Finite-Time Analysis of the Natural Policy Gradient in Finite-Horizon Markov Decision Processes

Asha Barua, Sajad Khodadadian

Keywords: Finite-Horizon Markov Decision Process, Natural Policy Gradient, Sublinear and Linear Convergence Rates.

Summary

The Natural Policy Gradient (NPG) is a well-established algorithm in Reinforcement Learning, with widely used practical variants such as Trust Region Policy Optimization and Proximal Policy Optimization, demonstrating strong empirical success. This paper studies exact NPG in finite-horizon Markov Decision Processes with known dynamics and horizon-dependent transition kernels. This setting differs from the discounted infinite-horizon case because optimal policies are generally nonstationary and discounted Bellman contraction arguments do not apply. We consider both constant and increasing step size regimes. For tabular MDPs with a constant step size $\eta_t = \eta$, we prove that NPG converges sublinearly with a rate of $\mathcal{O}(H^2/t)$ (where H is the horizon length and t is the number of iterations). We also extend this sublinear guarantee to linear MDPs under an oracle regime with exact action-value evaluation and an exact population-projection oracle with a full support projection distribution. Finally, for tabular MDPs with increasing problem-dependent step sizes, we establish a geometric convergence bound of order $\mathcal{O}((1 - 1/\vartheta_\rho)^t)$, where ϑ_ρ is a distribution-mismatch coefficient. Furthermore, we show that a horizon-only robust step size schedule of the form $\eta_t = \eta_0(H/(H-1))^t$ where $\eta_0 > 0$ and $H \geq 2$, achieves the same geometric convergence guarantee. To the best of our knowledge, these are the first finite-time guarantees for exact NPG in finite-horizon MDPs with horizon-dependent dynamics.

Contribution(s)

1. We study exact NPG for finite-horizon tabular MDPs with known dynamics and horizon-dependent transition probabilities. Under a constant step size $\eta_t = \eta$, we prove a finite-time suboptimality bound of order $\mathcal{O}(H^2/t)$ after t iterations, with explicit dependence on the horizon length H.
   Context: Prior finite-time convergence analyses of exact NPG primarily focus on infinite-horizon discounted MDPs with stationary, horizon-independent transition kernels (Agarwal et al., 2021; Khodadadian et al., 2022; Xiao, 2022; Liu et al., 2024), where the discount factor γ induces a Bellman contraction. By contrast, in finite-horizon MDPs, transition kernels may vary across decision epochs and no discount factor is available.
2. We identify an exact population-projection oracle under a full support projection distribution in which linear Q-NPG coincides in policy space with tabular NPG. Under exact linear realizability, the same sublinear rate follows from the tabular analysis, providing an exact-oracle baseline for future sample-based linear MDP extensions.
   Context: Existing finite-time convergence results for NPG with linear function approximation are established in the infinite-horizon discounted setting (Agarwal et al., 2021).
3. For finite-horizon tabular MDPs, we prove that an increasing step size schedule yields geometric convergence, with the optimality gap decaying at a rate of $\mathcal{O}\left(\left(1 - \frac{1}{\vartheta_\rho}\right)^t\right)$ for a problem-dependent constant $\vartheta_\rho > 1$. We further provide a horizon-only robust step size schedule $\eta_t = \eta_0(H/(H-1))^t$ where $\eta_0 > 0$ and $H \geq 2$, as a practical contribution that preserves the geometric rate.
   Context: Linear convergence of NPG is known in the infinite-horizon discounted setting, via increasing step sizes (Khodadadian et al., 2022; Xiao, 2022), entropy regularization (Cen et al., 2022), or constant step sizes (Liu et al., 2024).

Finite-Time Analysis of the Natural Policy Gradient in Finite-Horizon Markov Decision Processes

Asha Barua[1], **Sajad Khodadadian** [1]
ashabarua@vt.edu, sajadk@vt.edu

[1]**Grado Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University, USA**

Abstract

Natural Policy Gradient (NPG) is a well-established Reinforcement Learning algorithm that underlies widely used methods such as Trust Region Policy Optimization and Proximal Policy Optimization, both of which have demonstrated strong empirical success. In this paper, we study exact NPG in finite-horizon Markov Decision Processes with known dynamics and horizon-dependent transition kernels. We provide the first finite-time convergence guarantees for this algorithm in this setting, for which we consider both constant and increasing step size regimes. With a constant step size $\eta_t = \eta$, we prove that NPG converges sublinearly with a rate of $\mathcal{O}(H^2/t)$ after t iterations, where H is the horizon length. We also extend this constant step size analysis to linear MDPs in an exact population-projection oracle under a full support projection distribution, recovering the same sublinear rate as in the tabular setting. Furthermore, with increasing step sizes, we prove that this algorithm achieves a linear convergence rate of $\mathcal{O}\left(\left(1-\frac{1}{\vartheta_\rho}\right)^t\right)$ for a problem-dependent constant $\vartheta_\rho > 1$, and the horizon-only robust schedule of the form $\eta_t = \eta_0(H/(H-1))^t$ where $\eta_0 > 0$ and $H \geq 2$, attains this same geometric rate.

1 Introduction

Reinforcement Learning (RL) has achieved substantial empirical success in domains such as robotic control, game playing, and autonomous navigation (Singh et al., 2022; Silver et al., 2016; Arulkumaran et al., 2017). Among RL methods, policy gradient algorithms are widely used for high-dimensional control. The Natural Policy Gradient (NPG) (Kakade, 2001), in particular, accounts for the geometry of the policy space and underlies widely used practical methods such as Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) and Proximal Policy Optimization (PPO) (Schulman et al., 2017). Mathematically, MDPs characterize the underlying structure of RL algorithms.

MDPs are commonly formulated over either an infinite horizon, with average or discounted return objectives, or a finite horizon, with decisions made over a fixed number of time steps (Puterman, 2014; Jin et al., 2020). Existing theoretical analyses of policy gradient methods have primarily considered infinite-horizon discounted MDPs, for which both asymptotic convergence (Kakade, 2001) and finite-time guarantees (Fazel et al., 2018; Agarwal et al., 2021) are well developed. Finite-horizon MDPs (Puterman, 2014; Jin et al., 2020), however, arise naturally in sequential planning, robotic manipulation, and RL for large language models (Bai et al., 2022; Ouyang et al., 2022; Zhan et al., 2023; Shakya et al., 2023; Bhandari & Russo, 2024; Zhang et al., 2025). Despite their practical relevance, finite-time convergence guarantees for exact NPG in finite-horizon MDPs remain unavailable, even in the tabular setting.

The finite-horizon setting introduces two structural difficulties. First, an optimal policy is generally nonstationary and therefore consists of a sequence of horizon-dependent decision rules. Consequently, updating the horizon-h decision rule can alter the state distributions at subsequent horizons $h+1, \ldots, H$, coupling policy improvement across the horizon. Second, the analyses in the discounted setting commonly exploit the strict contraction induced by the discount factor. No analogous discount-induced contraction is available in the finite-horizon setting, where policies, transition kernels, and value functions may depend on the horizon index.

In this paper, we study the finite time convergence of NPG for finite horizon MDPs. To the best of our knowledge, this is the first finite-time analysis of exact NPG in finite-horizon MDPs with known, horizon-dependent transition kernels and nonstationary policies; the closest finite-horizon convergence analysis concerns vanilla softmax policy gradient (Klein et al., 2024), whose rates are sublinear, derived from smoothness and weak Polyak-Łojasiewicz inequalities with model-dependent constants. Our main contributions are as follows:

- **Constant step size:** For tabular MDPs, we prove that NPG with a constant step size attains an optimality gap of order $\mathcal{O}(H^2/t)$ after t iterations. Under the linear MDP assumption of Jin et al. (2020), the same rate holds in an oracle regime with exact action-value evaluation and exact population projection under a full support projection distribution.
- **Increasing step sizes:** For tabular MDPs, we prove that increasing step sizes yield an optimality gap of order $\mathcal{O}((1-1/\vartheta_\rho)^t)$, where $\vartheta_\rho > 1$ is a problem-dependent distribution-mismatch coefficient. We further show that the horizon-only schedule $\eta_t = \eta_0(H/(H-1))^t$, with $\eta_0 > 0$ and $H \geq 2$, satisfies the step size growth condition required for geometric convergence; in the best case $\vartheta_\rho = H$, it yields the rate $\mathcal{O}\Big(\big(1-\frac{1}{H}\big)^t\Big)$.
- **Numerical illustration:** Finally, we present simulations that illustrate the convergence behavior predicted by the theoretical bounds.

1.1 Related work

In this section, we provide a brief overview of the existing literature most relevant to our analysis. Table 1 compares our results with prior work.

Table 1: Finite-time convergence guarantees for NPG and PMD. The NPG update considered here is equivalent to KL-based PMD (Shani et al., 2020; Xiao, 2022). "Horizon-dep. P^h" indicates whether the analysis permits horizon-dependent transition kernels. Sublinear and linear denote, respectively, an $\mathcal{O}(1/t)$ optimality gap and a geometric $\mathcal{O}(c^t)$ optimality gap for some $c < 1$. Step size regimes and rates are paired within each row.

Paper	Algorithm	Step size	Horizon-dep. P^h	Rate
Agarwal et al. (2021)	NPG	Constant	✗	Sublinear
Khodadadian et al. (2022)	NPG	Constant & Adaptive	✗	Sublinear & Linear
Xiao (2022)	PMD	Constant & Increasing	✗	Sublinear & Linear
Lan (2023)	PMD	Constant & Increasing	✗	Sublinear & Linear
Liu et al. (2024)	NPG	Constant	✗	Linear
This paper	**NPG (KL-PMD)**	**Constant & Increasing**	✓	**Sublinear & Linear**

NPG was introduced by Kakade (2001), and its finite-time behavior, together with that of related policy gradient methods, is now well understood for tabular infinite-horizon discounted MDPs (Agarwal et al., 2021; Bhandari & Russo, 2024). Sublinear $\mathcal{O}(1/t)$ guarantees have been established under constant step sizes, whereas increasing or adaptive step sizes can yield geometric convergence (Khodadadian et al., 2022; Xiao, 2022; Shani et al., 2020; Lan, 2023). Geometric convergence has also been obtained under constant step sizes (Liu et al., 2024) and through regularization (Mei et al., 2020; Cen et al., 2022; Zhan et al., 2023). These analyses concern infinite-horizon discounted MDPs with stationary transition kernels and therefore do not cover finite-horizon MDPs with horizon-

dependent dynamics. We establish the corresponding sublinear and geometric guarantees for exact NPG in this setting, providing a baseline for regularized and sample-based finite-horizon extensions.

2 Preliminaries

2.1 Finite-Horizon Markov Decision Process

We consider a finite-horizon Markov Decision Process (MDP) represented by the tuple $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \mathcal{R})$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action sets, respectively, and H is the horizon length. The transition kernels are $\mathcal{P} = \{P^h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S}) \mid h = 1, 2, \ldots, H\}$, where $P^h(s' \mid s, a)$ denotes the probability of transitioning to s' from state-action pair (s, a) at horizon h, and $\Delta(\mathcal{S})$ is the probability simplex over $\mathcal{S}$. The reward functions are $\mathcal{R} = \{R^h : \mathcal{S} \times \mathcal{A} \to [0, 1] \mid h = 1, 2, \ldots, H\}$.

In the finite-horizon setting, an optimal policy is generally nonstationary. We therefore represent a policy as $\pi = (\pi^1, \ldots, \pi^H)$, where $\pi^h(\cdot \mid s) \in \Delta(\mathcal{A})$ is the action distribution at horizon h in state s. For $h \in [H]$ and $s \in \mathcal{S}$, we define

$$V^{\pi,h}(s) := \mathbb{E}\left[\sum_{i=h}^{H} R^i(S^i, A^i) \mid S^h = s,\ A^i \sim \pi^i(\cdot \mid S^i),\ S^{i+1} \sim P^i(\cdot \mid S^i, A^i),\ i = h, \ldots, H\right].$$

We adopt the convention $V^{\pi,H+1}(s) = 0$ for every policy π and $s \in \mathcal{S}$. Since $R^h(s, a) \in [0, 1]$, we have $0 \le V^{\pi,h}(s) \le H - h + 1$ for all (π, h, s). The corresponding action-value function is

$$Q^{\pi,h}(s, a) := R^h(s, a) + \sum_{s' \in \mathcal{S}} P^h(s' \mid s, a) V^{\pi,h+1}(s'),$$

and the advantage function is defined as $A^{\pi,h}(s, a) = Q^{\pi,h}(s, a) - V^{\pi,h}(s)$. The goal of the agent is to find an optimal policy $\pi^\star$ such that for every horizon h, every state s and any policy π,

$$V^{\pi^\star,h}(s) \ge V^{\pi,h}(s).$$

Throughout the paper, a superscript $\star$ denotes a quantity associated with the optimal policy $\pi^\star$.

We analyze the Natural Policy Gradient (NPG), an iterative algorithm that finds the optimal policy through a smooth form of policy iteration (Kakade, 2001). In iteration t, given a constant step size $\eta > 0$, the update of the NPG policy takes the form:

$$\pi^h_{t+1}(a \mid s) = \frac{\pi^h_t(a \mid s) \exp\big(\eta Q^{\pi_t,h}(s, a)\big)}{Z^h_t(s)}, \qquad h \in [H],\ s \in \mathcal{S},\ a \in \mathcal{A}, \tag{1}$$

where the normalization constant is $Z^h_t(s) = \sum_{a' \in \mathcal{A}} \pi^h_t(a' \mid s) \exp(\eta Q^{\pi_t,h}(s, a'))$. The NPG update (1) is equivalently expressed as the KL-based policy mirror descent (PMD) update (Shani et al., 2020; Lan, 2023)

$$\pi^h_{t+1}(\cdot \mid s) = \arg\max_{p \in \Delta(\mathcal{A})} \left\{\eta \left\langle Q^{\pi_t,h}(s, \cdot), p\right\rangle - D_{\mathrm{KL}}\big(p \,||\, \pi^h_t(\cdot \mid s)\big)\right\}. \tag{2}$$

Both forms of the update are applied independently to every horizon-state pair (h, s). We use the multiplicative form (1) in the sublinear analysis and the variational form (2), with an iteration-dependent step size η_t, in the geometric analysis.

Furthermore, for any horizon $h \in [H]$ and every $i \ge h$, we define the state visitation distribution induced by a policy π as

$$d_s^{h \to i,\pi}(\tilde{s}) := \mathbb{P}\Big(S^i = \tilde{s} \mid S^h = s,\ A^j \sim \pi^j(\cdot \mid S^j),\ S^{j+1} \sim P^j(\cdot \mid S^j, A^j),\ j = h, \ldots, i-1\Big). \tag{3}$$

2.2 Linear MDP

Beyond the tabular setting, we consider the linear MDP model of Jin et al. (2020), which represents transition kernels and rewards using a low-dimensional feature map.

Assumption 1. *An MDP is linear with respect to a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if, for every $h \in [H]$, there exist a vector-valued function $\omega^h = (\omega_1^h, \ldots, \omega_d^h) : \mathcal{S} \to \mathbb{R}^d$ and a vector $\zeta^h \in \mathbb{R}^d$ such that, for every $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$P^h(\cdot \mid s, a) = \left\langle \phi(s, a), \omega^h(\cdot) \right\rangle, \qquad R^h(s, a) = \left\langle \phi(s, a), \zeta^h \right\rangle,$$

where $\langle \phi(s, a), \omega^h(\cdot) \rangle := \sum_{j=1}^d \phi_j(s, a)\, \omega_j^h(\cdot)$. We further assume that $\|\phi(s, a)\| \le B$ for some $B > 0$ and every $(s, a) \in \mathcal{S} \times \mathcal{A}$, and that $\max\{\| \sum_{s \in \mathcal{S}} \omega^h(s)\|, \|\zeta^h\|\} \le \sqrt{d}$ for every $h \in [H]$[1].

