Sharp Convergence and Sample Complexity of Policy Mirror Descent for Average-Reward MDPs

Enes Arda
The Ohio State University
arda.2@osu.edu

Atilla Eryilmaz
The Ohio State University
eryilmaz.2@osu.edu

Abstract

Policy mirror descent (PMD) has a mature finite-time theory in discounted Markov decision processes (MDPs), but less is known in the average-reward setting, a more natural objective for many control applications. We give a finite-time, finite-sample analysis of PMD in ergodic average-reward MDPs built around a single master recursion that governs convergence for any critic, without external regularization. Its specializations yield linear rates for exact, inexact-tabular, and linear function approximation (LFA) updates. For exact PMD, the rate is tight in the worst case, and convergence can become superlinear. We complement these convergence results with critics whose end-to-end sample complexities improve the prior best policy-gradient dependence from $|\mathcal{S}||\mathcal{A}|\, t_{\text{mix}}^3$ to $|\mathcal{S}|\, t_{\text{mix}}^3 + |\mathcal{S}||\mathcal{A}|\, t_{\text{mix}}^2$ in the tabular setting and from $|\mathcal{S}||\mathcal{A}|\, t_{\text{mix}}^5$ to $d\, t_{\text{mix}}^3$ under LFA. Matching information-theoretic lower bounds establish that both critics are optimal up to logarithmic factors.

1 Introduction

Policy optimization, the process of updating an agent's decision rule to maximize expected long-term reward, is a workhorse of reinforcement learning applications [e.g., 1, 48]. It underlies classical policy-gradient [42] and natural-gradient methods [17] as well as widely used algorithms such as TRPO [36] and PPO [37]. A large and rapidly growing theory explains the behavior of these methods in *discounted* Markov decision processes (MDPs), including finite-time finite-sample guarantees [2, 20, 38], function approximation extensions [2, 3, 7, 47], and linear convergence results for policy mirror descent (PMD) and its natural-gradient instantiations [8, 16, 20, 45, 49]. However, many continuing-control applications are more naturally modeled by the *average-reward* criterion, where performance is determined by the long-run steady-state behavior rather than geometrically discounted rewards. This objective appears throughout ergodic control and operations research, including resource allocation, scheduling, queueing, and communication networks [6, 12, 19, 28, 32, 34, 39].

To make the distinction between discounted and average-reward MDPs concrete, let c_k denote the cost incurred by a policy at time k. The discounted and average objectives can be written informally as

$$\rho_\gamma = \lim_{H\to\infty} \sum_{k=0}^{H-1} \gamma^k\, c_k, \qquad \rho = \lim_{H\to\infty} \frac{1}{H} \sum_{k=0}^{H-1} c_k,$$

with $\gamma \in (0,1)$. The key difference is the weighting: discounting weights early costs more heavily, so *initial conditions* matter, while average reward weights all times equally and is governed by the *stationary* behavior. Two consequences follow. First, discounted PMD analyses exploit the initial state distribution to lower bound state visitation uniformly. In average reward, initial conditions are transient, so controlling the stationary distribution shift becomes a new challenge. Second, discounting yields a finite effective horizon $H_\gamma := 1/(1-\gamma)$ and admits unbiased value estimates via geometric stopping times. In average reward, all times are weighted equally, so there is no such stopping distribution, and truncating the horizon introduces a bias that must be controlled through mixing to stationarity.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

A standard relationship that links the two objectives is $\rho = \lim_{\gamma\uparrow 1}(1-\gamma)\rho_\gamma$ [6, 35]. This has led to the common workaround of analyzing a discounted surrogate with γ close to 1 and then sending $\gamma \uparrow 1$ [19, 30]. While analytically appealing, this reduction can yield bounds and algorithmic intuition that break down in the average-reward regime. In particular, discounted finite-time guarantees typically scale polynomially in H_γ (up to H_γ^5 for policy-gradient methods [2, 19, 45] and H_γ^2 for approximate policy iteration [30]), which make the resulting bounds in the average-reward limit blow up since $\lim_{\gamma\uparrow 1}(1-\gamma)H_\gamma^p = \infty$ for $p > 1$. Average-reward MDPs therefore need a theory of their own, developed directly rather than inherited from discounted analyses.

Contributions. In this paper, we develop a finite-time, finite-sample theory of policy mirror descent (PMD) in *average-reward* Markov decision processes (MDPs):

- **Master recursion.** We establish a master recursion (Lemma 3.7) that governs PMD for any critic, from which all downstream results follow as specializations.
- **Linear and asymptotic convergence.** We prove linear convergence for exact, inexact-tabular, and linear function approximation (LFA) updates (Corollary 4.3; Theorems 4.2, 5.6), at a rate that is tight in the worst case for exact PMD (Proposition B.4). For all three updates, we further sharpen prior worst-case rates with an asymptotic-rate analysis. Our LFA result is, to our knowledge, the first non-tabular guarantee of linear convergence to the optimum up to critic error.
- **End-to-end sample complexity.** For the tabular and LFA settings, we give critics whose end-to-end sample complexities (Thms. 4.9, 5.10) improve the prior best policy-gradient dependence from $|\mathcal{S}||\mathcal{A}|\, t_{\mathrm{mix}}^3$ to $|\mathcal{S}|\, t_{\mathrm{mix}}^3 + |\mathcal{S}||\mathcal{A}|\, t_{\mathrm{mix}}^2$ in the tabular setting and from $|\mathcal{S}||\mathcal{A}|\, t_{\mathrm{mix}}^5$ to $d\, t_{\mathrm{mix}}^3$ in the LFA setting. Both guarantees hold with high probability for the last iterate rather than in expectation for a random iterate. The LFA guarantee needs only a single trajectory, a strictly weaker oracle than the generative model the tabular guarantee uses.
- **Matching lower bounds.** We establish information-theoretic lower bounds via Assouad reductions (Propositions 4.8, 5.9), showing that both critics are optimal up to logarithmic factors.

2 Related work

Recent work on policy optimization for average-reward MDPs spans several partial results. Sun et al. [40] prove linear convergence of robust PMD for average-reward MDPs under squared Euclidean geometry. Specialized to zero robustness radius, their result partially overlaps with ours for an exact critic (Theorem 4.2 with $\varepsilon_Q = 0$) but uses a stricter step-size growth condition and does not address general Bregman geometry, inexact critics, function approximation, or sample complexity. Li et al. [25] give stochastic first-order methods with $O(1/\sqrt{T})$ unregularized rates and linear convergence up to an $\widetilde{O}(1/T)$ stochastic term only under strongly convex regularization. Their LFA sample complexity on a single trajectory (their Corollary 3) has $|\mathcal{S}||\mathcal{A}|\, t_{\mathrm{mix}}^5$ dependence. Kumar et al. [19], Murthy and Srikant [29] establish $O(1/T)$ rates for projected policy gradient and for natural policy gradient (NPG) and mirror-descent-like methods, Grosof et al. [13] obtain an $O(1/\sqrt{T})$ rate for NPG on countably-infinite-state queueing systems, and Murthy et al. [30] develop performance bounds for approximate policy iteration; their bound for a fixed-step mirror-descent update keeps an error floor that does not vanish with the critic error. Bai et al. [5], Ganesh et al. [10, 11] study general-parameterized average-reward policy gradient, achieving polynomial convergence and regret rates. In concurrent work, Lee and Ryu [22] extend policy gradient to multichain MDPs via α-clipped PMD, with sample complexity $\widetilde{O}(\varepsilon^{-6})$.[1] This extension is complementary to our sharper $\widetilde{O}(\varepsilon^{-2})$ analysis in the ergodic case.

Against this backdrop, we provide a unified PMD analysis under general Bregman geometry spanning exact, inexact-tabular, and LFA updates. We improve on prior tabular results along multiple axes (general Bregman geometry, inexact critics, asymptotic-rate analysis, sharper sample complexity, high-probability last-iterate guarantees) and, to our knowledge, establish the first linear convergence of non-tabular average-reward PMD to the optimum up to critic error, with a sharper sample complexity. Matching information-theoretic lower bounds establish that both critics are optimal up to logarithmic factors. Table 1 summarizes these comparisons, and Appendix A details them theorem by theorem.

[1] Throughout, $\widetilde{O}(\cdot)$ and $\widetilde{\Theta}(\cdot)$ hide polylogarithmic factors in the problem parameters, the target accuracy, and the confidence.

Table 1: Comparison with prior work. Rates concern the unregularized objective, up to critic-error floors. Sample complexities omit instance factors other than rate constants (App. A). Guarantees hold w.h.p. or in expectation (in exp.), for the last or a uniformly random (rand.) iterate.

Work	Update	Convergence	Sample complexity	Guarantee
Exact critic				
Murthy and Srikant [29]	PMD (KL)	$O(1/T)$	—	—
Kumar et al. [19]	Projected PG	$O(1/T)$	—	—
Sun et al. [40]	PMD (Euclidean)	$O((1-1/M_0)^T)$	—	—
Ours	PMD (Bregman)	$O((1-1/\bar{\chi})^T)$; $\limsup_t \Delta_t^{1/t} \le 1-1/\chi$	—	—
Tabular critic, generative-model oracle				
Lee and Ryu [22]†	α-clipped PMD (Bregman)	$O((1-1/C_\alpha)^T)$	$\widetilde{O}(t_{\mathrm{tar}}^3 \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert C_\alpha^6/\varepsilon^6)$	w.h.p., last
Li et al. [25]‡	PMD (Bregman)	$O(1/\sqrt{T})$	$\widetilde{O}(t_{\mathrm{mix}}^3 \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert/\varepsilon^2)$	in exp., rand.
Ours	PMD (Bregman)	$O((1-1/\bar{\chi})^T)$	$\widetilde{O}(t_{\mathrm{mix}}^2 \lvert\mathcal{S}\rvert(\lvert\mathcal{A}\rvert + t_{\mathrm{mix}})\bar{\chi}^3/\varepsilon^2)$	w.h.p., last
Lower bound (critic)			$\Omega(t_{\mathrm{mix}}^2 \lvert\mathcal{S}\rvert(\lvert\mathcal{A}\rvert + t_{\mathrm{mix}})/\varepsilon^2)$	
Linear function approximation, single-trajectory oracle				
Murthy et al. [30]	PMD (KL)	linear, non-vanishing floor	—	in exp., last
Li et al. [25]	PMD (KL)	$O(1/\sqrt{T})$	$\widetilde{O}(t_{\mathrm{mix}}^5 \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert/\varepsilon^2)$	in exp., rand.
Ours	PMD (Bregman)	$O((1-1/\bar{\chi})^T)$	$\widetilde{O}(t_{\mathrm{mix}}^3 d\,\bar{\chi}^3/\varepsilon^2)$	w.h.p., last
Lower bound (critic)			$\Omega(t_{\mathrm{mix}}^3 d/\varepsilon^2)$	

$M_0 \ge \bar{\chi}$ (Section 3.1). †Multichain, targeting the best α-clipped policy, not the optimum; t_{tar} and C_α are their analogues of t_{mix} and $\bar{\chi}$, respectively. ‡Mixing unichain; App. E discusses this extension of our results. Other rows are ergodic.

3 Problem setup and preliminaries

We consider a Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, P, c)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $P(\cdot \mid s, a)$ is the transition kernel, and $c : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the cost function. Let $\mathcal{P}(\mathcal{X})$ denote the probability simplex over the finite set $\mathcal{X}$. A policy $\pi \in \mathcal{P}(\mathcal{A})^{\mathcal{S}} =: \Pi$ is a collection of conditional distributions $\pi(\cdot \mid s) \in \mathcal{P}(\mathcal{A})$ for all $s \in \mathcal{S}$. Each policy π induces a Markov kernel $P_\pi(\cdot \mid s) := \sum_{a \in \mathcal{A}} \pi(a \mid s)\, P(\cdot \mid s, a)$. For all objects that live in finite-dimensional spaces, we freely switch between function and vector/matrix notation (e.g., $\pi(\cdot \mid s) = \pi_s \in \mathbb{R}^{|\mathcal{A}|}$ is a stochastic row vector and $P_\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ a row-stochastic matrix). We assume throughout:

Assumption 3.1 (Ergodicity)**.** For every policy $\pi \in \Pi$, the Markov chain with kernel P_π is irreducible and aperiodic, hence admits a unique stationary distribution $d^\pi \in \mathcal{P}(\mathcal{S})$ satisfying $d^\pi = d^\pi P_\pi$.

Assumption 3.1 is a standard regularity condition in average-reward settings, as it makes the steady-state objective well defined [19, 25, 29, 31]; App. E discusses its relaxation to unichain MDPs. A key consequence of Asm. 3.1 is that the stationary distributions are uniformly bounded away from zero:

Lemma 3.2. *The constant $\underline{d} := \inf_{\pi \in \Pi,\, s \in \mathcal{S}} d^\pi(s)$ is positive. Under a uniform minorization condition, $\underline{d}$ admits an explicit lower bound (Appendix E).*

Objective. The average cost of π is defined as $\rho(\pi) := \mathbb{E}_{s \sim d^\pi,\, a \sim \pi_s}\big[c(s,a)\big]$. Our goal is to find an optimal policy $\pi^\star \in \arg\min_{\pi \in \Pi} \rho(\pi)$ (fixed but arbitrary), with $\rho^\star := \rho(\pi^\star)$ and $d^\star := d^{\pi^\star}$.

Policy optimization requires knowing which actions are better than others in each state. In the average-reward setting, starting from any state–action pair (s,a) and following π afterwards leaves the long-run average at $\rho(\pi)$, so actions must be compared by the cost they incur along the way. This is measured by the *differential action-value function* [19], the expected total cost in excess of $\rho(\pi)$:

$$Q^\pi(s,a) \;:=\; \lim_{H \to \infty} \mathbb{E}\left[\sum_{k=0}^{H-1} \big(c(s_k,a_k) - \rho(\pi)\big) \;\middle|\; s_0 = s,\, a_0 = a \right]. \tag{1}$$

Under Assumption 3.1, P_π^k converges geometrically, so the limit exists (see Appendix B). Centering by $\rho(\pi)$ is necessary since the uncentered sum $\mathbb{E}\sum_{k<H} c(s_k,a_k)$ grows linearly in H at rate $\rho(\pi)$. In the discounted setting, geometric weighting plays the analogous role.

The following facts are the average-reward counterparts of two standard discounted tools, often used in the convergence analysis of policy gradient methods [2, 42, 45, 47]. See App. B for proofs.

Fact 3.3 (Performance difference lemma)**.** *For any policies π, π',*

$$\rho(\pi') - \rho(\pi) \;=\; \mathbb{E}_{s\sim d^{\pi'}}\Big[\big\langle Q^{\pi}_s,\, \pi'_s - \pi_s\big\rangle\Big].$$

Fact 3.4 (Policy gradient theorem)**.** *As a gradient on the simplex, $\nabla_{\pi_s}\rho(\pi)$ can be represented as*

$$\nabla_{\pi_s}\rho(\pi) \;=\; d^{\pi}(s)Q^{\pi}_s.$$

3.1 Policy mirror descent algorithms

Mirror descent [33] minimizes a function f through the steps $x^{t+1} \in \arg\min_x\big\{\eta_t\langle\nabla f(x^t), x\rangle + D(x, x^t)\big\}$, where $D(x, x') := F(x) - F(x') - \langle\nabla F(x'), x - x'\rangle$ is the Bregman divergence of a Legendre function F [21]. Following [38, 45], we apply it to ρ over Π with the state-weighted divergence $D^{d^t}(\pi, \pi^t)$, where $D^d(\pi, \pi') := \sum_s d(s)\, D(\pi_s, \pi'_s)$. With $Q^t := Q^{\pi^t}$ and $d^t := d^{\pi^t}$, Fact 3.4 gives $\nabla_{\pi_s}\rho(\pi^t) = d^t(s)\, Q^t_s$, so the update splits across states and the positive weights $d^t(s)$ drop out. Replacing Q^t by a critic $\tilde{Q}^t$ yields the *policy mirror descent* (PMD) update

$$\pi^{t+1}_s \;\in\; \arg\min_{p\in\mathcal{P}(\mathcal{A})}\Big\{\eta_t\,\langle\tilde{Q}^t_s,\, p\rangle + D\big(p, \pi^t_s\big)\Big\}. \tag{2}$$

Remark 3.5. When D is the KL divergence, which induces the local Fisher geometry on the simplex, (2) reduces to a natural policy gradient step [2, 45]. When D is the squared Euclidean distance, it reduces to projected Q-descent [40, 45]. We keep the general Bregman geometry in this paper.

Three instantiations of the critic $\tilde{Q}^t$ cover our analyses: $\tilde{Q}^t = Q^t$, the exact critic (Section 4); $\tilde{Q}^t = \widehat{Q}^t$, an estimate of each $Q^t(s, a)$ with a uniform error bound (Section 4); and $\tilde{Q}^t(s, a) = \phi^{\top}_{s,a} w_t$, a linear function of features $\phi_{s,a}$ with a mean-squared-error (MSE) bound (Section 5).

The most important property of mirror descent is the three-point inequality, which relates one-step progress to a difference of Bregman divergences and enables telescoping across iterations.

Fact 3.6 (Three-point inequality for PMD)**.** *For any $p \in \mathcal{P}(\mathcal{A})$, the PMD update (2) satisfies*

$$\eta_t\langle\tilde{Q}^t_s,\, \pi^{t+1}_s - p\rangle \le D(p, \pi^t_s) - D(p, \pi^{t+1}_s) - D(\pi^{t+1}_s, \pi^t_s).$$

Setting $p = \pi^t_s$ in Fact 3.6 yields, for any critic $\tilde{Q}^t$, the descent-direction property

$$\langle\tilde{Q}^t_s,\, \pi^{t+1}_s - \pi^t_s\rangle \;\le\; 0, \qquad \forall\, s \in \mathcal{S}. \tag{3}$$

Specializing to the exact case $\tilde{Q}^t = Q^t$, taking expectation over $s \sim d^{t+1}$, and applying Fact 3.3 gives the descent property $\rho(\pi^{t+1}) \le \rho(\pi^t)$, so each exact PMD update decreases the objective *without* requiring any restriction on the step size η_t. Moreover, (2) shows that larger η_t pushes π^{t+1}_s closer to a greedy update, recovering policy iteration in the limit. In the discounted setting, aggressively growing step sizes thus inherit the linear convergence of policy iteration [45, 47]. We show the same in the average-reward setting, at a rate governed by the distribution mismatch introduced next.

Stationary distribution mismatch. Let

$$\chi_t \;:=\; \left\|\frac{d^\star}{d^t}\right\|_\infty \;=\; \max_{s\in\mathcal{S}}\frac{d^\star(s)}{d^t(s)}$$

measure the mismatch between the stationary distributions of the PMD iterate π^t and an optimal policy $\pi^\star$. Its discounted analogue, the distribution-mismatch coefficient [2, 45, 47], is well-studied, but its analysis exploits a structural property that the average-reward setting lacks. In the discounted setting, the analogue of d^π is the discounted occupancy $d^\pi_\gamma := (1-\gamma)\sum_{k\ge 0}\gamma^k\mu^\pi_k$, with μ^π_k the state distribution at time k under π from an initial μ_0. Dropping all $k \ge 1$ terms gives $d^\pi_\gamma(s) \ge (1-\gamma)\mu_0(s)$, so the discounted distribution-mismatch coefficients are uniformly bounded via a freely chosen full-support μ_0. In the average-reward setting, this trick is not available: the objective depends on the stationary distribution d^π, which forgets μ_0, so no choice of μ_0 can bound it from below. Controlling χ_t along the iterates therefore becomes a central technical challenge. Nevertheless, by Lem. 3.2 we have the loose bound $\sup_t \chi_t \le \underline{d}^{-1}$, which characterizes the convergence rate in the analysis of Li et al. [25]. Sun et al. [40] use the tighter intermediate constant M_0 satisfying $\sup_t \chi_t \le M_0 \le \underline{d}^{-1}$. Our asymptotic analysis below is sharper than both, and we show empirically that its rate takes over after only a brief warm-up (Fig. 1).

3.2 Master recursion for inexact PMD

Write $\Delta_t := \rho(\pi^t) - \rho^\star$ for the suboptimality gap at iteration t and $D_t^\star := D^{d^\star}(\pi^\star, \pi^t)$ for the optimal-policy weighted divergence. The critic error in (2) enters the analysis through the quantities

$$A_t := \left|\mathbb{E}_{s\sim d^{t+1}}\left[\langle \tilde{Q}_s^t - Q_s^t, \pi_s^{t+1} - \pi_s^t\rangle\right]\right|, \quad B_t := \left|\mathbb{E}_{s\sim d^\star}\left[\langle \tilde{Q}_s^t - Q_s^t, \pi_s^t - \pi_s^\star\rangle\right]\right|. \tag{4}$$

The following lemma combines the descent-direction property (3) with the bounds on A_t, B_t to deliver both a non-asymptotic linear rate and a sharper asymptotic rate, governing all PMD specializations.

