all the details, including a proof sketch of Corollary 2. These dimensions are where an important part of the technical work was done. Therefore, outside of the tabular setting, and considering NPG methods that make use of a regression solver, our complexity analysis is currently the only analysis that is entirely correct that we are aware of.
- Finally we not only extend the work of Xiao (2022) to NPG for log-linear policy, but also consider the Q-NPG method and establish its linear convergence analysis. This is a method that is unique to log-linear policy and again had not been considered in Xiao (2022).

A.2 Finite-Time Analysis of the Natural Policy Gradient

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 of the unregularized NPG to $\mathcal{O}(1/k)$ with constant step sizes. Further, Khodadadian et al. (2021a) also achieves $\mathcal{O}(1/k)$ convergence rate for the off-policy natural actor-critic (NAC), and a slower sublinear result is established by Khodadadian et al. (2022a) for the two-time-scale NAC.

By using the entropy regularization, Cen et al. (2021) achieve a linear convergence rate for NPG. A similar linear convergence result has been 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). Such approach is also applied in the averaged MDP setting to achieve linear convergence for NPG (Li et al., 2022a). However, adding regularization might induce bias for the solution. Thus, Lan (2022) considers exponentially diminishing regularization to guarantee unbiased solution. Furthermore, by considering both the KL divergence and the diminishing entropy regularization, Li et al. (2022b) establish the linear convergence rate not only for the optimality gap but also for the policy. That is, the policy will converge to the fixed high entropy optimal policy. Consequently, Li et al. (2022b) show a local super-linear convergence of both the policy and optimality gap, as discussed in Xiao (2022, Section 4.3).

Recently, Bhandari & Russo (2021), Khodadadian et al. (2021b; 2022b) 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 & Russo (2021) propose to use an exact line search for the step size (Theorem 1 (a)) or to choose an adaptive step size (Theorem 1 (c)). Similar adaptive step size is proposed by Khodadadian et al. (2021b; 2022b). Notice that such adaptive step size requires complete knowledge about the environmental model. Instead, a sufficiently large step size might be

enough. In this paper, we extend the results of Xiao (2022) from the tabular setting to the log-linear policies, using *non-adaptive* geometrically increasing step size and obtaining a linear convergence rate for NPG without regularization.

NPG with function approximation. In the function approximation regime, there have been many works investigating the convergence rate of the NPG or NAC algorithms from different perspectives. Wang et al. (2020) establish the $\mathcal{O}(1/\sqrt{k})$ convergence rate for two-layer neural NAC with a projection step. The sublinear convergence results are also established by Zanette et al. (2021) and Hu et al. (2022) for the linear MDP (Jin et al., 2020). Agarwal et al. (2021) obtain the same $\mathcal{O}(1/\sqrt{k})$ convergence rate for the smooth policies with projections. This was 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, and it was also improved to $\mathcal{O}(1/k)$ by Xu et al. (2020) with NAC under Markovian sampling. The same $\mathcal{O}(1/k)$ convergence rate is established for log-linear policies by Chen et al. (2022) when considering the off-policy NAC.

With entropy regularization and a projection step, Cayci et al. (2021) obtain a linear convergence for log-linear policies. Same entropy regularization and a projection step are applied by Cayci et al. (2022) for the neural NAC to improve the $\mathcal{O}(1/\sqrt{k})$ convergence rate of Wang et al. (2020) to $\mathcal{O}(1/k)$. 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 additional regularization nor a projection step. We notice that Chen & Theja Maguluri (2022, Theorem 3.4)[4] also uses increasing step size and achieves linear convergence for log-linear policies without regularization. The main differences between our result and Theorem 3.4 in Chen & Theja Maguluri (2022) are fourfold. First, they rely on the contraction property of the generalized Bellman operator, while we consider the PMD analysis approach. So the proof techniques are completely different. Second, their parameter update results in the off-policy multi-step temporal difference learning, whereas we require to solve a linear regression problem to minimize the function approximation error. Third, their step size still depends on the iterates which is thus an adaptive step size and is proportional to the total number of iterations K, while ours is independent to the iterates nor to K. Finally, their assumption on the modeling error requires an L_∞ supremum norm, i.e., $\|Q_s(\theta^{(k)}) - \Phi w_\star^{(k)}\|_\infty \leq \epsilon_{\text{bias}}$ for all states s of the state space, our convergence guarantee depends on the expected error (e.g., Assumption 2, 5 or 8) which is a much weaker assumption. After our submission, we are aware of the concurrent work of Alfano & Rebeschini (2022). They only analyze the Q-NPG method and achieve similar linear convergence results as our Theorem 1. In particular, their result in Theorem 4.7 has a better concentrability coefficient compared to our Theorem 1. However, their Assumption 4.6 assumes that the relative condition number upper bounds a time-varying ratio which depends on the iterates, while our Assumption 3 is independent to the iterates, as defined in (25). Furthermore, they only consider the case when the initial state distribution is the same as the target state distribution, while our analysis generalizes with any target state distribution, which is extensively discussed on the distribution mismatch coefficients in Appendix H.

Fast linear convergence of other policy gradient methods. Different to the PMD analysis approach, by leveraging a gradient dominance property (Polyak, 1963; Łojasiewicz, 1963), fast linear convergence results have also been established for 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). Such gradient domination property is widely explored by Bhandari & Russo (2019) to identify more general structural MDP settings. Linear convergence of PG can also be obtained through exact line search (Bhandari & Russo, 2021, Theorem 1 (a)) or by exploiting non-uniform smoothness (Mei et al., 2021).

Alternatively, by considering a general strongly-concave utility function of the state-action occupancy measure and by exploiting the hidden convexity of the problem, Zhang et al. (2020) also achieve the linear convergence of a variational PG method. When the object is relaxed to a general concave utility function, Zhang et al. (2021) still achieve the linear convergence by leveraging the hidden convexity of the problem and by adding variance reduction to the PG method.

A.3 Future Work

The main focus of this paper was the theoretical analysis of NPG method. The results we have obtained open up several experimental questions related to parameter settings for NPG and Q-NPG. We leave such questions as an important future work to further support our theoretical findings.

An interesting application from our work is to investigate the sample complexity of natural actor-critic with our PMD analysis. Indeed, our paper obtains $w^{(k)}$ by a regression solver. One can also use temporal difference (TD) learning

[4] This result appears after conference proceedings and is available on https://arxiv.org/pdf/2208.03247.pdf.

(e.g., Cayci et al. (2021); Chen & Theja Maguluri (2022)) with Markovian sampling to achieve similar $O(1/\epsilon^2)$ sample complexity result. The performance analysis of TD learning will be expressed for ϵ_{stat}, which directly imply the total sample complexity results through our theorems.

One natural question is whether we can extend our analysis to the general policy classes. Here we provide one possible way. It can be extended by using a similar compatible function approximation framework. Concretely, consider the parameterized policy

$$\pi_{s,a}(\theta) = \frac{\exp(f_{s,a}(\theta))}{\sum_{a' \in \mathcal{A}} \exp(f_{s,a'}(\theta))},$$

where $f_{s,a}(\theta)$ is parameterized by $\theta \in \mathbb{R}^m$ and is differential. As Agarwal et al. (2021) mentioned, the gradient can be written as

$$\nabla_\theta \log \pi_{s,a}(\theta) = g_{s,a}(\theta) \quad \text{where} \quad g_{s,a}(\theta) = \nabla_\theta f_{s,a}(\theta) - \mathbb{E}_{a' \in \pi_s(\theta)} \left[\nabla_\theta f_{s,a'}(\theta) \right].$$

The NPG update is equivalent to the following compatible function approximation framework

$$\theta^{(k+1)} = \theta^{(k)} - \eta_k w_\star^{(k)}, \qquad w_\star^{(k)} \in \arg\min_w \mathbb{E}_{(s,a)\in \bar{d}^{(k)}} \left[\left(A_{s,a}(\theta^{(k)}) - w^\top g_{s,a}(\theta^{(k)}) \right)^2 \right].$$

As Alfano & Rebeschini (2022, Remark 4.8) mentioned, if we assume that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, function $f(\theta)$ satisfies

$$f_{s,a}(\theta^{(k+1)}) = f_{s,a}(\theta^{(k)}) - \eta_k (w_\star^{(k)})^\top g_{s,a}(\theta^{(k)}),$$

then one can easily verify that the NPG update resulted in a new policy is also equivalent to the policy mirror descent update

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

where $G_s^{(k)} \in \mathbb{R}^{|\mathcal{A}| \times m}$ is a matrix with rows $(g_{s,a}(\theta^{(k)}))^\top \in \mathbb{R}^{1 \times m}$ for $a \in \mathcal{A}$. Consequently, one can extend our work naturally in this general setting to derive linear convergence analysis for NPG.

Perhaps one can consider the *exponential tilting*, a generalization of Softmax to more general probability distributions. Another interesting venue of investigation is to consider the *generalized linear model* instead of linear function approximation for the Q function and the advantage function.

One interesting open question is that is there a way to increase stepsize when the discount factor is unknown? So far the PMD proof techniques used in Lan (2022); Xiao (2022) and ours require that the discount factor is known. Perhaps the work of Li et al. (2022a) can help to find a way to increase stepsize when the discount factor is unknown. Indeed, Li et al. (2022a) consider the averaged MDP setting. So there is no discount factor. They achieve linear convergence for NPG by increasing the stepsize with some regularization parameters. It will be interesting to investigate if the way of increasing stepsize in Li et al. (2022a) can be applied in our setting.

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 (12) and the NPG updates formalized as policy mirror descent ((17) and (18)). Then, we prove the performance difference lemma (Kakade & Langford, 2002), which is the first key ingredient for our PMD analysis. The three-point descent lemma (Lemma 11) is the second key ingredient for our PMD analysis.

Lemma 1 (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 (12)

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

where $w_\star^{(k)} \in \operatorname{argmin}_{w \in \mathbb{R}^m} L_A\big(w, \theta^{(k)}, \bar{d}^{(k)}\big)$*. 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, using the policy gradient (8) and 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)

$$\nabla_\theta V_\rho(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\theta,\, a\sim\pi_s(\theta)}\left[A_{s,a}(\theta)\,\nabla_\theta \log \pi_{s,a}(\theta)\right]. \tag{37}$$

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

$$\begin{aligned}
& \nabla_w L_A(w_\star^{(k)}, \theta^{(k)}, \bar{d}^{(k)}) = 0 \\
\iff\quad & \mathbb{E}_{(s,a)\sim\bar{d}^{(k)}}\left[\left((w_\star^{(k)})^\top \nabla_\theta \log \pi_{s,a}^{(k)} - A_{s,a}(\theta^{(k)})\right)\nabla_\theta \log \pi_{s,a}^{(k)}\right] = 0 \\
\iff\quad & \mathbb{E}_{(s,a)\sim\bar{d}^{(k)}}\left[\nabla_\theta \log \pi_{s,a}^{(k)}\left(\nabla_\theta \log \pi_{s,a}^{(k)}\right)^\top\right] w_\star^{(k)} = \mathbb{E}_{(s,a)\sim\bar{d}^{(k)}}\left[A_{s,a}(\theta^{(k)})\nabla_\theta \log \pi_{s,a}^{(k)}\right] \\
\overset{(9)+(37)}{\iff}\quad & F_\rho(\theta^{(k)})w_\star^{(k)} = (1-\gamma)\nabla_\theta V_\rho(\theta^{(k)}).
\end{aligned}$$