Under Assumption 1, for every policy π, there exist vectors $\{w^{\pi,h}\}_{h=1}^H \subset \mathbb{R}^d$ such that

$$Q^{\pi,h}(s, a) = \left\langle \phi(s, a), w^{\pi,h} \right\rangle \qquad \text{for every } (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{4}$$

The linear representation is most useful when $d \ll |\mathcal{S}||\mathcal{A}|$; choosing $\phi(s, a)$ as the standard basis of $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ recovers the tabular model. We parameterize the horizon-dependent policy by $\theta = (\theta^1, \ldots, \theta^H)$, where $\theta^h \in \mathbb{R}^d$, as

$$\pi_\theta^h(a \mid s) = \frac{\exp\left(\left\langle \phi(s, a), \theta^h \right\rangle\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\left\langle \phi(s, a'), \theta^h \right\rangle\right)}. \tag{5}$$

Writing $\pi_t := \pi_{\theta_t}$, in parameter space, the Q-NPG update (Agarwal et al., 2021) with constant step size $\eta > 0$ is

$$\theta_{t+1}^h = \theta_t^h + \eta\, w^{\pi_t,h}, \qquad h \in [H]. \tag{6}$$

Section 3.2 specifies an exact population-projection oracle under which (6) induces precisely the tabular NPG update (1) in policy space.

The remainder of the paper is organized as follows : Section 3 establishes sublinear convergence of NPG with a constant step size in the tabular (Section 3.1) and linear (Section 3.2) MDP settings. Section 4 establishes geometric convergence under increasing step sizes. Section 5 presents the simulation results, and Section 6 concludes. The main text contains the key lemmas, while detailed proofs and auxiliary results are deferred to the Appendix.

3 Sublinear Convergence

3.1 Tabular MDP

In this section, we first establish that NPG with a constant step size $\eta > 0$ in tabular MDPs achieves a finite-time global convergence rate of $\mathcal{O}((H - h + 1)^2/t)$ at any given horizon h after t iterations. The following lemmas provide the main ingredients for proving this result.

Lemma 1. *(Performance Difference Lemma) Consider a finite-horizon MDP and any pair of policies π and π'. For every horizon $h \in [H]$ and state $s \in \mathcal{S}$, we have*

$$V^{\pi,h}(s) - V^{\pi',h}(s) = \sum_{i=h}^{H} \sum_{\tilde{s} \in \mathcal{S},\, \tilde{a} \in \mathcal{A}} d_s^{h \to i, \pi}(\tilde{s})\, \pi^i(\tilde{a} \mid \tilde{s})\, A^{\pi',i}(\tilde{s}, \tilde{a}).$$

The next lemma lower bounds the one-step improvement of policies through the NPG update rule.

[1] Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm.

Lemma 2. *(Improvement Lower Bound for NPG) Consider the NPG update (1) with step size $\eta > 0$, and $t \geq 0$. Suppose that π_t has full support, i.e., $\pi_t^i(a \mid s) > 0$ for every $i \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$. Then, for every $h \in [H]$ and $s \in \mathcal{S}$,*

$$V^{\pi_{t+1},h}(s) - V^{\pi_t,h}(s) \geq \frac{1}{\eta} \log Z_t^h(s) - V^{\pi_t,h}(s) \geq 0.$$

Thus, the value functions are nondecreasing along the policy sequence. In particular, $V^{\pi^\star,h}(s) - V^{\pi_t,h}(s) \geq V^{\pi^\star,h}(s) - V^{\pi_T,h}(s)$ for every $t \leq T$; see Lemma 10 in the Appendix. We now state the first main result.

Theorem 1. *(Global Convergence of NPG) Consider a finite-horizon MDP and the NPG update (1) with constant step size $\eta > 0$ and uniform initialization $\pi_0^h(\cdot \mid s) = \mathrm{Unif}(\mathcal{A})$ for every $h \in [H]$ and $s \in \mathcal{S}$. Then, for every integer $T \geq 1$, horizon $h \in [H]$, and state $s \in \mathcal{S}$,*

$$V^{\pi^\star,h}(s) - V^{\pi_T,h}(s) \leq \frac{(H-h+1)\log|\mathcal{A}|}{\eta T} + \frac{(H-h+1)^2}{T}.$$

Consequently, for any fixed $h \in [H]$, if $\eta \geq \frac{\log|\mathcal{A}|}{H-h+1}$, then $V^{\pi^\star,h}(s) - V^{\pi_T,h}(s) \leq \frac{2(H-h+1)^2}{T}$ for every $s \in \mathcal{S}$, and an optimality gap of at most ε is attained at horizon h after $T \geq 2(H-h+1)^2/\varepsilon$ iterations.

The H^2 factor in Theorem 1 arises from Lemma 15, where the value potential is bounded by $\sum_{i=H-j}^{H}(H-i+1) = \frac{(j+1)(j+2)}{2} \leq (j+1)^2$. We do not know whether this quadratic dependence is unavoidable for constant step size NPG or is an artifact of our proof technique. In infinite-horizon discounted MDPs, Johnson et al. (2023) proved matching upper and lower bounds for exact PMD; a matching lower bound for the finite-horizon setting remains an open problem.

Because the tabular model is known, its action-value functions can be evaluated exactly. As $\eta \to \infty$, the NPG update approaches the greedy policy-improvement step of exact policy iteration with respect to $Q^{\pi_t,h}$, whereas a finite η yields a smooth policy-improvement step. We treat the tabular setting as an exact known-model baseline and next study the structured linear MDP setting as an exact-oracle benchmark for future sample-based extensions.

3.2 Linear MDP

We now identify an exact oracle regime in which the constant step size guarantee of Theorem 1 carries over to the linear MDP setting. Throughout this subsection, we work under Assumption 1 and use the linear representation (4), the softmax parametrization (5), and the Q-NPG update (6). Note that the softmax parametrization ensures that π_t has full support over actions, i.e., $\pi_t^i(a \mid s) > 0$ for every $i \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$, at every iteration $t \geq 0$.

To specify the oracle, fix a projection distribution v over $\mathcal{S} \times \mathcal{A}$ with full support:

$$v(s,a) > 0, \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{7}$$

Here v is a design choice rather than an environmental quantity; in the finite state-action setting, (7) can always be satisfied by taking, for example, $v = \mathrm{Unif}(\mathcal{S} \times \mathcal{A})$. We adopt this projection formulation because its empirical counterpart arises naturally in sample-based extensions, where coverage and estimation error must be addressed. For fixed policy parameters θ and horizon $h \in [H]$, define the population projection loss

$$L_h(w;\theta,v) := \mathbb{E}_{(s,a)\sim v}\left[\left(Q^{\pi_\theta,h}(s,a) - \langle w, \phi(s,a)\rangle\right)^2\right]. \tag{8}$$

By the linear realizability identity (4), the minimum of $L_h(\cdot;\theta,v)$ is attained and equals zero. Moreover, since v has full support, every exact minimizer represents the action-value function pointwise:

if $w \in \arg\min_{u \in \mathbb{R}^d} L_h(u; \theta, v)$, then

$$\sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} v(s,a) \big(Q^{\pi_\theta, h}(s,a) - \langle w, \phi(s,a) \rangle \big)^2 = 0,$$

and since each weight $v(s,a)$ is strictly positive and every summand is nonnegative, every summand must vanish. At iteration $t \geq 0$, the exact population-projection oracle returns

$$w^{\pi_t, h} \in \arg\min_{w \in \mathbb{R}^d} L_h(w; \theta_t, v), \qquad h \in [H], \tag{9}$$

so that

$$\langle w^{\pi_t, h}, \phi(s,a) \rangle = Q^{\pi_t, h}(s,a), \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{10}$$

The full support condition on v is used only to turn zero projection error into the pointwise identity (10), and non-uniqueness of the minimizer is harmless: all exact minimizers induce the same values $\langle w^{\pi_t, h}, \phi(s,a) \rangle$ on every state-action pair, and the policy update depends only on these values. full support is also not the only sufficient condition for (10): if $\Sigma_v := \mathbb{E}_{(s,a) \sim v}[\phi(s,a)\phi(s,a)^\top]$ is positive definite, then zero loss gives $(w - w^{\pi_\theta, h})^\top \Sigma_v (w - w^{\pi_\theta, h}) = 0$, hence $w = w^{\pi_\theta, h}$. We use the full support condition because it provides a direct finite-state argument without requiring additional assumptions on the feature map.

Consequently, the Q-NPG parameter update (6) induces the policy update

$$\pi^h_{t+1}(a \mid s) = \frac{\pi^h_t(a \mid s) \exp\big(\eta \langle w^{\pi_t, h}, \phi(s,a) \rangle\big)}{\sum_{a'} \pi^h_t(a' \mid s) \exp(\eta \langle w^{\pi_t, h}, \phi(s,a') \rangle)} = \frac{\pi^h_t(a \mid s) \exp\big(\eta Q^{\pi_t, h}(s,a)\big)}{\sum_{a'} \pi^h_t(a' \mid s) \exp(\eta Q^{\pi_t, h}(s,a'))}, \tag{11}$$

where the second equality uses (10). That is, under the oracle (9), Q-NPG induces exactly the tabular NPG update (1) in policy space, and the finite-time guarantee follows by applying Theorem 1 to the induced policy sequence.

Proposition 1. *Consider the Q-NPG update (6) with constant step size $\eta > 0$ under Assumption 1. Suppose that v satisfies the full support condition (7) and that, at every iteration $t \geq 0$ and horizon $h \in [H]$, the oracle returns an exact minimizer as in (9). Let $\theta^h_0 = \mathbf{0}$ for every $h \in [H]$, so that $\pi^h_0(\cdot \mid s) = \mathrm{Unif}(\mathcal{A})$ for every $h \in [H]$ and $s \in \mathcal{S}$. Then, for every integer $T \geq 1$, horizon $h \in [H]$, and state $s \in \mathcal{S}$,*

$$V^{\pi^\star, h}(s) - V^{\pi_T, h}(s) \leq \frac{(H - h + 1) \log |\mathcal{A}|}{\eta T} + \frac{(H - h + 1)^2}{T}.$$

Remark 1. *Proposition 1 identifies an exact oracle regime in which linear Q-NPG induces the same policy-space dynamics and achieves the same convergence order as tabular NPG. Within this regime, moving from the tabular model to the linear setting changes only the representation of the update direction. In sample-based extensions, $w^{\pi_t, h}$ is estimated from finite data, so the pointwise identity (10) may not hold exactly. The analysis must therefore account for estimation error and the coverage of v.*

Across the H horizons, storing the update directions $\{w^{\pi_t, h}\}_{h=1}^H$ uses $\mathcal{O}(dH)$ memory, compared with $\mathcal{O}(|\mathcal{S}||\mathcal{A}|H)$ for a full tabular action-value array. This comparison concerns only the representation of the update directions. Implementing exact action-value evaluation and the oracle (9) may still require access to the complete feature map, model dynamics, or state-action-level quantities. Therefore, Proposition 1 does not establish an end-to-end $\mathcal{O}(dH)$ memory or computational guarantee. Similarly, the absence of d from the bound follows from the exact-oracle assumption and does not imply a dimension-independent implementation guarantee. Estimating $w^{\pi_t, h}$ from data would introduce dependence on the feature dimension and data coverage, together with statistical error terms.

4 Geometric Convergence

In this section, we apply the PMD update (2) with an iteration-dependent step size $\eta_t > 0$ in place of the constant step size η, and fix an initial state distribution $\rho \in \Delta(\mathcal{S})$. For any horizon $h \in [H]$, define

$$V^{\pi,h}(\rho) \;:=\; \mathbb{E}_{S^h\sim\rho}\big[V^{\pi,h}(S^h)\big],$$

the expected return from horizon h when $S^h \sim \rho$ and policy $\pi = (\pi^1, \dots, \pi^H)$ is followed from horizon h onward; similarly, $d_\rho^{h\to i,\pi}(\tilde{s}) := \mathbb{E}_{S^h\sim\rho}[d_{S^h}^{h\to i,\pi}(\tilde{s})]$. We initialize the PMD iterates with any full support policy, i.e., $\pi_0^h(a \mid s) > 0$ for every $h \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$; the KL-based PMD update preserves full support for finite step sizes, since it takes the multiplicative form (1).

A key technical difficulty in the finite-horizon setting is that, for $h > 1$, the horizon-h visitation distribution induced by a policy π generally admits no policy-independent lower bound. To address this issue, we introduce an auxiliary objective that assigns policy-independent baseline mass to every horizon, analogous to the discounted infinite-horizon setting (Xiao, 2022).

Definition 1. *Let $H_0 \sim \mathrm{Unif}\{1, \dots, H\}$ be a random starting horizon, and let $V^{\pi,H_0}(\rho)$ be the return from reinitializing the process at horizon H_0 with $S^{H_0} \sim \rho$ and following policy π thereafter. The global multi-start objective is*

$$J(\pi) \;:=\; \mathbb{E}_{H_0\sim\mathrm{Unif}\{1,\dots,H\}}\big[V^{\pi,H_0}(\rho)\big] \;=\; \frac{1}{H}\sum_{h_0=1}^{H} V^{\pi,h_0}(\rho). \tag{12}$$

The optimality gap in the auxiliary objective $J(\pi)$ controls the horizon-1 optimality gap of the original objective up to a factor of H.

Lemma 3. *Let $\rho \in \Delta(\mathcal{S})$ and let $\pi^\star$ be an optimal nonstationary policy. Then for any policy π,*

$$V^{\pi^\star,1}(\rho) - V^{\pi,1}(\rho) \;\le\; H\big(J(\pi^\star) - J(\pi)\big). \tag{13}$$

To analyze the multi-start objective, we average the visitation distributions over all admissible starting horizons $h_0 \in \{1, \dots, i\}$, which yields the global multi-start horizon-i visitation measure[2]

$$\bar{d}_\rho^{i,\pi}(\tilde{s}) \;:=\; \frac{1}{H}\sum_{h_0=1}^{i} d_\rho^{h_0\to i,\pi}(\tilde{s}), \qquad i \in [H],\ \tilde{s} \in \mathcal{S}. \tag{14}$$

The next lemma verifies that this construction provides the required policy-independent lower bound.

Lemma 4. *Fix $\rho \in \Delta(\mathcal{S})$ and a policy π. Then for every $i \in [H]$ and $\tilde{s} \in \mathcal{S}$,*

$$\bar{d}_\rho^{i,\pi}(\tilde{s}) \;\ge\; \frac{1}{H}\rho(\tilde{s}). \tag{15}$$

We next derive the performance-difference identity for J, the multi-start analogue of Lemma 1.

Lemma 5 (Performance difference for the global multi-start objective)**.** *Let π and π' be nonstationary policies and fix $\rho \in \Delta(\mathcal{S})$. Then*

$$J(\pi') - J(\pi) = \sum_{i=1}^{H} \mathbb{E}_{S\sim\bar{d}_\rho^{i,\pi'}}\Big[\Big\langle Q^{\pi,i}(S,\cdot),\ \pi'^i(\cdot \mid S) - \pi^i(\cdot \mid S)\Big\rangle\Big]. \tag{16}$$

Lemma 5 expresses the multi-start objective difference in a form directly compatible with the KL-based PMD update; as a consequence, J is nondecreasing along the iterates.

[2]For any nonnegative finite measure μ on $\mathcal{S}$, we write $\mathbb{E}_{S\sim\mu}[f(S)] := \sum_{s\in\mathcal{S}} \mu(s)f(s)$, even when $\mu(\mathcal{S}) \neq 1$. In particular, $\bar{d}_\rho^{i,\pi}(\mathcal{S}) = i/H$.