Lemma 3.7 (Master convergence for inexact PMD)**.** *Let $\{\pi^t\}$ be generated by (2) with $\pi_s^0 \in \operatorname{rint}(\mathcal{P}(\mathcal{A}))$ for all $s \in \mathcal{S}$, and suppose $A_t \le \bar{A}$, $B_t \le \bar{B}$ for all $t \ge 0$ with constants $\bar{A}, \bar{B} \ge 0$.*

(i) Non-asymptotic linear rate. *Suppose $\sup_t \chi_t \le \bar{\chi}$ for some $\bar{\chi} \in (1, \infty)$ and the step sizes satisfy $\eta_t \ge \frac{\bar{\chi}}{\bar{\chi}-1}\eta_{t-1}$ for all $t \ge 0$ with some $\eta_{-1} > 0$. Then for all $t \ge 0$, with $(x)_+ := \max\{x, 0\}$,*

$$\left(\Delta_t - (\bar{\chi}\,\bar{A} + \bar{B})\right)_+ \;\le\; \left(1 - \tfrac{1}{\bar{\chi}}\right)^t \left(\Delta_0 + \tfrac{D_0^\star}{\eta_{-1}\bar{\chi}}\right). \tag{5}$$

(ii) Asymptotic rate. *Suppose $D_0^\star < \infty$ and the step sizes satisfy $\lim_{t\to\infty} \eta_t/\eta_{t-1} = \infty$ with some $\eta_{-1} > 0$, and let $\chi := \limsup_t \chi_t$. Then for every $\zeta > 0$,*

$$\limsup_{t\to\infty} \left(\Delta_t - ((\chi + \zeta)\,\bar{A} + \bar{B})\right)_+^{1/t} \;\le\; 1 - \tfrac{1}{\chi+\zeta}; \tag{6}$$

hence, $\limsup_{t\to\infty} \Delta_t \le \chi\,\bar{A} + \bar{B}$. If in addition $\bar{A} = 0$, then $\limsup_{t\to\infty}(\Delta_t - \bar{B})_+^{1/t} \le 1 - 1/\chi$.

Remark 3.8. Lemma 3.7(i) is informative whenever $D_0^\star < \infty$, which part (ii) assumes. Since the lemma assumes an interior π^0, $D_0^\star$ is finite whenever F is finite on the closed simplex, even for a deterministic $\pi^\star$. This covers the two standard choices: the KL divergence and the squared Euclidean distance. The exception is a geometry whose F diverges at the boundary, such as a log barrier. There, a deterministic $\pi^\star$ gives $D_0^\star = \infty$, and the bound of part (i), though still valid, is vacuous.

Both conclusions arise from the same one-step Lyapunov recursion $\Psi_{t+1} \le (1 - 1/\bar{\chi})\Psi_t + (\bar{\chi}A_t + B_t)/\bar{\chi}$ for $\Psi_t := \Delta_t + D_t^\star/(\eta_{t-1}\bar{\chi})$: part (i) iterates it from $t = 0$, and part (ii) on a tail with $\bar{\chi}$ replaced by $\chi + \zeta$ (Appendix B.6). Crucially, the lemma separates the *actor's machinery* (step-size schedule, mismatch coefficient) from the *critic's error* (the bounds $\bar{A}, \bar{B}$), so each of our convergence theorems reduces to instantiating $\tilde{Q}^t$ and bounding A_t, B_t.

On the step size schedule. A natural concern is that the aggressively growing step sizes of Lem. 3.7 would amplify critic errors. But step size and critic error decouple in PMD: unlike stochastic gradient descent (SGD), where each step injects noise scaled by η_t, here the critic's error is fixed, and η_t only controls how aggressively each update commits to that (possibly biased) direction. Larger η_t speeds up convergence but leaves the error floor unchanged. Lemma 3.7 formalizes this, and Fig. 4(a) (App. F.3) verifies it empirically. Numerical overflow is also mild: with $T = \Theta(\bar{\chi}\log(1/\varepsilon))$ iterations, the growth condition requires only $\eta_T = \operatorname{poly}(1/\varepsilon)$. With a constant step size, the one-step inequality behind Lem. 3.7 gives an $O(1/T)$ rate in expectation to the same error floor for a uniformly random iterate; with an exact critic, the same holds for the last iterate (App. B.7). This matches prior $O(1/T)$ rates [19, 29]. Growing step sizes instead make the rate linear.

Importantly, Lem. 3.7's guarantees do not require knowing χ or $\bar{\chi}$: doubling the step size every iteration gives part (i) with $\bar{\chi}$ replaced by $\max\{2, \bar{\chi}\}$, and part (ii) needs only $\eta_t/\eta_{t-1} \to \infty$. Constants enter only to state finite-time or finite-sample guarantees, and only as valid upper bounds.

4 Analysis in the tabular setting

We first instantiate Lem. 3.7 in the tabular setting, where the critic estimates each $Q^t(s, a)$ with a uniform error bound. Section 4.1 derives linear convergence of inexact PMD, including the exact case. Section 4.2 gives an end-to-end sample complexity and a matching lower bound for the critic.

4.1 Linear convergence with inexact updates

Since Q^t is not available in general, we study the PMD update (2) with $\tilde{Q}^t = \widehat{Q}^t \approx Q^t$, under the following critic-error assumption, which we satisfy with finite samples in Section 4.2:

Assumption 4.1 (Uniform critic error)**.** There exists $\varepsilon_Q \geq 0$ such that $\sup_{t\geq 0} \|\widehat{Q}^t - Q^t\|_\infty \leq \varepsilon_Q$.

Under Asm. 4.1, $\|\pi_s - \pi'_s\|_1 \leq 2$ gives $A_t, B_t \leq 2\varepsilon_Q$ in (4), so Lem. 3.7 with $\bar{A} = \bar{B} = 2\varepsilon_Q$ yields:

Theorem 4.2 (Linear convergence of tabular inexact PMD)**.** *Suppose Assumption 4.1 holds, and let* $\{\pi^t\}$ *be generated by* (2) *with* $\tilde{Q}^t = \widehat{Q}^t$*. Under the hypotheses of Lemma 3.7(i),*

$$(\Delta_t - 2(\bar{\chi}+1)\varepsilon_Q)_+ \;\leq\; \Big(1 - \tfrac{1}{\bar{\chi}}\Big)^t \Big(\Delta_0 + \tfrac{D_0^\star}{\eta_{-1}\bar{\chi}}\Big) \qquad \forall\, t \geq 0; \tag{7}$$

and under the hypotheses of Lemma 3.7(ii), for every $\zeta > 0$*,*

$$\limsup_{t\to\infty} \big(\Delta_t - 2(\chi+\zeta+1)\varepsilon_Q\big)_+^{1/t} \;\leq\; 1 - \tfrac{1}{\chi+\zeta}. \tag{8}$$

Corollary 4.3 (Exact PMD and superlinear convergence)**.** *For exact PMD (*$\tilde{Q}^t = Q^t$*), setting* $\varepsilon_Q = 0$ *in* (8) *removes the* ζ*-dependence from the floor, so taking* $\zeta \downarrow 0$ *gives the sharper asymptotic rate*

$$\limsup_{t\to\infty} \Delta_t^{1/t} \;\leq\; 1 - \tfrac{1}{\chi}.$$

Moreover, $\Delta_t \to 0$*, so by compactness of* Π *and continuity of* $\pi \mapsto \rho(\pi)$*, every limit point of* $\{\pi^t\}$ *is optimal. Consequently, if* $\{\pi^t\}$ *converges to an optimal policy (e.g., when the optimal policy is unique), then taking that limit as* $\pi^\star$ *gives* $\chi_t \to \chi = 1$*, so exact PMD is* superlinear*:* $\limsup_{t\to\infty} \Delta_t^{1/t} = 0$.

Remark 4.4. For the KL and squared Euclidean geometries, Corollary 4.3's convergence assumption can be dropped (App. B.8, adapting Lin and Zhang [26]): with KL and any nondecreasing $\eta_t \to \infty$, $\Delta_{t+1}/\Delta_t \to 0$ whenever $\Delta_0 > 0$, and with the squared Euclidean distance and any nondecreasing $\eta_t > 0$, exact PMD reaches an optimal policy in finitely many iterations and stays there.

We empirically validate that the asymptotic rate of Corollary 4.3 is typically much faster than the worst-case finite-time rate driven by $\bar{\chi}$ in (7). Across four diverse average-reward MDPs (random Dirichlet MDPs, access-control queuing, 5×5 grid-world, Jack's car rental, 80 runs each), the median $\chi_t/\bar{\chi}$ (Fig. 1) collapses by 10–1000$\times$ within roughly 20–40 iterations and stays there for the rest of the trajectory.

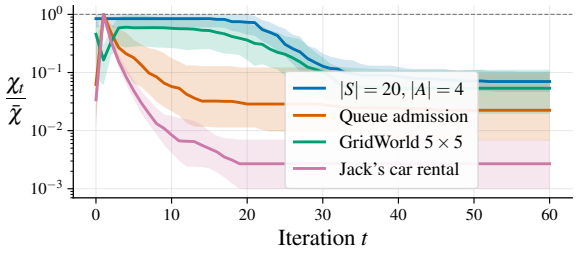

Figure 1: Median $\chi_t/\bar{\chi}$ trajectories. IQR shaded.

In other words, the trajectory enters Cor. 4.3's asymptotic regime (where the rate is governed by $\chi \ll \bar{\chi}$ rather than by $\bar{\chi}$ itself) after only a brief warm-up. The two-part decomposition of Lem. 3.7 therefore delivers a substantively sharper convergence guarantee than the worst-case bounds of Li et al. [25] and Sun et al. [40]: empirically, $\bar{\chi}$ controls only this warm-up, and the asymptotic rate governs convergence from then on. Details and further experiments are in App. F.

In the worst case, however, the rate of Lemma 3.7(i) is tight up to a constant factor over any fixed number of iterations: for every $\bar{\chi} > 1$ and $T \geq 1$, some MDP and initial policy meeting the hypotheses of Lemma 3.7(i) give $\Delta_t \geq \frac{1}{2}(1 - 1/\bar{\chi})^t \Delta_0 > 0$ for all $t \leq T$, for exact PMD with any Bregman divergence and any positive step sizes (App. B.9, adapting Johnson et al. [16, Theorem 4.2]).

4.2 Sample complexity

We now satisfy Assumption 4.1 with transitions sampled from a generative model. Our critic rests on the identity $Q^\pi(s,a) = c(s,a) - \rho(\pi) + P(\cdot \mid s,a)\, h^\pi$ (App. B.1), where $h^\pi(s) := \langle Q_s^\pi, \pi_s \rangle$ is the differential value function. The critic first estimates h^π at every state, then shares these estimates across all state–action pairs through one-step samples. For each state s, we draw B rollouts $\{(s_k^{(i)}, a_k^{(i)})\}_{k=0}^{H-1}$, $i = 1, \ldots, B$, of length H that start at $s_0^{(i)} = s$ and evolve under π.

Estimating h^π from such rollouts presents new challenges that have no discounted analog: in a discounted MDP, truncating the geometric sum $\sum_{k=0}^\infty \gamma^k c_k$ at horizon H yields a tail bias bounded by $\gamma^H/(1-\gamma)$, which decays geometrically with H, so larger H always helps [45, Lemma 15]. In contrast, h^π is an *equal-weighted* infinite-horizon sum of expected costs *centered by the unknown*

average cost $\rho(\pi)$, and these two features together make finite-horizon estimation nontrivial. Given an estimate $\hat{\rho}_s(\pi)$, the natural truncated centered-return estimator

$$\hat{h}^\pi(s) \;:=\; \frac{1}{B}\sum_{i=1}^{B}\sum_{k=0}^{H-1}\big(c(s_k^{(i)},a_k^{(i)}) - \hat{\rho}_s(\pi)\big) \tag{9}$$

decomposes as

$$\hat{h}^\pi(s) \;=\; \underbrace{\frac{1}{B}\sum_{i=1}^{B}\sum_{k=0}^{H-1}\big(c(s_k^{(i)},a_k^{(i)}) - \rho(\pi)\big)}_{\text{tail bias decays with } H \text{ via mixing}} \;-\; \underbrace{H\big(\hat{\rho}_s(\pi) - \rho(\pi)\big)}_{\text{grows linearly in } H},$$

revealing two bias sources that pull H in *opposite directions*: the tail bias of the first shrinks as $H \to \infty$ via mixing to d^π, while the second amplifies any error in $\hat{\rho}_s(\pi)$ by a factor of H. Taking H large without correspondingly sharpening $\hat{\rho}_s(\pi)$ therefore *worsens* $\hat{h}^\pi$, whereas in the discounted setting larger H always helps. A second, subtler pitfall: if $\hat{\rho}_s(\pi)$ is computed as the empirical mean of the same BH costs that appear in (9), the two sums cancel exactly and give $\hat{h}^\pi \equiv 0$. We resolve these issues by estimating $\rho(\pi)$ from the *tail* of the rollouts after a burn-in cutoff $\tau \in \{1,\dots,H-1\}$:

$$\hat{\rho}_s(\pi) \;:=\; \frac{1}{B(H-\tau)}\sum_{i=1}^{B}\sum_{k=\tau}^{H-1} c(s_k^{(i)},a_k^{(i)}), \tag{10}$$

which avoids the exact cancellation and concentrates around $\rho(\pi)$ once τ exceeds the mixing time. To quantify how large H and τ must be, we use the following uniform mixing property:

Lemma 4.5 (Uniform mixing)**.** *There exists an integer* $t_{\mathrm{mix}} \geq 1$ *such that* $\sup_{\mu_0 \in \mathcal{P}(\mathcal{S})} \|\mu_0 P_\pi^{t_{\mathrm{mix}}} - d^\pi\|_{\mathrm{TV}} \leq 1/4$ *for every* $\pi \in \Pi$*, and consequently*

$$\sup_{\mu_0 \in \mathcal{P}(\mathcal{S})} \big\|\mu_0 P_\pi^k - d^\pi\big\|_{\mathrm{TV}} \;\leq\; 2^{-\lfloor k/t_{\mathrm{mix}}\rfloor} \qquad \forall\, k \geq 0,\ \forall\, \pi \in \Pi. \tag{11}$$

Remark 4.6. Uniform mixing is usually posed as an assumption in average-reward reinforcement learning [e.g., 25, 44]. In our finite setting, it follows from Assumption 3.1 by compactness of Π. Appendix E gives the proof and a sufficient condition under which t_{mix} is explicit.

For each pair (s,a), we then draw n fresh one-step samples $s'_j \sim P(\cdot \mid s,a)$ and set

$$\widehat{Q}^\pi(s,a) \;:=\; c(s,a) - \hat{\rho}_s(\pi) + \frac{1}{n}\sum_{j=1}^{n} \hat{h}^\pi(s'_j), \tag{12}$$

where $\hat{\rho}_s(\pi)$ is the estimate (10) from the rollouts at s and $\hat{h}^\pi$ is clipped to $[-2t_{\mathrm{mix}}, 2t_{\mathrm{mix}}]$, which contains h^π by (11). Combined with a Bernstein-type bound for the rollouts, the mixing decay (11) yields the following finite-sample guarantee:

Lemma 4.7 (Tabular critic upper bound)**.** *Fix a policy* π*, and let* $\widehat{Q}^\pi$ *be computed via* (12)*. Then for any target critic accuracy* $\varepsilon_Q \in (0,1)$ *and confidence* $\delta \in (0,1)$*, the number of generative-model transitions*

$$N \;=\; \widetilde{O}\big((|\mathcal{S}|\, t_{\mathrm{mix}}^3 + |\mathcal{S}||\mathcal{A}|\, t_{\mathrm{mix}}^2)/\varepsilon_Q^2\big)$$

suffices to guarantee $\|\widehat{Q}^\pi - Q^\pi\|_\infty \leq \varepsilon_Q$ *with probability at least* $1-\delta$*.*

Rollouts started at every state–action pair would cost $t_{\mathrm{mix}}^3/\varepsilon_Q^2$ transitions per pair. Our critic instead runs them per state and shares the resulting estimates across all pairs through the one-step samples in (12), which cost only $t_{\mathrm{mix}}^2/\varepsilon_Q^2$ per pair. Both terms are information-theoretically necessary: an Assouad reduction perturbs the transitions of the action that π takes at every state, which forces the first term, and those of the other actions at every pair, which forces the second.

Proposition 4.8 (Tabular critic lower bound)**.** *For every* $|\mathcal{S}| \geq 2$*,* $p \in (0,1/4)$*,* $\varepsilon_Q \in (0, 1/(128p)]$*, and every (possibly adaptive) generative-model algorithm, there are an average-reward MDP with state space* $\mathcal{S}$ *and action space* $\mathcal{A}$ *satisfying Assumption 3.1 with* $t_{\mathrm{mix}} = \Theta(1/p)$ *and a deterministic policy* π *such that outputting* $\widehat{Q} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ *with* $\Pr(\|\widehat{Q} - Q^\pi\|_\infty \leq \varepsilon_Q) \geq 2/3$ *requires*

$$N \;=\; \Omega\big((|\mathcal{S}|\, t_{\mathrm{mix}}^3 + |\mathcal{S}||\mathcal{A}|\, t_{\mathrm{mix}}^2)/\varepsilon_Q^2\big)$$

generative-model transitions. The explicit construction and constants are given in Appendix C.

Lemma 4.7 and Proposition 4.8 match up to logarithmic factors for accuracies covered by both and confidence $\delta \le 1/3$, so the critic's worst-case sample complexity is $\widetilde{\Theta}\big((|\mathcal{S}|\, t_{\mathrm{mix}}^3 + |\mathcal{S}||\mathcal{A}|\, t_{\mathrm{mix}}^2)/\varepsilon_Q^2\big)$.

Composing Lem. 4.7's critic with Thm. 4.2's actor yields the following end-to-end sample complexity:

Theorem 4.9 (Tabular sample complexity)**.** *Suppose $D_0^\star < \infty$. For any target suboptimality $\varepsilon \in (0,1)$ and confidence $\delta \in (0,1)$, using the critic of Lemma 4.7 at accuracy $\varepsilon_Q = \varepsilon/(4(\bar{\chi}+1))$ together with the actor of Theorem 4.2 in (7) guarantees $\Delta_T \le \varepsilon$ with probability at least $1-\delta$ after $T = \widetilde{\Theta}(\bar{\chi})$ iterations, using a total of*

$$N \;=\; \widetilde{O}\big((|\mathcal{S}|\, t_{\mathrm{mix}}^3 + |\mathcal{S}||\mathcal{A}|\, t_{\mathrm{mix}}^2)\, \bar{\chi}^3/\varepsilon^2\big)$$

generative-model transitions.

Remark 4.10. With the same generative-model oracle as Li et al. [25, Cor. 1], Thm. 4.9 improves their $|\mathcal{S}||\mathcal{A}|\, t_{\mathrm{mix}}^3$ dependence to $|\mathcal{S}|\, t_{\mathrm{mix}}^3 + |\mathcal{S}||\mathcal{A}|\, t_{\mathrm{mix}}^2$, a factor of order $\min\{|\mathcal{A}|, t_{\mathrm{mix}}\}$, and strengthens their in-expectation, random-iterate guarantee to a high-probability, last-iterate one. By Prop. 4.8, this dependence of the critic's sample complexity is information-theoretically tight for any policy-gradient method whose critic satisfies Assumption 4.1 for every policy. In comparison, the $\widetilde{O}(|\mathcal{S}||\mathcal{A}|\, t_{\mathrm{mix}}/\varepsilon^2)$ rate of Wang et al. [44] is achieved by model-based planning, which does not need a critic satisfying Asm. 4.1 and is not directly applicable to the policy-gradient methods we analyze.

5 Analysis under linear function approximation (LFA)

Section 4's tabular sample complexity (Thm. 4.9) scales with $|\mathcal{S}|$ and $|\mathcal{A}|$. We now instantiate Lem. 3.7 with a linear critic whose error is bounded in MSE rather than uniformly (Section 5.1). Its d-dimensional weights replace the $|\mathcal{S}||\mathcal{A}|$ entries of a tabular critic, and an on-policy MSE bound can be met from a single trajectory instead of a generative model (Section 5.2).