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

$$\begin{aligned}
\pi_s^{(k+1)} &= \pi_s^{(k)} \odot \frac{\exp\left(-\eta_k \Phi_s w^{(k)}\right)}{\sum_{a\in\mathcal{A}} \pi_{s,a}^{(k)} \exp\left(-\eta_k \phi_{s,a}^\top w^{(k)}\right)} && (38)\\
&= \pi_s^{(k)} \odot \frac{\exp\left(-\eta_k \bar{\Phi}_s^{(k)} w^{(k)}\right)}{\sum_{a\in\mathcal{A}} \pi_{s,a}^{(k)} \exp\left(-\eta_k \left(\bar{\phi}_{s,a}(\theta^{(k)})\right)^\top w^{(k)}\right)} && (39)\\
&= \arg\min_{p\in\Delta(\mathcal{A})} \left\{\eta_k \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, p\right\rangle + D(p, \pi_s^{(k)})\right\}, \quad \forall s\in\mathcal{S}, && (40)
\end{aligned}$$

where $\odot$ is the element-wise product between vectors, and $\bar{\Phi}_s^{(k)} \in \mathbb{R}^{|\mathcal{A}|\times m}$ is defined in (18)*, i.e.*

$$\left(\bar{\Phi}_{s,a}^{(k)}\right)^\top \overset{def}{=} \bar{\phi}_{s,a}(\theta^{(k)}) \overset{(13)}{=} \phi_{s,a} - \mathbb{E}_{a'\sim\pi_s^{(k)}}\left[\phi_{s,a'}\right].$$

Such policy update coincides the inexact NPG updates (33) *of the log-linear policy, if $\theta^{(k+1)} = \theta^{(k)} - \eta_k w^{(k)}$ with $w^{(k)} \approx \operatorname{argmin}_w L_A(w, \theta^{(k)}, \tilde{d}^{(k)})$; and coincides the inexact Q-NPG updates* (19) *of the log-linear policy, if $\theta^{(k+1)} = \theta^{(k)} - \eta_k w^{(k)}$ with $w^{(k)} \approx \operatorname{argmin}_w L_Q(w, \theta^{(k)}, \tilde{d}^{(k)})$.*

Proof. For shorthand, let $g = \Phi_s w^{(k)}$. Thus, (17) fits the format of Lemma 10 in Appendix I where $q = \pi_s^{(k)}$. Consequently, the closed form solution is given by (99), that is

$$\begin{aligned}
\pi_s^{(k+1)} &= \frac{\pi_s^{(k)} \odot e^{-\eta_k g}}{\sum_{a\in\mathcal{A}} \pi_{s,a}^{(k)} e^{-\eta_k g_a}} = \frac{\pi_s^{(k)} \odot e^{-\eta_k \Phi_s w^{(k)}}}{\sum_{a\in\mathcal{A}} \pi_{s,a}^{(k)} e^{-\eta_k \phi_{s,a}^\top w^{(k)}}} \\
&= \pi_s^{(k)} \odot \frac{\exp\left(-\eta_k \bar{\Phi}_s(\theta^{(k)}) w^{(k)}\right)}{\sum_{a\in\mathcal{A}} \pi_{s,a}^{(k)} \exp\left(-\eta_k \left(\bar{\phi}_{s,a}(\theta^{(k)})\right)^\top w^{(k)}\right)}, && (41)
\end{aligned}$$

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 10 with $g = \bar{\Phi}_s^{(k)} w^{(k)}$, the closed form solution to (40) is (41).

As for the closed form updates of the policy for NPG (33) and Q-NPG (19) with the parameter updates $\theta^{(k+1)} = \theta^{(k)} - \eta_k w^{(k)}$, it is straightforward to verify that it coincides (38) and (39) given the specific structure of the log-linear policy (7), which concludes the proof. □

Lemma 3 (Performance difference lemma (Kakade & Langford, 2002)). *For any policy $\pi, \pi' \in \Delta(\mathcal{A})^{\mathcal{S}}$ and $\rho \in \Delta(\mathcal{S})$,*

$$V_\rho(\pi) - V_\rho(\pi') = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\bar{d}^\pi}\left[A_{s,a}(\pi')\right] \tag{42}$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi}\left[\langle Q_s(\pi'), \pi_s - \pi'_s\rangle\right], \tag{43}$$

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 (43), it suffices to show

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

Let $\mathbf{1}_n$ denote 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. □

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.

Algorithm 1: Natural policy gradient

Input: Initial state-action distribution ν, 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 (33) 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 η_k

Output: $\pi^{(K)}$

Similarly, 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 (33) or (19), 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 $(s_0, a_0, c_0, s_1, \cdots)$ to compute the values of $Q_{s,a}^{(k)}$ and $A_{s,a}^{(k)}$. 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 (19), we sample $(s,a) \sim \tilde{d}^{(k)}$ and $\widehat{Q}_{s,a}^{(k)}$ by a standard rollout, formalized in Algorithm 3. This sampling procedure is commonly used, for example in Agarwal et al. (2021, Algorithm 1).

Algorithm 2: Q-Natural policy gradient

Input: Initial state-action distribution ν, 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 (19) 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 η_k

Output: $\pi_{\theta^{(K)}}$

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 ν, 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 γ**:**
    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 γ**:**
    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 completeness.

Lemma 4. *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

$$\widehat{Q}_{s_h,a_h}(\theta) = \sum_{t=0}^{H} c(s_{t+h}, a_{t+h}), \tag{44}$$

where $(H+1)$ is the length of the horizon executed between lines 13 and 19 in Algorithm 3 for calculating $\widehat{Q}_{s_h,a_h}(\theta)$. To simplify notation, we consider the estimate of $\widehat{Q}_{s,a}$ for any $(s, a) \in \mathcal{S}\times\mathcal{A}$ following the same procedure starting from line 10 in Algorithm 3. Taking expectation, we have

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

The desired result is obtained by setting $s = s_h$ and $a = a_h$. □

Similar to Algorithm 3, to solve (33), 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 ν, 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 γ**:**
    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 γ**:**
    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 γ**:**
    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, since Algorithm 4 requires an additional estimation of $V_s(\theta)$ and thus doubles the number of samples to estimate $A_{s,a}(\theta)$. As in Lemma 4, we verify in the following lemma that the output (s_h, a_h) is sampled from the distribution $\tilde{d}^{\theta}$ 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 5. *Consider the output (s_h, a_h) and $\widehat{A}_{s_h,a_h}(\theta)$ of Algorithm 4. It follows that*

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

Proof. Since the procedure of sampling (s_h, a_h) in Algorithm 4 is identical to the one in Algorithm 3, from Lemma 4, 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 4, 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 4, 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 & Moulines (2013) to solve the regression problem (33) 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 (33)

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

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 (33) by

$$\widehat{\nabla}_w L_A(w, \theta, \tilde{d}^{\theta}) \overset{\text{def}}{=} 2\left(w^{\top}\bar{\phi}_{s,a}(\theta) - \widehat{A}_{s,a}(\theta)\right)\bar{\phi}_{s,a}(\theta). \tag{45}$$

Next, we show that (45) is an unbiased gradient estimator of the loss function L_A.

Lemma 6. *Consider the output (s, a) and $\widehat{A}_{s,a}(\theta)$ of Algorithm 4 and the stochastic gradient (45). 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

$$\begin{aligned}\mathbb{E}\left[\widehat{\nabla}_w L_A(w, \theta, \tilde{d}^{\theta})\right] &\overset{(45)}{=} \mathbb{E}_{s,\,a,\,\widehat{A}_{s,a}(\theta)}\left[2\left(w^{\top}\bar{\phi}_{s,a}(\theta) - \widehat{A}_{s,a}(\theta)\right)\bar{\phi}_{s,a}(\theta)\right] \\ &= \mathbb{E}_{(s,a)\sim\tilde{d}^{\theta},\,\widehat{A}_{s,a}(\theta)}\left[2\left(w^{\top}\bar{\phi}_{s,a}(\theta) - \widehat{A}_{s,a}(\theta)\right)\bar{\phi}_{s,a}(\theta) \mid s, a\right],\end{aligned} \tag{46}$$

where the second line is obtained by $(s,a) \sim \tilde{d}^\theta$ from Lemma 5.

From Lemma 5, we have

$$\mathbb{E}_{s_0,a_0,\cdots,s_t,a_t}\left[\widehat{A}_{s,a}(\theta) \mid s_0 = s, a_0 = a\right] = A_{s,a}(\theta). \tag{47}$$

Combining the above two equalities yield

$$\begin{aligned}
\mathbb{E}\left[\widehat{\nabla}_w L_A(w,\theta,\tilde{d}^\theta)\right] &\overset{(46)}{=} \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[2\left(w^\top \bar{\phi}_{s,a}(\theta) - \mathbb{E}\left[\widehat{A}_{s,a}(\theta) \mid s,a\right]\right)\bar{\phi}_{s,a}(\theta)\right] \\
&\overset{(47)}{=} \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[2\left(w^\top \bar{\phi}_{s,a}(\theta) - A_{s,a}(\theta)\right)\bar{\phi}_{s,a}(\theta)\right] \\
&= \nabla_w L_A(w,\theta,\tilde{d}^\theta),
\end{aligned}$$

as desired. □

Since (45) is unbiased shown in Lemma 6, 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 ν, 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 (45)

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 (19) of Q-NPG.

Recall the compatible function approximation L_Q in (19)

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

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 (19) by

$$\widehat{\nabla}_w L_Q(w,\theta,\tilde{d}^\theta) \overset{\text{def}}{=} 2\left(w^\top \phi_{s,a} - \widehat{Q}_{s,a}(\theta)\right)\phi_{s,a}, \tag{48}$$

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). Compared to (45), the cost of computing (48) is $|\mathcal{A}|$ times cheaper than that of computing (48). Indeed, to compute (48), we only need one single action for $\phi_{s,a}$, while to compute (45), one needs to go through all the actions to compute $\bar{\phi}_{s,a}(\theta)$. Thus, the computational cost of `Q-NPG-SGD` is $|\mathcal{A}|$ times cheaper than that of `NPG-SGD`.

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 ν, 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 (48)

Output: $w_{\text{out}} = \frac{1}{T}\sum_{t=1}^T w_t$

The estimator $\widehat{\nabla}_w L_Q(w,\theta,\tilde{d}^\theta)$ is also unbiased following the similar argument of the proof of Lemma 6. We formalize this in the following and omit the proof.

Lemma 7. *Consider the output (s,a) and $\widehat{Q}_{s,a}(\theta)$ of Algorithm 3 and the stochastic gradient (48). 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.

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 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

$$D_0^* \le \log|\mathcal{A}|, \tag{49}$$

for all $\rho\in\Delta(\mathcal{S})$ and for any $\pi^*\in\Delta(\mathcal{A})^{\mathcal{S}}$.

We first provide the one step analysis of the Q-NPG update, which will be helpful for proving Theorem 1 and 3 and the sublinear convergence result of Theorem 2 further in Appendix D.3.