Lemma 6. *Let $\{\pi_t\}_{t\geq 0}$ be generated by the KL-based PMD update (2) with step sizes $\eta_t > 0$. Then, for all $t \geq 0$,*

$$J(\pi_{t+1}) - J(\pi_t) = \frac{1}{H}\sum_{h_0=1}^{H}\Big(V^{\pi_{t+1},h_0}(\rho) - V^{\pi_t,h_0}(\rho)\Big) \;\geq\; 0.$$

For an optimal policy $\pi^\star$, define the multi-start optimality gap

$$\delta_t := J(\pi^\star) - J(\pi_t); \tag{17}$$

Following the monotonicity Lemma 6, $\delta_{t+1} \leq \delta_t$ for all $t \geq 0$.

Lemma 7. *Fix $t \geq 0$, and let π_{t+1} be generated from π_t by the KL-based PMD update (2) with step size $\eta_t > 0$. Suppose that $\pi_t^i(a \mid s) > 0$ for every $i \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$, so that all KL terms below are finite. Define the multi-start KL potential*

$$D_t^\star \;:=\; \sum_{i=1}^{H}\mathbb{E}_{S\sim\bar{d}_\rho^{i,\pi^\star}}\Big[D_{\mathrm{KL}}\big(\pi^{\star,i}(\cdot \mid S) \,||\, \pi_t^i(\cdot \mid S)\big)\Big]. \tag{18}$$

Then

$$\sum_{i=1}^{H}\mathbb{E}_{S\sim\bar{d}_\rho^{i,\pi^\star}}\Big[\big\langle Q^{\pi_t,i}(S,\cdot),\, \pi^{\star,i}(\cdot \mid S) - \pi_{t+1}^i(\cdot \mid S)\big\rangle\Big] \;\leq\; \frac{1}{\eta_t}\big(D_t^\star - D_{t+1}^\star\big). \tag{19}$$

Definition 2. *For each $t \geq 0$, the per-iteration multi-start distribution mismatch coefficient is*

$$\vartheta_t \;:=\; \max_{i\in[H]}\max_{\tilde{s}:\,\bar{d}_\rho^{i,\pi^\star}(\tilde{s})>0}\frac{\bar{d}_\rho^{i,\pi^\star}(\tilde{s})}{\bar{d}_\rho^{i,\pi_{t+1}}(\tilde{s})}, \tag{20}$$

with the convention $\vartheta_t = +\infty$ if $\bar{d}_\rho^{i,\pi^\star}(\tilde{s}) > 0$ but $\bar{d}_\rho^{i,\pi_{t+1}}(\tilde{s}) = 0$ for some pair $(i,\tilde{s})$.

By Lemma 4, if $\rho(\tilde{s}) > 0$ then $\bar{d}_\rho^{i,\pi_{t+1}}(\tilde{s}) \geq \rho(\tilde{s})/H$ for all i, so ϑ_t is finite whenever $\bar{d}_\rho^{i,\pi^\star}(\tilde{s}) > 0$ implies $\rho(\tilde{s}) > 0$.

Proposition 2. *Fix $\rho \in \Delta(\mathcal{S})$ and an optimal nonstationary policy $\pi^\star$. Let $\{\pi_t\}_{t\geq 0}$ be generated from a full support initial policy by the KL-based PMD update (2), with step sizes satisfying $0 < \eta_t < \infty$ for every $t \geq 0$. Then, for every $t \geq 0$ such that $\vartheta_t < \infty$,*

$$\vartheta_t(\delta_{t+1} - \delta_t) + \delta_t \leq \frac{1}{\eta_t}\big(D_t^\star - D_{t+1}^\star\big). \tag{21}$$

We further define the multi-start distribution mismatch coefficient

$$\vartheta_\rho := \begin{cases} H\max\limits_{i\in[H]}\max\limits_{s:\rho(s)>0}\dfrac{\bar{d}_\rho^{i,\pi^\star}(s)}{\rho(s)}, & \text{if } \bar{d}_\rho^{i,\pi^\star}(s) > 0 \Longrightarrow \rho(s) > 0 \\ & \text{for every } (i,s) \in [H]\times\mathcal{S}, \\ +\infty, & \text{otherwise.} \end{cases} \tag{22}$$

Thus, ϑ_ρ is finite if and only if, for every $i \in [H]$, the support of $\bar{d}_\rho^{i,\pi^\star}$ is contained in the support of ρ; in particular, ϑ_ρ is finite whenever ρ has full support on $\mathcal{S}$.

Lemma 8. *Fix $\rho \in \Delta(\mathcal{S})$ and let ϑ_ρ be as defined in (22). Then for every $t \geq 0$,*

$$\vartheta_t \;\leq\; \vartheta_\rho. \tag{23}$$

Theorem 2. *Fix $\rho \in \Delta(\mathcal{S})$ and consider the global multi-start objective $J(\pi)$ in (12). Let $\{\pi_t\}_{t\geq 0}$ be generated from a full support initial policy π_0 by the KL-based PMD update (2) with step sizes*

$\{\eta_t\}_{t\geq 0}$, and let $\pi^\star$ be an optimal nonstationary policy. Suppose $\vartheta_\rho \in (1,\infty)$ and the step sizes satisfy $\eta_0 > 0$ and

$$\eta_{t+1} \;\geq\; \frac{\vartheta_\rho}{\vartheta_\rho - 1}\,\eta_t, \qquad t = 0,1,2,\ldots \tag{24}$$

Then,

$$V^{\pi^\star,1}(\rho) - V^{\pi_t,1}(\rho) \;\leq\; H\Big(1 - \frac{1}{\vartheta_\rho}\Big)^t \left(\delta_0 + \frac{1}{\eta_0(\vartheta_\rho - 1)} D_0^\star\right), \qquad \forall t \geq 0. \tag{25}$$

Note that, unlike Theorem 1, which holds for all horizons $h \in [H]$, the bound (25) controls the suboptimality at the initial horizon $h = 1$ only. This is a consequence of the multi-start construction, which is essential for the geometric convergence argument; the reduction to the horizon-1 gap is carried out through Lemma 3.

Discussion: The role of the random start horizon H_0 is to create a policy-independent baseline mass at every horizon, and its uniform distribution is not arbitrary. To examine this, consider the weighted generalization $\mathbb{P}(H_0 = h) = m_h$ with $m_h > 0$ and $\sum_{h=1}^H m_h = 1$, the weighted objective $J_m(\pi) := \sum_{h=1}^H m_h\, V^{\pi,h}(\rho)$, and the weighted multi-start visitation measure

$$\bar{d}^{i,\pi}_{\rho,m}(s) \;:=\; \sum_{h_0=1}^{i} m_{h_0}\, d_\rho^{h_0 \to i,\pi}(s), \qquad i \in [H],\; s \in \mathcal{S}.$$

The summand $h_0 = i$ equals $m_i\,\rho(s)$ for every policy π, since initializing at horizon i involves no transitions; as the remaining summands are nonnegative, we obtain the policy-independent floor

$$\bar{d}^{i,\pi}_{\rho,m}(s) \;\geq\; m_i\,\rho(s), \qquad \forall i \in [H],\; \forall s \in \mathcal{S},$$

which is the only point in the analysis where the distribution of H_0 provides policy-independent control of the mismatch. Consequently, the floor-based mismatch bound degrades with $1/m_{\min}$, where $m_{\min} := \min_{i\in[H]} m_i$. On the other hand, by horizon-wise optimality of $\pi^\star$,

$$J_m(\pi^\star) - J_m(\pi) \;\geq\; m_1\big(V^{\pi^\star,1}(\rho) - V^{\pi,1}(\rho)\big),$$

so the reduction to the horizon-1 suboptimality gap incurs the factor $1/m_1$. Non-uniform weighting therefore trades these two worst-case bounds against each other: increasing m_1 beyond $1/H$ tightens the horizon-1 reduction, but since $\sum_{i=1}^H m_i = 1$, it forces some other weight below $1/H$ and thereby loosens the floor at that horizon. The uniform choice $m_i = 1/H$, which recovers J, $\bar{d}_\rho^{i,\pi}$, and ϑ_ρ, is the unique maximizer of $m_{\min}$ over the probability simplex; it treats all starting horizons symmetrically and prevents any single horizon from becoming a bottleneck through an arbitrarily small weight. We emphasize that this is a worst-case robustness rationale: it does not assert that uniform weighting minimizes the realized mismatch ratios for every MDP.

4.1 Horizon-Only Robust Step Size Schedule

Throughout this subsection, we assume $H \geq 2$. The case $H = 1$ is discussed at the end.

Lemma 9. *(Horizon lower bound on ϑ_ρ) Suppose that $\vartheta_\rho < \infty$, where ϑ_ρ is defined in (22). Then $\vartheta_\rho \geq H$. Consequently,*

$$\frac{\vartheta_\rho}{\vartheta_\rho - 1} \leq \frac{H}{H-1}. \tag{26}$$

Remark 2. *(Upper bounds on ϑ_ρ) Lemma 9 gives a lower bound on ϑ_ρ (see Appendix 8.2). For an upper bound, observe that*

$$\bar{d}_\rho^{i,\pi^\star}(s) \leq \sum_{\tilde{s}\in\mathcal{S}} \bar{d}_\rho^{i,\pi^\star}(\tilde{s}) = \frac{i}{H} \leq 1.$$

Definition (22) *therefore gives*

$$\vartheta_\rho \le \frac{H}{\min_{s:\,\rho(s)>0}\rho(s)}.$$

Moreover, ρ *is used only in the analysis and is not an input to the PMD update* (2)*, which is applied independently to every* (h, s)*. Thus, the same policy sequence can be analyzed using* $\rho = \mathrm{Unif}(\mathcal{S})$*. For this choice, the support condition in* (22) *holds automatically and* $\vartheta_\rho \le H|\mathcal{S}|$*. A bound under any initial distribution* μ *then follows from the uniform-*ρ *bound with an additional factor*

$$\max_{s\in\mathcal{S}} \frac{\mu(s)}{\rho(s)} \le |\mathcal{S}|,$$

because the per-state gaps $V^{\pi^\star,1}(s) - V^{\pi_t,1}(s)$ *are nonnegative. Thus,* ϑ_ρ *has a polynomial upper bound in* H *and* $|\mathcal{S}|$*. Under the uniform choice of* ρ*, the resulting worst-case contraction bound depends on* $|\mathcal{S}|$*. Obtaining sharper problem-dependent bounds on* ϑ_ρ *remains open.*

Corollary 1. *(Horizon-only robust step size schedule) Let* $\{\pi_t\}_{t\ge0}$ *be generated from a full support initial policy by the KL-based PMD update* (2)*, and let* $\pi^\star$ *be an optimal nonstationary policy. Let* δ_0 *and* $D_0^\star$ *be defined as in* (17) *and* (18)*, respectively. Suppose that* $\vartheta_\rho < \infty$*. For any* $\eta_0 > 0$*, the horizon-only step size schedule*

$$\eta_t = \eta_0 \left(\frac{H}{H-1}\right)^t, \qquad t = 0, 1, 2, \ldots, \tag{27}$$

satisfies the growth condition (24)*. Therefore, the geometric bound* (25) *in Theorem 2 holds for every* $t \ge 0$*. In particular, if* $\vartheta_\rho = H$*, the smallest value permitted by Lemma 9, then*

$$V^{\pi^\star,1}(\rho) - V^{\pi_t,1}(\rho) \le H\left(1 - \frac{1}{H}\right)^t \left(\delta_0 + \frac{D_0^\star}{\eta_0(H-1)}\right), \qquad t \ge 0. \tag{28}$$

The growth condition (24) depends on the generally unknown coefficient ϑ_ρ and therefore cannot be implemented directly. Corollary 1 removes this dependence. By Lemma 9, the multiplier $H/(H-1)$ upper-bounds $\vartheta_\rho/(\vartheta_\rho - 1)$ whenever $\vartheta_\rho < \infty$. Therefore, the schedule (27) depends only on the known horizon length and satisfies the required growth condition without problem-dependent tuning.

We next compare this result with classical dynamic programming. When the dynamics are known, backward induction computes $\pi^\star$ exactly in one pass over the horizons. Thus, Theorem 2, whose iteration complexity is of order $\vartheta_\rho \log(1/\varepsilon) \ge H \log(1/\varepsilon)$ (where $0 < \varepsilon < 1$), does not provide a computational advantage over backward induction. Indeed, since $\eta_t \to \infty$ under (27), each PMD update tends toward greedy policy improvement with respect to the current action-values.

Our analysis has a different purpose. NPG/PMD replaces the greedy maximization in dynamic programming with the smooth multiplicative update (1). For a finite step size, this update changes continuously with $Q^{\pi_t,h}$, while a small change in the action-values can change the greedy policy abruptly. This smooth policy update motivates TRPO/PPO-style methods and can be implemented using samples. In infinite-horizon discounted MDPs, inexact PMD analyses show how errors in estimating the action-values enter the convergence bound (Lan, 2023; Xiao, 2022). We analyze the iteration complexity of the exact update in finite-horizon MDPs and leave the corresponding inexact analysis for future work.

The assumption $H \ge 2$ is required only for the mismatch-based growth condition. When $H = 1$, the multi-start objective reduces to the original objective, $\bar{d}_\rho^{1,\pi} = \rho$, and $\vartheta_\rho = 1$, so the ratio $\vartheta_\rho/(\vartheta_\rho - 1)$ is undefined: in this single-step setting there is no coupling across horizons; the distribution-mismatch mechanism is absent, and an increasing step size regime is not the appropriate tool.

5 Simulation

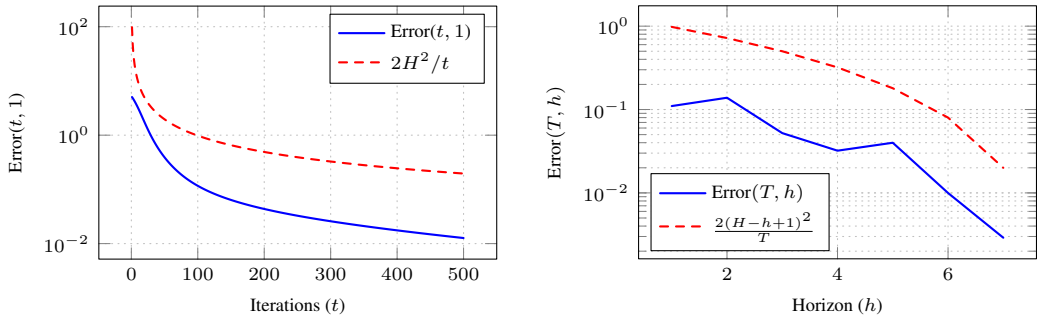
In this section, we perform simulations to illustrate the sublinear and linear convergence behavior of the NPG algorithm under constant and increasing step sizes, and to validate the predicted scaling

behavior of our theoretical bounds. For any horizon $h \in [H]$ and iteration t, we measure the suboptimality gap as

$$\text{Error}(t,h) \;=\; \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\left(V^{\pi^\star,h}(s) - V^{\pi_t,h}(s)\right).$$

5.1 Sublinear convergence with constant step size

We construct a randomly generated finite-horizon MDP with $|\mathcal{S}| = 15$, $|\mathcal{A}| = 4$, and $H = 7$.



(a) NPG sublinear convergence at $h = 1$.

(b) Horizon-wise error at fixed $T = 100$ (all h).

Figure 1: Constant step size NPG on a randomly generated finite-horizon tabular MDP: (a) error vs. iteration at $h = 1$; (b) error vs. all horizons at fixed iteration $T = 100$, where x-axis is on linear scale and y-axis is on logarithmic scale. In the plots, the solid (blue) curve shows the empirical error, and the dashed (red) curve shows the theoretical upper bound.