5.1 Linear convergence with inexact updates

To parameterize the critic, we assume access to a feature vector $\phi_{s,a} \in \mathbb{R}^d$ for every (s,a) and use a linear critic $\phi_{s,a}^\top w \approx Q^\pi(s,a)$. Since the features are fixed, we identify the critic with its weights $w \in \mathbb{R}^d$. Write $\nu^\pi(s,a) := d^\pi(s)\pi(a \mid s)$ for the stationary state–action distribution of π, and $\nu^t := \nu^{\pi^t}$, $\nu^\star := \nu^{\pi^\star}$. The critic's target at time t is the best linear fit to Q^t under ν^t:

$$w_t^\star \in \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \tfrac{1}{2}\, \mathbb{E}_{(s,a)\sim\nu^t}\Big[\big(w^\top \phi_{s,a} - Q^t(s,a)\big)^2\Big]. \tag{13}$$

Section 5.2 constructs an estimate $w_t \approx w_t^\star$ from samples, and the PMD update is then (2) with the LFA critic $\tilde{Q}^t(s,a) = \phi_{s,a}^\top w_t$.

Remark 5.1. Although (2) is written state by state, the actor in this section need not be tabular. For example, specializing D to KL divergence yields the multiplicative update $\pi^{t+1}(a \mid s) \propto \pi^t(a \mid s)\exp(-\eta_t \phi_{s,a}^\top w_t)$, so from a uniform π^0 every iterate is a log-linear policy in the critic's features, recovering the policy class of Yuan et al. [47]. Our results hold for any Bregman divergence D.

To bound A_t, B_t in Lemma 3.7, we make three assumptions. First, no linear critic can beat the best linear fit, so we bound its approximation error:

Assumption 5.2 (Approximation error)**.** There exists $\varepsilon_{\mathrm{app}} \ge 0$ such that $\mathbb{E}_{(s,a)\sim\nu^t}[(w_t^{\star\top}\phi_{s,a} - Q^t(s,a))^2] \le \varepsilon_{\mathrm{app}}^2$ for all $t \ge 0$.

Second, the critic's statistical error, measured in the feature-covariance norm $\|u\|_{\Sigma_t} := \sqrt{u^\top \Sigma_t u}$ with $\Sigma_t := \mathbb{E}_{(s,a)\sim\nu^t}[\phi_{s,a}\phi_{s,a}^\top]$, must be controllable (we show how in Section 5.2):

Assumption 5.3 (Statistical error)**.** There exists $\varepsilon_{\mathrm{stat}} \ge 0$ such that $\sup_{t\ge 0} \|w_t - w_t^\star\|_{\Sigma_t} \le \varepsilon_{\mathrm{stat}}$.

Third, alongside the uniform mismatch coefficient χ_t, the MSE analysis requires a second-moment form of distribution mismatch:

Assumption 5.4 (Second-moment distribution mismatch)**.** There exist finite deterministic constants $C_+, C_\star < \infty$ such that, almost surely for all $t \ge 0$, ν^{t+1} and $\nu^\star$ are absolutely continuous with respect to ν^t and

$$\mathbb{E}_{(s,a)\sim\nu^t}\left[\left(\frac{\nu^{t+1}(s,a)}{\nu^t(s,a)}\right)^2\right] \le C_+, \qquad \mathbb{E}_{(s,a)\sim\nu^t}\left[\left(\frac{\nu^\star(s,a)}{\nu^t(s,a)}\right)^2\right] \le C_\star.$$

Remark 5.5. A simple sufficient condition to satisfy Assumption 5.4 is to enforce an ε-exploration on every iterate so that $\pi^t(a \mid s) \geq \varepsilon/|\mathcal{A}|$ for all (s, a). Then $\nu^t(s, a) \geq (\underline{d}\varepsilon)/|\mathcal{A}|$ and the assumption holds with $C_+, C_\star = |\mathcal{A}|/(\underline{d}\varepsilon)$, which can be loose and incurs an $O(\varepsilon)$ optimization bias.

Under these assumptions, a change-of-measure argument (Appendix D) bounds the critic-error quantities (4) by $A_t \leq 2\sqrt{C_+}(\varepsilon_{\text{stat}} + \varepsilon_{\text{app}})$ and $B_t \leq 2\sqrt{C_\star}(\varepsilon_{\text{stat}} + \varepsilon_{\text{app}})$, so Lemma 3.7 with these bounds as $\bar{A}$ and $\bar{B}$ yields:

Theorem 5.6 (Linear convergence of inexact PMD under LFA)**.** *Suppose Assumptions 5.2–5.4 hold, and let $\{\pi^t\}$ be generated by (2) with $\tilde{Q}^t(s, a) = \phi_{s,a}^\top w_t$. Set $\varepsilon_{\text{floor}}(x) := 2(x\sqrt{C_+} + \sqrt{C_\star})(\varepsilon_{\text{stat}} + \varepsilon_{\text{app}})$. Under the hypotheses of Lemma 3.7(i),*

$$(\Delta_t - \varepsilon_{\text{floor}}(\bar{\chi}))_+ \;\leq\; \left(1 - \tfrac{1}{\bar{\chi}}\right)^t \left(\Delta_0 + \tfrac{D_0^\star}{\eta_{-1}\bar{\chi}}\right) \qquad \forall\, t \geq 0; \tag{14}$$

and under the hypotheses of Lemma 3.7(ii), for every $\zeta > 0$,

$$\limsup_{t\to\infty} \big(\Delta_t - \varepsilon_{\text{floor}}(\chi + \zeta)\big)_+^{1/t} \;\leq\; 1 - \tfrac{1}{\chi+\zeta}. \tag{15}$$

5.2 Sample complexity

We now satisfy Assumption 5.3 with transitions sampled along a single trajectory, by minimizing the objective in (13) with SGD. Two obstacles arise when adapting the closest prior work.

Obstacle 1: unbiased samples without a discount factor. In the closest discounted work, Yuan et al. [47] fit the critic by averaged SGD, which needs independent state–action pairs with unbiased Q-labels. They obtain both from geometric stopping, which samples exactly from the discounted occupancy. Average reward weights all times equally, so it has no such stopping distribution.

Our approach: sampling through mixing. We approximate independent pairs and unbiased labels on a single trajectory that alternates between burn-ins of m transitions and rollouts of H transitions. For the pairs, we take (s_k, a_k) at the end of the k-th burn-in, which makes each pair's distribution given the past close to ν^t by (11). A coupling argument then lets the analysis treat the pairs as independent, up to the burn-in error (App. D). For the labels, we reuse the truncated centered-return estimator of Section 4.2, applied to the rollout that starts at (s_k, a_k). By the bounds of Lem. C.2, each label's bias decays geometrically in H, like the burn-in error in m, so $m, H = \widetilde{O}(t_{\text{mix}})$ transitions per labeled pair make both errors arbitrarily small.

Obstacle 2: accurate critics for a linear-rate actor. In the closest average-reward work, the sublinear actor of Li et al. [25, Cor. 3] needs $\widetilde{O}(\kappa^2/\varepsilon^2)$ steps, where κ^2 bounds the second moment of their critic outputs and scales as t_{mix}^2. The many small actor steps average out the critic noise, so each critic call only needs to control bias. Each call of their exploratory variance-reduced temporal difference (EVRTD) critic does this on a single trajectory at a cost of order t_{mix}^3, adding up to t_{mix}^5. Our actor (Thm. 5.6) instead converges linearly, in $\widetilde{O}(\bar{\chi})$ steps with growing step sizes. To retain the linear rate with few, large actor steps, each critic call must also control its variance to meet Asm. 5.3.

Our approach: averaging inside the critic. We move the averaging from the actor into the critic. Each critic call runs SGD on the labeled pairs with the tail averaging of Jain et al. [15] instead of the full averaging used by Yuan et al. [47]. The labels' second moment is $\widetilde{O}(t_{\text{mix}}^2)$ by Lem. C.2 and plays the role of κ^2, so this tail-averaged SGD needs $\widetilde{O}(d\, t_{\text{mix}}^2/\varepsilon_{\text{stat}}^2)$ such pairs to leading order. Since a labeled pair's cost scales only as t_{mix} rather than the t_{mix}^3 of an EVRTD call, the end-to-end dependence (Thm. 5.10) drops from t_{mix}^5 to t_{mix}^3.

Algorithm 1 (App. D) combines the two approaches into our critic: tail-averaged SGD on the quadratic loss (13) over inner iterations $k = 1, \ldots, K$. Iteration k takes the next pair (s_k, a_k) and its truncated centered-return label $\widehat{Q}_k$ along the trajectory, then applies the stochastic gradient of (13) at that pair, $(w_{t,k-1}^\top \phi_{s_k,a_k} - \widehat{Q}_k)\, \phi_{s_k,a_k}$. The algorithm returns the average of the last $\lceil K/2 \rceil$ iterates, which has a small expected error. To make the error small with high probability, Algorithm 2 runs Algorithm 1 along the trajectory $O(\log(2/\delta))$ times and returns as w_t the output closest to a majority of the outputs in the empirical feature-covariance norm, the selection rule of Hsu and Sabato [14]. The same trajectory continues across outer iterations, each critic call following the current policy π^t. The resulting critic satisfies the following finite-sample guarantee:

Lemma 5.7 (LFA critic upper bound)**.** *Suppose Assumption 5.2 holds, $\|\phi_{s,a}\|_2 \le L$ for all (s,a), and $\Sigma_t \succeq \mu I_d$ for all t, for some $\mu > 0$. Let w_t be the output of Algorithm 2 with step size $\alpha = \frac{1}{4L^2}$. Then for any target critic accuracy $\varepsilon_{\mathrm{stat}} \in (0,1)$ and confidence $\delta \in (0,1)$, the number of transitions*

$$N \;=\; \widetilde{O}\left(\frac{d\, t_{\mathrm{mix}}^3}{\varepsilon_{\mathrm{stat}}^2} + \frac{L^2}{\mu}\, t_{\mathrm{mix}}\Big(1 + \frac{\varepsilon_{\mathrm{app}}^2}{\varepsilon_{\mathrm{stat}}^2}\Big)\right)$$

suffices to guarantee $\|w_t - w_t^\star\|_{\Sigma_t} \le \varepsilon_{\mathrm{stat}}$ with probability at least $1-\delta$.

Remark 5.8. The covariance condition $\Sigma_t \succeq \mu I_d$ of Lem. 5.7 can be guaranteed by the features alone. It holds with a single $\mu > 0$ for all policies if $\mathrm{span}\{\phi_{s,a_s} : s \in \mathcal{S}\} = \mathbb{R}^d$ for every choice of one action a_s per state. When $d \le |\mathcal{S}|$, this is true for almost every choice of features. If the features lack this property, the ε-exploration of Rem. 5.5 enforces the condition whenever $\mathrm{span}\{\phi_{s,a} : (s,a) \in \mathcal{S}\times\mathcal{A}\} = \mathbb{R}^d$. Appendix D.6 gives the details.

The $d\, t_{\mathrm{mix}}^3/\varepsilon_{\mathrm{stat}}^2$ term of Lemma 5.7 is information-theoretically tight. With one action and $2d$ states, the construction of Proposition 4.8 is linearly realizable in dimension d, and the same Assouad reduction in the covariance norm yields the dimension factor:

Proposition 5.9 (LFA critic lower bound)**.** *For every $d \ge 1$, $p \in (0, 1/4)$, $\varepsilon_{\mathrm{stat}} \in (0, 1/(256\,p)]$, and every (possibly adaptive) generative-model algorithm, there is a one-action average-reward MDP on $2d$ states satisfying Assumption 3.1 with $t_{\mathrm{mix}} = \Theta(1/p)$ and exactly realizable linear differential Q on which outputting $\widehat{w} \in \mathbb{R}^d$ with $\Pr(\|\widehat{w} - w^\star\|_\Sigma \le \varepsilon_{\mathrm{stat}}) \ge 2/3$ requires*

$$N \;=\; \Omega\big(d\, t_{\mathrm{mix}}^3/\varepsilon_{\mathrm{stat}}^2\big)$$

generative-model transitions. The explicit construction and constants are given in Appendix D.

The upper bound (Lem. 5.7) and the lower bound (Prop. 5.9) match up to logarithmic factors when $\varepsilon_{\mathrm{app}} \le \varepsilon_{\mathrm{stat}} \le t_{\mathrm{mix}}\sqrt{d\mu}/L$, for accuracies covered by both and confidence $\delta \le 1/3$. In this regime, the critic's worst-case sample complexity is $\widetilde{\Theta}(d\, t_{\mathrm{mix}}^3/\varepsilon_{\mathrm{stat}}^2)$, since the lower bound extends to every L and μ allowed by Lem. 5.7 and to a single trajectory (App. D.6). The ratio L^2/μ, which bounds the condition number of Σ_t, enters the upper bound only through a lower-order term.

Composing Lem. 5.7's critic with Thm. 5.6's actor yields the following end-to-end sample complexity:

Theorem 5.10 (LFA sample complexity)**.** *Suppose Assumption 5.4 and the hypotheses of Lemma 5.7 hold, and $D_0^\star < \infty$. For any target suboptimality $\varepsilon \in (0,1)$ with $\varepsilon \ge 8(\bar{\chi}\sqrt{C_+} + \sqrt{C_\star})\varepsilon_{\mathrm{app}}$ and confidence $\delta \in (0,1)$, using the critic of Lemma 5.7 at accuracy $\varepsilon_{\mathrm{stat}} = \varepsilon/(8(\bar{\chi}\sqrt{C_+} + \sqrt{C_\star}))$ together with the actor of Theorem 5.6 in (14) guarantees $\Delta_T \le \varepsilon$ with probability at least $1-\delta$ after $T = \widetilde{\Theta}(\bar{\chi})$ outer iterations, using a total of*

$$N \;=\; \widetilde{O}\left(\frac{d\, t_{\mathrm{mix}}^3\, \bar{\chi}\,(\bar{\chi}^2 C_+ + C_\star)}{\varepsilon^2} + \frac{L^2}{\mu}\, t_{\mathrm{mix}}\, \bar{\chi}\right)$$

transitions of a single trajectory.

Remark 5.11. With the same single-trajectory oracle as Li et al. [25, Corollary 3], Theorem 5.10 improves their $|\mathcal{S}||\mathcal{A}|\, t_{\mathrm{mix}}^5$ dependence to $d\, t_{\mathrm{mix}}^3$ through the trade-off of Obstacle 2: a linear-rate actor with few accurate critic calls in place of a sublinear actor with many bias-controlled ones. By Proposition 5.9, the $d\, t_{\mathrm{mix}}^3/\varepsilon^2$ dependence of the critic's sample complexity in Theorem 5.10 is information-theoretically tight for any policy-gradient method whose critic satisfies Assumption 5.3, even one with a generative model. The remaining factor $\bar{\chi}(\bar{\chi}^2 C_+ + C_\star)$ collects standard PMD-path and distribution-mismatch quantities outside the critic's sample complexity.

6 Conclusion

We developed a finite-time, finite-sample theory of policy mirror descent for ergodic average-reward MDPs around a master recursion (Lemma 3.7) whose two-part structure separates a transient $\bar{\chi}$-regime from an asymptotic χ-regime, the latter empirically reached after a brief warm-up. Our tabular and LFA sample complexities (Theorems 4.9, 5.10) improve the prior best policy-gradient dependence, with guarantees that hold with high probability for the last iterate. The tabular critic uses a generative model and pays t_{mix}^3 per state and t_{mix}^2 per state–action pair. The LFA critic uses a single trajectory and pays $d\, t_{\mathrm{mix}}^3$ in its leading term, free of $|\mathcal{S}|$ and $|\mathcal{A}|$. Matching lower bounds (Propositions 4.8, 5.9) establish that both critics are optimal up to logarithmic factors. Because the master recursion holds for any critic, neural critics are a natural next step, as are multichain MDPs and regret guarantees.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers and the area chair for their thoughtful comments and helpful suggestions, which improved the paper. This work was supported in part by the NSF AI Institute (AI-EDGE) 2112471 and the ARO Grant W911NF-24-1-0103.

References

[1] Suzan Ece Ada, Emre Ugur, and H. Levent Akin. Generalization in Transfer Learning: Robust Control of Robot Locomotion. *Robotica*, 40(11):3811–3836, 2022. doi: 10.1017/S0263574722000625.

[2] Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift, 2020. arXiv:1908.00261.

[3] Carlo Alfano, Rui Yuan, and Patrick Rebeschini. A Novel Framework for Policy Mirror Descent with General Parameterization and Linear Convergence. In *Advances in Neural Information Processing Systems*, 2023.

[4] Francis Bach and Eric Moulines. Non-Strongly-Convex Smooth Stochastic Approximation with Convergence Rate O(1/n). In *Advances in Neural Information Processing Systems*, 2013.

[5] Qinbo Bai, Washim Uddin Mondal, and Vaneet Aggarwal. Regret Analysis of Policy Gradient Algorithm for Infinite Horizon Average Reward Markov Decision Processes, 2024. arXiv:2309.01922.

[6] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, fourth edition, 2017.

[7] Semih Cayci, Niao He, and R. Srikant. Linear Convergence of Entropy-Regularized Natural Policy Gradient with Linear Function Approximation, 2024. arXiv:2106.04096.

[8] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization, 2021. arXiv:2007.06558.

[9] Brett Daley. Gym Classics. https://github.com/brett-daley/gym-classics, 2021.

[10] Swetha Ganesh, Washim Uddin Mondal, and Vaneet Aggarwal. Order-Optimal Regret with Novel Policy Gradient Approaches in Infinite-Horizon Average Reward MDPs. In *International Conference on Artificial Intelligence and Statistics*, 2025.

[11] Swetha Ganesh, Washim Uddin Mondal, and Vaneet Aggarwal. A Sharper Global Convergence Analysis for Average Reward Reinforcement Learning via an Actor-Critic Approach, 2025. arXiv:2407.18878.

[12] Abhijit Gosavi. A Reinforcement Learning Algorithm Based on Policy Iteration for Average Reward: Empirical Results with Yield Management and Convergence Analysis. *Machine Learning*, 55(1):5–29, 2004. doi: 10.1023/B:MACH.0000019802.64038.6c.

[13] Isaac Grosof, Siva Theja Maguluri, and R. Srikant. Convergence of Natural Policy Gradient for a Family of Infinite-State Queueing MDPs, 2025. arXiv:2402.05274.

[14] Daniel Hsu and Sivan Sabato. Loss Minimization and Parameter Estimation with Heavy Tails. *Journal of Machine Learning Research*, 2016.

[15] Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing Stochastic Gradient Descent for Least Squares Regression: Mini-Batching, Averaging, and Model Misspecification. *Journal of Machine Learning Research*, 2018.

[16] Emmeran Johnson, Ciara Pike-Burke, and Patrick Rebeschini. Optimal Convergence Rate for Exact Policy Mirror Descent in Discounted Markov Decision Processes, 2023. arXiv:2302.11381.

[17] Sham M. Kakade. A Natural Policy Gradient. In *Advances in Neural Information Processing Systems*, 2001.

[18] Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the Linear Convergence of Natural Policy Gradient Algorithm, 2021. arXiv:2105.01424.

[19] Navdeep Kumar, Yashaswini Murthy, Itai Shufaro, Kfir Y. Levy, R. Srikant, and Shie Mannor. On the Global Convergence of Policy Gradient in Average Reward Markov Decision Processes, 2024. arXiv:2403.06806.

[20] Guanghui Lan. Policy Mirror Descent for Reinforcement Learning: Linear Convergence, New Sampling Complexity, and Generalized Problem Classes, 2022. arXiv:2102.00135.

[21] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

[22] Jongmin Lee and Ernest K. Ryu. Policy Gradient Algorithms in Average-Reward Multichain MDPs, 2026. arXiv:2602.18003.

[23] Jongmin Lee, Ernest K. Ryu, and Vaneet Aggarwal. Learning Policy from a Single Trajectory in Average-Reward Markov Decision Process, 2026. arXiv:2606.16729.

[24] David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, second edition, 2017.

[25] Tianjiao Li, Feiyang Wu, and Guanghui Lan. Stochastic First-Order Methods for Average-Reward Markov Decision Processes, 2024. arXiv:2205.05800.

[26] Dachao Lin and Zhihua Zhang. On the Convergence of Policy in Unregularized Policy Mirror Descent, 2024. arXiv:2205.08176.

[27] Jiacai Liu, Wenye Li, Dachao Lin, Ke Wei, and Zhihua Zhang. On the Convergence of Projected Policy Gradient for Any Constant Step Sizes. *Journal of Machine Learning Research*, 2025.