D.1 The One Step Q-NPG Lemma

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

Lemma 8 (One step Q-NPG lemma)**.** *Fix a state distribution ρ; an initial state-action distribution ν; an arbitrary comparator policy π^*. Let $w_\star^{(k)} \in \operatorname{argmin}_w L_Q(w,\theta^{(k)},\tilde{d}^{(k)})$ denote the exact minimizer. Consider the $w^{(k)}$ and $\pi^{(k)}$ given in (19) and (17) respectively. We have that*

$$\begin{aligned}
&\vartheta_\rho(1-\gamma)\left(V_\rho^{(k+1)} - V_\rho^{(k)}\right) + (1-\gamma)\left(V_\rho^{(k)} - V_\rho(\pi^*)\right)\\
&+\vartheta_\rho\Bigg(\underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k+1)}\phi_{s,a}^\top\left(w^{(k)} - w_\star^{(k)}\right)}_{①} + \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k+1)}\left(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right)}_{②}\\
&+\underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k)}\phi_{s,a}^\top\left(w_\star^{(k)} - w^{(k)}\right)}_{③} + \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k)}\left(Q_{s,a}^{(k)} - \phi_{s,a}^\top w_\star^{(k)}\right)}_{④}\Bigg)\\
&+\underbrace{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^*\pi_{s,a}^{(k)}\phi_{s,a}^\top\left(w^{(k)} - w_\star^{(k)}\right)}_{ⓐ} + \underbrace{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^*\pi_{s,a}^{(k)}\left(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right)}_{ⓑ}\\
&+\underbrace{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^*\pi_{s,a}^*\phi_{s,a}^\top\left(w_\star^{(k)} - w^{(k)}\right)}_{ⓒ} + \underbrace{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^*\pi_{s,a}^*\left(Q_{s,a}^{(k)} - \phi_{s,a}^\top w_\star^{(k)}\right)}_{ⓓ}\\
\le\ &\frac{1}{\eta_k}D_k^* - \frac{1}{\eta_k}D_{k+1}^*.
\end{aligned} \tag{50}$$

Proof. In the context of the PMD method (17), we apply the three-point descent lemma (Lemma 11) 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 11, 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

$$\left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - p \right\rangle + \frac{1}{\eta_k} D(\pi_s^{(k+1)}, \pi_s^{(k)}) \leq \frac{1}{\eta_k} D(p, \pi_s^{(k)}) - \frac{1}{\eta_k} D(p, \pi_s^{(k+1)}). \tag{51}$$

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

$$\left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \leq -\frac{1}{\eta_k} D(\pi_s^{(k+1)}, \pi_s^{(k)}) - \frac{1}{\eta_k} D(\pi_s^{(k)}, \pi_s^{(k+1)}) \leq 0. \tag{52}$$

Letting $p = \pi_s^*$ and subtract and add $\pi_s^{(k)}$ within the inner product term in (51) 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

$$\mathbb{E}_{s \sim d^*} \left[\left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \right] + \mathbb{E}_{s \sim d^*} \left[\left\langle \Phi_s w^{(k)}, \pi_s^{(k)} - \pi_s^* \right\rangle \right] \leq \frac{1}{\eta_k} D_k^* - \frac{1}{\eta_k} D_{k+1}^*. \tag{53}$$

For the first expectation in (53), we have

$$\begin{aligned}
& \mathbb{E}_{s \sim d^*} \left[\left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \right] \\
= & \sum_{s \in \mathcal{S}} d_s^* \left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
= & \sum_{s \in \mathcal{S}} \frac{d_s^*}{d_s^{(k+1)}} d_s^{(k+1)} \left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
\geq & \vartheta_{k+1} \sum_{s \in \mathcal{S}} d_s^{(k+1)} \left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
\geq & \vartheta_{\rho} \sum_{s \in \mathcal{S}} d_s^{(k+1)} \left\langle \Phi_s w^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
= & \vartheta_{\rho} \sum_{s \in \mathcal{S}} d_s^{(k+1)} \left\langle Q_s^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle + \vartheta_{\rho} \sum_{s \in \mathcal{S}} d_s^{(k+1)} \left\langle \Phi_s w^{(k)} - Q_s^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
= & \vartheta_{\rho} (1 - \gamma) \left(V_{\rho}^{(k+1)} - V_{\rho}^{(k)} \right) + \vartheta_{\rho} \sum_{s \in \mathcal{S}} d_s^{(k+1)} \left\langle \Phi_s w^{(k)} - Q_s^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle,
\end{aligned} \tag{54}$$

where the last equality is due to the performance difference lemma (43) in Lemma 3 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 (52) and by using the following inequality

$$\frac{d_s^*}{d_s^{(k+1)}} \overset{(21)}{\leq} \vartheta_{k+1} \overset{(21)}{\leq} \vartheta_{\rho}.$$

The second term of (54) can be decomposed into four terms. That is,

$$
\begin{aligned}
&\sum_{s\in\mathcal{S}} d_s^{(k+1)} \left\langle \Phi_s w^{(k)} - Q_s^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k+1)} \left(\phi_{s,a}^\top w^{(k)} - Q_{s,a}^{(k)}\right) + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k)} \left(Q_{s,a}^{(k)} - \phi_{s,a}^\top w^{(k)}\right) \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k+1)} \phi_{s,a}^\top \left(w^{(k)} - w_\star^{(k)}\right) + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k+1)} \left(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right) \\
&\quad + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k)} \phi_{s,a}^\top \left(w_\star^{(k)} - w^{(k)}\right) + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k)} \left(Q_{s,a}^{(k)} - \phi_{s,a}^\top w_\star^{(k)}\right) \\
&= ① + ② + ③ + ④,
\end{aligned} \tag{55}
$$

where ①, ②, ③ and ④ are defined in (50).

For the second expectation in (53), by applying again the performance difference lemma (43), we have

$$
\begin{aligned}
&\mathbb{E}_{s\sim d^*}\left[\left\langle \Phi_s w^{(k)}, \pi_s^{(k)} - \pi_s^* \right\rangle\right] \\
=\ &\mathbb{E}_{s\sim d^*}\left[\left\langle Q_s^{(k)}, \pi_s^{(k)} - \pi_s^* \right\rangle\right] + \mathbb{E}_{s\sim d^*}\left[\left\langle \Phi_s w^{(k)} - Q_s^{(k)}, \pi_s^{(k)} - \pi_s^* \right\rangle\right] \\
\overset{(43)}{=}\ &(1-\gamma)\left(V_\rho^{(k)} - V_\rho(\pi^*)\right) + \mathbb{E}_{s\sim d^*}\left[\left\langle \Phi_s w^{(k)} - Q_s^{(k)}, \pi_s^{(k)} - \pi_s^* \right\rangle\right].
\end{aligned} \tag{56}
$$

Similarly, we decompose the second term of (56) into four terms. That is,

$$
\begin{aligned}
&\mathbb{E}_{s\sim d^*}\left[\left\langle \Phi_s w^{(k)} - Q_s^{(k)}, \pi_s^{(k)} - \pi_s^* \right\rangle\right] \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^* \pi_{s,a}^{(k)} \left(\phi_{s,a}^\top w^{(k)} - Q_{s,a}^{(k)}\right) + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^* \pi_{s,a}^* \left(Q_{s,a}^{(k)} - \phi_{s,a}^\top w^{(k)}\right) \\
&= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \pi_{s,a}^{(k)} \phi_{s,a}^\top \left(w^{(k)} - w_\star^{(k)}\right) + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \pi_{s,a}^{(k)} \left(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right) \\
&\quad + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \pi_{s,a}^* \phi_{s,a}^\top \left(w_\star^{(k)} - w^{(k)}\right) + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \pi_{s,a}^* \left(Q_{s,a}^{(k)} - \phi_{s,a}^\top w_\star^{(k)}\right) \\
&= ⓐ + ⓑ + ⓒ + ⓓ,
\end{aligned} \tag{57}
$$

where ⓐ, ⓑ, ⓒ and ⓓ are defined in (50).

Plugging (54) with the decomposition (55) and (56) with the decomposition (57) into (53) concludes the proof. □

Consequently, the convergence analysis of Q-NPG (Theorem 1, 3 and Theorem 2 further in Appendix D.3) will be obtained by upper bounding the absolute values of ①, ②, ③, ④, ⓐ, ⓑ, ⓒ, ⓓ in (50) with different set of assumptions (assumptions in Theorem 1 or assumptions in Theorem 3) and with different step size scheme (geometrically increasing step size for Theorem 1 and 3 or constant step size for Theorem 2).

D.2 Proof of Theorem 1

Here we provide the general Q-NPG convergence analysis with any initial policy, which is in contrast to Theorem 1 with the uniform initial policy.

Theorem 5. *Fix a state distribution ρ, an state-action distribution ν and a comparator policy π^*. We consider the Q-NPG method (19) 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 $k \geq 0$,*

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

Remark. A deviation from the setting of Theorem 1 is that here we require the step sizes satisfying $\eta_0 \geq \frac{1-\gamma}{\gamma} D_0^*$ instead of $\eta_0 \geq \frac{1-\gamma}{\gamma} \log |\mathcal{A}|$. This is because when using uniform initial policy in Theorem 1, from (49), we know that $\eta_0 \geq \frac{1-\gamma}{\gamma} \log |\mathcal{A}|$ is a sufficient condition for $\eta_0 \geq \frac{1-\gamma}{\gamma} D_0^*$. Consequently, to prove Theorem 1, it suffices to prove the more general Theorem 5.

Proof. From (50) in Lemma 8, we will upper bound $|①|$ and $|③|$ by the statistical error assumption (20) and upper bound $|②|$ and $|④|$ by using the transfer error assumption (23).

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

$$\begin{aligned}
|①| &\leq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k+1)} \left|\phi_{s,a}^\top \left(w^{(k)} - w_\star^{(k)}\right)\right| \\
&\leq \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(d_s^{(k+1)}\right)^2 \left(\pi_{s,a}^{(k+1)}\right)^2}{d_s^* \cdot \mathrm{Unif}_{\mathcal{A}}(a)} \cdot \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \cdot \mathrm{Unif}_{\mathcal{A}}(a) \left(\phi_{s,a}^\top \left(w^{(k)} - w_\star^{(k)}\right)\right)^2} \\
&\overset{(24)}{=} \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(d_s^{(k+1)}\right)^2 \left(\pi_{s,a}^{(k+1)}\right)^2}{d_s^* \cdot \mathrm{Unif}_{\mathcal{A}}(a)} \left\|w^{(k)} - w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^*}}^2} \\
&\leq \sqrt{\mathbb{E}_{s\sim d^*}\left[\left(\frac{d_s^{(k+1)}}{d_s^*}\right)^2\right] |\mathcal{A}| \left\|w^{(k)} - w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^*}}^2} \\
&\overset{(26)}{\leq} \sqrt{C_\rho |\mathcal{A}| \left\|w^{(k)} - w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^*}}^2},
\end{aligned} \tag{58}$$

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

$$\sum_{a\in\mathcal{A}} \left(\pi_{s,a}^{(k+1)}\right)^2 \leq \sum_{a\in\mathcal{A}} \pi_{s,a}^{(k+1)} = 1. \tag{59}$$

Then, by using Assumption 3 with the definition of κ_ν, (58) is upper bounded by

$$\begin{aligned}
|①| &\overset{(25)}{\leq} \sqrt{C_\rho |\mathcal{A}| \kappa_\nu \left\|w^{(k)} - w_\star^{(k)}\right\|_{\Sigma_\nu}^2} \\
&\overset{(6)}{\leq} \sqrt{\frac{C_\rho |\mathcal{A}| \kappa_\nu}{1-\gamma} \left\|w^{(k)} - w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^{(k)}}}^2},
\end{aligned} \tag{60}$$

where we use the shorthand

$$\Sigma_{\tilde{d}^{(k)}} \overset{\text{def}}{=} \mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\phi_{s,a}\phi_{s,a}^\top\right]. \tag{61}$$

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

$$(w - w_\star^{(k)})^\top \nabla_w L_Q(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}) \geq 0, \qquad \text{for all } w \in \mathbb{R}^m. \tag{62}$$

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

$$\begin{aligned}
& L_Q(w, \theta^{(k)}, \tilde{d}^{(k)}) - L_Q(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}) \\
= \; & \mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\phi_{s,a}^\top w - \phi_{s,a}^\top w_\star^{(k)} + \phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right)^2\right] - L_Q(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}) \\
= \; & \mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[(\phi_{s,a}^\top w - \phi_{s,a}^\top w_\star^{(k)})^2\right] + 2(w - w_\star^{(k)})^\top \mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)})\phi_{s,a}\right] \\
= \; & \left\|w - w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^{(k)}}}^2 + (w - w_\star^{(k)})^\top \nabla_w L_Q(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}) \\
\overset{(62)}{\geq} \; & \left\|w - w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^{(k)}}}^2 .
\end{aligned} \tag{63}$$

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 (20), we have

$$\mathbb{E}\left[\epsilon_{\text{stat}}^{(k)}\right] \leq \epsilon_{\text{stat}}. \tag{64}$$

Plugging (63) into (60), we have

$$|①| \leq \sqrt{\frac{C_\rho |\mathcal{A}| \kappa_\nu}{1-\gamma} \epsilon_{\text{stat}}^{(k)}}. \tag{65}$$