In Figure 1a, we fix $h = 1$, initialize π_0 uniformly, and run NPG for $T = 500$ iterations with constant step size $\eta = \log(|\mathcal{A}|)/H$. We plot the empirical gap together with the theoretical upper bound $2H^2/t$ from Theorem 1. The empirical error decreases monotonically and remains below the theoretical curve, exhibiting the predicted $\mathcal{O}(1/t)$ scaling; the gap between the two is expected, as the bound is a worst-case guarantee.

In Figure 1b, we instead fix the iteration to $T = 100$ and evaluate horizon-wise performance. Starting from the same uniform initialization, we run NPG with the per-horizon step sizes $\eta^h = \log(|\mathcal{A}|)/(H-h+1)$ of Theorem 1 and plot $\text{Error}(T,h)$ against h, together with the corresponding bound $2(H-h+1)^2/T$. As h increases, both the empirical gap and the bound shrink, since fewer future rewards remain to be optimized and the bound scales with $(H-h+1)^2$ at fixed T.

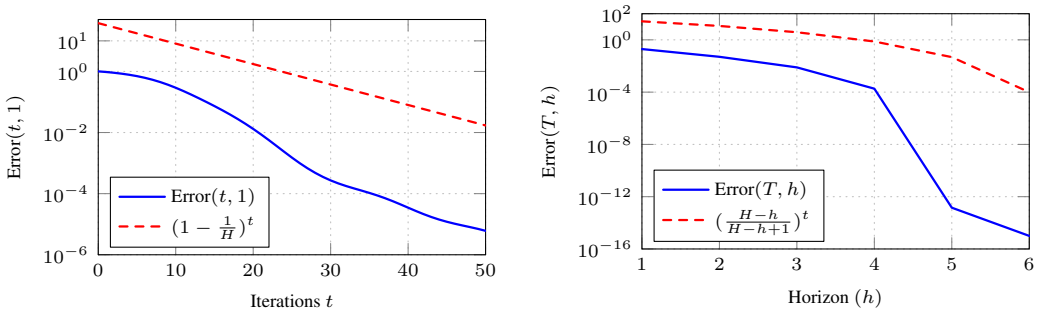
5.2 Geometric convergence with increasing step size

We next examine the geometric convergence guarantee of Theorem 2. To obtain an instance with a mismatch coefficient known in closed form, we consider an MDP with $|\mathcal{S}| = 15$, $|\mathcal{A}| = 5$, and $H = 7$. The rewards are drawn independently from the uniform distribution on $[0,1]$. Across all horizons, we use the same action-independent transition kernel,

$$P^h(s' \mid s,a) \;=\; 0.7\cdot\mathbf{1}\{s' = s\} \;+\; 0.3\cdot\mathbf{1}\{s' = s+1 \;(\text{mod } |\mathcal{S}|)\},$$

This kernel is doubly stochastic, and hence the uniform initial-state distribution ρ is invariant under every policy. Consequently, the optimal multi-start visitation measures satisfy $\bar{d}_\rho^{h,\pi^\star} = \frac{h}{H}\rho, h \in [H]$ which implies $\vartheta_\rho = H$. Thus, this instance attains the lower bound in Lemma 9.

In Figure 2a, we consider the objective starting at $h = 1$. We initialize π_0 uniformly and run NPG for $T = 50$ iterations using the horizon-only schedule from Corollary 1. Because $\vartheta_\rho = H$, the



(a) NPG linear convergence at $h = 1$ ($\vartheta_\rho = H$).

(b) Horizon-wise error at fixed $T = 20$ (all h).

Figure 2: Increasing step size NPG on the structured instance with $\vartheta_\rho = H$, under the horizon-only robust schedule of Corollary 1: (a) error vs. iteration at $h = 1$; (b) error vs. all horizons at fixed iteration $T = 20$. Both plots use linear x-axis and log-scale y-axis. The solid (blue) curve shows the empirical error, and the dashed (red) curve shows the corresponding geometric reference curve, which is undefined at $h = H$ where $\vartheta_\rho = 1$.

contraction factor in Theorem 2 becomes $1 - 1/H$. The empirical error decreases geometrically and remains below the theoretical bound evaluated in this instance, consistent with the geometric convergence guaranteed by the theorem.

In Figure 2b, we fix $T = 20$ and compare Error(T, h) across all horizons $h \in [H - 1]$. For each starting horizon h, we regard the tail beginning at h as a separate finite-horizon MDP with effective horizon $H - h + 1$. We initialize a separate uniform policy on each tail and use the corresponding horizon-only robust schedule $\eta^h_{t+1} = \frac{H-h+1}{H-h}\eta^h_t$ of Corollary 1. Because every tail retains the same uniform-invariant, action-independent transition kernel, its multi-start visitation measures satisfy $\bar{d}^{h,\pi^\star}_\rho = \frac{h}{H-h+1}\rho, h \in [H-h+1]$, and its mismatch coefficient is therefore exactly $\vartheta_\rho = H - h + 1$. Each data point is directly covered by the best case of Theorem 2 applied to the corresponding tail MDP. The empirical errors and their theoretical upper bounds become smaller for later starting horizons, reflecting the shorter remaining horizon. We omit $h = H$ because the corresponding tail has effective horizon 1. In this case, $\vartheta_\rho = 1$, and the growth factor is undefined; thus, the increasing step size regime does not apply (see Section 4.1). We repeat both experiments on a randomly generated MDP, where ϑ_ρ is evaluated directly from the multi-start visitation measures and exceeds its horizon-based lower bound in the generated instance, confirming that the guarantee is not an artifact of the structured instance (see Appendix 8.1).

6 Conclusion

In this paper, we have presented the finite-time analysis for exact NPG in finite-horizon MDPs. For tabular MDPs, we have proved that NPG with a constant step size achieves the sublinear rate $\mathcal{O}(H^2/t)$. The same rate holds for linear MDPs under an exact population-projection oracle with a full support projection distribution, where linear Q-NPG induces the same policy-space update as tabular NPG. For tabular MDPs, we have established geometric convergence of the horizon-1 optimality gap under increasing step sizes, for both the problem-dependent schedule and the horizon-only robust schedule.

Regarding future directions, an important next step is to develop sample-based finite-time guarantees for these results. Another direction is to establish lower bounds that determine whether the H^2/t rate in Theorem 1 and the H-factor losses in the multi-start reduction are necessary.

Broader Impact Statement

This work is theoretical, and we do not foresee any negative societal impacts.

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.

Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017. DOI: 10.1109/MSP.2017.2743240.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Thomas Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022. URL https://api.semanticscholar.org/CorpusID:248118878.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927, 2024.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning (ICML)*, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pp. 2137–2143. PMLR, 2020.

Emmeran Johnson, Ciara Pike-Burke, and Patrick Rebeschini. Optimal convergence rate for exact policy mirror descent in discounted markov decision processes. *Advances in Neural Information Processing Systems*, 36:76496–76524, 2023.

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

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

Sara Klein, Simon Weissmann, and Leif Döring. Beyond stationarity: Convergence analysis of stochastic softmax policy gradient methods. In *International Conference on Learning Representations (ICLR)*, 2024.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106, 2023.

Jiacai Liu, Wenye Li, and Ke Wei. Elementary analysis of policy gradient methods. *arXiv preprint arXiv:2404.03372*, 2024.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*, pp. 6820–6829. PMLR, 2020.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

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

Ashish Kumar Shakya, Gopinatha Pillai, and Sohom Chakrabarty. Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications*, 231:120495, 2023. ISSN 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2023.120495. URL https://www.sciencedirect.com/science/article/pii/S0957417423009971.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5668–5675, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Bharat Singh, Rajesh Kumar, and Vinay Pratap Singh. Reinforcement learning in robotic applications: a comprehensive survey. *Artificial Intelligence Review*, 55(2):945–990, 2022.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091, 2023.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025.

Supplementary Materials

The following content was not necessarily subject to peer review.

We recall that for a finite-horizon MDP, the state visitation distribution captures the probability of visiting state $\tilde{s}$ at horizon i given that the process starts from s at horizon h and follows policy π thereafter in Eq. (3). This distribution satisfies the following basic properties:

- Initialization: Since the process must be in the state s at the horizon h,

$$d_s^{h\to h,\pi}(\tilde{s}) = \mathbb{1}\{s = \tilde{s}\}.$$

- Normalization: At every horizon $i \in [h, H]$,

$$\sum_{\tilde{s}} d_s^{h\to i,\pi}(\tilde{s}) \;=\; 1, \qquad \forall i \geq h.$$

7 Sublinear Convergence: Supporting Lemmas and Proofs

7.1 MDP Setting Proofs

In this appendix, we provide detailed algebraic derivations and proofs that support Theorem 1. First, we give the detailed proof of Performance Difference Lemma 1.

Proof of Lemma 1

Proof. Using the telescoping argument together with the tower property of conditional expectations, we derive

$$\begin{aligned}
V^{\pi,h}(s) - V^{\pi',h}(s) &= \mathbb{E}\left[\sum_{i=h}^{H} R^i(S^i, A^i) \mid S^h = s,\ A^j \sim \pi^j(\cdot \mid S^j),\ \forall j = h, \ldots, H\right] - V^{\pi',h}(s) \\
&= \mathbb{E}\left[\sum_{i=h}^{H} R^i(S^i, A^i) - V^{\pi',h}(S^h) \mid S^h = s,\ A^j \sim \pi^j(\cdot \mid S^j)\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\sum_{i=h}^{H} \Big(R^i(S^i, A^i) + V^{\pi',i+1}(S^{i+1}) - V^{\pi',i}(S^i)\Big) \mid S^h = s,\ A^j \sim \pi^j(\cdot \mid S^j)\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\sum_{i=h}^{H} \Big(R^i(S^i, A^i) + \mathbb{E}\big[V^{\pi',i+1}(S^{i+1}) \mid S^i, A^i\big] - V^{\pi',i}(S^i)\Big) \mid S^h = s,\ A^j \sim \pi^j(\cdot \mid S^j)\right] \\
&\overset{(c)}{=} \mathbb{E}\left[\sum_{i=h}^{H} A^{\pi',i}(S^i, A^i) \mid S^h = s,\ A^j \sim \pi^j(\cdot \mid S^j)\right] \\
&= \sum_{i=h}^{H} \sum_{\tilde{s},\tilde{a}} \mathbb{P}\left(S^i = \tilde{s},\ A^i = \tilde{a} \mid S^h = s,\ A^j \sim \pi^j(\cdot \mid S^j),\ \forall j = h, \ldots, i\right) A^{\pi',i}(\tilde{s}, \tilde{a}) \\
&\overset{(d)}{=} \sum_{i=h}^{H} \sum_{\tilde{s},\tilde{a}} \mathbb{P}\left(S^i = \tilde{s} \mid S^h = s,\ A^j \sim \pi^j(\cdot \mid S^j),\ \forall j = h, \ldots, i-1\right) \pi^i(\tilde{a} \mid \tilde{s})\, A^{\pi',i}(\tilde{s}, \tilde{a}) \\
&\overset{(e)}{=} \sum_{i=h}^{H} \sum_{\tilde{s},\tilde{a}} d_s^{h\to i,\pi}(\tilde{s})\, \pi^i(\tilde{a} \mid \tilde{s})\, A^{\pi',i}(\tilde{s}, \tilde{a}).
\end{aligned}$$

Here (a) uses the telescoping identity $\sum_{i=h}^{H} \big(V^{\pi',i+1}(S^{i+1}) - V^{\pi',i}(S^i)\big) = -V^{\pi',h}(S^h) + V^{\pi',H+1}(S^{H+1})$ together with the convention $V^{\pi',H+1} \equiv 0$, (b) applies the tower property of conditional expectations, (c) uses $A^{\pi',i}(s,a) = Q^{\pi',i}(s,a) - V^{\pi',i}(s)$ with $Q^{\pi',i}(s,a) = R^i(s,a) +$

$\mathbb{E}\big[V^{\pi',i+1}(S^{i+1}) \mid S^i = s, A^i = a\big]$, (d) factorizes the joint law of (S^i, A^i) as $\mathbb{P}(S^i = \tilde{s})\,\pi^i(\tilde{a} \mid \tilde{s})$, and (e) substitutes the definition of the state visitation distribution in Eq. (3). □

Lemma 10. *Fix a horizon $h \in [H]$. Let $\{\pi_t\}_{t=0}^T$ be a sequence of policies such that*

$$V^{\pi_{t+1},h}(s) \;\geq\; V^{\pi_t,h}(s) \quad \textit{for all } s \in \mathcal{S} \textit{ and } t = 0, \ldots, T-1,$$

and let $\pi^\star$ be an optimal policy, i.e., $V^{\pi^\star,h}(s) \geq V^{\pi,h}(s)$ for all $s \in \mathcal{S}$ and all policies π. Then, for every $t \leq T$ and every $s \in \mathcal{S}$,

$$V^{\pi^\star,h}(s) - V^{\pi_t,h}(s) \;\geq\; V^{\pi^\star,h}(s) - V^{\pi_T,h}(s).$$

Proof of Lemma 10. Applying the monotonicity assumption repeatedly along $t, t+1, \ldots, T$ gives $V^{\pi_T,h}(s) \geq V^{\pi_t,h}(s)$. Multiplying (-1) in both sides of the inequality and adding $V^{\pi^\star,h}(s)$ to both sides yields

$$V^{\pi^\star,h}(s) - V^{\pi_t,h}(s) \;\geq\; V^{\pi^\star,h}(s) - V^{\pi_T,h}(s);$$

that is, the suboptimality gap is nonincreasing along the sequence $\{\pi_t\}_{t=0}^T$, which proves the claim. □

Lemma 11 (Nonnegativity of the KL divergence)**.** *For any distributions $\pi(\cdot \mid s)$ and $\pi'(\cdot \mid s)$ over a finite action set $\mathcal{A}$, the KL divergence satisfies*

$$D_{\mathrm{KL}}\big(\pi(\cdot \mid s) \,||\, \pi'(\cdot \mid s)\big) \geq 0.$$

If $\pi(\cdot \mid s)$ is not absolutely continuous with respect to $\pi'(\cdot \mid s)$, then the KL divergence is defined as $+\infty$, and the inequality is immediate.

Proof of Lemma 11. If $\pi(\cdot \mid s)$ is not absolutely continuous with respect to $\pi'(\cdot \mid s)$, then there exists an action $a \in \mathcal{A}$ such that

$$\pi(a \mid s) > 0 \qquad \text{and} \qquad \pi'(a \mid s) = 0.$$

In this case,

$$D_{\mathrm{KL}}\big(\pi(\cdot \mid s) \,||\, \pi'(\cdot \mid s)\big) = +\infty,$$

and the claim is immediate.

Otherwise, assume

$$\pi(\cdot \mid s) \ll \pi'(\cdot \mid s),$$

meaning that $\pi(a \mid s) > 0$ implies $\pi'(a \mid s) > 0$ for every $a \in \mathcal{A}$. Hence all ratios below are well-defined on the support of $\pi(\cdot \mid s)$. We also use the standard convention

$$0 \log \frac{0}{q} = 0, \qquad q \geq 0,$$

so actions with $\pi(a \mid s) = 0$ contribute zero to the KL divergence.