[28] Sridhar Mahadevan. Average Reward Reinforcement Learning: Foundations, Algorithms, and Empirical Results. *Machine Learning*, 22(1-3):159–195, 1996. doi: 10.1007/BF00114727.

[29] Yashaswini Murthy and R. Srikant. On the Convergence of Natural Policy Gradient and Mirror Descent-Like Policy Methods for Average-Reward MDPs. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 1979–1984, 2023. doi: 10.1109/CDC49753.2023.10383691.

[30] Yashaswini Murthy, Mehrdad Moharrami, and R. Srikant. Performance Bounds for Policy-Based Average Reward Reinforcement Learning Algorithms. In *Advances in Neural Information Processing Systems*, 2023.

[31] Yashaswini Murthy, Isaac Grosof, Siva Theja Maguluri, and R. Srikant. Performance of NPG in Countable State-Space Average-Cost RL, 2024. arXiv:2405.20467.

[32] Michael J. Neely. *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan & Claypool Publishers, 2010.

[33] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[34] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994. doi: 10.1002/9780470316887.

[35] Sheldon M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, 1983.

[36] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust Region Policy Optimization, 2017. arXiv:1502.05477.

[37] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, 2017. arXiv:1707.06347.

[38] Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive Trust Region Policy Optimization: Global Convergence and Faster Rates for Regularized MDPs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5668–5675, 2020. doi: 10.1609/aaai.v34i04.6021.

[39] R. Srikant and Lei Ying. *Communication Networks: An Optimization, Control, and Stochastic Networks Perspective*. Cambridge University Press, 2014.

[40] Zhongchang Sun, Sihong He, Fei Miao, and Shaofeng Zou. Policy Optimization for Robust Average Cost MDPs. In *Advances in Neural Information Processing Systems*, 2024.

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

[42] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, 1999.

[43] Joel A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

[44] Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal Sample Complexity for Average Reward Markov Decision Processes, 2024. arXiv:2310.08833.

[45] Lin Xiao. On the Convergence Rates of Policy Gradient Methods. *Journal of Machine Learning Research*, 2022.

[46] Bin Yu. Assouad, Fano, and Le Cam. In David Pollard, Erik Torgersen, and Grace L. Yang, editors, *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

[47] Rui Yuan, Simon S. Du, Robert M. Gower, Alessandro Lazaric, and Lin Xiao. Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies, 2023. arXiv:2210.01400.

[48] Nida Zamir and I.-Hong Hou. Deep Index Policy for Multi-Resource Restless Matching Bandit and Its Application in Multi-Channel Scheduling, 2024. arXiv:2408.07205.

[49] 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, 2023. arXiv:2105.11066.

Supplementary Material

Organization. This appendix contains detailed related-work comparisons, deferred proofs, and additional technical details. Appendix A provides detailed comparisons to closely related work. Appendix B collects the background facts of Section 3, with references, and proves Lemma 3.7. Appendix C contains the convergence and sample complexity proofs for the tabular case. Appendix D contains the convergence and sample complexity proofs under linear function approximation. Appendix E proves Lemma 4.5 and provides additional details on mixing times. Appendix F presents numerical experiments.

A Detailed related work

This appendix expands on the landscape summarized in Section 2. We discuss three closest comparators in Appendix A.1 (Li et al., Sun et al., Lee and Ryu) and broader policy-optimization lines in Appendix A.2. Each entry uses the same two-part format: a statement of the work's results, followed by a comparison to ours.

A.1 Closest comparators

Li, Wu, Lan [25]. *Results.* Li et al. [25] analyze stochastic policy mirror descent for average-reward MDPs. Their Theorem 1 gives an $O(1/\sqrt{K})$ rate for a uniformly random iterate of the *unregularized* objective with a constant step size. Under strongly convex regularization, their Theorem 2 gives $O(1/K)$ with decaying step sizes, and their Theorem 3, with a constant step size, bounds the expected divergence of the last iterate from the regularized optimum by a geometrically decaying term, an $\widetilde{O}(1/K)$ stochastic term, and a bias floor. With a generative model, Corollary 1 translates Theorem 1

into a tabular sample-complexity guarantee of $\widetilde{O}(|\mathcal{S}||\mathcal{A}|\, t_{\text{mix}}^3/\varepsilon^2)$ in expectation. In the unregularized LFA case, Corollary 3 uses a single trajectory and runs $\widetilde{O}(\kappa^2/\varepsilon^2)$ actor iterations, where κ^2 bounds the second moment of the critic outputs (their Assumption 3), each with an EVRTD critic call whose bias control costs $\widetilde{O}(t_{\text{mix}}^3/\bar{\mu})$ transitions ($\bar{\mu}$ is their lower bound, over all policies, on the smallest eigenvalue of the feature covariance), for $\widetilde{O}(t_{\text{mix}}^3\kappa^2/(\bar{\mu}\,\varepsilon^2))$ in total; their bound $\kappa^2 \lesssim t_{\text{mix}}^2$, up to conditioning and approximation terms, gives an effective t_{mix}^5 mixing dependence. Its conditioning term divides by their smallest exploratory state–action occupancy over all policies (their (4.22) and Corollary 3), which is at most $1/(|\mathcal{S}||\mathcal{A}|)$, so their stated bound contains a factor of at least $|\mathcal{S}||\mathcal{A}|$, the value Table 1 lists.

Comparison. Our analysis is complementary along three axes. First, we obtain linear convergence to the optimum up to critic error, without external regularization, using geometric step-size growth rather than constant steps (Theorem 4.2, Corollary 4.3, Theorem 5.6); the geometric term of their Theorem 3 concerns the regularized objective. Second, with the same generative-model oracle as their Corollary 1, our tabular sample complexity improves their $|\mathcal{S}||\mathcal{A}|\, t_{\text{mix}}^3$ dependence to $|\mathcal{S}|\, t_{\text{mix}}^3 + |\mathcal{S}||\mathcal{A}|\, t_{\text{mix}}^2$ (Theorem 4.9): our critic pays t_{mix}^3 per state, for the rollouts that estimate h^π, and only t_{mix}^2 per state–action pair (Section 4.2). Our LFA sample complexity, also on a single trajectory, has t_{mix}^3 rather than their t_{mix}^5 dependence, through the trade-off described in Obstacle 2 of Section 5.2 (Theorem 5.10): our linear-rate actor needs only $\widetilde{O}(\bar{\chi})$ critic calls, each made accurate with high probability by tail-averaged SGD with confidence boosting on $\widetilde{O}(d\, t_{\text{mix}}^2/\varepsilon_{\text{stat}}^2)$ rollout labels when $\varepsilon_{\text{app}} \le \varepsilon_{\text{stat}}$, up to a lower-order term (Lemma D.2), costing $\widetilde{O}(t_{\text{mix}})$ transitions each, whereas their sublinear actor needs $\widetilde{O}(\kappa^2/\varepsilon^2)$ bias-controlled calls costing $\widetilde{O}(t_{\text{mix}}^3/\bar{\mu})$ transitions each. Our labels pay the same quadratic second moment as their κ^2, but inside the critic. Our bound carries no such occupancy factor. Both our sample complexities, with the same oracles as theirs, are also high-probability and last-iterate rather than in expectation for a random iterate (Theorems 4.9, 5.10). Third, Propositions 4.8 and 5.9 establish that both critics are optimal up to logarithmic factors, the LFA one in its leading term.

Sun, He, Miao, Zou [40]. *Results.* Sun et al. [40] study robust PMD for ergodic average-cost MDPs under rectangular transition uncertainty, with Bregman divergence specialized to squared Euclidean. Their Theorem 4.6 establishes linear convergence at rate $(1 - 1/M_0)^t$ for an increasing step-size schedule $\eta_t/\eta_{t-1} \ge M_0^2/(M_0 - 1)$, where $M_0 := \sup_{\pi\in\Pi} \|d^{\pi^\star}/d^\pi\|_\infty$ is the global worst-case distribution mismatch. Specialized to robustness radius $R = 0$, this becomes a linear-convergence result for standard (non-robust) ergodic exact PMD.

Comparison. At $R = 0$ their result partially overlaps with the exact case ($\varepsilon_Q = 0$) of our Theorem 4.2. Our analysis admits any upper bound $\bar{\chi}$ on $\sup_t \chi_t$ (with $\sup_t \chi_t \le M_0 \le \underline{d}^{-1}$), and tighter $\bar{\chi}$ yield strictly sharper rates than Sun's worst-case M_0 in path-dependent regimes. Our analysis holds for general Legendre-Bregman divergences, not only squared Euclidean. Taking $\bar{\chi} = M_0$ matches Sun's rate $1 - 1/M_0$ at growth ratio $M_0/(M_0 - 1)$, a factor M_0 milder than Sun's $M_0^2/(M_0 - 1)$; tighter $\bar{\chi}$ trade larger growth ratios for faster path-dependent rates, a trade-off Sun's worst-case analysis cannot expose. Our Corollary 4.3 additionally provides an asymptotic rate $\limsup_t \Delta_t^{1/t} \le 1 - 1/\chi$ with a superlinear regime when $\chi = 1$. Finally, we extend the analysis to inexact critics (Theorem 4.2), linear function approximation (Theorem 5.6), and end-to-end sample complexity with matching information-theoretic lower bounds on the critic's sample complexity (Propositions 4.8, 5.9). Sun's framework in turn handles rectangular robust uncertainty, which ours does not.

Lee and Ryu [22]. *Results.* Concurrent work by Lee and Ryu [22] extends policy gradient analysis to multichain average-reward MDPs, a strict generalization of our ergodic assumption. They introduce recurrent and transient visitation measures, establish a multichain policy gradient theorem, and analyze an α-clipped policy mirror ascent algorithm restricted to the set $\Pi_\alpha := \{\pi : \pi(a \mid s) \ge \alpha\}$. Their Theorem 10 establishes linear convergence of the optimality gap with respect to the best α-clipped policy. For any fixed $\alpha > 0$ this leaves a clipping gap to the positive-policy optimum $J_{+,\mu}^\star := \sup_{\pi\in\Pi_+} \mathbb{E}_{s\sim\mu}[J^\pi(s)]$, which is nonzero in general and vanishes as $\alpha \to 0$; here Π_+ is the simplex interior. This restriction is justified by the discontinuity of $\pi \mapsto J^\pi$ on the simplex boundary in multichain MDPs. Their Theorem 13 gives sample complexity $\widetilde{O}(t_{\text{tar}}^3\, \|Q\|_\infty^4\, R^3\, C_\alpha^6\, B_\alpha^6\, |\mathcal{S}||\mathcal{A}|/\varepsilon^6)$ under an adaptive step-size schedule (the constant step-size variant has a worse ε^{-12} dependence).

Comparison. Their axis of generalization (ergodic $\rightarrow$ multichain) and ours (sharper ergodic analysis) are complementary. Their sample complexity is $\widetilde{O}(\varepsilon^{-6})$ versus our $\widetilde{O}(\varepsilon^{-2})$ (with the critic's sample complexity tight by Proposition 4.8), and their optimality notion is positive-policy rather than global. Within the ergodic setting, we additionally obtain asymptotic and superlinear rates (Corollary 4.3) and linear function approximation results (Section 5), neither of which the multichain analysis addresses. Concurrent work by Lee et al. [23] extends this α-clipped scheme to a single trajectory in weakly communicating MDPs, with high-probability tabular sample complexities of $\widetilde{O}(\varepsilon^{-4})$ for that policy-based method and $\widetilde{O}(\varepsilon^{-2})$ for a value-based one. Their bounds are stated through hitting, cover, and target times; on a single trajectory, ours give a policy-based $\widetilde{O}(\varepsilon^{-2})$ sample complexity with high probability under linear function approximation, with the leading term of the critic's sample complexity tight by Proposition 5.9.

A.2 Broader policy-optimization lines

The remaining works address average-reward policy optimization in parallel directions (sublinear rates, general parameterization, or infinite-state extensions). We use a uniform single-paragraph format below.

Kumar et al. [19]. *Results.* Kumar et al. [19] establish a global $O(1/T)$ suboptimality rate for projected policy gradient in ergodic tabular average-reward MDPs. *Comparison.* Their algorithm (projected PG in Euclidean geometry) and sublinear rate differ from our linear PMD result (Theorem 4.2 with $\varepsilon_Q = 0$) under general Bregman geometry.

Murthy and Srikant [29], Murthy et al. [30]. *Results.* Murthy and Srikant [29] establish $O(1/T)$ convergence for natural policy gradient and mirror-descent-like methods in ergodic average-reward MDPs under irreducibility, with discussion of non-tabular settings. Murthy et al. [30] develop approximate-policy-iteration performance bounds with step-size-dependent tradeoffs (their Remark 4.6 notes that larger steps reduce some terms while increasing others, making the overall effect non-transparent). For a mirror-descent update with a TD critic and linear function approximation (their Corollary 4.5), the bound keeps a term proportional to $\log(1/\omega)/\beta$, where β is the fixed step size and ω the smallest probability the iterates assign to an optimal action, so it does not vanish as the critic becomes exact. *Comparison.* Our master-lemma analysis (Lemma 3.7) makes the step-size effect explicit and upgrades the $O(1/T)$ rate of Murthy and Srikant [29] to linear via geometric step-size growth. Our bound has no such term: with growing step sizes, the error floor of Theorem 5.6 vanishes with the critic error.

Bai et al. [5]. *Results.* Bai et al. [5] analyze general-parameterized policy gradient for average-reward MDPs, obtaining $\widetilde{O}(T^{-1/4})$ optimality error and $\widetilde{O}(T^{3/4})$ regret. *Comparison.* Their work is broader in parameterization but weaker in rate; our PMD analysis is restricted to tabular and linear function approximation but yields linear convergence.

Ganesh et al. [10, 11]. *Results.* Ganesh et al. [10] improve regret for general-parameterized average-reward policy gradient via variance-reduction techniques, achieving near order-optimal $\widetilde{O}(\sqrt{T})$ rates. Ganesh et al. [11] develop a multi-level-Monte-Carlo natural actor-critic attaining $\widetilde{O}(T^{-1/2})$ global convergence without requiring knowledge of mixing or hitting times. *Comparison.* Both works are broader in parameterization than ours but do not establish linear rates or the t_{mix}^3 mixing dependence we obtain.

Grosof et al. [13]. *Results.* Grosof et al. [13] establish $O(1/\sqrt{T})$ global convergence for natural policy gradient in a family of countably-infinite-state queueing MDPs, initialized from the MaxWeight policy. *Comparison.* Extends NPG beyond finite state spaces; complementary to our finite-state PMD analysis.

B Preliminaries

B.1 Existence of the Poisson equation and differential values

Fact B.1 (Poisson equation and differential values)**.** *For each policy π, there exist a scalar $\rho(\pi)$ and a function $h^\pi : \mathcal{S} \to \mathbb{R}$ such that*

$$\rho(\pi) + h^\pi \;=\; c_\pi + P_\pi h^\pi. \tag{16}$$

The pair $(\rho(\pi), h^\pi)$ is unique up to adding a constant to h^π. Throughout the paper, we fix the centered *convention $d^\pi h^\pi = 0$, under which $h^\pi(s) = \lim_{H\to\infty} \mathbb{E}[\sum_{k=0}^{H-1}(c(s_k,a_k) - \rho(\pi)) \mid s_0 = s]$ and $Q^\pi(s,a) = c(s,a) - \rho(\pi) + \mathbb{E}[h^\pi(s_1) \mid s_0 = s, a_0 = a]$.*

Proof. Fix a policy π, let $\mathbf{1} \in \mathbb{R}^{|\mathcal{S}|}$ denote the all-ones vector, and write $g \coloneqq c_\pi - \rho(\pi)\mathbf{1}$. For an ergodic chain, (16) has a solution, unique up to adding a constant to h^π; see, e.g., Puterman [34] and Li et al. [25, Section 2]. Applying d^π to (16) and using $d^\pi P_\pi = d^\pi$ gives $\rho(\pi) = d^\pi c_\pi$, so $d^\pi g = 0$, and the centered convention $d^\pi h^\pi = 0$ selects a unique h^π.

Finally, we verify the limit representation. Under Assumption 3.1, $P_\pi^k \to \mathbf{1}d^\pi$ geometrically fast (Lemma E.1), and $d^\pi g = 0$, so $P_\pi^k g = (P_\pi^k - \mathbf{1}d^\pi)\, g \to 0$ geometrically and $\tilde{h} \coloneqq \sum_{k\ge 0} P_\pi^k g$ converges. Telescoping gives $(I - P_\pi)\tilde{h} = g$, and $d^\pi \tilde{h} = \sum_{k\ge 0} d^\pi P_\pi^k g = \sum_{k\ge 0} d^\pi g = 0$, so $\tilde{h} = h^\pi$ by uniqueness under the centered convention. Since $(P_\pi^k g)(s) = \mathbb{E}[c(s_k,a_k) - \rho(\pi) \mid s_0 = s]$, this is the stated limit for h^π; the expression for Q^π follows by conditioning on the first transition in (1). □

B.2 Fact 3.3 (Performance difference lemma)

Fact 3.3 is the unregularized case of Li et al. [25, Lemma 2]. In our conventions it reads, for any policies π, π' under Assumption 3.1,

$$\rho(\pi') - \rho(\pi) \;=\; \mathbb{E}_{s\sim d^{\pi'}}\Big[\big\langle Q_s^\pi,\, \pi'_s - \pi_s \big\rangle\Big]. \tag{17}$$

Proof. By the definition of Q^π and (16), $\langle Q_s^\pi, \pi_s\rangle = h^\pi(s)$ and $\langle Q_s^\pi, \pi'_s\rangle = \langle c_s, \pi'_s\rangle - \rho(\pi) + \mathbb{E}[h^\pi(s_1) \mid s_0 = s, a_0 \sim \pi'_s]$. Subtracting, averaging over $s \sim d^{\pi'}$, and using the stationarity of $d^{\pi'}$ under $P_{\pi'}$ cancels the h^π terms; $\rho(\pi') = \mathbb{E}_{s\sim d^{\pi'}}[\langle c_s, \pi'_s\rangle]$ then gives (17). □

B.3 Fact 3.4 (Policy gradient theorem)

Fact 3.4 is the average-reward policy gradient theorem [42, Theorem 1] for the direct parameterization; see also the unregularized case of Li et al. [25, Lemma 1]. It also follows from Fact 3.3: perturbing π_s alone to $\pi_s + \varepsilon u \in \mathcal{P}(\mathcal{A})$ with $\sum_a u(a) = 0$ gives a policy π_ε with $\rho(\pi_\varepsilon) - \rho(\pi) = \varepsilon\, d^{\pi_\varepsilon}(s)\, \langle u, Q_s^\pi\rangle$, and $d^{\pi_\varepsilon}(s) \to d^\pi(s)$ as $\varepsilon \to 0$ by the continuity of $\pi \mapsto d^\pi$ shown in the proof of Lemma 3.2.

B.4 Fact 3.6 (Three-point inequality for PMD)

Fact 3.6 is the three-point descent lemma of Xiao [45, Lemma 6], applied on $\mathcal{P}(\mathcal{A})$ with the linear function $\varphi(q) = \eta_t\langle \tilde{Q}_s^t, q\rangle$ and reference point π_s^t: for every $p \in \mathcal{P}(\mathcal{A})$ it gives $\varphi(\pi_s^{t+1}) + D(\pi_s^{t+1}, \pi_s^t) \le \varphi(p) + D(p, \pi_s^t) - D(p, \pi_s^{t+1})$, which rearranges to the stated inequality. The lemma also keeps π_s^{t+1} in the relative interior of the domain of F whenever π_s^t is, so it applies at every iteration from an interior π^0.

B.5 Proof of Lemma 3.2 (Uniform boundedness of stationary distributions)

Proof. Fix π. By irreducibility, $d^\pi(s) > 0$ for every s: if $d^\pi(s_0) = 0$, stationarity, $0 = \sum_s d^\pi(s) P_\pi(s_0 \mid s)$, forces $d^\pi(s) = 0$ wherever $P_\pi(s_0 \mid s) > 0$, and iterating along the chain gives $d^\pi \equiv 0$, a contradiction. Next, $\pi \mapsto d^\pi$ is continuous: $\pi \mapsto P_\pi$ is continuous, and d^π is the unique solution of the linear system obtained from $d(P_\pi - I) = 0$ by replacing its last equation with $d\,\mathbf{1} = 1$; since 1 is a simple eigenvalue of the ergodic matrix P_π, this system is nonsingular, and

matrix inversion is continuous. Finally, Π is compact, so the continuous function $(\pi, s) \mapsto d^\pi(s)$ attains its minimum on $\Pi \times \mathcal{S}$, which is positive; this minimum is $\underline{d}$. Appendix E derives the explicit lower bound under a uniform minorization condition. □

B.6 Proof of Lemma 3.7 (Master convergence for inexact PMD)

Proof. The proof has two parts. We first derive a one-step Lyapunov recursion that holds whenever the step-size and mismatch conditions are satisfied at time t with some parameter $\bar{\chi} > 1$; then we iterate this recursion from $t = 0$ to obtain (i), and iterate it on a tail with $\bar{\chi} = \chi + \zeta$ (for any fixed $\zeta > 0$) to obtain (ii).