Similar to (58), we get the same upper bound for |③| by just replacing $\pi_{s,a}^{(k+1)}$ into $\pi_{s,a}^{(k)}$. That is,

$$|③| \leq \sqrt{\frac{C_\rho |\mathcal{A}| \kappa_\nu}{1-\gamma} \epsilon_{\text{stat}}^{(k)}}. \tag{66}$$

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 (23), we have

$$\mathbb{E}\left[\epsilon_{\text{bias}}^{(k)}\right] \leq \epsilon_{\text{bias}}. \tag{67}$$

By Cauchy-Schwartz's inequality, we have

$$\begin{aligned}
|②| \; & \leq \; \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d_s^{(k+1)} \pi_{s,a}^{(k+1)} \left|\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right| \\
& \leq \; \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(d_s^{(k+1)}\right)^2 \left(\pi_{s,a}^{(k+1)}\right)^2}{d_s^* \cdot \text{Unif}_{\mathcal{A}}(a)} \cdot \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \cdot \text{Unif}_{\mathcal{A}}(a) \left(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right)^2} \\
& = \; \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(d_s^{(k+1)}\right)^2 \left(\pi_{s,a}^{(k+1)}\right)^2}{d_s^* \cdot \text{Unif}_{\mathcal{A}}(a)} \cdot \epsilon_{\text{bias}}^{(k)}} \\
& \overset{(59)}{\leq} \; \sqrt{\mathbb{E}_{s\sim d^*}\left[\left(\frac{d_s^{(k+1)}}{d_s^*}\right)^2\right] |\mathcal{A}| \epsilon_{\text{bias}}^{(k)}} \overset{(26)}{\leq} \sqrt{C_\rho |\mathcal{A}| \epsilon_{\text{bias}}^{(k)}}.
\end{aligned} \tag{68}$$

Similar to (68), we get the same upper bound for $|④|$ by just replacing $\pi_{s,a}^{(k+1)}$ into $\pi_{s,a}^{(k)}$. That is,

$$|④| \leq \sqrt{C_\rho |\mathcal{A}| \epsilon_{\text{bias}}^{(k)}}. \tag{69}$$

Next, we will upper bound the absolute values of ⓐ, ⓑ, ⓒ and ⓓ of (50) separately by using again the statistical error (20) and by using the transfer error assumption (23).

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

$$\begin{aligned}
|ⓐ| &\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \pi_{s,a}^{(k)} \left| \phi_{s,a}^\top \left(w^{(k)} - w_\star^{(k)} \right) \right| \\
&\leq \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{(d_s^*)^2 \left(\pi_{s,a}^{(k)}\right)^2}{d_s^* \cdot \text{Unif}_{\mathcal{A}}(a)} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \cdot \text{Unif}_{\mathcal{A}}(a) \left(\phi_{s,a}^\top \left(w^{(k)} - w_\star^{(k)} \right) \right)^2} \\
&\overset{(24)}{=} \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{(d_s^*)^2 \left(\pi_{s,a}^{(k)}\right)^2}{d_s^* \cdot \text{Unif}_{\mathcal{A}}(a)} \left\| w^{(k)} - w_\star^{(k)} \right\|_{\Sigma_{\tilde{d}^*}}^2} \\
&\overset{(59)}{\leq} \sqrt{|\mathcal{A}| \left\| w^{(k)} - w_\star^{(k)} \right\|_{\Sigma_{\tilde{d}^*}}^2}.
\end{aligned}$$

From the definition of κ_ν, we further obtain

$$\begin{aligned}
|ⓐ| &\overset{(25)}{\leq} \sqrt{|\mathcal{A}| \kappa_\nu \left\| w^{(k)} - w_\star^{(k)} \right\|_{\Sigma_\nu}^2} \\
&\overset{(6)}{\leq} \sqrt{\frac{|\mathcal{A}| \kappa_\nu}{1-\gamma} \left\| w^{(k)} - w_\star^{(k)} \right\|_{\Sigma_{\tilde{d}^{(k)}}}^2} \\
&\overset{(63)}{\leq} \sqrt{\frac{|\mathcal{A}| \kappa_\nu}{1-\gamma} \epsilon_{\text{stat}}^{(k)}}.
\end{aligned} \tag{70}$$

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

$$|ⓒ| \leq \sqrt{\frac{|\mathcal{A}| \kappa_\nu}{1-\gamma} \epsilon_{\text{stat}}^{(k)}}. \tag{71}$$

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

$$\begin{aligned}
|ⓑ| &\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \pi_{s,a}^{(k)} \left| \left(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)} \right) \right| \\
&\leq \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{(d_s^*)^2 \left(\pi_{s,a}^{(k)}\right)^2}{d_s^* \cdot \text{Unif}_{\mathcal{A}}(a)} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_s^* \cdot \text{Unif}_{\mathcal{A}}(a) \left(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)} \right)^2} \\
&= \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{(d_s^*)^2 \left(\pi_{s,a}^{(k)}\right)^2}{d_s^* \cdot \text{Unif}_{\mathcal{A}}(a)} \epsilon_{\text{bias}}^{(k)}} \\
&\overset{(59)}{\leq} \sqrt{|\mathcal{A}| \epsilon_{\text{bias}}^{(k)}}.
\end{aligned} \tag{72}$$

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

$$|ⓓ| \leq \sqrt{|\mathcal{A}| \epsilon_{\text{bias}}^{(k)}}. \tag{73}$$

Plugging all the upper bounds (65) of |①|, (68) of |②|, (66) of |③|, (69) of |④|, (70) of |ⓐ|, (72) of |ⓑ|, (71) of |ⓒ| and (73) of |ⓓ| into (50) yields

$$\vartheta_\rho\left(\delta_{k+1}-\delta_k\right)+\delta_k \le \frac{D_k^*}{(1-\gamma)\eta_k}-\frac{D_{k+1}^*}{(1-\gamma)\eta_k}+\frac{2\sqrt{|\mathcal{A}|}\left(\vartheta_\rho\sqrt{C_\rho}+1\right)}{1-\gamma}\left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}^{(k)}}+\sqrt{\epsilon_{\text{bias}}^{(k)}}\right), \tag{74}$$

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

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

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 (21), 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

$$(1-\gamma)(\vartheta_\rho-1)\overset{(21)}{\ge}(1-\gamma)\left(\frac{1}{1-\gamma}-1\right)=\gamma,$$

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

$$
\begin{aligned}
& \mathbb{E}\left[V_\rho(\pi^{(k)})\right] - V_\rho(\pi^*) \\
\leq \quad & \left(1 - \frac{1}{\vartheta_\rho}\right)^k \frac{2}{1-\gamma} \\
& + \frac{2\sqrt{|\mathcal{A}|}\left(\sqrt{C_\rho} + \frac{1}{\vartheta_\rho}\right)}{1-\gamma} \sum_{t=0}^{k-1} \left(1 - \frac{1}{\vartheta_\rho}\right)^{k-1-t} \left(\mathbb{E}\left[\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}^{(t)}}\right] + \mathbb{E}\left[\sqrt{\epsilon_{\text{bias}}^{(t)}}\right]\right) \\
\leq \quad & \left(1 - \frac{1}{\vartheta_\rho}\right)^k \frac{2}{1-\gamma} \\
& + \frac{2\sqrt{|\mathcal{A}|}\left(\sqrt{C_\rho} + \frac{1}{\vartheta_\rho}\right)}{1-\gamma} \sum_{t=0}^{k-1} \left(1 - \frac{1}{\vartheta_\rho}\right)^{k-1-t} \left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\mathbb{E}\left[\epsilon_{\text{stat}}^{(t)}\right]} + \sqrt{\mathbb{E}\left[\epsilon_{\text{bias}}^{(t)}\right]}\right) \\
\overset{(64)+(67)}{\leq} \quad & \left(1 - \frac{1}{\vartheta_\rho}\right)^k \frac{2}{1-\gamma} \\
& + \frac{2\sqrt{|\mathcal{A}|}\left(\sqrt{C_\rho} + \frac{1}{\vartheta_\rho}\right)}{1-\gamma} \sum_{t=0}^{k-1} \left(1 - \frac{1}{\vartheta_\rho}\right)^{k-1-t} \left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{bias}}}\right) \\
\leq \quad & \left(1 - \frac{1}{\vartheta_\rho}\right)^k \frac{2}{1-\gamma} + \frac{2\sqrt{|\mathcal{A}|}\left(\vartheta_\rho\sqrt{C_\rho} + 1\right)}{1-\gamma} \left(\sqrt{\frac{\kappa_\nu}{1-\gamma}\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{bias}}}\right),
\end{aligned}
$$

where the second inequality is obtained by Jensen's inequality. This concludes the proof. □

D.3 Proof of Theorem 2

Similar to Theorem 5, to prove Theorem 2, it suffices to prove the following sublinear convergence of Q-NPG with any initial policy.

Theorem 6. *Fix a state distribution ρ, an state-action distribution ν and an* optimal policy π^*. *We consider the Q-NPG method* (19) *with any constant step size $\eta_k = \eta > 0$. Suppose that Assumptions 1, 2, 3 and 4 all hold. Then we have for all $k \geq 0$,*

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

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

Proof. By (74) and using a constant step size η, we have

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

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

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

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

where we use the following inequalities

$$\mathbb{E}\left[\sqrt{\epsilon_{\text{stat}}^{(t)}}\right] \leq \sqrt{\mathbb{E}\left[\epsilon_{\text{stat}}^{(t)}\right]} \overset{(64)}{\leq} \sqrt{\epsilon_{\text{stat}}},$$

$$\mathbb{E}\left[\sqrt{\epsilon_{\text{bias}}^{(t)}}\right] \leq \sqrt{\mathbb{E}\left[\epsilon_{\text{bias}}^{(t)}\right]} \overset{(67)}{\leq} \sqrt{\epsilon_{\text{bias}}}.$$

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

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

□

D.4 Proof of Theorem 3

Similar to Theorem 5, to prove Theorem 3, it suffices to prove the following Q-NPG result with any initial policy.

Theorem 7. *Fix a state distribution ρ, an state-action distribution ν and a comparator policy π^*. We consider the Q-NPG method (19) 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, 5 and 6 hold. Then we have for all $k \geq 0$,*

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

Proof. Similar to the proof of Theorem 1, by Lemma 8, we upper bound the absolute values of ①, ②, ③, ④, ⓐ, ⓑ, ⓒ, ⓓ introduced in (50), separately, with the set of assumptions in Theorem 3.

In comparison with the proof of Theorem 1, we will also upper bound |①|, |③|, |ⓐ| and |ⓒ| by the statistical error assumption (20) as in the proof of Theorem 1. However, we will upper bound |②|, |④|, |ⓑ| and |ⓓ| by using the approximation error assumption (28) instead of the transfer error assumption (23).

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

$$\begin{aligned}
|①| &\leq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k+1)}\left|\phi_{s,a}^\top\left(w^{(k)} - w_\star^{(k)}\right)\right| \\
&\leq \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(d_s^{(k+1)}\right)^2\left(\pi_{s,a}^{(k+1)}\right)^2}{\tilde{d}_{s,a}^{(k)}} \cdot \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \tilde{d}_{s,a}^{(k)}\left(\phi_{s,a}^\top\left(w^{(k)} - w_\star^{(k)}\right)\right)^2} \\
&\overset{(61)}{=} \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{d_s^{(k+1)}\pi_{s,a}^{(k+1)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right]\left\|w^{(k)} - w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^{(k)}}}^2} \\
&\overset{(29)}{\leq} \sqrt{C_\nu\left\|w^{(k)} - w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^{(k)}}}^2} \\
&\overset{(63)}{\leq} \sqrt{C_\nu \epsilon_{\text{stat}}^{(k)}}.
\end{aligned}$$

Similar to |①|, by using Assumption 6 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

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

Next, we define

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

By Assumption 5, 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

$$\begin{aligned}
|②| \quad &\leq \quad \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{(k+1)}\pi_{s,a}^{(k+1)}\left|\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right| \\
&\leq \quad \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(d_s^{(k+1)}\right)^2\left(\pi_{s,a}^{(k+1)}\right)^2}{\tilde{d}_{s,a}^{(k)}} \cdot \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \tilde{d}_{s,a}^{(k)}\left(\phi_{s,a}^\top w_\star^{(k)} - Q_{s,a}^{(k)}\right)^2} \\
&= \quad \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{d_s^{(k+1)}\pi_{s,a}^{(k+1)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right] \cdot \epsilon_{\text{approx}}^{(k)}} \\
&\stackrel{(29)}{\leq} \quad \sqrt{C_\nu \epsilon_{\text{approx}}^{(k)}}.
\end{aligned}$$

Similar to |②|, by using Assumption 5 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

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

Consequently, plugging all these upper bounds into (50) 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 (74), we obtain the final performance bound with the linear convergence rate

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

□

D.5 Proof of Corollary 1

In order to better understand our proof, we first identify an issue appeared in the sample complexity analysis of Q-NPG in Agarwal et al. (2021, Corollay 26). Agarwal et al. (2021) adopts the optimization results of Shalev-Shwartz & Ben-David (2014, Theorem 14.8) where the stochastic gradient $\widehat{\nabla} L_Q(w, \theta, \tilde{d}^\theta)$ in (48) needs to be bounded [6].