Let

$$\mathrm{supp}(\pi(\cdot \mid s)) := \{a \in \mathcal{A} : \pi(a \mid s) > 0\}.$$

From the definition of the KL divergence,

$$\begin{aligned} D_{\mathrm{KL}}\big(\pi(\cdot \mid s) \,||\, \pi'(\cdot \mid s)\big) &= \sum_{a \in \mathrm{supp}(\pi(\cdot|s))} \pi(a \mid s) \log \frac{\pi(a \mid s)}{\pi'(a \mid s)} \\ &= \sum_{a \in \mathrm{supp}(\pi(\cdot|s))} \pi(a \mid s) \left(-\log \frac{\pi'(a \mid s)}{\pi(a \mid s)}\right). \end{aligned}$$

Since $x \mapsto -\log x$ is convex, Jensen's inequality gives

$$D_{\mathrm{KL}}\big(\pi(\cdot \mid s) \,||\, \pi'(\cdot \mid s)\big) \geq -\log\left(\sum_{a \in \mathrm{supp}(\pi(\cdot|s))} \pi(a \mid s) \frac{\pi'(a \mid s)}{\pi(a \mid s)}\right)$$
$$= -\log\left(\sum_{a \in \mathrm{supp}(\pi(\cdot|s))} \pi'(a \mid s)\right).$$

Since

$$\sum_{a \in \mathrm{supp}(\pi(\cdot|s))} \pi'(a \mid s) \leq \sum_{a \in \mathcal{A}} \pi'(a \mid s) = 1,$$

and $x \mapsto -\log x$ is decreasing on $(0, \infty)$, it follows that

$$-\log\left(\sum_{a \in \mathrm{supp}(\pi(\cdot|s))} \pi'(a \mid s)\right) \geq -\log(1) = 0.$$

Therefore,

$$D_{\mathrm{KL}}\big(\pi(\cdot \mid s) \,||\, \pi'(\cdot \mid s)\big) \geq 0,$$

which proves the claim. □

Lemma 12. *Fix an iteration $t \geq 0$, a horizon $h \in [H]$, a state $s \in \mathcal{S}$, and any step size $\eta > 0$. Then the following inequality holds:*

$$\frac{1}{\eta} \log Z_t^h(s) - V^{\pi_t,h}(s) \;\geq\; 0, \quad \textit{where} \quad V^{\pi_t,h}(s) = \sum_{a' \in \mathcal{A}} \pi_t^h(a' \mid s)\, Q^{\pi_t,h}(s, a').$$

Proof of Lemma 12. Substituting the definition of the normalization constant $Z_t^h(s)$ from Eq. (1) and applying Jensen's inequality to the concave function $\log$, we obtain

$$\log Z_t^h(s) = \log\left[\sum_{a'} \pi_t^h(a' \mid s)\, \exp\big(\eta\, Q^{\pi_t,h}(s, a')\big)\right] \;\geq\; \sum_{a'} \pi_t^h(a' \mid s)\, \eta\, Q^{\pi_t,h}(s, a') \;=\; \eta\, V^{\pi_t,h}(s).$$

Dividing both sides by $\eta > 0$ yields

$$\frac{1}{\eta} \log Z_t^h(s) \;\geq\; V^{\pi_t,h}(s),$$

which is the claim. □

Lemma 13. *Fix an iteration $t \geq 0$, a horizon $h \in [H]$, a state $s \in \mathcal{S}$, and a step size $\eta > 0$. For each $l \in \{h, \ldots, H\}$, define*

$$F_l := \sum_{\tilde{s} \in \mathcal{S}} d_s^{h \to l, \pi_{t+1}}(\tilde{s}) \left[\frac{1}{\eta} \log Z_t^l(\tilde{s}) - V^{\pi_t,l}(\tilde{s})\right].$$

Then

$$\sum_{l=h}^{H} F_l \geq \frac{1}{\eta} \log Z_t^h(s) - V^{\pi_t,h}(s).$$

Proof of Lemma 13. We consider,

$$\sum_{l=h}^{H} F_l = \sum_{l=h}^{H} \sum_{\tilde{s}} d_s^{h \to l, \pi_{t+1}}(\tilde{s}) \left[\frac{1}{\eta} \log Z_t^l(\tilde{s}) - V^{\pi_t,l}(\tilde{s})\right]$$

$$\geq \sum_{\tilde{s}} d_s^{h \to h, \pi_{t+1}}(\tilde{s}) \left[\frac{1}{\eta} \log Z_t^h(\tilde{s}) - V^{\pi_t, h}(\tilde{s}) \right] + \sum_{l=h+1}^{H} \sum_{\tilde{s}} d_s^{h \to l, \pi_{t+1}}(\tilde{s}) \left[\frac{1}{\eta} \log Z_t^l(\tilde{s}) - V^{\pi_t, l}(\tilde{s}) \right]$$

$$= F_h + \sum_{l=h+1}^{H} F_l.$$

For case, $l = h$:

$$F_h = \sum_{\tilde{s}} d_s^{h \to h, \pi_{t+1}}(\tilde{s}) \left[\frac{1}{\eta} \log Z_t^h(\tilde{s}) - V^{\pi_t, h}(\tilde{s}) \right] = \frac{1}{\eta} \log Z_t^h(s) - V^{\pi_t, h}(s).$$

For case, $l > h$:

$$\sum_{l=h+1}^{H} F_l = \sum_{l=h+1}^{H} \sum_{\tilde{s}} d_s^{h \to l, \pi_{t+1}}(\tilde{s}) \left[\frac{1}{\eta} \log Z_t^l(\tilde{s}) - V^{\pi_t, l}(\tilde{s}) \right] \geq 0. \qquad \text{(Lemma 12)}$$

Thus, considering both of the cases, we can conclude as follows,

$$\sum_{l=h}^{H} F_l \geq \frac{1}{\eta} \log Z_t^h(s) - V^{\pi_t, h}(s).$$

□

Next, we present the lemma for the improved lower bound, which serves as a crucial component throughout.

Proof of Lemma 2

Proof. Since the rewards are bounded and the horizon is finite, $Q^{\pi_t, i}(s, a)$ is finite for every $i \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$. Therefore, the multiplicative update (1) and the full support assumption imply

$$\pi_{t+1}^i(a \mid s) > 0, \qquad \forall i \in [H],\ s \in \mathcal{S},\ a \in \mathcal{A}.$$

Consequently, all logarithmic ratios below are well-defined, and

$$D_{\mathrm{KL}}\big(\pi_{t+1}^i(\cdot \mid s) \,||\, \pi_t^i(\cdot \mid s)\big) < \infty.$$

Applying Lemma 1 with $(\pi, \pi') = (\pi_{t+1}, \pi_t)$ gives

$$\begin{aligned}
V^{\pi_{t+1}, h}(s) - V^{\pi_t, h}(s) &= \sum_{l=h}^{H} \sum_{\tilde{s}, \tilde{a}} d_s^{h \to l, \pi_{t+1}}(\tilde{s})\, \pi_{t+1}^l(\tilde{a} \mid \tilde{s}) A^{\pi_t, l}(\tilde{s}, \tilde{a}) \\
&= \sum_{l=h}^{H} \sum_{\tilde{s}, \tilde{a}} d_s^{h \to l, \pi_{t+1}}(\tilde{s})\, \pi_{t+1}^l(\tilde{a} \mid \tilde{s}) \left[Q^{\pi_t, l}(\tilde{s}, \tilde{a}) - V^{\pi_t, l}(\tilde{s}) \right] \\
&= \sum_{l=h}^{H} \sum_{\tilde{s}, \tilde{a}} d_s^{h \to l, \pi_{t+1}}(\tilde{s})\, \pi_{t+1}^l(\tilde{a} \mid \tilde{s}) \left[\frac{1}{\eta} \log \frac{\pi_{t+1}^l(\tilde{a} \mid \tilde{s}) Z_t^l(\tilde{s})}{\pi_t^l(\tilde{a} \mid \tilde{s})} - V^{\pi_t, l}(\tilde{s}) \right] \\
&\qquad \text{(Eq.(1))} \\
&= \sum_{l=h}^{H} \sum_{\tilde{s}} d_s^{h \to l, \pi_{t+1}}(\tilde{s}) \left[\frac{1}{\eta} D_{\mathrm{KL}}(\pi_{t+1}^l(\cdot \mid \tilde{s}) \,||\, \pi_t^l(\cdot \mid \tilde{s})) + \frac{1}{\eta} \log Z_t^l(\tilde{s}) - V^{\pi_t, l}(\tilde{s}) \right] \\
&\geq \sum_{l=h}^{H} \sum_{\tilde{s}} d_s^{h \to l, \pi_{t+1}}(\tilde{s}) \left[\frac{1}{\eta} \log Z_t^l(\tilde{s}) - V^{\pi_t, l}(\tilde{s}) \right] \qquad \text{(Lemma 11)}
\end{aligned}$$

$$\geq \frac{1}{\eta} \log Z_t^h(s) - V^{\pi_t,h}(s) \qquad \text{(Lemma 13)}$$
$$\geq 0. \qquad \text{(Lemma 12)}$$

Hence,

$$V^{\pi_{t+1},h}(s) - V^{\pi_t,h}(s) \geq \frac{1}{\eta} \log Z_t^h(s) - V^{\pi_t,h}(s) \geq 0.$$

□

Lemma 14. *Let $\pi_0^i(\cdot \mid s) = \mathrm{Unif}(\mathcal{A})$ for all $i \in [H]$ and $s \in \mathcal{S}$. Then for every $i \in [H]$ and every $s \in \mathcal{S}$,*

$$D_{\mathrm{KL}}\big(\pi^{\star,i}(\cdot \mid s) \,||\, \pi_0^i(\cdot \mid s)\big) \;\leq\; \log|\mathcal{A}|,$$

and consequently, for any $j \in \{0, \ldots, H-1\}$ and any $s \in \mathcal{S}$,

$$m_0^{H-j\to H}(s) \;=\; \sum_{i=H-j}^{H} \sum_{\tilde{s}} d_s^{H-j\to i,\pi^\star}(\tilde{s})\, D_{\mathrm{KL}}\big(\pi^{\star,i}(\cdot \mid \tilde{s}) \,||\, \pi_0^i(\cdot \mid \tilde{s})\big) \;\leq\; (j+1)\log|\mathcal{A}|,$$

where the second inequality uses $\sum_{\tilde{s}} d_s^{H-j\to i,\pi^\star}(\tilde{s}) = 1$.

Proof of Lemma 14. Applying the definition of the KL divergence and substituting the uniform initial policy $\pi_0^i(a \mid s) = \frac{1}{|\mathcal{A}|}$, we have

$$\begin{aligned}
D_{\mathrm{KL}}\big(\pi^{\star,i}(\cdot \mid s) \,||\, \pi_0^i(\cdot \mid s)\big) &= \sum_{a\in\mathcal{A}} \pi^{\star,i}(a \mid s)\, \log \frac{\pi^{\star,i}(a \mid s)}{\pi_0^i(a \mid s)} \\
&= \sum_{a\in\mathcal{A}} \pi^{\star,i}(a \mid s)\, \log\big[\pi^{\star,i}(a \mid s) \cdot |\mathcal{A}|\big] \\
&= \sum_{a\in\mathcal{A}} \pi^{\star,i}(a \mid s) \log \pi^{\star,i}(a \mid s) \;+\; \log|\mathcal{A}| \\
&\leq \log|\mathcal{A}|,
\end{aligned}$$

where the last step uses $\sum_a \pi^{\star,i}(a \mid s) \log \pi^{\star,i}(a \mid s) \leq 0$, since each term is nonpositive, and $\sum_a \pi^{\star,i}(a \mid s) = 1$ in the second summand.

For the second claim, fix $j \in \{0, \ldots, H-1\}$ and $s \in \mathcal{S}$. Applying the per-state bound above to every $\tilde{s}$ in the inner sum,

$$\begin{aligned}
m_0^{H-j\to H}(s) &= \sum_{i=H-j}^{H} \sum_{\tilde{s}} d_s^{H-j\to i,\pi^\star}(\tilde{s})\, D_{\mathrm{KL}}\big(\pi^{\star,i}(\cdot \mid \tilde{s}) \,||\, \pi_0^i(\cdot \mid \tilde{s})\big) \\
&\leq \sum_{i=H-j}^{H} \sum_{\tilde{s}} d_s^{H-j\to i,\pi^\star}(\tilde{s})\, \log|\mathcal{A}| \\
&= \sum_{i=H-j}^{H} \log|\mathcal{A}| \qquad \big(\textstyle\sum_{\tilde{s}} d_s^{H-j\to i,\pi^\star}(\tilde{s}) = 1\big) \\
&= (j+1)\log|\mathcal{A}|,
\end{aligned}$$

since the sum over i ranges over $H - (H-j) + 1 = j+1$ terms. This completes the proof. □

Lemma 15. *For any $j \in \{0, \ldots, H-1\}$, $s \in \mathcal{S}$, and $t \geq 0$, define*

$$n_t^{H-j\to H}(s) := \sum_{i=H-j}^{H} n_t^i(s) = \sum_{i=H-j}^{H} \sum_{\tilde{s}\in\mathcal{S}} d_s^{H-j\to i,\pi^\star}(\tilde{s})\, V^{\pi_t,i}(\tilde{s}).$$

Then

$$n_t^{H-j\to H}(s) \leq \frac{(j+1)(j+2)}{2} \leq (j+1)^2.$$

Proof of Lemma 15. Since the rewards take values in $[0, 1]$, the value function satisfies

$$0 \le V^{\pi_t, i}(\tilde{s}) \le H - i + 1$$

for every $t \ge 0$, $i \in [H]$, and $\tilde{s} \in \mathcal{S}$. Therefore,

$$\begin{aligned} n_t^{H-j \to H}(s) &= \sum_{i=H-j}^{H} \sum_{\tilde{s} \in \mathcal{S}} d_s^{H-j \to i, \pi^\star}(\tilde{s})\, V^{\pi_t, i}(\tilde{s}) \\ &\le \sum_{i=H-j}^{H} \sum_{\tilde{s} \in \mathcal{S}} d_s^{H-j \to i, \pi^\star}(\tilde{s})\, (H - i + 1) \\ &= \sum_{i=H-j}^{H} (H - i + 1) \\ &= \frac{(j+1)(j+2)}{2} \\ &\le (j+1)^2. \end{aligned}$$

The third line uses

$$\sum_{\tilde{s} \in \mathcal{S}} d_s^{H-j \to i, \pi^\star}(\tilde{s}) = 1, \qquad i = H - j, \ldots, H,$$

and the final inequality follows from $j \ge 0$. This proves the claim. □

Proof of Theorem 1

Proof. Fix an integer $T \ge 1$. The uniform initial policy π_0 has full support. We next verify that the NPG update preserves this property. Suppose that π_t has full support. Since the rewards are bounded and the horizon is finite, $Q^{\pi_t, i}(s, a)$ is finite for every $i \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$. Therefore,

$$\pi_{t+1}^i(a \mid s) = \frac{\pi_t^i(a \mid s) \exp\big(\eta Q^{\pi_t, i}(s, a)\big)}{Z_t^i(s)} > 0.$$

Thus, by induction, π_t has full support for every $t = 0, \ldots, T$. Consequently, Lemma 2 applies at every iteration, and all KL terms appearing below are finite.