One-step recursion. Fix $t \geq 0$ and suppose $\chi_{t+1} \leq \bar{\chi}$ and $\eta_t/\eta_{t-1} \geq \bar{\chi}/(\bar{\chi} - 1)$ for some $\bar{\chi} > 1$. Applying Fact 3.6 with $p = \pi_s^\star$ to the generalized PMD update (2) and dropping the nonnegative term $D(\pi_s^{t+1}, \pi_s^t) \geq 0$ gives $\eta_t \langle \tilde{Q}_s^t, \pi_s^{t+1} - \pi_s^\star \rangle \leq D(\pi_s^\star, \pi_s^t) - D(\pi_s^\star, \pi_s^{t+1})$. Decompose $\pi_s^{t+1} - \pi_s^\star = (\pi_s^{t+1} - \pi_s^t) + (\pi_s^t - \pi_s^\star)$ and take expectation over $s \sim d^\star$:

$$\eta_t \, \mathbb{E}_{s\sim d^\star}\big[\langle \tilde{Q}_s^t, \pi_s^t - \pi_s^\star \rangle\big] + \eta_t \, \mathbb{E}_{s\sim d^\star}\big[\langle \tilde{Q}_s^t, \pi_s^{t+1} - \pi_s^t \rangle\big] \leq D_t^\star - D_{t+1}^\star. \tag{18}$$

By Fact 3.3 and the definition of B_t, $\mathbb{E}_{s\sim d^\star}[\langle \tilde{Q}_s^t, \pi_s^t - \pi_s^\star \rangle] \geq \Delta_t - B_t$. By (3) the integrand $\langle \tilde{Q}_s^t, \pi_s^{t+1} - \pi_s^t \rangle$ is pointwise nonpositive, so the change-of-measure ratio $d^\star(s)/d^{t+1}(s) \leq \bar{\chi}$ gives $\mathbb{E}_{s\sim d^\star}[\langle \tilde{Q}_s^t, \pi_s^{t+1} - \pi_s^t \rangle] \geq \bar{\chi} \, \mathbb{E}_{s\sim d^{t+1}}[\langle \tilde{Q}_s^t, \pi_s^{t+1} - \pi_s^t \rangle] \geq \bar{\chi}\big((\Delta_{t+1} - \Delta_t) - A_t\big)$, where the second inequality uses Fact 3.3 and the definition of A_t. Substituting both into (18) and rearranging,

$$\Delta_{t+1} + \frac{D_{t+1}^\star}{\eta_t \bar{\chi}} \leq (1 - 1/\bar{\chi})\Delta_t + \frac{D_t^\star}{\eta_t \bar{\chi}} + \frac{\bar{\chi} A_t + B_t}{\bar{\chi}}.$$

The step-size growth condition gives $1/(\eta_t \bar{\chi}) \leq (1 - 1/\bar{\chi}) \cdot 1/(\eta_{t-1}\bar{\chi})$, so defining the Lyapunov sequence $\Psi_t := \Delta_t + D_t^\star/(\eta_{t-1}\bar{\chi})$ yields the one-step recursion

$$\Psi_{t+1} \leq (1 - 1/\bar{\chi}) \, \Psi_t + \frac{\bar{\chi} A_t + B_t}{\bar{\chi}}. \tag{19}$$

Part (i): non-asymptotic rate. Under the hypotheses of (i), the one-step recursion (19) holds at every $t \geq 0$ with $\bar{\chi}$ the global uniform bound. Using $A_t \leq \bar{A}$, $B_t \leq \bar{B}$ and iterating from $t = 0$,

$$\Psi_t \leq (1 - 1/\bar{\chi})^t \Psi_0 + \frac{\bar{\chi}\bar{A} + \bar{B}}{\bar{\chi}} \sum_{i=0}^{t-1} (1 - 1/\bar{\chi})^i = (1 - 1/\bar{\chi})^t \Psi_0 + (\bar{\chi}\bar{A} + \bar{B})\big(1 - (1 - 1/\bar{\chi})^t\big).$$

Since $\Delta_t \leq \Psi_t$, subtracting $\bar{\chi}\bar{A} + \bar{B}$ and taking the positive part yields (5).

Part (ii): asymptotic rate. Let $\chi := \limsup_t \chi_t$ and fix $\zeta > 0$, setting $\hat{\chi} := \chi + \zeta$. By the $\limsup$ definition, $\chi_{t+1} \leq \hat{\chi}$ for all t beyond some $t_0(\zeta)$. By $\eta_t/\eta_{t-1} \to \infty$, there exists $t_1(\zeta)$ such that $\eta_t/\eta_{t-1} \geq \hat{\chi}/(\hat{\chi} - 1)$ for $t \geq t_1(\zeta)$. Set $t_\zeta := \max\{t_0(\zeta), t_1(\zeta)\}$; both hypotheses of the one-step recursion hold with $\bar{\chi} = \hat{\chi}$ for all $t \geq t_\zeta$. Iterating (19) from t_ζ gives

$$\Psi_t \leq (1 - 1/\hat{\chi})^{t - t_\zeta} \, \Psi_{t_\zeta} + (\hat{\chi}\bar{A} + \bar{B})\big(1 - (1 - 1/\hat{\chi})^{t - t_\zeta}\big) \qquad \forall \, t \geq t_\zeta.$$

Hence $(\Delta_t - (\hat{\chi}\bar{A} + \bar{B}))_+ \leq (1 - 1/\hat{\chi})^{t - t_\zeta} \, \Psi_{t_\zeta}$. Here $\Psi_{t_\zeta} < \infty$: $D_0^\star < \infty$, and the one-step inequality, which holds before the growth condition is invoked, keeps every $D_t^\star$ finite. Taking t-th roots, $\limsup$, and noting $\Psi_{t_\zeta}^{1/t} \to 1$ yields (6): $\limsup_{t\to\infty} (\Delta_t - ((\chi+\zeta)\bar{A} + \bar{B}))_+^{1/t} \leq 1 - 1/(\chi+\zeta)$. The same bound also gives $\limsup_t \Delta_t \leq \hat{\chi}\bar{A} + \bar{B}$; letting $\zeta \downarrow 0$ yields $\limsup_t \Delta_t \leq \chi\bar{A} + \bar{B}$. When $\bar{A} = 0$ the floor $\hat{\chi}\bar{A} + \bar{B}$ reduces to $\bar{B}$ independently of ζ, so (6) becomes $\limsup_{t\to\infty} (\Delta_t - \bar{B})_+^{1/t} \leq 1 - 1/(\chi + \zeta)$; letting $\zeta \downarrow 0$ gives the sharper rate $1 - 1/\chi$. □

B.7 Constant step sizes

The one-step inequality in the proof of Lemma 3.7 holds before the growth condition is invoked: if $\sup_t \chi_t \leq \bar{\chi}$, then for every $t \geq 0$,

$$\Delta_{t+1} + \frac{D_{t+1}^\star}{\eta_t \bar{\chi}} \;\leq\; \Big(1 - \frac{1}{\bar{\chi}}\Big)\Delta_t + \frac{D_t^\star}{\eta_t \bar{\chi}} + \frac{\bar{\chi} A_t + B_t}{\bar{\chi}}.$$

With a constant step size $\eta_t \equiv \eta$, multiplying by $\bar{\chi}$ and summing over $t = 0, \dots, T-1$ telescopes the divergence terms. Dropping $D_T^\star \geq 0$ and using $A_t \leq \bar{A}$, $B_t \leq \bar{B}$ gives

$$\bar{\chi} \sum_{t=1}^{T} \Delta_t - (\bar{\chi} - 1) \sum_{t=0}^{T-1} \Delta_t \;\leq\; \frac{D_0^\star}{\eta} + T\,(\bar{\chi}\bar{A} + \bar{B}).$$

The left side equals $\sum_{t=1}^{T} \Delta_t + (\bar{\chi} - 1)(\Delta_T - \Delta_0) \geq \sum_{t=1}^{T} \Delta_t - (\bar{\chi} - 1)\Delta_0$, so

$$\frac{1}{T} \sum_{t=1}^{T} \Delta_t \;\leq\; \frac{D_0^\star}{\eta T} + \frac{(\bar{\chi} - 1)\,\Delta_0}{T} + \bar{\chi}\bar{A} + \bar{B},$$

an $O(1/T)$ rate for the average suboptimality of the iterates, and hence for the best iterate, to the same error floor as Lemma 3.7(i). In the exact case ($\bar{A} = \bar{B} = 0$), the descent property $\rho(\pi^{t+1}) \leq \rho(\pi^t)$ of exact PMD (Section 3) makes Δ_t nonincreasing, so the same bound holds for the last iterate Δ_T.

What the growth condition buys. The growth condition is what makes the contraction of Lemma 3.7(i) uniform: rate $1 - 1/\bar{\chi}$ from $t = 0$ on every instance, with no dependence on advantage gaps. In the proof of Lemma 3.7, this requires the term $D_t^\star/(\eta_{t-1}\bar{\chi})$ to contract along with Δ_t, which the growth condition provides. Constant steps can still converge linearly, but at an instance-dependent rate: in a two-armed bandit, constant-step KL-PMD shrinks the odds of the suboptimal action by the same factor at every step, a factor set by η and the gap between the two actions' costs. On every instance, the cumulative bound of Proposition B.3(i) below gives a linear rate for constant step sizes, with an instance-dependent constant. The χ-adaptive rate of Lemma 3.7(ii) and the superlinear regime of Corollary 4.3 also rest on growing step sizes.

B.8 Exact PMD without a convergence assumption

Corollary 4.3 obtains a superlinear rate when $\{\pi^t\}$ converges to an optimal policy. This subsection removes that assumption for the KL divergence and the squared Euclidean distance. Both arguments rest on the optimal advantage $g^\star(s,a) := Q^{\pi^\star}(s,a) - h^{\pi^\star}(s)$, where $\pi^\star$ is any optimal policy.

Lemma B.2 (Optimal advantage)**.** *The optimal advantage $g^\star$ is nonnegative and the same for every optimal $\pi^\star$, and for every policy π,*

$$\rho(\pi) - \rho^\star \;=\; \mathbb{E}_{s\sim d^\pi}\big[\langle g_s^\star,\, \pi_s\rangle\big]. \tag{20}$$

In particular, π is optimal if and only if each π_s is supported on $\mathcal{A}_s^\star := \{a : g^\star(s,a) = 0\}$.

Proof. By (16), $\langle Q_s^{\pi^\star}, \pi_s^\star\rangle = h^{\pi^\star}(s)$, so $\langle g_s^\star, \pi_s^\star\rangle = 0$, and Fact 3.3 with (π, π') replaced by $(\pi^\star, \pi)$ gives (20). If $g^\star(s_0,a_0) < 0$, the policy π that plays a_0 at s_0 and agrees with $\pi^\star$ elsewhere has $\rho(\pi) - \rho^\star = d^\pi(s_0)\, g^\star(s_0,a_0) < 0$ by (20) and Lemma 3.2, contradicting the optimality of $\pi^\star$. Hence $g^\star \geq 0$, and since $d^\pi \geq \underline{d} > 0$, the right side of (20) vanishes exactly when each π_s is supported on $\mathcal{A}_s^\star$. Finally, if π' is another optimal policy, then each π'_s is supported on $\mathcal{A}_s^\star$, so $\langle Q_s^{\pi^\star}, \pi'_s\rangle = h^{\pi^\star}(s)$ for all s. As $\rho(\pi') = \rho^\star$, this says that $h^{\pi^\star}$ solves the Poisson equation (16) of π', so $h^{\pi^\star} - h^{\pi'}$ is a constant by Fact B.1. Then $Q^{\pi^\star} - Q^{\pi'}$ is the same constant, and π' defines the same $g^\star$. □

Proposition B.3 (Exact PMD in the KL and squared Euclidean geometries)**.** *Suppose not every policy is optimal. Let $\{\pi^t\}$ be generated by (2) with $\tilde{Q}^t = Q^t$, $\pi_s^0 \in \operatorname{rint}(\mathcal{P}(\mathcal{A}))$ for all s, and nondecreasing step sizes $\eta_t > 0$. Let $g_{\min} := \min\{g^\star(s,a) : g^\star(s,a) > 0\}$ and $g_{\max} := \max_{s,a} g^\star(s,a)$.*

(i) KL divergence. *For all sufficiently large t,*

$$\Delta_{t+1} \;\leq\; \frac{2 g_{\max}}{\underline{d}\, g_{\min}}\, e^{-\eta_t g_{\min}/2}\, \Delta_t,$$

and there is a constant $C < \infty$ such that $\Delta_t \leq C \exp\big(-\frac{g_{\min}}{2}\sum_{j=0}^{t-1}\eta_j\big)$ for all $t \geq 0$. In particular, if $\eta_t \to \infty$, then $\Delta_{t+1}/\Delta_t \to 0$.

(ii) Squared Euclidean distance. *There exists t_0 such that $\pi^t = \pi^{t_0}$ for all $t \geq t_0$, and π^{t_0} is optimal.*

Proof. The proof has four steps. Once Δ_t is small, Q^t separates optimal from suboptimal actions by at least $g_{\min}/2$, so every update moves probability away from suboptimal actions; Steps 3 and 4 quantify this for each geometry.

Step 1: $\Delta_t \to 0$. By Lemma 3.2, $\sup_t \chi_t \le \bar{\chi} := 1/\underline{d}$, and $D_0^\star < \infty$ for both geometries (Remark 3.8). For nondecreasing η_t, summation by parts gives $\sum_{t=0}^{T-1}(D_t^\star - D_{t+1}^\star)/\eta_t \le D_0^\star/\eta_0$, so the argument of Appendix B.7 goes through with $\eta = \eta_0$ and gives $\bar{\Delta}_T \le (D_0^\star/\eta_0 + (\bar{\chi}-1)\Delta_0)/T \to 0$.

Step 2: Suboptimal mass and action contrasts. Let $\mathcal{B}_s := \mathcal{A} \setminus \mathcal{A}_s^\star$ and $m_t := \max_s \pi^t(\mathcal{B}_s \mid s)$. Since $g^\star = 0$ on $\mathcal{A}_s^\star$ and $g_{\min} \le g^\star \le g_{\max}$ on $\mathcal{B}_s$, (20) and $d^t \ge \underline{d}$ give

$$\underline{d}\, g_{\min}\, m_t \;\le\; \Delta_t \;\le\; g_{\max}\, m_t, \tag{21}$$

so $m_t \to 0$ by Step 1. Moreover, every limit point of $\{\pi^t\}$ is optimal, as in Corollary 4.3, and $\pi \mapsto Q^\pi$ is continuous on the compact set Π: by Fact B.1 it suffices that $h^\pi = (I - P_\pi + \mathbf{1}d^\pi)^{-1}(c_\pi - \rho(\pi)\mathbf{1})$ is continuous, which holds because the matrix is invertible under Assumption 3.1 and continuous in π by the proof of Lemma 3.2. By Lemma B.2, every optimal policy has the action contrasts $Q^{\pi^\star}(s,b) - Q^{\pi^\star}(s,a) = g^\star(s,b) \ge g_{\min}$ for $a \in \mathcal{A}_s^\star$ and $b \in \mathcal{B}_s$. If some contrast of Q^t stayed below $g_{\min}/2$ along a subsequence, the corresponding policies would have an optimal limit point, which continuity rules out. Since also $m_t \to 0$, there is t_1 such that for all $t \ge t_1$, all s, $a \in \mathcal{A}_s^\star$ and $b \in \mathcal{B}_s$,

$$Q^t(s,b) - Q^t(s,a) \;\ge\; g_{\min}/2 \qquad \text{and} \qquad m_t \;\le\; 1/2. \tag{22}$$

Step 3: Contraction under the KL divergence. Here (2) gives $\pi^{t+1}(a \mid s) = \pi^t(a \mid s)\, e^{-\eta_t Q^t(s,a)}/Z_t(s)$ with $Z_t(s) := \sum_{a'} \pi^t(a' \mid s)\, e^{-\eta_t Q^t(s,a')}$, so all iterates are interior and $\Delta_t > 0$ by (21). Fix $t \ge t_1$ and s, and let $q := \max_{a \in \mathcal{A}_s^\star} Q^t(s,a)$. Then $Z_t(s) \ge (1 - \pi^t(\mathcal{B}_s \mid s))\, e^{-\eta_t q}$, and $Q^t(s,b) \ge q + g_{\min}/2$ for $b \in \mathcal{B}_s$ by (22), so $\pi^{t+1}(b \mid s) \le \pi^t(b \mid s)\, e^{-\eta_t g_{\min}/2}/(1 - \pi^t(\mathcal{B}_s \mid s))$. Summing over $b \in \mathcal{B}_s$ and maximizing over s,

$$m_{t+1} \;\le\; \frac{e^{-\eta_t g_{\min}/2}}{1 - m_t}\, m_t \;\le\; 2e^{-\eta_t g_{\min}/2}\, m_t, \tag{23}$$

and (21) gives $\Delta_{t+1} \le g_{\max} m_{t+1} \le 2 g_{\max} e^{-\eta_t g_{\min}/2} m_t \le \frac{2 g_{\max}}{\underline{d}\, g_{\min}} e^{-\eta_t g_{\min}/2} \Delta_t$. For the cumulative bound, $m_t \to 0$ makes $1/(1 - m_t) \le e^{\eta_0 g_{\min}/4}$ eventually, so (23) and $\eta_t \ge \eta_0$ make m_t decay geometrically. Hence $\sum_t m_t < \infty$ and $\prod_{t \ge t_1}(1 - m_t)^{-1} < \infty$, and iterating (23) from t_1 and applying (21) gives the bound for $t \ge t_1$; enlarging C covers $t < t_1$.

Step 4: Finite termination under the squared Euclidean distance. With $D(p,p') = \frac{1}{2}\|p - p'\|_2^2$ (other scalings rescale η_t), π_s^{t+1} is the Euclidean projection of $y := \pi_s^t - \eta_t Q_s^t$ onto $\mathcal{P}(\mathcal{A})$, so $\pi^{t+1}(a \mid s) = (y_a - \lambda)_+$ with λ chosen so that these sum to 1. Since $\sum_{a \in \mathcal{A}_s^\star}(y_a - \lambda) \le 1$ and $\sum_{a \in \mathcal{A}_s^\star} \pi^t(a \mid s) = 1 - \pi^t(\mathcal{B}_s \mid s)$, we have $\lambda \ge -\pi^t(\mathcal{B}_s \mid s)/|\mathcal{A}_s^\star| - \eta_t \bar{q}$, where $\bar{q}$ is the average of $Q^t(s,\cdot)$ over $\mathcal{A}_s^\star$. For $t \ge t_1$ and $b \in \mathcal{B}_s$, (22) and $\eta_t \ge \eta_0$ then give

$$y_b - \lambda \;\le\; \pi^t(b \mid s) + \pi^t(\mathcal{B}_s \mid s) - \eta_t\big(Q^t(s,b) - \bar{q}\big) \;\le\; 2m_t - \eta_0 g_{\min}/2.$$

Take $t \ge t_1$ with $m_t < \eta_0 g_{\min}/4$. Then $\pi^{t+1}(b \mid s) = 0$ for all s and $b \in \mathcal{B}_s$, so π^{t_0} with $t_0 := t + 1$ is optimal by Lemma B.2. An optimal π is a fixed point of the update: by Lemma B.2, $Q_s^\pi = h^\pi(s)\mathbf{1} + g_s^\star$, so for $y := \pi_s - \eta_t Q_s^\pi$ the threshold $\lambda = -\eta_t h^\pi(s)$ gives $(y_a - \lambda)_+ = (\pi(a \mid s) - \eta_t g^\star(s,a))_+$, which equals $\pi(a \mid s)$ on $\mathcal{A}_s^\star$ and 0 on $\mathcal{B}_s$ and thus sums to 1. Hence $\pi^t = \pi^{t_0}$ for all $t \ge t_0$. □

With the doubling schedule of Section 3.2, $\sum_{j<t} \eta_j = \eta_0(2^t - 1)$, so the bound of part (i) decays doubly exponentially in t. Proposition B.3 is the average-reward counterpart of Lin and Zhang [26, Theorems 4.3 and 4.5], who prove both parts for discounted MDPs, whereas the earlier superlinear rates of Khodadadian et al. [18] and Xiao [45, Section 4.3] assume additional conditions on the iterates' transition matrices. Its proof uses the same two mechanisms: the contraction of the suboptimal action mass, by which Khodadadian et al. [18] obtain an asymptotic linear rate for NPG with constant step sizes, and the simplex projection argument behind the finite termination of projected Q-descent [27]. In average reward, the new ingredients are Step 1, which gives $\Delta_t \to 0$ for nondecreasing step sizes, and Lemma B.2, which defines the optimal advantage through the Poisson equation.