[6] We are aware that Agarwal et al. (2021, Corollary 6.10) also use Bach & Moulines (2013, Theorem 1) in an early version https://arxiv.org/pdf/1908.00261v2.pdf to obtain $\epsilon_{\text{stat}} = \mathcal{O}(1/T)$. With further regularity (32), Agarwal et al. (2021) mentioned that $\epsilon_{\text{stat}} = \mathcal{O}(1/T)$ can also be achieved through Hsu et al. (2012, Theorem 16).

However, although they consider a projection step for the iterate w_t and assume that the feature map $\phi_{s,a}$ is bounded, $\widehat{\nabla} L_Q(w, \theta, \tilde{d}^\theta)$ is still not guaranteed to be bounded. Indeed, recall the stochastic gradient of the function L_Q in (48)

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

They incorrectly use the argument that $w, \phi_{s,a}$ and $\widehat{Q}_{s,a}(\theta)$ are bounded to imply that $\left\|\widehat{\nabla}_w L_Q(w, \theta, \tilde{d}^\theta)\right\|$ is bounded. In fact, $\widehat{Q}_{s,a}(\theta)$ can be unbounded even though $\mathbb{E}\left[\widehat{Q}_{s,a}(\theta)\right] = Q_{s,a}(\theta) \in \left[0, \frac{1}{1-\gamma}\right]$ is bounded. To see this, we can rewrite $\widehat{Q}_{s,a}(\theta)$ from (44) 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 sampled trajectory for estimating $Q_{s,a}(\theta)$ in Algorithm 3. From Algorithm 3 and from the proof of Lemma 4, we know that the probability of $H = k+1$ is that

$$\Pr(H = k+1) = (1-\gamma)\gamma^k.$$

So, with exponentially decreasing low probability, H can be unbounded. Consequently, $|\widehat{Q}_{s,a}(\theta)|$ upper bounded by H is not guaranteed to be bounded.

Proof sketch. Instead, we adopt the optimization results of Bach & Moulines (2013, Theorem 1) (see also Theorem 12), which does not require the boundedness of the stochastic gradient. However, in our following proof, we can verify that $\mathbb{E}\left[\widehat{Q}_{s,a}(\theta)^2\right]$ is bounded even though $\widehat{Q}_{s,a}(\theta)$ is unbounded. As to verify the condition (vi) in Theorem 12 in our proof, i.e., the covariance of the stochastic gradient at the optimum is upper bounded by the covariance of the feature map up to a finite constant, we use a conditional expectation argument to separate the correlated random variables $(s, a) \sim \tilde{d}^\theta$ and $\widehat{Q}_{s,a}(\theta)$ appeared in the stochastic gradient.

Non-singularity of the covariance matrix Σ_ν. As for the condition (32), it is shown in Cayci et al. (2021, Proposition 3) that with ν chosen as uniform distribution over $\mathcal{S}\times\mathcal{A}$ and $\phi_{s,a} \sim \mathcal{N}(0, \mathbf{I}_m)$ sampled as Gaussian random features, (32) is guaranteed with high probability. More generally, with $m \ll |\mathcal{S}||\mathcal{A}|$, it is easy to find m linearly independent $\phi_{s,a}$ among all $|\mathcal{S}||\mathcal{A}|$ features such that the covariance matrix Σ_ν has full rank.

Proof. From Theorem 3, 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 (20), we aim to apply the standard analysis for the averaged SGD, i.e., Theorem 12. Now we verify all the assumptions in order for `Q-NPG-SGD`.

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

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 $\|\phi_{s,a}\| \le B$, we have $\mathbb{E}\left[\|\phi_{s,a}\|^2\right]$ finite. From (32), 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 $\|\phi_{s,a}\| \le B$, we have

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

Thus, it remains to show $\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta)\right)^2\right]$ finite for (iii). From (44), 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

$$\begin{aligned}
\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta)\right)^2\right] &= \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\sum_{k=0}^{\infty}\Pr(H=k)\mathbb{E}\left[\left(\sum_{t=0}^{k} c(s_t,a_t)\right)^2 \mid H=k, s_0=s, a_0=a\right]\right] \\
&= \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[(1-\gamma)\sum_{k=0}^{\infty}\gamma^k\mathbb{E}\left[\left(\sum_{t=0}^{k} c(s_t,a_t)\right)^2 \mid H=k, s_0=s, a_0=a\right]\right] \\
&\leq \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[(1-\gamma)\sum_{k=0}^{\infty}\gamma^k(k+1)^2\right] \leq \frac{2}{(1-\gamma)^2},
\end{aligned} \tag{75}$$

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

$$\xi \overset{\text{def}}{=} \left(\widehat{Q}_{s,a}(\theta) - w_\star^\top \phi_{s,a}\right)\phi_{s,a} \overset{(48)}{=} \frac{1}{2}\widehat{\nabla}_w L_Q(w_\star, \theta, \tilde{d}^\theta). \tag{76}$$

From Lemma 7, 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

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

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 48, 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

$$\begin{aligned}
\mathbb{E}\left[\xi\xi^\top\right] &\overset{(76)}{=} \mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta) - w_\star^\top\phi_{s,a}\right)^2 \phi_{s,a}\phi_{s,a}^\top\right] \\
&= \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\left(\widehat{Q}_{s,a}(\theta) - w_\star^\top\phi_{s,a}\right)^2 \phi_{s,a}\phi_{s,a}^\top \mid s, a\right] \\
&= \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta) - w_\star^\top\phi_{s,a}\right)^2 \mid s, a\right]\phi_{s,a}\phi_{s,a}^\top\right].
\end{aligned}$$

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

$$\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta) - w_\star^\top\phi_{s,a}\right)^2 \mid s, a\right] = \mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta)\right)^2 \mid s, a\right] - 2Q_{s,a}(\theta)w_\star^\top\phi_{s,a} + \left(w_\star^\top\phi_{s,a}\right)^2 \leq \sigma^2 \tag{77}$$

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

$$\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta)\right)^2 \mid s, a\right] \overset{(75)}{\leq} \frac{2}{(1-\gamma)^2}.$$

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

$$
\begin{aligned}
w_\star &= \left(\mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\phi_{s,a}\phi_{s,a}^\top\right]\right)^\dagger \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[Q_{s,a}(\theta)\phi_{s,a}\right] \\
&\overset{(6)}{\leq} \frac{1}{1-\gamma}\left(\mathbb{E}_{(s,a)\sim\nu}\left[\phi_{s,a}\phi_{s,a}^\top\right]\right)^\dagger \mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[Q_{s,a}(\theta)\phi_{s,a}\right].
\end{aligned}
$$

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 (32), we have the minimizer $w_\star$ bounded by

$$
\|w_\star\| \overset{(32)}{\leq} \frac{B}{\mu(1-\gamma)^2}.
$$

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 (77) can be upper bounded by

$$
\begin{aligned}
\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta) - w_\star^\top\phi_{s,a}\right)^2 \mid s,a\right] &\leq \frac{2}{(1-\gamma)^2} + \frac{2B^2}{\mu(1-\gamma)^3} + \frac{B^4}{\mu^2(1-\gamma)^4} \\
&= \frac{1}{(1-\gamma)^2}\left(\left(\frac{B^2}{\mu(1-\gamma)}+1\right)^2+1\right) \\
&\leq \frac{2}{(1-\gamma)^2}\left(\frac{B^2}{\mu(1-\gamma)}+1\right)^2.
\end{aligned}
$$

Thus, in order to satisfy (77), we choose

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

Now all the conditions (i) - (vi) in Theorem 12 are verified. With step size $\alpha = \frac{1}{2B^2}$, the initialization $w_0 = 0$ and T steps of `Q-NPG-SGD` updates (48), we have

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

Consequently, Assumption 1 is verified by

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

The proof is completed by replacing the above upper bound of $\sqrt{\epsilon_{\text{stat}}}$ in the results of Theorem 3. □

E Proof of Section 5

E.1 The One Step NPG Lemma

To prove Theorem 4 and the sublinear convergence result of Theorem 9 further in Appendix F, we start from providing the one step analysis of the NPG update.

Lemma 9 (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, let $w_\star^{(k)} \in \operatorname{argmin}_w L_A(w,\theta^{(k)},\tilde{d}^{(k)})$ denote the exact minimizer. Consider the $w^{(k)}$ and $\pi^{(k)}$ NPG iterates given in (33) and (18) respectively. Note*

$$
\epsilon_{\text{stat}}^{(k)} \overset{def}{=} L_A(w^{(k)},\theta^{(k)},\tilde{d}^{(k)}) - L_A(w_\star^{(k)},\theta^{(k)},\tilde{d}^{(k)}), \tag{78}
$$

$$
\epsilon_{\text{approx}}^{(k)} \overset{def}{=} L_A(w_\star^{(k)},\theta^{(k)},\tilde{d}^{(k)}), \tag{79}
$$

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

If Assumptions 7, 8 and 9 hold for all $k \geq 0$, then we have that

$$\vartheta_\rho \left(\delta_{k+1} - \delta_k\right) + \delta_k \leq \frac{D_k^*}{(1-\gamma)\eta_k} - \frac{D_{k+1}^*}{(1-\gamma)\eta_k} + \frac{\sqrt{C_\nu}\left(\vartheta_\rho + 1\right)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}^{(k)}} + \sqrt{\epsilon_{\text{approx}}^{(k)}}\right). \tag{80}$$

Proof. From the three-point descent lemma (Lemma 11) and (18), 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

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

which is due to (13), we have

$$\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)} \right\rangle \leq -\frac{1}{\eta_k} D(\pi_s^{(k+1)}, \pi_s^{(k)}) - \frac{1}{\eta_k} D(\pi_s^{(k)}, \pi_s^{(k+1)}) \leq 0. \tag{81}$$

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

$$\mathbb{E}_{s\sim d^*}\left[\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)} \right\rangle\right] - \mathbb{E}_{s\sim d^*}\left[\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{*} \right\rangle\right] \leq \frac{1}{\eta_k} D_k^* - \frac{1}{\eta_k} D_{k+1}^*. \tag{82}$$

For the first expectation in (82), we have

$$\begin{aligned}
& \mathbb{E}_{s\sim d^*}\left[\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)} \right\rangle\right] \\
= \quad & \sum_{s\in\mathcal{S}} d_s^* \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)} \right\rangle \\
= \quad & \sum_{s\in\mathcal{S}} \frac{d_s^*}{d_s^{(k+1)}} d_s^{(k+1)} \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)} \right\rangle \\
\overset{(21)+(81)}{\geq} \quad & \vartheta_{k+1} \sum_{s\in\mathcal{S}} d_s^{(k+1)} \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)} \right\rangle \\
\overset{(21)+(81)}{\geq} \quad & \vartheta_{\rho} \sum_{s\in\mathcal{S}} d_s^{(k+1)} \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)} \right\rangle \\
= \quad & \vartheta_\rho \mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[(\bar{\phi}_{s,a}^{(k)})^\top w^{(k)}\right] \\
= \quad & \vartheta_\rho \mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[A_{s,a}^{(k)}\right] + \vartheta_\rho \mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[(\bar{\phi}_{s,a}^{(k)})^\top w^{(k)} - A_{s,a}^{(k)}\right] \\
= \quad & \vartheta_\rho (1-\gamma)\left(V_\rho^{(k+1)} - V_\rho^{(k)}\right) + \vartheta_\rho \mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[(\bar{\phi}_{s,a}^{(k)})^\top w^{(k)} - A_{s,a}^{(k)}\right],
\end{aligned} \tag{83}$$

where the last line is obtained by the performance difference lemma (42), and we use the shorthand $\bar{\phi}_{s,a}^{(k)}$ as $\bar{\phi}_{s,a}(\theta^{(k)})$.