Applying the performance difference Lemma 1, we obtain

$$\begin{aligned} V^{\pi^\star, H-j}(s) - V^{\pi_t, H-j}(s) &= \sum_{i=H-j}^{H} \sum_{\tilde{s}, \tilde{a}} d_s^{H-j \to i, \pi^\star}(\tilde{s}) \pi^{\star, i}(\tilde{a} \mid \tilde{s}) A^{\pi_t, i}(\tilde{s}, \tilde{a}) \\ &= \sum_{i=H-j}^{H} \sum_{\tilde{s}, \tilde{a}} d_s^{H-j \to i, \pi^\star}(\tilde{s}) \pi^{\star, i}(\tilde{a} \mid \tilde{s}) \Big[Q^{\pi_t, i}(\tilde{s}, \tilde{a}) - V^{\pi_t, i}(\tilde{s}) \Big] \\ &= \sum_{i=H-j}^{H} \sum_{\tilde{s}, \tilde{a}} d_s^{H-j \to i, \pi^\star}(\tilde{s}) \pi^{\star, i}(\tilde{a} \mid \tilde{s}) \left[\frac{1}{\eta} \log \frac{\pi_{t+1}^i(\tilde{a} \mid \tilde{s}) Z_t^i(\tilde{s})}{\pi_t^i(\tilde{a} \mid \tilde{s})} - V^{\pi_t, i}(\tilde{s}) \right] \\ &\qquad \text{(Eq.(1))} \\ &= \sum_{i=H-j}^{H} \sum_{\tilde{s}} d_s^{H-j \to i, \pi^\star}(\tilde{s}) \Bigg[\frac{1}{\eta} \Big[D_{\mathrm{KL}}(\pi^{\star, i}(\cdot \mid \tilde{s}) \,||\, \pi_t^i(\cdot \mid \tilde{s})) - D_{\mathrm{KL}}(\pi^{\star, i}(\cdot \mid \tilde{s}) \,||\, \pi_{t+1}^i(\cdot \mid \tilde{s})) \Big] \\ &\qquad + \frac{1}{\eta} \log Z_t^i(\tilde{s}) - V^{\pi_t, i}(\tilde{s}) \Bigg] \\ &\le \sum_{i=H-j}^{H} \sum_{\tilde{s}} d_s^{H-j \to i, \pi^\star}(\tilde{s}) \left[\frac{1}{\eta} \Big[D_{\mathrm{KL}}(\pi^{\star, i}(\cdot \mid \tilde{s}) \,||\, \pi_t^i(\cdot \mid \tilde{s})) - D_{\mathrm{KL}}(\pi^{\star, i}(\cdot \mid \tilde{s}) \,||\, \pi_{t+1}^i(\cdot \mid \tilde{s})) \Big] \right. \end{aligned}$$

$$+ V^{\pi_{t+1},i}(\tilde{s}) - V^{\pi_t,i}(\tilde{s})\Big] \qquad \text{(Lemma 2)}$$
$$= \sum_{i=H-j}^{H}\left[\frac{1}{\eta}(m_t^i(s) - m_{t+1}^i(s)) + n_{t+1}^i(s) - n_t^i(s)\right]$$
$$= \frac{1}{\eta}(m_t^{H-j\to H}(s) - m_{t+1}^{H-j\to H}(s)) + n_{t+1}^{H-j\to H}(s) - n_t^{H-j\to H}(s),$$

where for $j \in \{0, 1, \cdots, H-1\}$, we assume that

$$m_t^{H-j\to H}(s) = \sum_{i=H-j}^{H} m_t^i(s) = \sum_{i=H-j}^{H}\sum_{\tilde{s}\in\mathcal{S}} d_s^{H-j\to i,\pi^\star}(\tilde{s}) D_{\mathrm{KL}}(\pi^{\star,i}(\cdot \mid \tilde{s}) \,||\, \pi_t^i(\cdot \mid \tilde{s})),$$

and

$$n_t^{H-j\to H}(s) = \sum_{i=H-j}^{H} n_t^i(s) = \sum_{i=H-j}^{H}\sum_{\tilde{s}\in\mathcal{S}} d_s^{H-j\to i,\pi^\star}(\tilde{s}) V^{\pi_t,i}(\tilde{s}).$$

Applying the above improvement bound and telescopic argument, for all states $s \in \mathcal{S}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big[V^{\pi^\star,H-j}(s) - V^{\pi_t,H-j}(s)\Big]$$
$$\le \frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{1}{\eta}\big(m_t^{H-j\to H}(s) - m_{t+1}^{H-j\to H}(s)\big) + n_{t+1}^{H-j\to H}(s) - n_t^{H-j\to H}(s)\right]$$
$$= \frac{1}{T}\left[\frac{1}{\eta}\big(m_0^{H-j\to H}(s) - m_T^{H-j\to H}(s)\big) + n_T^{H-j\to H}(s) - n_0^{H-j\to H}(s)\right]$$
$$\overset{(a)}{\le} \frac{1}{T}\left[\frac{1}{\eta}\sum_{i=H-j}^{H} m_0^i(s) + n_T^{H-j\to H}(s)\right]$$
$$\le \frac{1}{T}\left[\frac{1}{\eta}(j+1)\log|\mathcal{A}| + (j+1)^2\right], \qquad \text{(Lemmas 14 and 15)}$$

where in (a), we have used that $m_T^{H-j\to H}(s) \ge 0$ and $n_0^{H-j\to H}(s) \ge 0$.

Applying Lemma 10 we obtain,

$$V^{\pi^\star,H-j}(s) - V^{\pi_T,H-j}(s) \le \frac{(j+1)\log|\mathcal{A}|}{\eta T} + \frac{(j+1)^2}{T}.$$

Substituting $j = H - h$, we obtain

$$V^{\pi^\star,h}(s) - V^{\pi_T,h}(s) \le \frac{(H-h+1)\log|\mathcal{A}|}{\eta T} + \frac{(H-h+1)^2}{T},$$

which concludes the result. □

7.2 Linear MDP Proofs

In this appendix, whenever $w^{\pi_t,h}$ denotes the output of the full support exact population projection oracle, we use the pointwise identity (10).

Lemma 16 (Jin et al. (2020)). *Under Assumption 1, for any policy π, there exist vectors $\{w^{\pi,h}\}_{h=1}^H \subset \mathbb{R}^d$ such that, for every $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$,*

$$Q^{\pi,h}(s,a) = \langle w^{\pi,h}, \phi(s,a)\rangle.$$

Proof of Lemma 16. Fix a policy π and a horizon $h \in [H]$. By the Bellman equation and the linear MDP structure in Assumption 1, for every $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned} Q^{\pi,h}(s,a) &= R^h(s,a) + (P^h V^{\pi,h+1})(s,a) \\ &= \langle \phi(s,a), \zeta^h \rangle + \sum_{s' \in \mathcal{S}} P^h(s' \mid s,a) V^{\pi,h+1}(s') \\ &= \langle \phi(s,a), \zeta^h \rangle + \sum_{s' \in \mathcal{S}} \langle \phi(s,a), \omega^h(s') \rangle V^{\pi,h+1}(s') \\ &= \left\langle \phi(s,a), \zeta^h + \sum_{s' \in \mathcal{S}} \omega^h(s') V^{\pi,h+1}(s') \right\rangle. \end{aligned}$$

Therefore, defining

$$w^{\pi,h} := \zeta^h + \sum_{s' \in \mathcal{S}} \omega^h(s') V^{\pi,h+1}(s')$$

gives

$$Q^{\pi,h}(s,a) = \langle w^{\pi,h}, \phi(s,a) \rangle, \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

This proves the claim. □

For the auxiliary lemmas below, define for each iteration t, horizon h, and state s,

$$\widetilde{Z}_t^h(s) := \sum_{a' \in \mathcal{A}} \pi_t^h(a' \mid s) \exp\left(\eta \, \langle w^{\pi_t,h}, \phi(s,a') \rangle\right).$$

Lemma 17. *Fix an iteration $t \geq 0$, a horizon $h \in [H]$, a state $s \in \mathcal{S}$, and a step size $\eta > 0$. Under an exact population-projection oracle with a full support projection distribution,*

$$\frac{1}{\eta} \log \widetilde{Z}_t^h(s) - V^{\pi_t,h}(s) \geq 0.$$

Proof of Lemma 17. Using the definition of $\widetilde{Z}_t^h(s)$ and applying Jensen's inequality to the concave function $\log$, we obtain

$$\begin{aligned} \log \widetilde{Z}_t^h(s) &= \log \left[\sum_{a' \in \mathcal{A}} \pi_t^h(a' \mid s) \exp\left(\eta \, \langle w^{\pi_t,h}, \phi(s,a') \rangle\right) \right] \\ &\geq \sum_{a' \in \mathcal{A}} \pi_t^h(a' \mid s) \log \exp\left(\eta \, \langle w^{\pi_t,h}, \phi(s,a') \rangle\right) \\ &= \eta \sum_{a' \in \mathcal{A}} \pi_t^h(a' \mid s) \langle w^{\pi_t,h}, \phi(s,a') \rangle \\ &= \eta \sum_{a' \in \mathcal{A}} \pi_t^h(a' \mid s) Q^{\pi_t,h}(s,a') \\ &= \eta V^{\pi_t,h}(s). \end{aligned}$$

The fourth line uses the pointwise identity (10), and the final line uses the definition of the value function. Dividing by $\eta > 0$ gives

$$\frac{1}{\eta} \log \widetilde{Z}_t^h(s) \geq V^{\pi_t,h}(s),$$

which proves the claim. □

The next lemma aggregates the nonnegative local terms across future horizons.

Lemma 18. *Fix an iteration $t \geq 0$, a horizon $h \in [H]$, a state $s \in \mathcal{S}$, and a step size $\eta > 0$. For each $l \in \{h, \dots, H\}$, define*

$$F'_l := \sum_{\tilde{s} \in \mathcal{S}} d_s^{h \to l, \pi_{t+1}}(\tilde{s}) \left[\frac{1}{\eta} \log \widetilde{Z}_t^l(\tilde{s}) - V^{\pi_t, l}(\tilde{s}) \right].$$

Then

$$\sum_{l=h}^{H} F'_l \geq \frac{1}{\eta} \log \widetilde{Z}_t^h(s) - V^{\pi_t, h}(s).$$

Proof of Lemma 18. By Lemma 17, for every $l \in \{h, \dots, H\}$ and every $\tilde{s} \in \mathcal{S}$,

$$\frac{1}{\eta} \log \widetilde{Z}_t^l(\tilde{s}) - V^{\pi_t, l}(\tilde{s}) \geq 0.$$

Since $d_s^{h \to l, \pi_{t+1}}(\tilde{s}) \geq 0$, it follows that $F'_l \geq 0$ for every $l \in \{h, \dots, H\}$.

For $l = h$, the visitation distribution satisfies

$$d_s^{h \to h, \pi_{t+1}}(\tilde{s}) = \mathbb{1}\{\tilde{s} = s\}.$$

Therefore,

$$\begin{aligned} F'_h &= \sum_{\tilde{s} \in \mathcal{S}} \mathbb{1}\{\tilde{s} = s\} \left[\frac{1}{\eta} \log \widetilde{Z}_t^h(\tilde{s}) - V^{\pi_t, h}(\tilde{s}) \right] \\ &= \frac{1}{\eta} \log \widetilde{Z}_t^h(s) - V^{\pi_t, h}(s). \end{aligned}$$

Combining the nonnegativity of F'_l for all $l > h$ with the expression for F'_h gives

$$\sum_{l=h}^{H} F'_l = F'_h + \sum_{l=h+1}^{H} F'_l \geq F'_h = \frac{1}{\eta} \log \widetilde{Z}_t^h(s) - V^{\pi_t, h}(s).$$

This proves the claim. □

Lemma 19. *(Improvement Lower Bound in the Linear MDP Setting) Fix an iteration $t \geq 0$ and a horizon $h \in [H]$. Suppose the Q-NPG update (6) is run under an exact population-projection oracle with a full support projection distribution of Section 3.2. Then, for every $s \in \mathcal{S}$,*

$$V^{\pi_{t+1}, h}(s) - V^{\pi_t, h}(s) \geq \frac{1}{\eta} \log \widetilde{Z}_t^h(s) - V^{\pi_t, h}(s) \geq 0.$$

Proof of Lemma 19. By the softmax parametrization (5) and the Q-NPG parameter update (6), for every $l \in [H]$, $\tilde{s} \in \mathcal{S}$, and $\tilde{a} \in \mathcal{A}$, the induced policy update satisfies

$$\pi_{t+1}^l(\tilde{a} \mid \tilde{s}) = \frac{\pi_t^l(\tilde{a} \mid \tilde{s}) \exp\big(\eta \langle w^{\pi_t, l}, \phi(\tilde{s}, \tilde{a}) \rangle\big)}{\widetilde{Z}_t^l(\tilde{s})}.$$

Since the softmax parametrization assigns strictly positive probability to every action, we may rearrange the preceding display as

$$\langle w^{\pi_t, l}, \phi(\tilde{s}, \tilde{a}) \rangle = \frac{1}{\eta} \log \frac{\pi_{t+1}^l(\tilde{a} \mid \tilde{s}) \, \widetilde{Z}_t^l(\tilde{s})}{\pi_t^l(\tilde{a} \mid \tilde{s})}.$$

Using the pointwise identity (10), this gives

$$Q^{\pi_t, l}(\tilde{s}, \tilde{a}) = \frac{1}{\eta} \log \frac{\pi_{t+1}^l(\tilde{a} \mid \tilde{s}) \, \widetilde{Z}_t^l(\tilde{s})}{\pi_t^l(\tilde{a} \mid \tilde{s})}. \tag{29}$$

Applying the performance difference lemma 1 with $(\pi, \pi') = (\pi_{t+1}, \pi_t)$ yields

$$\begin{aligned}
V^{\pi_{t+1},h}(s) - V^{\pi_t,h}(s) &= \sum_{l=h}^{H} \sum_{\tilde{s},\tilde{a}} d_s^{h\to l,\pi_{t+1}}(\tilde{s})\, \pi_{t+1}^l(\tilde{a} \mid \tilde{s})\, A^{\pi_t,l}(\tilde{s},\tilde{a}) \\
&= \sum_{l=h}^{H} \sum_{\tilde{s},\tilde{a}} d_s^{h\to l,\pi_{t+1}}(\tilde{s})\, \pi_{t+1}^l(\tilde{a} \mid \tilde{s}) \left[Q^{\pi_t,l}(\tilde{s},\tilde{a}) - V^{\pi_t,l}(\tilde{s})\right] \\
&= \sum_{l=h}^{H} \sum_{\tilde{s},\tilde{a}} d_s^{h\to l,\pi_{t+1}}(\tilde{s})\, \pi_{t+1}^l(\tilde{a} \mid \tilde{s}) \left[\frac{1}{\eta} \log \frac{\pi_{t+1}^l(\tilde{a} \mid \tilde{s})\, \widetilde{Z}_t^l(\tilde{s})}{\pi_t^l(\tilde{a} \mid \tilde{s})} - V^{\pi_t,l}(\tilde{s})\right] \\
&= \sum_{l=h}^{H} \sum_{\tilde{s}} d_s^{h\to l,\pi_{t+1}}(\tilde{s}) \left[\frac{1}{\eta} D_{\mathrm{KL}}\big(\pi_{t+1}^l(\cdot \mid \tilde{s}) \,||\, \pi_t^l(\cdot \mid \tilde{s})\big) + \frac{1}{\eta} \log \widetilde{Z}_t^l(\tilde{s}) - V^{\pi_t,l}(\tilde{s})\right] \\
&\geq \sum_{l=h}^{H} \sum_{\tilde{s}} d_s^{h\to l,\pi_{t+1}}(\tilde{s}) \left[\frac{1}{\eta} \log \widetilde{Z}_t^l(\tilde{s}) - V^{\pi_t,l}(\tilde{s})\right] \\
&\geq \frac{1}{\eta} \log \widetilde{Z}_t^h(s) - V^{\pi_t,h}(s) \\
&\geq 0.
\end{aligned}$$

The first inequality uses the nonnegativity of the KL divergence. The second inequality follows from Lemma 18, and the final inequality follows from Lemma 17. Therefore,

$$V^{\pi_{t+1},h}(s) - V^{\pi_t,h}(s) \geq \frac{1}{\eta} \log \widetilde{Z}_t^h(s) - V^{\pi_t,h}(s) \geq 0,$$

which proves the claim. □

The preceding lemmas show that, under an exact population-projection oracle with a full support projection distribution, the linear MDP setting inherits the same one-step improvement structure as the tabular setting. Since the projection is exact, no approximation, projection, or regression-error term appears in the lower bound. We now prove Proposition 1 directly from the induced policy space identity (11).

Proof of Proposition 1

Proof. By the pointwise identity (10) and the induced update identity (11), the policy sequence generated by the Q-NPG parameter update (6) satisfies exactly the tabular NPG update (1) with constant step size η.