B.9 Tightness of the non-asymptotic rate

Under its growth condition on the step sizes, Lemma 3.7(i) gives the rate $1-1/\bar{\chi}$ whenever $\sup_t \chi_t \le \bar{\chi}$. The following proposition matches this rate up to a constant factor over any fixed number of iterations, for any divergence and any positive step sizes.

Proposition B.4 (Lower bound for exact PMD)**.** *For every $\bar{\chi} > 1$ and every integer $T \ge 1$, there are an MDP with costs in $[0,1]$ satisfying Assumption 3.1 and a policy π^0 with $\pi^0_s \in \mathrm{rint}(\mathcal{P}(\mathcal{A}))$ for all s, such that $\Delta_0 > 0$, $\chi_0 \ge \bar{\chi}/2$, $d^\star(s) \le \bar{\chi}\, d^\pi(s)$ for every policy π and every state s (so $\chi_t \le \bar{\chi}$ for all t), and the iterates of (2) with $\tilde{Q}^t = Q^t$, any Bregman divergence D and any step sizes $\eta_t > 0$ satisfy*

$$\Delta_t \;\ge\; \frac{1}{2}\Big(1 - \frac{1}{\bar{\chi}}\Big)^t \Delta_0 \qquad \textit{for all } 0 \le t \le T.$$

Proof. The instance adds a uniform reset to the hard chain of Johnson et al. [16, Appendix E]. The reset makes every policy ergodic and represents the average cost as a scaled discounted cost with discount factor $\gamma := 1 - 1/\bar{\chi}$, so exact PMD runs as on their discounted chain, where the optimal action can become greedy in at most one new state per iteration.

Step 1: The instance. Let $\gamma := 1 - 1/\bar{\chi} \in (0,1)$, $n := T + \lceil \log 4/\log(1/\gamma) \rceil$ and $\delta := \min\{(1-\gamma)\gamma^n/2,\ \gamma^{n+1}/(8n)\}$. The states are $s_0, s_1, \dots, s_n, s'_1, \dots, s'_n$, so $|\mathcal{S}| = 2n+1$, and the actions are a_1, a_2. Let P^{ch} be the following deterministic chain: from s_i with $1 \le i \le n$, action a_1 moves to s_{i-1} at cost 1 and action a_2 moves to s'_i at cost $1 - r_i$, where $r_i := \gamma^{i+1} + \delta$; the state s_0 is absorbing at cost 0, and s'_i is absorbing at cost $1 - r_i$. The MDP has the transition kernel $P(\cdot \mid s,a) := \gamma P^{\mathrm{ch}}(\cdot \mid s,a) + (1-\gamma)u$, where u is the uniform distribution on $\mathcal{S}$. Then $P_\pi \ge (1-\gamma)/|\mathcal{S}|$ entrywise for every policy π, so Assumption 3.1 holds. The initial policy π^0 plays a_1 with probability $(1-\gamma)\delta$ in every state.

Step 2: Reduction to discounted costs. For a policy π, let $W^\pi := (I - \gamma P^{\mathrm{ch}}_\pi)^{-1} c_\pi$ be its discounted cost on the chain and $Q^\pi_\gamma(s,a) := c(s,a) + \gamma P^{\mathrm{ch}}(\cdot \mid s,a)\, W^\pi$ its action values. Since $P_\pi = \gamma P^{\mathrm{ch}}_\pi + (1-\gamma)\mathbf{1}u$, we have $c_\pi + P_\pi W^\pi = W^\pi + (1-\gamma)(uW^\pi)\mathbf{1}$, so W^π solves the Poisson equation (16) with

$$\rho(\pi) \;=\; (1-\gamma)\, uW^\pi. \tag{24}$$

By Fact B.1, $h^\pi - W^\pi$ is a constant, so $Q^\pi - Q^\pi_\gamma$ is the same constant at every (s,a). Adding a constant to $\tilde{Q}^t_s$ leaves the argmin set in (2) unchanged, so exact PMD has the same possible iterates as with $Q^t_\gamma := Q^{\pi^t}_\gamma$ in place of Q^t, which is the discounted exact PMD of Johnson et al. [16].

Let $\pi^\star$ play a_1 at every s_i, and let $W^\star := W^{\pi^\star}$. Then $W^\star(s_0) = 0$, $W^\star(s_i) = (1-\gamma^i)/(1-\gamma)$ and $W^\star(s'_i) = (1-r_i)/(1-\gamma)$, so $Q^{\pi^\star}_\gamma(s_i, a_1) = W^\star(s_i)$ and

$$Q^{\pi^\star}_\gamma(s_i, a_2) - W^\star(s_i) \;=\; \gamma^i - \frac{\delta}{1-\gamma} \;=:\; g_i \;>\; 0,$$

because $\delta < (1-\gamma)\gamma^n$. Hence $W^\star$ satisfies the Bellman optimality equation of the chain, so $W^\star \le W^\pi$ for every policy π, and $\pi^\star$ is optimal by (24). The optimal advantage of Lemma B.2 is $g^\star(s,a) = Q^{\pi^\star}_\gamma(s,a) - W^\star(s)$, which equals g_i at (s_i, a_2) and vanishes elsewhere. Since $W^\pi = W^\star$ at the states that are absorbing under P^{ch}, every policy π satisfies, for $1 \le i \le n$,

$$W^\pi(s_i) - W^\star(s_i) \;=\; \pi(a_1 \mid s_i)\,\gamma\big(W^\pi(s_{i-1}) - W^\star(s_{i-1})\big) + \pi(a_2 \mid s_i)\, g_i, \tag{25}$$

and (24) gives $\Delta_t = \frac{1-\gamma}{|\mathcal{S}|}\sum_{i=1}^n \big(W^t(s_i) - W^\star(s_i)\big)$ with $W^t := W^{\pi^t}$. In particular, $\Delta_0 > 0$ by (25), since $g_i > 0$ and $\pi^0(a_2 \mid s_i) > 0$.

Step 3: The optimal action propagates at most one state per iteration. We show by induction that $\pi^t(a_1 \mid s_i) \le (1-\gamma)\delta$ for all $i > t$. This holds at $t = 0$. Suppose it holds at t, and let $i > t+1$. By (25) at s_{i-1}, the induction hypothesis and $W^t \ge W^\star$, we have $W^t(s_{i-1}) - W^\star(s_{i-1}) \ge (1 - (1-\gamma)\delta)\, g_{i-1}$. Since $Q^t_\gamma(s_i, a_1) = 1 + \gamma W^t(s_{i-1})$ and $Q^t_\gamma(s_i, a_2) = W^\star(s_i) + g_i$, and since $\gamma g_{i-1} - g_i = \delta$,

$$Q^t_\gamma(s_i, a_1) - Q^t_\gamma(s_i, a_2) \;=\; \gamma\big(W^t(s_{i-1}) - W^\star(s_{i-1})\big) - g_i \;\ge\; \delta - (1-\gamma)\delta\,\gamma\, g_{i-1} \;>\; 0.$$

The last inequality holds because $(1-\gamma)\gamma\, g_{i-1} < 1$. With two actions, the sign property (3) at s_i reads $\big(\pi^{t+1}(a_1 \mid s_i) - \pi^t(a_1 \mid s_i)\big)\big(Q^t_\gamma(s_i,a_1) - Q^t_\gamma(s_i,a_2)\big) \le 0$, so $\pi^{t+1}(a_1 \mid s_i) \le \pi^t(a_1 \mid s_i) \le (1-\gamma)\delta$.

Step 4: Value gaps. For $i > t$, (25), $W^t \ge W^\star$ and Step 3 give $W^t(s_i) - W^\star(s_i) \ge (1-(1-\gamma)\delta)\, g_i \ge \gamma^i - 2\delta/(1-\gamma)$. At $t = 0$, (25), $g_i \le \gamma^i$ and $W^0(s_0) = W^\star(s_0)$ give $W^0(s_i) - W^\star(s_i) \le \gamma^i$ by induction on i.

Step 5: The lower bound. Let $0 \le t \le T$. By Step 2, dropping the states s_i with $i \le t$, and by Step 4,

$$\Delta_t \;\ge\; \frac{1-\gamma}{|\mathcal{S}|}\sum_{i=t+1}^{n}\Big(\gamma^i - \frac{2\delta}{1-\gamma}\Big) \;\ge\; \frac{\gamma^{t+1} - \gamma^{n+1} - 2n\delta}{|\mathcal{S}|}, \qquad \Delta_0 \;\le\; \frac{1-\gamma}{|\mathcal{S}|}\sum_{i=1}^{n}\gamma^i \;\le\; \frac{\gamma}{|\mathcal{S}|}.$$

By the choice of n, $\gamma^{n+1} \le \gamma^{t+1}/4$, and $2n\delta \le \gamma^{n+1}/4$, so $\Delta_t \ge \gamma^{t+1}/(2|\mathcal{S}|) \ge \frac{1}{2}\gamma^t\Delta_0$.

Step 6: Mismatch. For every policy π, stationarity and the reset give $d^\pi(s) \ge (1-\gamma)/|\mathcal{S}|$ for all s, and $d^\pi(s) \ge \gamma d^\pi(s) + (1-\gamma)/|\mathcal{S}|$, hence $d^\pi(s) \ge 1/|\mathcal{S}|$, at s_0 and s'_i, which are absorbing under P^{ch}. Under $\pi^\star$, the states s'_i receive mass from other states only through the reset, so $d^\star(s'_i) = 1/|\mathcal{S}|$; the chain states satisfy $d^\star(s_i) = \gamma d^\star(s_{i+1}) + (1-\gamma)/|\mathcal{S}|$ with $d^\star(s_{n+1}) := 0$, so $d^\star(s_i) = (1-\gamma^{n-i+1})/|\mathcal{S}|$; and the remaining mass is $d^\star(s_0) = (1-\gamma^{n+1})/((1-\gamma)|\mathcal{S}|)$. Hence $d^\star(s) \le d^\pi(s)/(1-\gamma) = \bar\chi\, d^\pi(s)$ for every s. Under π^0, stationarity at s_0 gives $(1-\gamma)\, d^0(s_0) = \gamma(1-\gamma)\delta\, d^0(s_1) + (1-\gamma)/|\mathcal{S}| \le \gamma(1-\gamma)\delta + (1-\gamma)/|\mathcal{S}|$, that is, $d^0(s_0) \le \delta + 1/|\mathcal{S}|$. As $\gamma^{n+1} \le 1/4$ and $|\mathcal{S}|\,\delta \le 3\gamma^{n+1}/8 \le 3/32$,

$$\chi_0 \;\ge\; \frac{d^\star(s_0)}{d^0(s_0)} \;\ge\; \frac{1-\gamma^{n+1}}{(1-\gamma)\,(1+|\mathcal{S}|\,\delta)} \;\ge\; \frac{3/4}{1+3/32}\,\bar\chi \;\ge\; \frac{\bar\chi}{2}. \qquad \square$$

Proposition B.4 adapts Johnson et al. [16, Theorem 4.2] to average reward, using their discounted chain and the propagation argument of their Appendix E in our cost convention. The reset makes the instance ergodic and gives (24); Step 5 aggregates the statewise gaps, and Step 6 calibrates $\gamma = 1 - 1/\bar\chi$ through the mismatch bound. Since the instance can be built for every T, no bound $\Delta_t \le Cr^t\Delta_0$ with $r < 1 - 1/\bar\chi$ and one constant C for all MDPs with $\sup_t \chi_t \le \bar\chi$ holds for all $t \ge 0$, whatever the divergence and the step sizes. The comparison with the prefactor of Lemma 3.7(i), the initial Lyapunov value $\Psi_0 = \Delta_0 + D^\star_0/(\eta_{-1}\bar\chi)$, is direct: if $D^\star_0 < \infty$, the step sizes $\eta_t := \eta_{-1}\gamma^{-(t+1)}$ with $\eta_{-1} \ge D^\star_0/(\bar\chi\Delta_0)$ satisfy its growth condition and give $\Psi_0 \le 2\Delta_0$, so $\frac{1}{4}\gamma^t\Psi_0 \le \Delta_t \le \gamma^t\Psi_0$ for $0 \le t \le T$. As in their result, the instance must depend on T: by Proposition B.3(ii), exact PMD with the squared Euclidean distance and nondecreasing step sizes reaches an optimal policy in finitely many iterations.

C Proofs in the tabular setting

C.1 Proof of Theorem 4.2 (Linear convergence of tabular inexact PMD)

Proof. Here, $\tilde Q^t = \widehat Q^t$. Under Assumption 4.1, $\|\widehat Q^t - Q^t\|_\infty \le \varepsilon_Q$; combined with $\|\pi_s - \pi'_s\|_1 \le 2$ for any two policies, $|\langle \widehat Q^t_s - Q^t_s, \pi_s - \pi'_s\rangle| \le 2\varepsilon_Q$. Taking expectations preserves this bound, so $A_t, B_t \le 2\varepsilon_Q$ for all t.

Applying Lemma 3.7(i) with $\bar A = \bar B = 2\varepsilon_Q$ yields directly the non-asymptotic bound with floor $\bar\chi\bar A + \bar B = 2(\bar\chi+1)\varepsilon_Q$; part (ii) yields, for every $\zeta > 0$, $\limsup_t(\Delta_t - ((\chi+\zeta)\bar A + \bar B))_+^{1/t} \le 1 - 1/(\chi+\zeta)$ with floor $(\chi+\zeta)\bar A + \bar B = 2(\chi+\zeta+1)\varepsilon_Q$. $\square$

C.2 Proof of Lemma 4.7 (Tabular critic upper bound)

Lemma C.1 (Tabular critic upper bound, detailed form)**.** *Fix a policy π, a target critic accuracy $\varepsilon_Q \in (0,1)$, and a confidence $\delta \in (0,1)$, and set*

$$H := \left\lceil 4\, t_{\mathrm{mix}} \log \frac{72\, t_{\mathrm{mix}}}{\varepsilon_Q} \right\rceil, \qquad \tau := \lceil 2H/3 \rceil,$$
$$B := \left\lceil \frac{3600\, H\, t_{\mathrm{mix}}}{\varepsilon_Q^2} \log \frac{8\,|\mathcal{S}|}{\delta} \right\rceil, \qquad n := \left\lceil \frac{72\, t_{\mathrm{mix}}^2}{\varepsilon_Q^2} \log \frac{4\,|\mathcal{S}||\mathcal{A}|}{\delta} \right\rceil.$$

Let $\widehat{Q}^\pi$ be computed as in Section 4.2 with these parameters: for each state s, $\hat{h}^\pi(s)$ and $\hat{\rho}_s(\pi)$ are given by (9) and (10) from B independent rollouts that start at s, and for each pair (s,a),

$$\widehat{Q}^\pi(s,a) \;:=\; c(s,a) - \hat{\rho}_s(\pi) + \frac{1}{n}\sum_{j=1}^{n} \mathrm{clip}\big(\hat{h}^\pi(s'_j)\big),$$
$$\mathrm{clip}(x) \;:=\; \max\big\{-2t_{\mathrm{mix}},\, \min\{2t_{\mathrm{mix}},\, x\}\big\},$$

where $s'_1, \dots, s'_n$ are drawn independently from $P(\cdot \mid s,a)$, independently of the rollouts. Then $\|\widehat{Q}^\pi - Q^\pi\|_\infty \le \varepsilon_Q$ with probability at least $1-\delta$, and the critic uses

$$N \;=\; |\mathcal{S}|\, B\, H + |\mathcal{S}||\mathcal{A}|\, n$$

generative-model transitions, counting H transitions per rollout.

The proof bounds the error of each term of the identity $Q^\pi(s,a) = c(s,a) - \rho(\pi) + P(\cdot \mid s,a)\, h^\pi$ (Fact B.1): the rollouts make $\hat{h}^\pi$ and every $\hat{\rho}_s(\pi)$ accurate to $\varepsilon_Q/3$, and the one-step averages add at most $\varepsilon_Q/3$. The rollout bound is the following lemma, which also supplies the labels of Section 5.2:

Lemma C.2 (Truncated centered-return estimator)**.** *Fix a policy π, a state–action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, a target critic accuracy $\varepsilon_Q \in (0,1)$, and a confidence $\delta \in (0,1)$. For $H \ge 4$, set $\tau := \lceil 2H/3 \rceil$ and $H_\tau := H - \tau$. Generate B independent length-H trajectories $\{x_k^{(i)}\}_{k=0}^{H-1}$ by setting $x_0^{(i)} = (s,a)$ and following π, and define*

$$\hat{\rho}_{\tau,H} \;:=\; \frac{1}{B\, H_\tau} \sum_{i=1}^{B} \sum_{k=\tau}^{H-1} c(x_k^{(i)}), \qquad \widehat{Q}_{H,\tau}(s,a) \;:=\; \frac{1}{B} \sum_{i=1}^{B} \sum_{k=0}^{H-1} \big(c(x_k^{(i)}) - \hat{\rho}_{\tau,H}\big).$$

There exist universal constants $C_1, C_2, C_3 > 0$ such that whenever

$$H \;\ge\; C_2\, t_{\mathrm{mix}}\, \log(C_3\, t_{\mathrm{mix}}/\varepsilon_Q) \qquad \text{and} \qquad B \;\ge\; C_1\, \frac{H\, t_{\mathrm{mix}}}{\varepsilon_Q^2}\, \log \frac{2}{\delta},$$

we have $|\widehat{Q}_{H,\tau}(s,a) - Q^\pi(s,a)| \le \varepsilon_Q$ with probability at least $1-\delta$. The same holds for $\hat{h}^\pi(s)$ and $h^\pi(s)$ in place of $\widehat{Q}_{H,\tau}(s,a)$ and $Q^\pi(s,a)$ when each trajectory instead starts at s with an independent first action $a_0^{(i)} \sim \pi_s$, so that the same two formulas give (10) and (9).

The proof of Lemma C.2 rests on a sharp variance bound for partial sums of a uniformly mixing Markov chain applied to the independent trajectories, followed by a Bernstein concentration argument. The key input is the following covariance summability bound:

Lemma C.3 (Covariance summability under uniform mixing)**.** *Fix a policy π and a function $g : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ with $\mathbb{E}_{\nu^\pi}[g] = 0$, where $\nu^\pi(s,a) := d^\pi(s)\pi(a \mid s)$. Let $M := \|g\|_\infty$ and $\mathrm{osc}(g) := \sup g - \inf g$. Then for any initial joint distribution $\xi_0 \in \mathcal{P}(\mathcal{S}\times\mathcal{A})$ and every $H \ge 1$,*

$$\mathbb{E}_{\xi_0}\left[\left(\sum_{k=0}^{H-1} g(x_k)\right)^2\right] \;\le\; M^2\, H \;+\; 4\, M\, \mathrm{osc}(g)\, H\, t_{\mathrm{mix}}.$$

Proof. Write $S_H := \sum_{k=0}^{H-1} g(x_k)$. Expanding the square,

$$\mathbb{E}_{\xi_0}[S_H^2] \;=\; \sum_{k=0}^{H-1} \mathbb{E}_{\xi_0}\big[g(x_k)^2\big] + 2 \sum_{0\le i<j\le H-1} \mathbb{E}_{\xi_0}[g(x_i)\, g(x_j)].$$

Since $|g| \le M$, the diagonal satisfies $\sum_{k=0}^{H-1} \mathbb{E}_{\xi_0}[g(x_k)^2] \le M^2 H$.