The second term of (83) can be lower bounded. To do it, we first decompose it into two terms. That is,

$$\mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[(\bar{\phi}_{s,a}^{(k)})^\top w^{(k)} - A_{s,a}^{(k)}\right] = \underbrace{\mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[(\bar{\phi}_{s,a}^{(k)})^\top (w^{(k)} - w_\star^{(k)})\right]}_{①} + \underbrace{\mathbb{E}_{(s,a)\sim\bar{d}^{(k+1)}}\left[(\bar{\phi}_{s,a}^{(k)})^\top w_\star^{(k)} - A_{s,a}^{(k)}\right]}_{②}. \tag{84}$$

We will upper bound the absolute values of the above two terms $|①|$ and $|②|$ separately. More precisely, similar to the proof of Theorem 3, we will upper bound the first term $|①|$ by the statistical error assumption (34) and upper bound the second term $|②|$ by using the approximation error assumption (35).

To upper bound ①, we first define the following covariance matrix of the centered feature map

$$\Sigma_{\tilde{d}^{(k)}}^{(k)} \stackrel{\text{def}}{=} \mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\bar{\phi}_{s,a}^{(k)}(\bar{\phi}_{s,a}^{(k)})^\top\right]. \tag{85}$$

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 (61) in the proof of Theorem 1, as the centered feature map $\bar{\phi}_{s,a}^{(k)}$ depends on the iterates $\theta^{(k)}$.

By Cauchy-Schwartz's inequality, we have

$$\begin{aligned}
|①| &\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bar{d}_{s,a}^{(k+1)} \left|(\bar{\phi}_{s,a}^{(k)})^\top (w^{(k)} - w_\star^{(k)})\right| \\
&\leq \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(\bar{d}_{s,a}^{(k+1)}\right)^2}{\tilde{d}_{s,a}^{(k)}} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \tilde{d}_{s,a}^{(k)} \left((\bar{\phi}_{s,a}^{(k)})^\top (w^{(k)} - w_\star^{(k)})\right)^2} \\
&\stackrel{(85)}{=} \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{\bar{d}_{s,a}^{(k+1)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right] \left\|w^{(k)} - w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^{(k)}}^{(k)}}^2}.
\end{aligned}$$

By further using the concentrability assumption 9, we have

$$\begin{aligned}
|①| &\stackrel{(36)}{\leq} \sqrt{C_\nu \left\|w^{(k)} - w_\star^{(k)}\right\|_{\Sigma_{\tilde{d}^{(k)}}^{(k)}}^2} \\
&\leq \sqrt{C_\nu \left(L_A(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}) - L_A(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)})\right)} \qquad (86) \\
&\stackrel{(78)}{=} \sqrt{C_\nu \epsilon_{\text{stat}}^{(k)}}, \qquad (87)
\end{aligned}$$

where (86) uses that $w_\star^{(k)}$ is a minimizer of L_A and $w_\star^{(k)}$ is feasible (see the same arguments of (63) in the proof of Theorem 1).

For the second term |②| in (84), by Cauchy-Schwartz's inequality, we have

$$
\begin{aligned}
|②| &\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bar{d}_{s,a}^{(k+1)} \left|(\bar{\phi}_{s,a}^{(k)})^\top w_\star^{(k)} - A_{s,a}^{(k)}\right| \\
&\leq \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(\bar{d}_{s,a}^{(k+1)}\right)^2}{\tilde{d}_{s,a}^{(k)}} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \tilde{d}_{s,a}^{(k)} \left((\bar{\phi}_{s,a}^{(k)})^\top w_\star^{(k)} - A_{s,a}^{(k)}\right)^2} \\
&= \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{\bar{d}_{s,a}^{(k+1)}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right] L_A(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)})} \\
&\overset{(36)+(79)}{\leq} \sqrt{C_\nu \epsilon_{\text{approx}}^{(k)}}.
\end{aligned} \tag{88}
$$

Plugging (87) and (88) into (83) yields

$$
\mathbb{E}_{s\sim d^*}\left[\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^{(k+1)}\right\rangle\right] \geq \vartheta_\rho(1-\gamma)\left(V_\rho^{(k+1)} - V_\rho^{(k)}\right) - \vartheta_\rho\sqrt{C_\nu}\left(\sqrt{\epsilon_{\text{stat}}^{(k)}} + \sqrt{\epsilon_{\text{approx}}^{(k)}}\right). \tag{89}
$$

Now for the second expectation in (82), by using the performance difference lemma (42) in Lemma 3, we have

$$
\begin{aligned}
-\mathbb{E}_{s\sim d^*}\left[\left\langle \bar{\Phi}_s^{(k)} w^{(k)}, \pi_s^*\right\rangle\right] &= -\mathbb{E}_{(s,a)\sim\bar{d}^{\pi^*}}\left[A_{s,a}^{(k)}\right] + \mathbb{E}_{(s,a)\sim\bar{d}^{\pi^*}}\left[A_{s,a}^{(k)} - (\bar{\phi}_{s,a}^{(k)})^\top w^{(k)}\right] \\
&= (1-\gamma)\left(V_\rho^{(k)} - V_\rho(\pi^*)\right) + \mathbb{E}_{(s,a)\sim\bar{d}^{\pi^*}}\left[A_{s,a}^{(k)} - (\bar{\phi}_{s,a}^{(k)})^\top w^{(k)}\right].
\end{aligned} \tag{90}
$$

The second term of (90) can be lower bounded. We first decompose it into two terms. That is,

$$
\begin{aligned}
\mathbb{E}_{(s,a)\sim\bar{d}^{\pi^*}}\left[A_{s,a}^{(k)} - (\bar{\phi}_{s,a}^{(k)})^\top w^{(k)}\right] &= \underbrace{\mathbb{E}_{(s,a)\sim\bar{d}^{\pi^*}}\left[A_{s,a}^{(k)} - (\bar{\phi}_{s,a}^{(k)})^\top w_\star^{(k)}\right]}_{ⓐ} \\
&\quad + \underbrace{\mathbb{E}_{(s,a)\sim\bar{d}^{\pi^*}}\left[(\bar{\phi}_{s,a}^{(k)})^\top (w_\star^{(k)} - w^{(k)})\right]}_{ⓑ}.
\end{aligned} \tag{91}
$$

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

For the first one |ⓐ|, by Cauchy-Schwartz's inequality, we have

$$
\begin{aligned}
|ⓐ| &\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bar{d}_{s,a}^{\pi^*} \left|A_{s,a}^{(k)} - (\bar{\phi}_{s,a}^{(k)})^\top w_\star^{(k)}\right| \\
&\leq \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(\bar{d}_{s,a}^{\pi^*}\right)^2}{\tilde{d}_{s,a}^{(k)}} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \tilde{d}_{s,a}^{(k)} \left((\bar{\phi}_{s,a}^{(k)})^\top w_\star^{(k)} - A_{s,a}^{(k)}\right)^2} \\
&= \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}}\left[\left(\frac{\bar{d}_{s,a}^{\pi^*}}{\tilde{d}_{s,a}^{(k)}}\right)^2\right] L_A(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)})} \\
&\overset{(36)+(79)}{\leq} \sqrt{C_\nu \epsilon_{\text{approx}}^{(k)}}.
\end{aligned} \tag{92}
$$

For the second term $|ⓑ|$ in (91), by Cauchy-Schwartz's inequality, we have

$$
\begin{aligned}
|ⓑ| &\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bar{d}^{\pi^*}_{s,a} \left|(\bar{\phi}^{(k)}_{s,a})^\top (w^{(k)}_\star - w^{(k)})\right| \\
&\leq \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(\bar{d}^{\pi^*}_{s,a}\right)^2}{\tilde{d}^{(k)}_{s,a}} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \tilde{d}^{(k)}_{s,a} \left((\bar{\phi}^{(k)}_{s,a})^\top (w^{(k)} - w^{(k)}_\star)\right)^2} \\
&\overset{(85)}{=} \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}^{(k)}} \left[\left(\frac{\bar{d}^{\pi^*}_{s,a}}{\tilde{d}^{(k)}_{s,a}}\right)^2\right] \left\|w^{(k)} - w^{(k)}_\star\right\|^2_{\Sigma^{(k)}_{\tilde{d}^{(k)}}}} \\
&\overset{(36)}{\leq} \sqrt{C_\nu \left\|w^{(k)} - w^{(k)}_\star\right\|^2_{\Sigma^{(k)}_{\tilde{d}^{(k)}}}} \\
&\overset{(86)}{\leq} \sqrt{C_\nu \left(L_A(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}) - L_A(w^{(k)}_\star, \theta^{(k)}, \tilde{d}^{(k)})\right)} \\
&\overset{(78)}{=} \sqrt{C_\nu \epsilon^{(k)}_{\text{stat}}}.
\end{aligned} \tag{93}
$$

Thus, we lower bound (91) by

$$
-\mathbb{E}_{s\sim d^*}\left[\left\langle \bar{\Phi}^{(k)}_s w^{(k)}, \pi^*_s \right\rangle\right] \overset{(92)+(93)}{\geq} (1-\gamma)\left(V^{(k)}_\rho - V_\rho(\pi^*)\right) - \sqrt{C_\nu}\left(\sqrt{\epsilon^{(k)}_{\text{stat}}} + \sqrt{\epsilon^{(k)}_{\text{approx}}}\right). \tag{94}
$$

Substituting (89) and (94) into (82), dividing both side by $1-\gamma$ and rearranging terms, we get

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

□

E.2 Proof of Theorem 4

To prove Theorem 4, we prove the following general result with any initial policy by knowing that $D^*_0 \leq \log|\mathcal{A}|$ when using the uniform initial policy in (49).

Theorem 8. *Fix a state distribution ρ, a state-action distribution ν, and a comparator policy π^*. We consider the NPG method (33) 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 7, 8 and 9 hold. Then we have for all $k \geq 0$,*

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

Proof. From (80) in Lemma 9, by using the same increasing step size as in Theorem 5, i.e. $\eta_0 \geq \frac{1-\gamma}{\gamma} D^*_0$ and $\eta_{k+1} \geq \eta_k/\gamma$, and following the same arguments in the proof of Theorem 5 after (74), we obtain the final performance bound with the linear convergence rate

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

□

F Sublinear Convergence of NPG with Constant Step Size

In this section, we provide the sublinear convergence of NPG with arbitrary constant step size.

Theorem 9. *Fix a state distribution ρ, an state-action distribution ν and an* optimal policy π^*. *We consider the NPG method* (33) *with any constant step size $\eta_k = \eta > 0$. Suppose that Assumptions 7, 8 and 9 hold. Then we have for all $k \geq 0$,*

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

Compared to Theorem 2, again here we require π^* to be an optimal policy for the same reason as indicated in Footnote 5. Furthermore our sublinear convergence guarantees for both Q-NPG and NPG are the same. Compared to Theorem 29 in Agarwal et al. (2021), the main differences are also similar to those for Q-NPG as summarized right after Theorem 2: our convergence rate improves from $\mathcal{O}(1/\sqrt{k})$ to $\mathcal{O}(1/k)$; they use a diminishing step size of order $\mathcal{O}(1/\sqrt{k})$ while we can take any constant step size we want.