Moreover, $\theta_0^h = \mathbf{0}$ for every $h \in [H]$ implies, by the softmax parametrization (5), that

$$\pi_0^h(\cdot \mid s) = \mathrm{Unif}(\mathcal{A}), \qquad \forall h \in [H],\ s \in \mathcal{S}.$$

Therefore, the induced policy sequence satisfies precisely the assumptions of Theorem 1. Applying Theorem 1, for every integer $T \geq 1$, for all horizons $h \in [H]$ and states $s \in \mathcal{S}$, we obtain

$$V^{\pi^\star,h}(s) - V^{\pi_T,h}(s) \leq \frac{(H-h+1)\log|\mathcal{A}|}{\eta T} + \frac{(H-h+1)^2}{T}.$$

This completes the proof. □

8 Geometric Convergence Proofs

Proof of Lemma 3. By the definition of the global multi-start objective $J(\cdot)$, we have

$$H\big(J(\pi^\star) - J(\pi)\big) = \sum_{h_0=1}^{H} \Big(V^{\pi^\star,h_0}(\rho) - V^{\pi,h_0}(\rho)\Big).$$

Now, we use horizon-wise optimality of $\pi^\star$, for each horizon $h_0 \in \{1, \dots, H\}$ and each state $s \in \mathcal{S}$ to have, $V^{\pi^\star,h_0}(s) \geq V^{\pi,h_0}(s)$. Taking expectation over $S \sim \rho$ yields $V^{\pi^\star,h_0}(\rho) - V^{\pi,h_0}(\rho) \geq 0$, for every h_0. Therefore, each term in the above sum is nonnegative. In particular, the sum of these nonnegative terms dominates the $h_0 = 1$ term:

$$\begin{aligned} V^{\pi^\star,1}(\rho) - V^{\pi,1}(\rho) \;&\leq\; \sum_{h_0=1}^{H} \Big(V^{\pi^\star,h_0}(\rho) - V^{\pi,h_0}(\rho) \Big) \\ &= H\big(J(\pi^\star) - J(\pi)\big) \end{aligned}$$

which completes the proof. □

Proof of Lemma 4. Fix $i \in \{1, \dots, H\}$ and $\tilde{s} \in \mathcal{S}$. By construction, the term in (14) corresponding to $h_0 = i$ is

$$d_\rho^{i \to i,\pi}(\tilde{s}) = \mathbb{P}_{S^i \sim \rho}(S^i = \tilde{s}) = \rho(\tilde{s}),$$

which holds because the process is initialized at horizon i with $S^i \sim \rho$, and the case $h_0 = i$ involves no transitions.
Since each summand $d_\rho^{h_0 \to i,\pi}(\tilde{s})$ is a probability and hence nonnegative, we obtain

$$\bar{d}_\rho^{i,\pi}(\tilde{s}) = \frac{1}{H} \sum_{h_0=1}^{i} d_\rho^{h_0 \to i,\pi}(\tilde{s}) \;\geq\; \frac{1}{H}\, d_\rho^{i \to i,\pi}(\tilde{s}) = \frac{1}{H} \rho(\tilde{s}),$$

which proves (15). □

Proof of Lemma 5. Starting from the definition of the global multi-start objective, $J(\cdot)$:

$$J(\pi') - J(\pi) = \frac{1}{H} \sum_{h_0=1}^{H} \Big(V^{\pi',h_0}(\rho) - V^{\pi,h_0}(\rho) \Big).$$

Fix a start horizon $h_0 \in [H]$. Apply the finite-horizon performance difference lemma (Lemma 1) at horizon h_0 with π and π' in the appropriate order gives

$$V^{\pi',h_0}(\rho) - V^{\pi,h_0}(\rho) = \sum_{i=h_0}^{H} \sum_{\tilde{s} \in \mathcal{S}} d_\rho^{h_0 \to i,\pi'}(\tilde{s}) \Big\langle Q^{\pi,i}(\tilde{s}, \cdot),\; \pi'^i(\cdot \mid \tilde{s}) - \pi^i(\cdot \mid \tilde{s}) \Big\rangle.$$

Substituting this identity into the average over h_0 and exchanging the finite sums, we obtain

$$\begin{aligned} J(\pi') - J(\pi) &= \frac{1}{H} \sum_{h_0=1}^{H} \sum_{i=h_0}^{H} \sum_{\tilde{s} \in \mathcal{S}} d_\rho^{h_0 \to i,\pi'}(\tilde{s}) \Big\langle Q^{\pi,i}(\tilde{s}, \cdot),\; \pi'^i(\cdot \mid \tilde{s}) - \pi^i(\cdot \mid \tilde{s}) \Big\rangle \\ &= \sum_{i=1}^{H} \sum_{\tilde{s} \in \mathcal{S}} \left(\frac{1}{H} \sum_{h_0=1}^{i} d_\rho^{h_0 \to i,\pi'}(\tilde{s}) \right) \Big\langle Q^{\pi,i}(\tilde{s}, \cdot),\; \pi'^i(\cdot \mid \tilde{s}) - \pi^i(\cdot \mid \tilde{s}) \Big\rangle. \end{aligned}$$

By the definition of the global multi-start horizon-i visitation measure in (14),

$$\frac{1}{H} \sum_{h_0=1}^{i} d_\rho^{h_0 \to i,\pi'}(\tilde{s}) = \bar{d}_\rho^{i,\pi'}(\tilde{s}).$$

Therefore,

$$J(\pi') - J(\pi) = \sum_{i=1}^{H} \sum_{\tilde{s} \in \mathcal{S}} \bar{d}_\rho^{i,\pi'}(\tilde{s}) \Big\langle Q^{\pi,i}(\tilde{s}, \cdot),\; \pi'^i(\cdot \mid \tilde{s}) - \pi^i(\cdot \mid \tilde{s}) \Big\rangle,$$

which is exactly the expectation form in (16). □

Proof of Lemma 6. We first establish a pointwise (local) nonnegativity property of the KL-based PMD update.

Fix any horizon $h \in [H]$ and state $s \in \mathcal{S}$. By optimality of $\pi^h_{t+1}(\cdot \mid s)$ in the update (2) and the three-point inequality for KL-based mirror descent (see, e.g., Lan, 2023; Xiao, 2022), applied to the linear objective $p \mapsto \eta_t \langle Q^{\pi_t,h}(s,\cdot), p\rangle$, for any $p \in \Delta(\mathcal{A})$ we have

$$\eta_t \big\langle Q^{\pi_t,h}(s,\cdot), \pi^h_{t+1}(\cdot \mid s)\big\rangle - D_{\mathrm{KL}}\big(\pi^h_{t+1}(\cdot \mid s) \,||\, \pi^h_t(\cdot \mid s)\big) \geq \eta_t \big\langle Q^{\pi_t,h}(s,\cdot), p\big\rangle - D_{\mathrm{KL}}\big(p \,||\, \pi^h_t(\cdot \mid s)\big) + D_{\mathrm{KL}}\big(p \,||\, \pi^h_{t+1}(\cdot \mid s)\big)\,.$$

Rearranging and dividing by $\eta_t > 0$ gives

$$\big\langle Q^{\pi_t,h}(s,\cdot),\, \pi^h_{t+1}(\cdot \mid s) - p\big\rangle \;\geq\; \frac{1}{\eta_t} D_{\mathrm{KL}}\big(\pi^h_{t+1}(\cdot \mid s) \,||\, \pi^h_t(\cdot \mid s)\big) + \frac{1}{\eta_t} D_{\mathrm{KL}}\big(p \,||\, \pi^h_{t+1}(\cdot \mid s)\big) - \frac{1}{\eta_t} D_{\mathrm{KL}}\big(p \,||\, \pi^h_t(\cdot \mid s)\big)\,.$$

Choosing $p = \pi^h_t(\cdot \mid s)$ and using $D_{\mathrm{KL}}(\pi^h_t(\cdot \mid s) \,||\, \pi^h_t(\cdot \mid s)) = 0$, we obtain

$$\begin{aligned}\big\langle Q^{\pi_t,h}(s,\cdot),\, \pi^h_{t+1}(\cdot \mid s) - \pi^h_t(\cdot \mid s)\big\rangle &\geq \frac{1}{\eta_t} D_{\mathrm{KL}}\big(\pi^h_{t+1}(\cdot \mid s) \,||\, \pi^h_t(\cdot \mid s)\big) + \frac{1}{\eta_t} D_{\mathrm{KL}}\big(\pi^h_t(\cdot \mid s) \,||\, \pi^h_{t+1}(\cdot \mid s)\big) \\ &\geq 0,\end{aligned}$$

where the last step uses the nonnegativity of the KL divergence (Lemma 11). Thus for every (h, s),

$$\big\langle Q^{\pi_t,h}(s,\cdot),\, \pi^h_{t+1}(\cdot \mid s) - \pi^h_t(\cdot \mid s)\big\rangle \;\geq\; 0. \tag{30}$$

Next, fix a starting horizon $h_0 \in [H]$. Applying the finite-horizon performance difference lemma (Lemma 1) at horizon h_0 with initial distribution ρ to the pair $(\pi, \pi') = (\pi_{t+1}, \pi_t)$, and using $V^{\pi_t,i}(\tilde{s}) = \langle Q^{\pi_t,i}(\tilde{s},\cdot), \pi^i_t(\cdot \mid \tilde{s})\rangle$ to rewrite the advantage sum as an inner product of differences, gives

$$V^{\pi_{t+1},h_0}(\rho) - V^{\pi_t,h_0}(\rho) = \sum_{i=h_0}^{H} \sum_{\tilde{s}\in\mathcal{S}} d_\rho^{h_0 \to i, \pi_{t+1}}(\tilde{s}) \big\langle Q^{\pi_t,i}(\tilde{s},\cdot),\, \pi^i_{t+1}(\cdot \mid \tilde{s}) - \pi^i_t(\cdot \mid \tilde{s})\big\rangle.$$

By (30) with $(h, s) = (i, \tilde{s})$ and the fact that $d_\rho^{h_0 \to i, \pi_{t+1}}(\tilde{s}) \geq 0$, every summand is nonnegative, so

$$V^{\pi_{t+1},h_0}(\rho) - V^{\pi_t,h_0}(\rho) \;\geq\; 0, \qquad \forall h_0 \in [H].$$

Averaging these inequalities over $h_0 \in [H]$ yields

$$J(\pi_{t+1}) - J(\pi_t) = \frac{1}{H} \sum_{h_0=1}^{H} \Big(V^{\pi_{t+1},h_0}(\rho) - V^{\pi_t,h_0}(\rho)\Big) \;\geq\; 0,$$

which proves the monotonicity of the global multi-start objective $J(\cdot)$. □

Proof of Lemma 7. We apply the one-step KL mirror-descent inequality (see, e.g., Lan, 2023; Xiao, 2022) at the coordinate (i, s): for any $q \in \Delta(\mathcal{A})$,

$$\big\langle Q^{\pi_t,i}(s,\cdot),\, q(\cdot) - \pi^i_{t+1}(\cdot \mid s)\big\rangle \;\leq\; \frac{1}{\eta_t}\Big(D_{\mathrm{KL}}\big(q \,||\, \pi^i_t(\cdot \mid s)\big) - D_{\mathrm{KL}}\big(q \,||\, \pi^i_{t+1}(\cdot \mid s)\big)\Big). \tag{31}$$

Choosing $q(\cdot) = \pi^{\star,i}(\cdot \mid s)$ gives

$$\big\langle Q^{\pi_t,i}(s,\cdot),\, \pi^{\star,i}(\cdot \mid s) - \pi^i_{t+1}(\cdot \mid s)\big\rangle \;\leq\; \frac{1}{\eta_t}\Big(D_{\mathrm{KL}}\big(\pi^{\star,i}(\cdot \mid s) \,||\, \pi^i_t(\cdot \mid s)\big) - D_{\mathrm{KL}}\big(\pi^{\star,i}(\cdot \mid s) \,||\, \pi^i_{t+1}(\cdot \mid s)\big)\Big). \tag{32}$$

Taking expectation of both sides with respect to $S \sim \bar{d}_\rho^{i,\pi^\star}$ and summing over $i = 1, \dots, H$, we obtain

$$\begin{aligned}\sum_{i=1}^{H} \mathbb{E}_{S\sim\bar{d}_\rho^{i,\pi^\star}} \Big[\big\langle Q^{\pi_t,i}(S,\cdot),\ \pi^{\star,i}(\cdot \mid S) - \pi_{t+1}^i(\cdot \mid S)\big\rangle\Big] \le \frac{1}{\eta_t} \sum_{i=1}^{H} \Big(&\mathbb{E}_{S\sim\bar{d}_\rho^{i,\pi^\star}} D_{\mathrm{KL}}\big(\pi^{\star,i}(\cdot \mid S) \,||\, \pi_t^i(\cdot \mid S)\big) \\ &- \mathbb{E}_{S\sim\bar{d}_\rho^{i,\pi^\star}} D_{\mathrm{KL}}\big(\pi^{\star,i}(\cdot \mid S) \,||\, \pi_{t+1}^i(\cdot \mid S)\big)\Big) \\ = \frac{1}{\eta_t}\big(D_t^\star - D_{t+1}^\star\big),&\end{aligned}$$

where the last equality follows from the definition of $D_t^\star$ in (18). This proves (19). □

Proof of Proposition 2

Proof. Fix $t \ge 0$ such that $\vartheta_t < \infty$. Because π_0 has full support and each step size η_t is finite, the multiplicative form of the KL-based PMD update preserves full support. Consequently, all KL terms appearing below are finite, and Lemma 7 applies.

Applying that lemma gives

$$\sum_{i=1}^{H} \mathbb{E}_{S\sim\bar{d}_\rho^{i,\pi^\star}} \left[\big\langle -Q^{\pi_t,i}(S,\cdot), \pi_{t+1}^i(\cdot \mid S) - \pi^{\star,i}(\cdot \mid S)\big\rangle\right] \le \frac{1}{\eta_t}\big(D_t^\star - D_{t+1}^\star\big). \tag{33}$$

For every $i \in [H]$ and $S \in \mathcal{S}$, decompose the inner product as

$$\begin{aligned}\big\langle -Q^{\pi_t,i}(S,\cdot), \pi_{t+1}^i(\cdot \mid S) - \pi^{\star,i}(\cdot \mid S)\big\rangle = &\big\langle -Q^{\pi_t,i}(S,\cdot), \pi_{t+1}^i(\cdot \mid S) - \pi_t^i(\cdot \mid S)\big\rangle \\ &+ \big\langle Q^{\pi_t,i}(S,\cdot), \pi^{\star,i}(\cdot \mid S) - \pi_t^i(\cdot \mid S)\big\rangle.\end{aligned}$$

Accordingly, denote the two resulting sums in the left-hand side of (33) by (I) and (II), respectively.