For the cross terms, fix $i < j$ and let $\ell := j - i \ge 1$. We bound $\mathbb{E}[g(x_\ell) \mid x_0 = (s,a)]$ uniformly over (s,a). Conditional on $x_0 = (s,a)$, the law of s_ℓ is $P(\cdot \mid s,a)\, P_\pi^{\ell-1}$ (the first transition uses the fixed action a, after which π acts). By the submultiplicative decay (11),

$$\big\| P(\cdot \mid s,a)\, P_\pi^{\ell-1} - d^\pi \big\|_{\mathrm{TV}} \;\le\; 2^{-\lfloor (\ell-1)/t_{\mathrm{mix}} \rfloor}.$$

Since $a_\ell \mid s_\ell \sim \pi(\cdot \mid s_\ell)$, the law of $x_\ell = (s_\ell, a_\ell)$ satisfies the same total variation (TV) bound relative to ν^π. Using $\mathbb{E}_{\nu^\pi}[g] = 0$ and the standard total-variation duality $|\mathbb{E}_{\nu'} f - \mathbb{E}_{\nu^\pi} f| \le (\sup f - \inf f)\, \|\nu' - \nu^\pi\|_{\mathrm{TV}}$ with $\sup g - \inf g = \mathrm{osc}(g)$,

$$\big|\mathbb{E}[g(x_\ell) \mid x_0 = (s,a)]\big| \;\le\; \mathrm{osc}(g)\, 2^{-\lfloor (\ell-1)/t_{\mathrm{mix}} \rfloor} \qquad \forall\, (s,a),\ \ell \ge 1.$$

By the Markov property (tower rule over x_i), for $i < j$ and $\ell = j - i$,

$$\big|\mathbb{E}_{\xi_0}[g(x_i)\, g(x_j)]\big| \;=\; \big|\mathbb{E}_{\xi_0}\big[g(x_i)\, \mathbb{E}[g(x_j) \mid x_i]\big]\big| \;\le\; M\, \mathrm{osc}(g)\, 2^{-\lfloor (j-i-1)/t_{\mathrm{mix}} \rfloor}.$$

Summing over cross terms and using the block decomposition $\sum_{\ell=0}^{\infty} 2^{-\lfloor \ell/t_{\mathrm{mix}} \rfloor} = t_{\mathrm{mix}} \sum_{q \ge 0} 2^{-q} = 2\, t_{\mathrm{mix}}$,

$$\left| 2 \sum_{0 \le i < j \le H-1} \mathbb{E}_{\xi_0}[g(x_i)\, g(x_j)] \right| \;\le\; 2M\, \mathrm{osc}(g) \sum_{\ell=1}^{H-1} (H - \ell)\, 2^{-\lfloor (\ell-1)/t_{\mathrm{mix}} \rfloor} \;\le\; 4M\, \mathrm{osc}(g)\, H\, t_{\mathrm{mix}}.$$

Combining,

$$\mathbb{E}_{\xi_0}[S_H^2] \;\le\; M^2\, H \;+\; 4\, M\, \mathrm{osc}(g)\, H\, t_{\mathrm{mix}}. \qquad \square$$

Proof of Lemma C.2. Throughout, write $\rho := \rho(\pi)$, $Q := Q^\pi$, and $g(x) := c(x) - \rho$. Since $c(x) \in [0,1]$ and $\rho \in [0,1]$, we have $M = \|g\|_\infty \le 1$ and $\mathrm{osc}(g) = \mathrm{osc}(c) \le 1$; in particular $\mathbb{E}_{\nu^\pi}[g] = 0$, so Lemma C.3 applies and its conclusion $M^2 J + 4M\, \mathrm{osc}(g)\, J\, t_{\mathrm{mix}} \le J\,(1 + 4\, t_{\mathrm{mix}})$ yields, for any initial distribution ξ_0 and any $J \ge 1$,

$$\mathbb{E}_{\xi_0}\left[\left(\sum_{k=0}^{J-1} g(x_k) \right)^2 \right] \;\le\; J\,(1 + 4\, t_{\mathrm{mix}}). \tag{26}$$

Trajectory-level decomposition. Define, per trajectory i,

$$U_i \;:=\; \sum_{k=0}^{\tau-1} g\big(x_k^{(i)}\big), \qquad T_i \;:=\; \sum_{k=\tau}^{H-1} g\big(x_k^{(i)}\big),$$

and $Y_i := U_i - (\tau/H_\tau)\, T_i$. A short calculation gives

$$\widehat{Q}_{H,\tau}(s,a) \;=\; \frac{1}{B} \sum_{i=1}^{B} \sum_{k=0}^{H-1} \big(c(x_k^{(i)}) - \rho\big) \;-\; H\big(\hat{\rho}_{\tau,H} - \rho\big) \;=\; \frac{1}{B} \sum_{i=1}^{B} Y_i,$$

since $H\,(\hat{\rho}_{\tau,H} - \rho) = (H/H_\tau) \sum_i T_i / B$ and $A_i := U_i + T_i = \sum_{k=0}^{H-1} g(x_k^{(i)})$ implies $A_i - (H/H_\tau)\, T_i = U_i - (\tau/H_\tau)\, T_i = Y_i$. The B trajectories are independent and identically distributed, so $\{Y_i\}$ are i.i.d.

Variance of Y_i. By (26) with $\xi_0 = \delta_{(s,a)}$ and horizon τ,

$$\mathbb{E}[U_i^2] \;\le\; \tau\,(1 + 4\, t_{\mathrm{mix}}).$$

Conditional on $x_\tau^{(i)} = (s', a')$, the chain $\{x_{\tau+j}^{(i)}\}_{j \ge 0}$ is a state–action chain under π with initial distribution $\delta_{(s',a')}$. By Lemma C.3 (which handles any initial joint distribution) with horizon H_τ,

$$\mathbb{E}[T_i^2 \mid x_\tau^{(i)}] \;\le\; H_\tau\,(1 + 4\, t_{\mathrm{mix}}) \quad \text{a.s.}, \qquad \text{hence} \qquad \mathbb{E}[T_i^2] \;\le\; H_\tau\,(1 + 4\, t_{\mathrm{mix}}).$$

By $(X - Y)^2 \le 2X^2 + 2Y^2$,

$$\mathrm{Var}(Y_i) \;\le\; \mathbb{E}[Y_i^2] \;\le\; 2\,\mathbb{E}[U_i^2] + 2\,\frac{\tau^2}{H_\tau^2}\,\mathbb{E}[T_i^2] \;\le\; 2\,(1 + 4\, t_{\mathrm{mix}})\left(\tau + \frac{\tau^2}{H_\tau}\right).$$

For $\tau = \lceil 2H/3 \rceil$ and $H \ge 4$, a direct check gives $\tau + \tau^2/H_\tau \le 4H$ (e.g., $H = 4$: $\tau = 3, H_\tau = 1, \tau + \tau^2/H_\tau = 12 = 3H$; $H = 5$: $\tau = 4, H_\tau = 1, \tau + \tau^2/H_\tau = 20 = 4H$; asymptotically $\tau \approx 2H/3, \tau^2/H_\tau \approx 4H/3$, total $\approx 2H$). Hence, using $1 + 4\, t_{\mathrm{mix}} \le 5\, t_{\mathrm{mix}}$ for $t_{\mathrm{mix}} \ge 1$,

$$\mathrm{Var}(Y_i) \;\le\; 8\, H\,(1 + 4\, t_{\mathrm{mix}}) \;\le\; 40\, H\, t_{\mathrm{mix}}.$$

Range of Y_i. Since $|g| \le 1$: $|U_i| \le \tau$, $|T_i| \le H_\tau$, and $|(\tau/H_\tau)\,T_i| \le \tau$, so $|Y_i| \le 2\tau \le 2H$. Hence the centered range is $|Y_i - \mathbb{E}[Y_i]| \le 4H$.

Bernstein concentration. By i.i.d. Bernstein's inequality, with $\sigma^2 := \mathrm{Var}(Y_i) \le 40\,H\,t_{\mathrm{mix}}$ and centered range $4H$, for any $u > 0$,

$$\Pr\left(\left|\frac{1}{B}\sum_{i=1}^{B}\big(Y_i - \mathbb{E}[Y_i]\big)\right| \ge u\right) \;\le\; 2\exp\left(-\frac{B\,u^2}{2\,\sigma^2 + \frac{8}{3}\,H\,u}\right).$$

Setting $u = \varepsilon_Q/2$, the right-hand side is $\le \delta$ when

$$B \;\ge\; \frac{4\,\log(2/\delta)}{\varepsilon_Q^2}\,\left(80\,H\,t_{\mathrm{mix}} + \tfrac{4\,H\,\varepsilon_Q}{3}\right).$$

For $\varepsilon_Q \le 1$ and $t_{\mathrm{mix}} \ge 1$, the variance term dominates ($80\,H\,t_{\mathrm{mix}} + \frac{4}{3}H\,\varepsilon_Q \le \frac{244}{3}\,H\,t_{\mathrm{mix}}$), so an absolute constant $C_1 > 0$ (e.g., $C_1 = 400$) suffices:

$$B \;\ge\; C_1\,\frac{H\,t_{\mathrm{mix}}}{\varepsilon_Q^2}\,\log\frac{2}{\delta} \qquad\Longrightarrow\qquad \Pr\Big(\big|\widehat{Q}_{H,\tau}(s,a) - \mathbb{E}[\widehat{Q}_{H,\tau}(s,a)]\big| \le \varepsilon_Q/2\Big) \ge 1-\delta. \tag{27}$$

Bias bound. By Lemma 4.5, starting from $x_0 = (s,a)$, the state marginal at time $k \ge 1$ is $P(\cdot \mid s,a)\,P_\pi^{k-1}$, and since $a_k \mid s_k \sim \pi$, the total-variation duality together with (11) gives

$$\big|\mathbb{E}[c(x_k) - \rho \mid x_0 = (s,a)]\big| \;\le\; 2^{-\lfloor (k-1)/t_{\mathrm{mix}}\rfloor}, \qquad k \ge 1, \tag{28}$$

using $\mathrm{osc}(c) \le 1$. In particular, the series $Q(s,a) = \sum_{k=0}^{\infty}\mathbb{E}[g(x_k) \mid x_0 = (s,a)]$ converges absolutely.

Define $q_H := \mathbb{E}[A_1 \mid x_0 = (s,a)] = \sum_{k=0}^{H-1}\mathbb{E}[g(x_k) \mid x_0]$. By (28) and the block-decomposition bound $\sum_{\ell=\ell_0}^{\infty} 2^{-\lfloor \ell/t_{\mathrm{mix}}\rfloor} \le 2\,t_{\mathrm{mix}}\cdot 2^{-\lfloor \ell_0/t_{\mathrm{mix}}\rfloor}$,

$$|q_H - Q(s,a)| \;=\; \left|\sum_{k=H}^{\infty}\mathbb{E}[g(x_k) \mid x_0]\right| \;\le\; \sum_{k=H}^{\infty} 2^{-\lfloor (k-1)/t_{\mathrm{mix}}\rfloor} \;\le\; 2\,t_{\mathrm{mix}}\,2^{-\lfloor (H-1)/t_{\mathrm{mix}}\rfloor}.$$

Similarly, $|\mathbb{E}[T_1]| \le 2\,t_{\mathrm{mix}}\,2^{-\lfloor (\tau-1)/t_{\mathrm{mix}}\rfloor}$. Using $\mathbb{E}[\widehat{Q}_{H,\tau}(s,a)] = \mathbb{E}[Y_1] = \mathbb{E}[A_1] - (H/H_\tau)\,\mathbb{E}[T_1]$,

$$\big|\mathbb{E}[\widehat{Q}_{H,\tau}(s,a)] - Q(s,a)\big| \;\le\; |q_H - Q(s,a)| + \frac{H}{H_\tau}\,|\mathbb{E}[T_1]|.$$

For any τ with $H_\tau = H - \tau \ge H/5$ (e.g., $\tau = \lceil 2H/3\rceil$ or $\tau = \lceil H/2\rceil$) and $H \ge 4$: $H/H_\tau \le 5$. Using $2^{-\lfloor (H-1)/t_{\mathrm{mix}}\rfloor} \le 2^{-\lfloor (\tau-1)/t_{\mathrm{mix}}\rfloor}$ (since $H \ge \tau$),

$$\big|\mathbb{E}[\widehat{Q}_{H,\tau}(s,a)] - Q(s,a)\big| \;\le\; 12\,t_{\mathrm{mix}}\,2^{-\lfloor (\tau-1)/t_{\mathrm{mix}}\rfloor}. \tag{29}$$

Combining. Choose H so that $12\,t_{\mathrm{mix}}\,2^{-\lfloor (\tau-1)/t_{\mathrm{mix}}\rfloor} \le \varepsilon_Q/2$, i.e., $\lfloor (\tau-1)/t_{\mathrm{mix}}\rfloor \ge \log_2(24\,t_{\mathrm{mix}}/\varepsilon_Q)$. With $\tau \ge 2H/3$ and $H \ge 4$, this is implied by

$$H \;\ge\; C_2\,t_{\mathrm{mix}}\,\log(C_3\,t_{\mathrm{mix}}/\varepsilon_Q)$$

for suitable universal constants (e.g., $C_2 = 4, C_3 = 24$, which absorbs the integer-floor correction $\lfloor (\tau-1)/t_{\mathrm{mix}}\rfloor \ge (\tau-1)/t_{\mathrm{mix}} - 1$). On the high-probability event of (27),

$$\begin{aligned}\big|\widehat{Q}_{H,\tau}(s,a) - Q(s,a)\big| \;&\le\; \big|\widehat{Q}_{H,\tau}(s,a) - \mathbb{E}[\widehat{Q}_{H,\tau}(s,a)]\big| + \big|\mathbb{E}[\widehat{Q}_{H,\tau}(s,a)] - Q(s,a)\big| \\ &\le\; \varepsilon_Q/2 + \varepsilon_Q/2 \;=\; \varepsilon_Q,\end{aligned}$$

which holds with probability at least $1-\delta$.

Initial action drawn from π_s. If each trajectory instead starts at s with an independent first action $a_0^{(i)} \sim \pi_s$, the trajectories remain independent and identically distributed, and Lemma C.3 holds for every initial joint distribution, so the variance, range, and Bernstein steps apply verbatim with the initial distribution $\delta_s \otimes \pi_s$ in place of $\delta_{(s,a)}$. The mean of the estimate is the average over $a \sim \pi_s$ of $\mathbb{E}[\widehat{Q}_{H,\tau}(s,a)]$, and $h^\pi(s) = \sum_a \pi(a \mid s)\,Q^\pi(s,a)$ by definition, so averaging (29) over a bounds the bias by the same quantity. The same H and B therefore give $|\hat{h}^\pi(s) - h^\pi(s)| \le \varepsilon_Q$ with probability at least $1-\delta$. □

Proof of Lemma C.1. Write $\rho := \rho(\pi)$ and $h := h^\pi$. We allot failure probability $\delta/4$ to the state values, $\delta/4$ to the average costs, and $\delta/2$ to the one-step averages.

Horizon. With $x := 72\,t_{\mathrm{mix}}/\varepsilon_Q \geq 72$, the choices $\tau \geq 2H/3$ and $H \geq 4\,t_{\mathrm{mix}} \log x$ give

$$\left\lfloor \frac{\tau - 1}{t_{\mathrm{mix}}} \right\rfloor \;\geq\; \frac{2H}{3\,t_{\mathrm{mix}}} - 2 \;\geq\; \frac{8}{3}\log x - 2 \;\geq\; \log_2 x,$$

where the last step holds because $(8/3 - 1/\log 2)\log 72 > 2$. Hence

$$12\,t_{\mathrm{mix}}\,2^{-\lfloor(\tau-1)/t_{\mathrm{mix}}\rfloor} \;\leq\; \frac{12\,t_{\mathrm{mix}}}{x} \;=\; \frac{\varepsilon_Q}{6}, \tag{30}$$

which is the horizon condition in the proof of Lemma C.2 at accuracy $\varepsilon_Q/3$. Also $H \geq 4$.

State values. Fix a state s. Apply Lemma C.2 to the rollouts that start at s, at accuracy $\varepsilon_Q/3$ and confidence $\delta/(4|\mathcal{S}|)$: its horizon condition is (30), and its Bernstein condition (27) with $C_1 = 400$ reads $B \geq 3600\,H\,t_{\mathrm{mix}}\,\varepsilon_Q^{-2}\log(8|\mathcal{S}|/\delta)$. Hence $|\hat{h}^\pi(s) - h(s)| \leq \varepsilon_Q/3$ except on an event of probability at most $\delta/(4|\mathcal{S}|)$. Clipping cannot increase this distance, because $|h(s)| \leq 2\,t_{\mathrm{mix}}$. Indeed, by Fact B.1, $h(s) = \sum_{k\geq 0}(\delta_s P_\pi^k - d^\pi)\,c_\pi$ with $c_\pi(s') := \sum_a \pi(a \mid s')\,c(s',a) \in [0,1]$, and the total-variation duality in the proof of Lemma C.3 together with (11) bounds the k-th term by $2^{-\lfloor k/t_{\mathrm{mix}}\rfloor}$, whose sum over $k \geq 0$ is $2\,t_{\mathrm{mix}}$.

Average costs. Fix a state s. The estimate $\hat{\rho}_s(\pi)$ averages B independent tail averages with values in $[0,1]$. As for $h(s)$, $|\mathbb{E}[c(s_k^{(i)}, a_k^{(i)})] - \rho| \leq 2^{-\lfloor k/t_{\mathrm{mix}}\rfloor}$, so by the block-decomposition bound in the proof of Lemma C.2 and (30), each tail average has mean within $2\,t_{\mathrm{mix}}\,2^{-\lfloor \tau/t_{\mathrm{mix}}\rfloor} \leq \varepsilon_Q/36$ of ρ. Hoeffding's inequality and $B \geq 18\,\varepsilon_Q^{-2}\log(8|\mathcal{S}|/\delta)$ then give $|\hat{\rho}_s(\pi) - \rho| \leq \varepsilon_Q/6 + \varepsilon_Q/36 \leq \varepsilon_Q/3$ except on an event of probability at most $2\exp(-B\varepsilon_Q^2/18) \leq \delta/(4|\mathcal{S}|)$.

One-step averages. Condition on the rollouts. For each pair (s,a), the values $\mathrm{clip}(\hat{h}^\pi(s'_j))$, $j = 1,\ldots,n$, are independent, lie in $[-2t_{\mathrm{mix}}, 2t_{\mathrm{mix}}]$, and have mean $P(\cdot \mid s,a)\,\mathrm{clip}(\hat{h}^\pi)$. Hoeffding's inequality bounds the conditional probability that their average deviates from this mean by more than $\varepsilon_Q/3$ by $2\exp(-n\varepsilon_Q^2/(72\,t_{\mathrm{mix}}^2)) \leq \delta/(2|\mathcal{S}||\mathcal{A}|)$, by the choice of n. Since this bound holds for every outcome of the rollouts, a union bound over the pairs costs at most $\delta/2$.

Combining. Outside the exceptional events, whose probabilities sum to at most $|\mathcal{S}|\cdot\delta/(4|\mathcal{S}|) + |\mathcal{S}|\cdot\delta/(4|\mathcal{S}|) + \delta/2 = \delta$, the identity of Fact B.1 gives, for every pair (s,a),

$$\begin{aligned}\left|\widehat{Q}^\pi(s,a) - Q^\pi(s,a)\right| \;\leq\; |\hat{\rho}_s(\pi) - \rho| &+ \left|\frac{1}{n}\sum_{j=1}^{n}\mathrm{clip}\big(\hat{h}^\pi(s'_j)\big) - P(\cdot \mid s,a)\,\mathrm{clip}(\hat{h}^\pi)\right| \\ &+ \left|P(\cdot \mid s,a)\big(\mathrm{clip}(\hat{h}^\pi) - h\big)\right| \;\leq\; \frac{\varepsilon_Q}{3} + \frac{\varepsilon_Q}{3} + \frac{\varepsilon_Q}{3} \;=\; \varepsilon_Q,\end{aligned}$$

where the last term uses $\|\mathrm{clip}(\hat{h}^\pi) - h\|_\infty \leq \varepsilon_Q/3$ and that $P(\cdot \mid s,a)$ is a probability distribution. The critic draws B rollouts of length H at each of the $|\mathcal{S}|$ states and n next states at each of the $|\mathcal{S}||\mathcal{A}|$ pairs, so $N = |\mathcal{S}|BH + |\mathcal{S}||\mathcal{A}|n$. Since $H = O(t_{\mathrm{mix}}\log(t_{\mathrm{mix}}/\varepsilon_Q))$, $B = O(H\,t_{\mathrm{mix}}\,\varepsilon_Q^{-2}\log(|\mathcal{S}|/\delta))$, and $n = O(t_{\mathrm{mix}}^2\,\varepsilon_Q^{-2}\log(|\mathcal{S}||\mathcal{A}|/\delta))$, this is $N = \widetilde{O}((|\mathcal{S}|\,t_{\mathrm{mix}}^3 + |\mathcal{S}||\mathcal{A}|\,t_{\mathrm{mix}}^2)/\varepsilon_Q^2)$, as stated in Lemma 4.7. □

C.3 Proof of Theorem 4.9 (Tabular sample complexity)

Theorem C.4 (Tabular sample complexity, detailed form)**.** *Suppose $D_0^\star < \infty$, and recall the initial Lyapunov value $\Psi_0 = \Delta_0 + D_0^\star/(\eta_{-1}\bar{\chi})$. Fix a target suboptimality $\varepsilon \in (0,1)$ and a confidence $\delta \in (0,1)$, and set $\varepsilon_Q := \varepsilon/(4(\bar{\chi}+1))$ and $T := \lceil \bar{\chi}\log\max\{e, 2\,\Psi_0/\varepsilon\}\rceil$. Run inexact PMD (Theorem 4.2) for T iterations under the hypotheses of Lemma 3.7(i), computing each $\widehat{Q}^t$ by the critic of Lemma C.1 for π^t at accuracy ε_Q and confidence δ/T, from fresh generative-model transitions. Then $\Delta_T \leq \varepsilon$ with probability at least $1-\delta$, using*

$$N \;=\; T\left(|\mathcal{S}|\,B\,H + |\mathcal{S}||\mathcal{A}|\,n\right)$$

generative-model transitions, where H, B, and n are the parameters of Lemma C.1 at accuracy ε_Q and confidence δ/T.