Despite the difference of using $\tilde{d}^{(k)}$ instead of $\bar{d}^{(k)}$ for the compatible function approximation $L_A(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)})$, notice that same sublinear convergence rate $\mathcal{O}(1/k)$ is established by Liu et al. (2020) for NPG with constant step size, while their step size is bounded by the inverse of a smoothness constant and they further require that the feature map is bounded and the Fisher information matrix (10) is strictly lower bounded for all parameters $\theta \in \mathbb{R}^m$ (see this condition later in (95)). With such additional conditions, we are able to provide a $\tilde{\mathcal{O}}(\frac{1}{(1-\gamma)^5\epsilon^2})$ sample complexity result of NPG in Appendix G.

Proof. From (80) in Lemma 9 with the constant step size, we have

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

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

$$\begin{aligned}
\vartheta_\rho\left(\mathbb{E}\left[\delta_{k+1}\right] - \mathbb{E}\left[\delta_k\right]\right) + \mathbb{E}\left[\delta_k\right] \quad &\leq \quad \frac{\mathbb{E}\left[D_k^*\right]}{(1-\gamma)\eta} - \frac{\mathbb{E}\left[D_{k+1}^*\right]}{(1-\gamma)\eta} \\
&\quad + \frac{\sqrt{C_\nu}\left(\vartheta_\rho + 1\right)}{1-\gamma}\left(\mathbb{E}\left[\sqrt{\epsilon_{\text{stat}}^{(k)}}\right] + \mathbb{E}\left[\sqrt{\epsilon_{\text{approx}}^{(k)}}\right]\right) \\
&\leq \quad \frac{\mathbb{E}\left[D_k^*\right]}{(1-\gamma)\eta} - \frac{\mathbb{E}\left[D_{k+1}^*\right]}{(1-\gamma)\eta} \\
&\quad + \frac{\sqrt{C_\nu}\left(\vartheta_\rho + 1\right)}{1-\gamma}\left(\sqrt{\mathbb{E}\left[\epsilon_{\text{stat}}^{(k)}\right]} + \sqrt{\mathbb{E}\left[\epsilon_{\text{approx}}^{(k)}\right]}\right) \\
&\overset{(34)+(35)}{\leq} \quad \frac{\mathbb{E}\left[D_k^*\right]}{(1-\gamma)\eta} - \frac{\mathbb{E}\left[D_{k+1}^*\right]}{(1-\gamma)\eta} + \frac{\sqrt{C_\nu}\left(\vartheta_\rho + 1\right)}{1-\gamma}\left(\sqrt{\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{approx}}}\right).
\end{aligned}$$

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

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

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

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

□

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 2. *Consider the setting of Theorem 8. 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*

$$\mathbb{E}_{(s,a)\sim\tilde{d}^\theta}\left[\bar{\phi}_{s,a}(\theta)(\bar{\phi}_{s,a}(\theta))^\top\right] \;\geq\; \mu \mathbf{I}_m, \tag{95}$$

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

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

Now we compare our Corollary 2 with Corollary 33 in Agarwal et al. (2021), which is their corresponding sample complexity results for NPG. The main differences between Corollary 2 and Corollary 33 in Agarwal et al. (2021) are similar to those for Q-NPG as summarized right after Corollary 1: Their sample complexity is $\mathcal{O}\big(\frac{1}{(1-\gamma)^{11}\epsilon^6}\big)$ while ours is $\tilde{\mathcal{O}}\big(\frac{1}{(1-\gamma)^5\epsilon^2}\big)$; they consider a projection step for the iterates and incorrectly bound the stochastic gradient due to a similar error indicated in Footnote 3 (and see Appendix G.1 for more details), while we assume Fisher-non-degeneracy (95).

Compared to Corollary 1, the sample complexities for both Q-NPG and NPG are the same. The assumption (95) on the Fisher information matrix is much stronger than (32), as (32) is independent to the iterates. However, despite the difference of using ν instead of ρ, the Fisher-non-degeneracy (95) is commonly used in the optimization literature (Byrd et al., 2016; Gower et al., 2016; Wang et al., 2017) and in the RL literature (Liu et al., 2020; Ding et al., 2022; Yuan et al., 2022). It characterizes that the Fisher information matrix behaves well as a preconditioner in the NPG update (9). Indeed, (95) is directly assumed to be positive definite in the pioneering NPG work (Kakade, 2001) and in the follow-up works on natural actor-critic algorithms (Peters & Schaal, 2008; Bhatnagar et al., 2009). It is satisfied by a wide families of policies, including the Gaussian policy (Duan et al., 2016; Papini et al., 2018; Huang et al., 2020) and certain neural policy with log-linear policy as a special case. We refer to Liu et al. (2020, Section B.2) and Ding et al. (2022, Section 8) for more discussions on the Fisher-non-degenerate setting.

To prove Corollary 2, our approach is inspired from the proof of the sample complexity analysis of Liu et al. (2020, Theorem 4.9). That is, we require the Fisher-non-degeneracy (95) and apply Theorem 12 to the minimization of function $L_A(w,\theta,\tilde{d}^\theta)$ without relying on the boundedness of the stochastic gradient. A proof sketch is provided in Appendix G.1. Compared to their result, they obtain worse $\mathcal{O}\big(\frac{1}{(1-\gamma)^7\epsilon^3}\big)$ sample complexity for NPG due to a slower $\mathcal{O}(1/k)$ convergence rate.

G.1 PROOF OF COROLLARY 2

There is a similar remark for the proof of Corollary 2 to the one right before the proof of Corollary 1 in Appendix D.5. We notice that there is the same error occurred for the proof of NPG sample complexity analysis in Agarwal et al. (2021). Recall the stochastic gradient of L_A in (45)

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

It turns out that $\widehat{\nabla}_w L_A(w,\theta,\tilde{d}^\theta)$ is unbounded, since the estimate $\widehat{A}_{s,a}(\theta)$ of $A_{s,a}(\theta)$ can be unbounded due to the unbounded length of the trajectory sampled in the sampling procedure, Algorithm 4. Thus, Agarwal et al. (2021) incorrectly verify $\widehat{\nabla} L_A(w,\theta,\tilde{d}^\theta)$ bounded by claiming that $\widehat{A}_{s,a}(\theta)$ is bounded by $\frac{2}{1-\gamma}$.

Proof sketch. Despite the difference of using either $\tilde{d}^\theta$ or $\bar{d}^\theta$ in the loss function L_A, we use the same assumptions of Liu et al. (2020), i.e., the Fisher-non-degeneracy (95) and the boundedness of the feature map, and verify all the conditions of Theorem 12 *without* relying on the boundedness of the stochastic gradient. In particular, similar to the proof of Corollary 1, we verify that $\mathbb{E}\big[\widehat{A}_{s,a}(\theta)^2\big]$ is bounded even though $\widehat{A}_{s,a}(\theta)$ is unbounded. To verify the condition (vi) in Theorem 12 in our proof, we use the same conditional expectation argument as in the proof of Corollary 1 to separate the correlated random variables $\widehat{A}_{s,a}(\theta)$ and $\bar{\phi}_{s,a}(\theta)$ with $(s,a) \sim \tilde{d}^\theta$ appeared in the stochastic gradient. Thanks to this argument, we fix a flaw in the previous proof of Liu et al. (2020, Proposition G.1) [7].

Proof. Similar to the proof of Corollary 1, we suppress the subscript k. First, the centered feature map is bounded by $\big\|\bar{\phi}_{s,a}(\theta)\big\| \leq 2B$. In order to apply Theorem 12, it remains to upper bound $\mathbb{E}\big[\|\widehat{A}_{s,a}(\theta)\bar{\phi}_{s,a}(\theta)\|^2\big]$ and $\|w_\star\|$ with $w_\star \in \operatorname{argmin}_w L_A(w,\theta,\tilde{d}^\theta)$, and find $\sigma > 0$ such that

$$\mathbb{E}\left[\left(\widehat{A}_{s,a}(\theta) - w_\star^\top \bar{\phi}_{s,a}(\theta)\right)^2 \mid s,a\right] = \mathbb{E}\left[\left(\widehat{A}_{s,a}(\theta)\right)^2 \mid s,a\right] - 2A_{s,a}(\theta) w_\star^\top \bar{\phi}_{s,a}(\theta) + \left(w_\star^\top \bar{\phi}_{s,a}(\theta)\right)^2 \leq \sigma^2 \quad (96)$$

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

Similar to the proof of Corollary 1, 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 (95), 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 (96). Indeed, by using $\widehat{A}_{s,a}(\theta) = \widehat{Q}_{s,a}(\theta) - \widehat{V}_s(\theta)$, we have

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{A}_{s,a}(\theta)\right)^2 \mid s,a\right] &\leq 2\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta)\right)^2 \mid s,a\right] + 2\mathbb{E}\left[\left(\widehat{V}_{s,a}(\theta)\right)^2 \mid s,a\right] \\ &\overset{(75)}{\leq} \frac{8}{(1-\gamma)^2},\end{aligned} \quad (97)$$

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 (75) of $\mathbb{E}\left[\left(\widehat{Q}_{s,a}(\theta)\right)^2 \mid s,a\right]$ by using the similar argument.

From (97) 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.

[7] In a previous version of the proof in Section G, Liu et al. (2020, Proposition G.1) use the inequality

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

which is incorrect since $\widehat{A}_{s,a}(\theta)$ and $\bar{\phi}_{s,a}(\theta)$ are correlated random variables. To fix it, we use the following conditional expectation computation argument

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

and bound the term $\mathbb{E}\left[\left(\widehat{A}_{s,a}(\theta) - w_\star^\top \bar{\phi}_{s,a}(\theta)\right)^2 \mid s,a\right]$ in (96).

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 (96) 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 12 are verified. The reminder of the proof follows that of Corollary 1. □

H Discussion on the Distribution Mismatch Coefficients, the Concentrability Coefficients and the Increasing Step Size

We have already mentioned in the comparison with Agarwal et al. (2021) right after Theorem 1 that, although we have linear convergence rates, the magnitude of our error floor is worse (larger) by a factor of $\vartheta_\rho\sqrt{C_\rho}$ ($\vartheta_\rho\sqrt{C_\nu}$ for Theorem 3 and 4), due to the concentrability C_ρ and the distribution mismatch coefficients ϑ_ρ used in our proof. Such difference comes from different nature of the proof techniques. Here the distribution mismatch coefficients ϑ_ρ and the concentrability coefficients C_ρ and C_ν are potentially large in our convergence theories. We give extensive discussions on them, respectively. We also provide further noteworthy observations about the connection between the use of the increasing step size and the policy iteration.

Distribution mismatch coefficients ϑ_ρ. Our distribution mismatch coefficient ϑ_ρ in (21) is the same as the one in Xiao (2022). It contains both an upper bound and a lower bound. The linear convergence rate in our theories is $1-\frac{1}{\vartheta_\rho} > 0$. Thus, the smaller ϑ_ρ is, the faster the resulting linear convergence rate. The best linear convergence rate is achieved when ϑ_ρ achieves its lower bound. Here our analysis is general that it includes all the distribution mismatch coefficient ϑ_ρ induced by any target state distribution ρ. Our results generalizes and sometimes also improves with respect to prior results.