For term (II), Lemma 5, applied with $(\pi, \pi') = (\pi_t, \pi^\star)$, gives

$$\begin{aligned}\text{(II)} &= \sum_{i=1}^{H} \mathbb{E}_{S\sim\bar{d}_\rho^{i,\pi^\star}} \left[\big\langle Q^{\pi_t,i}(S,\cdot), \pi^{\star,i}(\cdot \mid S) - \pi_t^i(\cdot \mid S)\big\rangle\right] \\ &= J(\pi^\star) - J(\pi_t) = \delta_t. \qquad \text{(From (17))}\end{aligned}$$

For term (I), define

$$f_i(s) := \big\langle -Q^{\pi_t,i}(s,\cdot), \pi_{t+1}^i(\cdot \mid s) - \pi_t^i(\cdot \mid s)\big\rangle.$$

The local monotonicity property (30) implies

$$f_i(s) \le 0, \qquad \forall i \in [H],\ s \in \mathcal{S}.$$

If $\bar{d}_\rho^{i,\pi^\star}(s) > 0$, the definition of ϑ_t and the assumption $\vartheta_t < \infty$ give

$$\bar{d}_\rho^{i,\pi^\star}(s) \le \vartheta_t\, \bar{d}_\rho^{i,\pi_{t+1}}(s).$$

If $\bar{d}_\rho^{i,\pi^\star}(s) = 0$, the same inequality holds trivially because $\vartheta_t \ge 0$ and $\bar{d}_\rho^{i,\pi_{t+1}}(s) \ge 0$. Hence the inequality holds for every $s \in \mathcal{S}$. Since $f_i(s) \le 0$, multiplying by $f_i(s)$ reverses the inequality:

$$\bar{d}_\rho^{i,\pi^\star}(s) f_i(s) \ge \vartheta_t\, \bar{d}_\rho^{i,\pi_{t+1}}(s) f_i(s).$$

Summing over $s \in \mathcal{S}$ and $i \in [H]$ yields

$$\text{(I)} \ge \vartheta_t \sum_{i=1}^{H} \mathbb{E}_{S\sim\bar{d}_\rho^{i,\pi_{t+1}}}\big[f_i(S)\big]. \tag{34}$$

Applying Lemma 5 with $(\pi, \pi') = (\pi_t, \pi_{t+1})$ gives

$$\sum_{i=1}^{H} \mathbb{E}_{S\sim\bar{d}_\rho^{i,\pi_{t+1}}}\big[f_i(S)\big] = J(\pi_t) - J(\pi_{t+1}) \\ = \delta_{t+1} - \delta_t.$$

Therefore,

$$\text{(I)} \geq \vartheta_t(\delta_{t+1} - \delta_t).$$

Combining this inequality with $\text{(II)} = \delta_t$ and (33) gives

$$\vartheta_t(\delta_{t+1} - \delta_t) + \delta_t \leq \frac{1}{\eta_t}\big(D_t^\star - D_{t+1}^\star\big),$$

which proves (21). □

Proof of Lemma 8. If $\vartheta_\rho = +\infty$, then (23) holds trivially. Hence, assume $\vartheta_\rho < \infty$. By the definition of ϑ_ρ, this implies the support condition

$$\bar{d}_\rho^{i,\pi^\star}(s) > 0 \implies \rho(s) > 0 \qquad \text{for all } i \in [H],\ s \in \mathcal{S}.$$

Now fix any $i \in [H]$ and any state $s \in \mathcal{S}$ such that $\bar{d}_\rho^{i,\pi^\star}(s) > 0$. Then $\rho(s) > 0$, and applying Lemma 4 to the policy π_{t+1} yields

$$\bar{d}_\rho^{i,\pi_{t+1}}(s) \ \geq\ \frac{1}{H}\rho(s). \tag{35}$$

Therefore,

$$\frac{\bar{d}_\rho^{i,\pi^\star}(s)}{\bar{d}_\rho^{i,\pi_{t+1}}(s)} \ \leq\ \frac{\bar{d}_\rho^{i,\pi^\star}(s)}{\rho(s)/H} = H\,\frac{\bar{d}_\rho^{i,\pi^\star}(s)}{\rho(s)} \ \leq\ H \max_{j\in[H]} \max_{x:\rho(x)>0} \frac{\bar{d}_\rho^{j,\pi^\star}(x)}{\rho(x)} = \vartheta_\rho.$$

Since the inequality holds for every admissible pair (i, s) with $\bar{d}_\rho^{i,\pi^\star}(s) > 0$, taking the maxima in the definition of ϑ_t yields $\vartheta_t \leq \vartheta_\rho$. □

Proof of Theorem 2

Proof. Recall that $\delta_t := J(\pi^\star) - J(\pi_t)$. By Lemma 6, $\{\delta_t\}_{t\geq 0}$ is nonincreasing, so $\delta_{t+1} - \delta_t \leq 0$. Since $\vartheta_\rho < \infty$ and Lemma 8 gives $\vartheta_t \leq \vartheta_\rho$, we have $\vartheta_t < \infty$ for every $t \geq 0$. Therefore,

$$\vartheta_\rho(\delta_{t+1} - \delta_t) + \delta_t \ \leq\ \vartheta_t(\delta_{t+1} - \delta_t) + \delta_t.$$

Combining this with Proposition 2 yields

$$\vartheta_\rho(\delta_{t+1} - \delta_t) + \delta_t \ \leq\ \frac{1}{\eta_t}\big(D_t^\star - D_{t+1}^\star\big). \tag{36}$$

Rearranging (36) gives

$$\vartheta_\rho\delta_{t+1} - (\vartheta_\rho - 1)\delta_t \ \leq\ \frac{1}{\eta_t}D_t^\star - \frac{1}{\eta_t}D_{t+1}^\star.$$

Moving the $D_{t+1}^\star$ term to the left and dividing by ϑ_ρ to obtain

$$\delta_{t+1} + \frac{1}{\eta_t\vartheta_\rho}D_{t+1}^\star \ \leq\ \Big(1 - \frac{1}{\vartheta_\rho}\Big)\delta_t + \frac{1}{\eta_t\vartheta_\rho}D_t^\star$$

$$
\begin{aligned}
&= \Big(1 - \frac{1}{\vartheta_\rho}\Big)\delta_t + \frac{(\vartheta_\rho - 1)}{\eta_t \vartheta_\rho (\vartheta_\rho - 1)} D_t^\star \\
&= \Big(1 - \frac{1}{\vartheta_\rho}\Big)\delta_t + \Big(1 - \frac{1}{\vartheta_\rho}\Big)\frac{1}{\eta_t(\vartheta_\rho - 1)} D_t^\star \\
&= \Big(1 - \frac{1}{\vartheta_\rho}\Big)\Big[\delta_t + \frac{1}{\eta_t(\vartheta_\rho - 1)} D_t^\star\Big].
\end{aligned}
$$

If the step size satisfies (24), i.e. $\eta_{t+1}(\vartheta_\rho - 1) \geq \eta_t \vartheta_\rho$, then we have,

$$
\delta_{t+1} + \frac{1}{\eta_{t+1}(\vartheta_\rho - 1)} D_{t+1}^\star \leq \Big(1 - \frac{1}{\vartheta_\rho}\Big)\Big[\delta_t + \frac{1}{\eta_t(\vartheta_\rho - 1)} D_t^\star\Big]
$$

This forms the following recursion, for all $t \geq 0$,

$$
\delta_t + \frac{1}{\eta_t(\vartheta_\rho - 1)} D_t^\star \leq \Big(1 - \frac{1}{\vartheta_\rho}\Big)^t \Big[\delta_0 + \frac{1}{\eta_0(\vartheta_\rho - 1)} D_0^\star\Big]
$$

Finally, since $D_t^\star \geq 0$ we have $\delta_t \leq \delta_t + \frac{1}{\eta_t(\vartheta_\rho - 1)} D_t^\star$, hence

$$
\delta_t = J(\pi^\star) - J(\pi_t) \;\leq\; \Big(1 - \frac{1}{\vartheta_\rho}\Big)^t \left(\delta_0 + \frac{1}{\eta_0(\vartheta_\rho - 1)} D_0^\star\right), \qquad \forall t \geq 0.
$$

By Lemma 3, for any policy π, start at horizon-1, we know

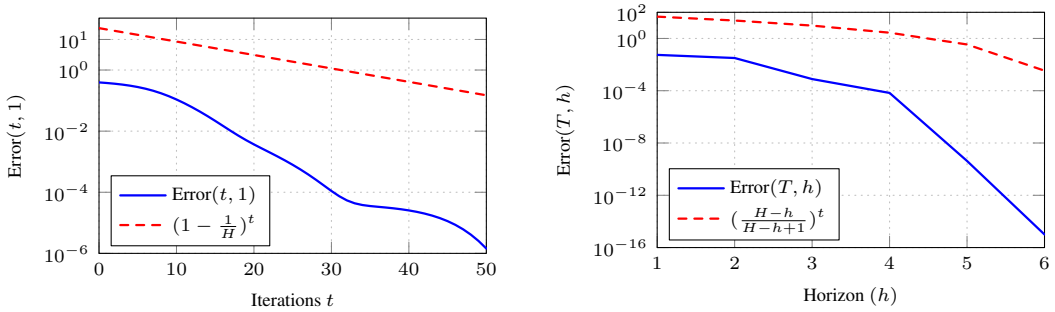
$$
V^{\pi^\star,1}(\rho) - V^{\pi,1}(\rho) \;\leq\; H\big(J(\pi^\star) - J(\pi)\big).
$$

Apply this inequality with $\pi = \pi_t$ to obtain

$$
V^{\pi^\star,1}(\rho) - V^{\pi_t,1}(\rho) \;\leq\; H\Big(1 - \frac{1}{\vartheta_\rho}\Big)^t \left(\delta_0 + \frac{1}{\eta_0(\vartheta_\rho - 1)} D_0^\star\right), \qquad \forall t \geq 0.
$$

This completes the proof. □

8.1 Further Simulations



(a) NPG linear convergence at $h = 1$ ($\vartheta_\rho > 1$).

(b) Horizon-wise error at fixed $T = 20$ (all h).

Figure 3: Increasing step size NPG on the randomly generated MDP under the problem-dependent $\vartheta_\rho > 1$ bound of Theorem 2: (a) error vs. iteration at $h = 1$; (b) error vs. all horizons at fixed iteration $T = 20$. Both plots use linear x-axis and log-scale y-axis. The solid (blue) curve shows the empirical error, and the dashed (red) curve shows the corresponding geometric reference curve, which is undefined at $h = H$ where $\vartheta_\rho = 1$.

We repeat the experiments of Section 5.2 on a randomly generated a finite-horizon MDP with the same dimensions ($|\mathcal{S}| = 15$, $|\mathcal{A}| = 5$, $H = 7$): each transition distribution $P^h(\cdot \mid s, a)$ is drawn

independently from the symmetric Dirichlet distribution with unit concentration, and rewards are i.i.d. uniform on $[0, 1]$. We use random seed 42 for both the transitions and rewards. Unlike the structured instance in Section 5.2, the uniform distribution need not be invariant under the transitions of this MDP. We therefore evaluate the mismatch coefficient directly. Specifically, we compute an optimal policy $\pi^\star$ by backward induction, construct its multi-start visitation measures, and evaluate ϑ_ρ from these measures.

Figure 3a plots $\mathrm{Error}(t, 1)$ using problem-dependent bound of Theorem 2, with contraction factor $1 - 1/\vartheta_\rho$ and prefactor evaluated at the measured ϑ_ρ. As in the structured experiment, we initialize π_0 uniformly.

Figure 3b repeats the horizon-wise experiment: for each $h \in [H - 1]$, we run NPG on the tail MDP with effective horizon $H - h + 1$ and evaluate ϑ_ρ directly on each tail. In both panels the empirical error remains below the predicted bound.

8.2 Linear Convergence with Robust Step Size Proofs

Proof of Lemma 9. Recall that $\bar{d}_\rho^{i,\pi^\star}$ is defined by

$$\bar{d}_\rho^{i,\pi^\star}(s) = \frac{1}{H} \sum_{h_0=1}^{i} d_\rho^{h_0 \to i,\pi^\star}(s).$$

Since each $d_\rho^{h_0 \to i,\pi^\star}(\cdot)$ is a probability distribution, we have

$$\sum_{s \in \mathcal{S}} \bar{d}_\rho^{i,\pi^\star}(s) = \frac{1}{H} \sum_{h_0=1}^{i} \sum_{s \in \mathcal{S}} d_\rho^{h_0 \to i,\pi^\star}(s) = \frac{i}{H}.$$

Because $\vartheta_\rho < \infty$, the support condition holds:

$$\bar{d}_\rho^{i,\pi^\star}(s) > 0 \implies \rho(s) > 0.$$

Thus the ratio $\bar{d}_\rho^{i,\pi^\star}(s)/\rho(s)$ is well-defined on $\{s : \rho(s) > 0\}$, and

$$\mathbb{E}_{S \sim \rho}\left[\frac{\bar{d}_\rho^{i,\pi^\star}(S)}{\rho(S)}\right] = \sum_{s:\rho(s)>0} \rho(s) \frac{\bar{d}_\rho^{i,\pi^\star}(s)}{\rho(s)} = \sum_{s \in \mathcal{S}} \bar{d}_\rho^{i,\pi^\star}(s) = \frac{i}{H}.$$

Therefore,

$$\max_{s:\rho(s)>0} \frac{\bar{d}_\rho^{i,\pi^\star}(s)}{\rho(s)} \geq \frac{i}{H}.$$

Taking the maximum over $i \in [H]$ yields

$$\max_{i \in [H]} \max_{s:\rho(s)>0} \frac{\bar{d}_\rho^{i,\pi^\star}(s)}{\rho(s)} \geq 1,$$

and multiplying by H gives $\vartheta_\rho \geq H$ by (22).

Finally, since the function $x \mapsto x/(x-1)$ is decreasing on $(1, \infty)$ and $\vartheta_\rho \geq H \geq 2$, we obtain

$$\frac{\vartheta_\rho}{\vartheta_\rho - 1} \leq \frac{H}{H - 1},$$

which is (26). □

Lemma 20. *Fix $H \geq 2$ and $\rho \in \Delta(\mathcal{S})$. Let $\{\pi_t\}_{t\geq 0}$ be generated by the KL-based PMD update (2), with a positive step size sequence $\{\eta_t\}_{t\geq 0}$, from a full support initial policy. Let $\pi^\star$ be an optimal nonstationary policy. Let δ_0 and $D_0^\star$ be as defined in (17) and (18) respectively. Suppose $\vartheta_\rho < \infty$. If the positive step sizes satisfy*

$$\eta_{t+1} \geq \frac{H}{H-1}\eta_t, \qquad t = 0, 1, 2, \ldots. \tag{37}$$

then the problem-dependent growth condition (24) is satisfied. Consequently, for every $t \geq 0$,

$$V^{\pi^\star,1}(\rho) - V^{\pi_t,1}(\rho) \leq H\left(1 - \frac{1}{\vartheta_\rho}\right)^t \left(\delta_0 + \frac{D_0^\star}{\eta_0(\vartheta_\rho - 1)}\right). \tag{38}$$

Proof of Lemma 20. Since $H \geq 2$ and $\vartheta_\rho < \infty$, Lemma 9 gives

$$\vartheta_\rho \geq H \geq 2,$$

and (26) gives

$$\frac{\vartheta_\rho}{\vartheta_\rho - 1} \leq \frac{H}{H-1}.$$

Since $\eta_t > 0$, multiplying this inequality by η_t yields, for every $t \geq 0$,

$$\frac{\vartheta_\rho}{\vartheta_\rho - 1}\eta_t \leq \frac{H}{H-1}\eta_t.$$

Combining this inequality with (37), we obtain

$$\eta_{t+1} \geq \frac{H}{H-1}\eta_t \geq \frac{\vartheta_\rho}{\vartheta_\rho - 1}\eta_t, \qquad t \geq 0.$$

Therefore, the step size condition (24) required by Theorem 2 is satisfied.

Moreover, $\vartheta_\rho \geq H \geq 2$ and $\vartheta_\rho < \infty$ imply $\vartheta_\rho \in (1, \infty)$, while the positivity of the step size sequence implies $\eta_0 > 0$. Together with the remaining assumptions in the lemma statement, these facts verify all the hypotheses of Theorem 2. Applying that theorem gives

$$V^{\pi^\star,1}(\rho) - V^{\pi_t,1}(\rho) \leq H\left(1 - \frac{1}{\vartheta_\rho}\right)^t \left(\delta_0 + \frac{D_0^\star}{\eta_0(\vartheta_\rho - 1)}\right), \qquad t \geq 0,$$

which proves (38). □