A very pessimistic and trivial upper bound on ϑ_ρ is

$$\vartheta_\rho \leq \frac{1}{(1-\gamma)\rho_{\min}}.$$

However, if the target state distribution $\rho \in \Delta(\mathcal{S})$ does not have full support, i.e., $\rho_s = 0$ for some $s \in \mathcal{S}$, then ϑ_ρ might be infinite from this upper bound. Xiao (2022) just assumes that ϑ_ρ is finite. We further propose a solution to this particular issue. Indeed, if ρ does not have full support, consider π^* as an optimal policy. We can always convert the convergence guarantees for some state distribution $\rho' \in \Delta(\mathcal{S})$ with full support, i.e., $\rho'_s > 0$ for all $s \in \mathcal{S}$ as follows:

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

Then we only need convergence guarantees of $V_{\rho'}(\pi^{(k)}) - V_{\rho'}(\pi^*)$ for arbitrary ρ' obtained from all our convergence analysis above. In this case, the linear convergence rate depends on

$$\vartheta_{\rho'} \stackrel{\text{def}}{=} \frac{1}{1-\gamma}\left\|\frac{d^{\pi^*}(\rho')}{\rho'}\right\|_\infty < \infty.$$

Equation (21) provides the lower bound $\frac{1}{1-\gamma}$ for ϑ_ρ. Such lower bound can be achieved when the target state distribution ρ satisfies that $\rho = d^{\pi^*}(\rho)$ where π^* is an optimal policy. The advantage of this case is that, not only it implies the best linear convergence rate, more importantly, the fast linear convergence rate is known to be γ. So we know the convergence rate explicitly without any estimation, even though the optimal policy or the policy iterates are unknown before training. Hence, we know when to stop running the algorithm. Lan (2022) only considers the case when $\rho = d^{\pi^*}(\rho)$ and we are able to recover the same linear convergence rate γ in their result.

Furthermore, the convergence performance $V_\rho(\pi^{(k)}) - V_\rho(\pi^*)$ depends on the target state distribution ρ. If the optimal policy π^* is independent to the target state distribution ρ which is usually the case in RL problems, then we are always allowed to fix $\rho = d^{\pi^*}(\rho)$ for the analysis without knowing ρ and π^* and derive this best linear convergence performance with rate γ, because we use the initial state-action distribution ν in training which is independent to ρ.

Finally, from (21), if $d^{(k)}$ converges to d^*, then ϑ_k converges to 1. This might imply superlinear convergence results as Section 4.3 in Xiao (2022). In this case, the notion of the distribution mismatch coefficients ϑ_ρ no longer exists for the superlinear convergence analysis. In other words, it is no longer concerned.

Concentrability coefficients C_ν. The issue of having (potentially large) concentrability coefficients is unavoidable in all the fast linear convergence analysis of the inexact NPG that we are aware of, including even the tabular setting (e.g., Lan (2022) and Xiao (2022)) and the log-linear policy setting (Cayci et al. (2021), Chen & Theja Maguluri (2022) and ours).

First, in the fast linear convergence analysis of inexact NPG, the concentrability coefficients appear from the errors, including the statistical error and the approximation error. Thus, one way to avoid having the concentrability coefficients appear is to consider the exact NPG in the tabular setting (See Theorem 10 in Xiao (2022)). Because the tabular setting makes no approximation error and the exact NPG makes no statistical error. We consider the *inexact* NPG with the *log-linear* policy. Consequently, we have the concentrability coefficients multiplied by both the statistical error ϵ_{stat} and the approximation error (ϵ_{bias} in Assumption 2 or ϵ_{approx} in Assumption 5 and 8).

To remove the concentrability coefficients, one has to make strong assumptions on the errors with the L_∞ supremum norm. In the tabular setting, Lan (2022) and Xiao (2022) assume that $\|\widehat{Q}(\pi) - Q(\pi)\|_\infty \leq \epsilon_{\text{stat}}$. The cons of such strong assumption requires high sample complexity and is already explained in Appendix A.1. In the log-linear policy setting, Chen & Theja Maguluri (2022) assume that $\|Q_s(\theta^{(k)}) - \Phi w_\star^{(k)}\|_\infty \leq \epsilon_{\text{bias}}$ for the approximation error, which is a very strong assumption in the function approximation regime. Due to the supremum norm, ϵ_{bias} is unlikely to be small, especially for large action spaces. Under this strong assumption, Lan (2022), Xiao (2022) and Chen & Theja Maguluri (2022) are able to eliminate the concentrability coefficients. To avoid assuming such strong assumptions, Cayci et al. (2021) and our paper consider the expected L_2 errors in the log-linear policy setting, which are much weaker assumptions, especially much more reasonable for the approximation error ϵ_{bias} compared to the one in Chen & Theja Maguluri (2022). The tradeoff is that, the concentrability coefficients can not be eliminated in this case both in Cayci et al. (2021) and our results.

Furthermore, as mentioned right after Theorem 4, under the expected error assumptions (Assumption 7 and 8), our concentrability coefficient C_ν is better presented than the one in Assumption 2 in Cayci et al. (2021) in the sense that it is independent to the policies throughout the iterations thanks to the use of $\tilde{d}^{(k)}$ instead of $\bar{d}^{(k)}$ (which is mentioned in Remark 1 as well) and is controllable to be finite by ν, while the one in Cayci et al. (2021) depends on the iterates, thus is unknown and is not guaranteed to be finite.

Finally, like the distribution mismatch coefficient, the upper bound of C_ν in (31) is very pessimistic. By the definition of C_ν in (29), one can expect that C_ν is closed to 1, when $\pi^{(k)}$ and $\pi^{(k+1)}$ converge to π^* with π^* the optimal policy.

So our concentrability coefficient C_ν is the "best" one among all concentrability coefficients in the sense that, it takes the weakest assumptions on errors compared to Lan (2022), Xiao (2022) and Chen & Theja Maguluri (2022), it does not impose any restrictions on the MDP dynamics compared to Cayci et al. (2021) and it can be controlled to be finite by ν when other concentrability coefficients are infinite (Scherrer, 2014).

It is still an open question whether we can obtain fast linear convergence results of the inexact NPG in the log-linear policy setting, with small error floor and a much improved concentrability coefficient, e.g., as the same magnitude as the one in Agarwal et al. (2021).

Increasing step size and the connection with policy iteration. As for the use of the increasing step size, intuitively, the reason is that both NPG in (18) and Q-NPG in (17) behave more and more like policy iteration. For instance, when $\eta_k \to \infty$ and we replace the linear approximation $\Phi_s w^{(k)}$ by $Q_s(\theta^{(k)})$, (17) becomes

$$\pi_s^{(k+1)} = \arg\min_{p \in \Delta(\mathcal{A})} \left\{ \left\langle Q_s(\theta^{(k)}), p \right\rangle \right\}, \quad \forall s \in \mathcal{S},$$

which is exactly the classical Policy Iteration method (e.g., Puterman, 1994; Bertsekas, 2012). We refer to Xiao (2022, Section 4.4) for more discussion on the connection with policy iteration.

I Standard Optimization Results

In this section, we present the standard optimization results from Beck (2017); Xiao (2022); Bach & 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 10 (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*

$$\min_{p \in \Delta^n} \ \eta \langle g, p \rangle + D(p, q) \tag{98}$$

is given by

$$p = \frac{q \odot e^{-\eta g}}{\sum_{i=1}^n q_i e^{-\eta g_i}}, \tag{99}$$

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

Proof. The Lagrangian of (98) 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 (99). □

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 & Teboulle (1993). First, we need some technical conditions.

Definition 10 (Legendre function, Section 26 in Rockafellar (1970))**.** *We say a function h is of* Legendre type *or a* Legendre function *if the following properties are satisfied:*

(i) *h is strictly convex in the relative interior of* $\operatorname{dom} h$*, denoted as* $\operatorname{rint}\operatorname{dom} h$*.*

(ii) *h is essentially smooth, i.e., h is differentiable in* $\operatorname{rint}\operatorname{dom} h$ *and, for any boundary point x_b of* $\operatorname{rint}\operatorname{dom} h$*,* $\lim_{x\to x_b}\|\nabla h(x)\| \to \infty$ *where* $x \in \operatorname{rint}\operatorname{dom} h$*.*

Definition 11 (Bregman divergence (Bregman, 1967; Censor & Zenios, 1997))**.** *Let $h : \operatorname{dom} h \to \mathbb{R}$ be a Legendre function and assume that* $\operatorname{rint}\operatorname{dom} h$ *is nonempty. The* Bregman divergence $D_h(\cdot,\cdot) : \operatorname{dom} h \times \operatorname{rint}\operatorname{dom} h \to [0,\infty)$ *generated by h is a distance-like function defined as*

$$D_h(p,p') \stackrel{def}{=} h(p) - h(p') - \langle \nabla h(p'), p - p' \rangle . \tag{100}$$

Under the above conditions, we have the following result. We also provide its proof for self-containment. (Xiao (2022) does not provide a formal proof.)

Lemma 11 (Three-point decent lemma, Lemma 6 in Xiao (2022))**.** *Suppose that $\mathcal{C} \subset \mathbb{R}^m$ is a closed convex set, $f : \mathcal{C} \to \mathbb{R}$ is a proper, closed* [8] *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*

$$x^+ \in \arg\min_{u \in \operatorname{dom} h \cap \mathcal{C}} \{f(u) + D_h(u,x)\}.$$

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

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

Proof. First, we prove that for any $a, b \in \operatorname{rint}\operatorname{dom} h$ and $c \in \operatorname{dom} h$, the following identity holds:

$$D_h(c,a) + D_h(a,b) - D_h(c,b) = \langle \nabla h(b) - \nabla h(a), c - a \rangle . \tag{101}$$

Indeed, using the definition of D_h in (100), we have

$$\langle \nabla h(a), c - a \rangle = h(c) - h(a) - D_h(c,a), \tag{102}$$
$$\langle \nabla h(b), a - b \rangle = h(a) - h(b) - D_h(a,b), \tag{103}$$
$$\langle \nabla h(b), c - b \rangle = h(c) - h(b) - D_h(c,b). \tag{104}$$

Subtracting (102) and (103) from (104) yields (101).

Next, since h is of Legendre type, we have $x^+ \in \operatorname{rint}\operatorname{dom} h \cap \mathcal{C}$. Otherwise, x^+ is a boundary point of $\operatorname{dom} h$. From the definition of Legendre function, $\|\nabla h(x^+)\| = \infty$ which is not possible, as x^+ is also the minimum point of $f(u) + D_h(u,x)$. By the first-order optimality condition, we have

$$\left\langle u - x^+, g^+ + \nabla_y D_h(y,x)|_{y=x^+} \right\rangle \ge 0,$$

where $g^+ \in \partial f(x^+)$ is the subdifferential of f at x^+. From the definition of D_h, the above inequality is equivalent to

$$\left\langle u - x^+, \nabla h(x^+) - \nabla h(x) \right\rangle \ge \left\langle x^+ - u, g^+ \right\rangle . \tag{105}$$

Besides, plugging $c = u, a = x^+$ and $b = x$ into (101), we obtain

$$\left\langle u - x^+, \nabla h(x^+) - \nabla h(x) \right\rangle = D_h(u,x) - D_h(u,x^+) - D_h(x^+,x) \stackrel{(105)}{\ge} \left\langle x^+ - u, g^+ \right\rangle .$$

Rearranging terms and adding $f(u)$ on both sides, we have

$$\begin{aligned} D_h(u,x) - D_h(u,x^+) + f(u) &\ge f(u) + \left\langle x^+ - u, g^+ \right\rangle + D_h(x^+,x) \\ &\ge f(x^+) + D_h(x^+,x), \end{aligned}$$

which concludes the proof. The last inequality is obtained by the convexity of f and $g^+ \in \partial f(x^+)$. □

[8] A convex function f is proper if $\operatorname{dom} f$ is nonempty and for all $x \in \operatorname{dom} f$, $f(x) > -\infty$. A convex function is closed, if it is lower semi-continuous.

Finally, we use the following linear regression analysis for the proof of our sample complexity results, i.e., Corollary 1 and 2.

Theorem 12 (Theorem 1 in Bach & 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* $\theta_* \in \mathcal{H}$*. Let* $\xi_n = z_n - \langle\theta_*, x_n\rangle x_n$ *denote the residual. We have* $\mathbb{E}[\xi_n] = 0$*.*

(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*

$$\mathbb{E}\left[f(\theta_{\text{out}}) - f(\theta_*)\right] \leq \frac{2}{n}\left(\sigma\sqrt{m} + R\|\theta_0 - \theta_*\|\right)^2. \tag{106}$$