Proof. The proof has two ingredients: a union bound converting the per-call guarantee of Lemma C.1 into critic accuracy along the PMD trajectory, and a deterministic composition with Theorem 4.2.

Critic accuracy along the trajectory. Let $\mathcal{F}_{t-1}$ denote the σ-algebra generated by the transitions used in iterations $0, \dots, t-1$; the policy π^t is $\mathcal{F}_{t-1}$-measurable, and the critic call uses fresh generative-model randomness, independent of $\mathcal{F}_{t-1}$. Applying Lemma C.1 conditionally on $\mathcal{F}_{t-1}$ gives

$$\Pr\Big(\big\| \widehat{Q}^t - Q^{\pi^t} \big\|_\infty > \varepsilon_Q \;\Big|\; \mathcal{F}_{t-1} \Big) \;\le\; \frac{\delta}{T},$$

and a union bound over $t \in \{0, \dots, T-1\}$ bounds the probability that some iteration fails by δ. On the complementary good event, the bound of Assumption 4.1 holds at level ε_Q for every $t < T$, which is all that the bound of Theorem 4.2 at time T uses.

Composition with Theorem 4.2. Condition on the good event. Since $\sup_t \chi_t \le \bar{\chi}$, Theorem 4.2 gives

$$\big(\Delta_T - 2(\bar{\chi}+1)\varepsilon_Q\big)_+ \;\le\; \Big(1 - \tfrac{1}{\bar{\chi}}\Big)^T \Psi_0.$$

The choice $\varepsilon_Q = \varepsilon/(4(\bar{\chi}+1))$ makes the error floor equal to $\varepsilon/2$. Using $(1-1/\bar{\chi})^T \le \exp(-T/\bar{\chi})$, the choice $T = \lceil \bar{\chi} \log \max\{e, 2\,\Psi_0/\varepsilon\} \rceil$ ensures $(1-1/\bar{\chi})^T \Psi_0 \le \varepsilon/2$. Therefore

$$\Delta_T \;\le\; \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} \;=\; \varepsilon,$$

with probability at least $1-\delta$ over the sampling randomness. □

The total number of generative-model transitions is $N = T\,(|\mathcal{S}|BH + |\mathcal{S}||\mathcal{A}|n)$. At accuracy $\varepsilon_Q = \varepsilon/(4(\bar{\chi}+1))$ and confidence δ/T, the parameters of Lemma C.1 satisfy

$$B\,H \;=\; \widetilde{O}\!\left(\frac{H^2\, t_{\mathrm{mix}}\,(\bar{\chi}+1)^2}{\varepsilon^2}\right), \qquad n \;=\; \widetilde{O}\!\left(\frac{t_{\mathrm{mix}}^2\,(\bar{\chi}+1)^2}{\varepsilon^2}\right), \qquad T = \widetilde{\Theta}(\bar{\chi}),$$

with $H = \Theta(t_{\mathrm{mix}} \log(t_{\mathrm{mix}}/\varepsilon_Q)) = \widetilde{O}(t_{\mathrm{mix}})$, so $H^2\, t_{\mathrm{mix}} = \widetilde{O}(t_{\mathrm{mix}}^3)$. Combining, and using $\bar{\chi} \ge 1$ (so $(\bar{\chi}+1)^2 \le 4\bar{\chi}^2$),

$$N \;=\; \widetilde{O}\!\left(\frac{(|\mathcal{S}|\, t_{\mathrm{mix}}^3 + |\mathcal{S}||\mathcal{A}|\, t_{\mathrm{mix}}^2)\,\bar{\chi}^3}{\varepsilon^2}\right),$$

recovering Theorem 4.9.

C.4 Proof of Proposition 4.8 (Tabular critic lower bound)

We first restate the proposition with the explicit construction used in the proof.

Proposition C.5 (Tabular critic lower bound, detailed form)**.** *For every $|\mathcal{S}| \ge 2$, $p \in (0, 1/4)$, and $\varepsilon_Q \in (0, 1/(128p)]$, there are a finite family of MDPs with state space $\mathcal{S}$ and action space $\mathcal{A}$ and a deterministic policy π such that every member satisfies Assumption 3.1 and the bound of Lemma 4.5 with $t_{\mathrm{mix}} = \lceil \log 4/p \rceil < 2/p$, and the $1/4$-mixing time of π in every member is at least $1/(16p)$. Any (possibly adaptive) generative-model algorithm that outputs $\widehat{Q} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ with $\Pr(\|\widehat{Q} - Q^\pi\|_\infty \le \varepsilon_Q) \ge 2/3$ on every member, where Q^π is the differential action-value function of π in that member, must use*

$$N \;\ge\; \frac{d}{2^{12}\cdot 9\, p^3\, \varepsilon_Q^2} + \frac{d\,(|\mathcal{A}|-1)}{2^8 \cdot 9\, p^2\, \varepsilon_Q^2} \;\ge\; \frac{|\mathcal{S}|\,\big(t_{\mathrm{mix}}^3 + |\mathcal{A}|\, t_{\mathrm{mix}}^2\big)}{2^{13}\cdot 135\, \varepsilon_Q^2}$$

generative-model transitions, where $d := \lfloor |\mathcal{S}|/2 \rfloor$.

Here, an algorithm is adaptive if it chooses each query based on the samples seen so far.

Proof. The family hides one bit in the transitions of the action that π takes at each pair of states and one bit in the transitions of every other action at each paired state. A critic that is accurate to ε_Q in sup norm reveals every bit, the bits change disjoint groups of transition rows, and a Le Cam two-point test for each bit then forces enough queries to each group. This is Assouad's hypercube reduction in coordinatewise form. With one action and $2d$ states, the construction is that of Proposition D.5.

Construction. Write $\mathcal{A} = \{a_1, \dots, a_{|\mathcal{A}|}\}$, and identify $\mathcal{S}$ with $[d] \times \{-1, +1\}$ if $|\mathcal{S}| = 2d$, and with $([d] \times \{-1, +1\}) \cup \{o\}$, for one more state o, if $|\mathcal{S}| = 2d + 1$. Let u be the uniform distribution on $\mathcal{S}$, set $c((j, \sigma), a) := (1 + \sigma)/2$ and $c(o, a) := 1/2$ for every action a, and let π always take a_1. The members are indexed by state bits $\theta \in \{0, 1\}^d$ and action bits $\vartheta_{j,\sigma,a} \in \{-1, +1\}$ for $(j, \sigma) \in [d] \times \{-1, +1\}$ and $a \neq a_1$, of which there are none if $|\mathcal{A}| = 1$. With

$$\Delta \;:=\; 64\, p^2\, \varepsilon_Q, \qquad \Delta' \;:=\; 8\, p\, \varepsilon_Q, \qquad q_j \;:=\; p + \theta_j\, \Delta,$$

the local kernel $R_{\theta,\vartheta}(\cdot \mid (j, \sigma), a)$ moves (j, σ) to $(j, -\sigma)$ with probability q_j if $a = a_1$ and with probability $1/2 + \vartheta_{j,\sigma,a}\Delta'$ if $a \neq a_1$, and otherwise leaves it in place. Set $R_{\theta,\vartheta}(\cdot \mid o, a) := \delta_o$. The transition kernel is

$$P_{\theta,\vartheta}(\cdot \mid s, a) \;:=\; (1 - p)\, R_{\theta,\vartheta}(\cdot \mid s, a) + p\, u(\cdot).$$

The restriction $\varepsilon_Q \leq 1/(128p)$ gives $\Delta \leq p/2$ and $\Delta' \leq 1/16$.

Mixing for every policy. Every row satisfies $P_{\theta,\vartheta}(\cdot \mid s, a) \geq p\, u(\cdot)$, which is the minorization condition (41) with $\alpha = p$ and $\nu = u$. Every policy therefore induces an irreducible and aperiodic chain, so Assumption 3.1 holds, and Appendix E gives $\sup_{\mu_0} \|\mu_0 P_{\pi'}^k - d^{\pi'}\|_{\mathrm{TV}} \leq (1 - p)^k \leq e^{-pk}$ for every policy π'. At $k = \lceil \log 4/p \rceil$ this is at most $1/4$, which gives the bound of Lemma 4.5 with $t_{\mathrm{mix}} = \lceil \log 4/p \rceil \leq \log 4/p + 1 < 2/p$, where the last step uses $1 < 1/(4p)$.

The evaluated policy. Under π, each pair $\{(j, +1), (j, -1)\}$ moves as in the construction of Proposition D.5, and o moves only by the reset. Steps 1–4 of its proof therefore apply: the uniform distribution is stationary, so $d^\pi = u$ and $\rho(\pi) = 1/2$; the eigenfunction argument shows that the $1/4$-mixing time of π is at least $1/(16p)$; the centered Poisson solution is $h^\pi(j, \sigma) = \beta_j\, \sigma$ with $\beta_j := \beta(q_j) \in [1/(8p), 1/(5p)]$, where $\beta(q) := 1/(2p + 4q(1 - p))$; and $\beta(p) - \beta(p + \Delta) \geq \Delta/(16p^2) = 4\, \varepsilon_Q$. The state o does not change these steps: u remains stationary because the local kernel of π is symmetric on each pair and fixes o, the eigenfunctions vanish at o, and $h^\pi(o) := 0$ solves the Poisson equation at o because $c(o, a_1) = \rho(\pi)$ and $u\, h^\pi = 0$.

Decoding. By Fact B.1, $Q^\pi(s, a) = c(s, a) - \rho(\pi) + P_{\theta,\vartheta}(\cdot \mid s, a)\, h^\pi$, in which the reset contributes $p\, u\, h^\pi = 0$. For a row at (j, σ) that flips the sign with probability f, the local part of $P_{\theta,\vartheta}(\cdot \mid s, a)\, h^\pi$ is $(1 - p)(1 - 2f)\, \beta_j \sigma$, so

$$\begin{aligned} Q^\pi((j, \sigma), a_1) &= \beta_j\, \sigma, \qquad Q^\pi(o, a) = 0, \\ \sigma\, Q^\pi((j, \sigma), a) &= \frac{1}{2} - 2(1 - p)\, \beta_j\, \Delta'\, \vartheta_{j,\sigma,a} \qquad (a \neq a_1), \end{aligned}$$

where the first identity is the Poisson equation at (j, σ). Since $(1 - p)\, \beta_j\, \Delta' \geq \frac{3}{4} \cdot \frac{1}{8p} \cdot 8p\, \varepsilon_Q$, an action bit moves $\sigma Q^\pi((j, \sigma), a)$ away from $1/2$ by at least $\frac{3}{2}\varepsilon_Q$, and a state bit moves $Q^\pi((j, +1), a_1)$ between $\beta(p)$ and $\beta(p + \Delta)$, which differ by at least $4\varepsilon_Q$. Hence, whenever $\|\widehat{Q} - Q^\pi\|_\infty \leq \varepsilon_Q$, the decoders

$$\widehat{\theta}_j \;:=\; \mathbf{1}\Big\{\widehat{Q}((j, +1), a_1) < \frac{\beta(p) + \beta(p + \Delta)}{2}\Big\}, \qquad \hat{\vartheta}_{j,\sigma,a} \;:=\; \mathrm{sign}\Big(\frac{1}{2} - \sigma\, \widehat{Q}((j, \sigma), a)\Big)$$

return every bit of the member.

Per-query KL divergence. Changing θ_j changes only the rows $((j, \pm 1), a_1)$, and changing $\vartheta_{j,\sigma,a}$ only the row $((j, \sigma), a)$, so different bits change disjoint groups of rows. For a row of θ_j, Step 5 in the proof of Proposition D.5 bounds the KL divergence between the two members' next-state laws by $2\Delta^2/p$ in either direction, as in (38); its lower bounds on the two changing atoms ignore the reset mass, so they hold here too. For a row of $\vartheta_{j,\sigma,a}$, revealing whether the reset occurred, which has the same law under both members, can only increase the KL divergence, which is then at most

$$\mathrm{KL}\big(\mathrm{Bern}(\tfrac{1}{2} + \Delta') \,\big\|\, \mathrm{Bern}(\tfrac{1}{2} - \Delta')\big) \;=\; 2\Delta' \log \frac{1 + 2\Delta'}{1 - 2\Delta'} \;\leq\; \frac{8\, \Delta'^2}{1 - 2\Delta'} \;\leq\; 16\, \Delta'^2,$$

using $\log((1 + x)/(1 - x)) \leq 2x/(1 - x)$ for $x \in [0, 1)$.

Assembly. Fix a member (θ, ϑ) and one of its bits, and let $\mathbb{P}$ and $\mathbb{P}'$ be the laws of the transcript under this member and under the member in which only that bit is flipped. As in Step 6 of the proof of Proposition D.5, the query rule has the same conditional law under both members, so the KL chain rule gives $\mathrm{KL}(\mathbb{P} \,\|\, \mathbb{P}') \le (2\Delta^2/p)\,\mathbb{E}[N_{\mathrm{bit}}]$ for a state bit and $\mathrm{KL}(\mathbb{P} \,\|\, \mathbb{P}') \le 16\Delta'^2\,\mathbb{E}[N_{\mathrm{bit}}]$ for an action bit, where N_{bit} is the number of queries to the rows that the bit changes and the expectation is under $\mathbb{P}$. The decoder of this bit returns the member's bit with probability at least $2/3$ under $\mathbb{P}$ and at most $1/3$ under $\mathbb{P}'$, so $\|\mathbb{P} - \mathbb{P}'\|_{\mathrm{TV}} \ge 1/3$, and Pinsker's inequality $\|\mathbb{P} - \mathbb{P}'\|_{\mathrm{TV}} \le \sqrt{\mathrm{KL}(\mathbb{P} \,\|\, \mathbb{P}')/2}$ gives $\mathrm{KL}(\mathbb{P} \,\|\, \mathbb{P}') \ge 2/9$. Hence $\mathbb{E}[N_{\mathrm{bit}}] \ge p/(9\Delta^2)$ for a state bit and $\mathbb{E}[N_{\mathrm{bit}}] \ge 1/(72\Delta'^2)$ for an action bit. The groups are disjoint, so summing over the d state bits and the $2d(|\mathcal{A}| - 1)$ action bits of the fixed member gives

$$N \;\ge\; d\,\frac{p}{9\,\Delta^2} + 2d\,(|\mathcal{A}| - 1)\,\frac{1}{72\,\Delta'^2} \;=\; \frac{d}{2^{12} \cdot 9\,p^3\,\varepsilon_Q^2} + \frac{d\,(|\mathcal{A}| - 1)}{2^8 \cdot 9\,p^2\,\varepsilon_Q^2}.$$

Final form. We have $d \ge |\mathcal{S}|/3$ for $|\mathcal{S}| \ge 2$, and the right-hand side equals $d\,(1 + 16(|\mathcal{A}| - 1)p)/(2^{12} \cdot 9\,p^3\varepsilon_Q^2)$ with $1 + 16(|\mathcal{A}| - 1)p \ge \frac{4}{5}(1 + |\mathcal{A}|\,p)$ for $p \le 1/4$, so $N \ge |\mathcal{S}|\,(p^{-3} + |\mathcal{A}|\,p^{-2})/(2^{10} \cdot 135\,\varepsilon_Q^2)$. Since $t_{\mathrm{mix}} < 2/p$, we have $p^{-3} > t_{\mathrm{mix}}^3/8$ and $p^{-2} > t_{\mathrm{mix}}^2/4$, which gives the second inequality of the proposition. □

The assumption $|\mathcal{S}| \ge 2$ is necessary: with a single state, $Q^\pi(s,a) = c(s,a) - \sum_{a'} \pi(a' \mid s)\,c(s,a')$ is determined by the costs, without any transitions.

D Proofs under linear function approximation

D.1 Proof of Theorem 5.6 (Linear convergence of inexact PMD under LFA)

Proof. PMD is the update (2) with the LFA critic $\tilde{Q}^t(s,a) = \phi_{s,a}^\top w_t$. We apply Lemma 3.7 after bounding the critic-error quantities A_t, B_t; all bounds below hold on the probability-one event of Assumption 5.4.

Bounding A_t and B_t via state–action concentrability. Define the scalar error function $e_t(s,a) := w_t^\top \phi_{s,a} - Q^{\pi^t}(s,a)$. Expanding the inner products against $\pi^{t+1} - \pi^t$ and $\pi^t - \pi^\star$ into expectations over state–action pairs,

$$A_t = \Big|\mathbb{E}_{\nu^{t+1}}[e_t] - \mathbb{E}_{d^{t+1} \times \pi^t}[e_t]\Big|, \qquad B_t = \Big|\mathbb{E}_{d^\star \times \pi^t}[e_t] - \mathbb{E}_{\nu^\star}[e_t]\Big|,$$

where $\nu^t := d^t \times \pi^t$ denotes the stationary state–action distribution of π^t.

We use the generic change-of-measure bound: for any two distributions ν, ν' on $\mathcal{S} \times \mathcal{A}$ with $\nu' \ll \nu$ and any measurable f,

$$\big|\mathbb{E}_{(s,a)\sim\nu'}[f(s,a)]\big| \;\le\; \sqrt{\mathbb{E}_{(s,a)\sim\nu}\left[\left(\frac{\nu'(s,a)}{\nu(s,a)}\right)^2\right]}\,\sqrt{\mathbb{E}_{(s,a)\sim\nu}[f(s,a)^2]}. \tag{31}$$

First, apply (31) with $\nu = \nu^t$ and $\nu' = \nu^{t+1}$, and use Assumption 5.4 to get

$$\big|\mathbb{E}_{\nu^{t+1}}[e_t]\big| \;\le\; \sqrt{C_+}\,\sqrt{\mathbb{E}_{\nu^t}[e_t^2]}.$$

Next, note that the mixed distribution $d^{t+1} \times \pi^t(\cdot \mid s)$ satisfies

$$\frac{d^{t+1}(s)\pi^t(a \mid s)}{\nu^t(s,a)} = \frac{d^{t+1}(s)}{d^t(s)} = \sum_{a'} \pi^t(a' \mid s)\,\frac{\nu^{t+1}(s,a')}{\nu^t(s,a')},$$

so by Jensen's inequality,

$$\left(\frac{d^{t+1}(s)}{d^t(s)}\right)^2 \le \sum_{a'} \pi^t(a' \mid s)\left(\frac{\nu^{t+1}(s,a')}{\nu^t(s,a')}\right)^2.$$

Taking expectation over $s \sim d^t$ yields

$$\mathbb{E}_{(s,a)\sim\nu^t}\left[\left(\frac{d^{t+1}(s)\pi^t(a \mid s)}{\nu^t(s,a)}\right)^2\right] = \mathbb{E}_{s\sim d^t}\left[\left(\frac{d^{t+1}(s)}{d^t(s)}\right)^2\right] \le \mathbb{E}_{(s,a)\sim\nu^t}\left[\left(\frac{\nu^{t+1}(s,a)}{\nu^t(s,a)}\right)^2\right] \le C_+.$$

Applying (31) with $\nu = \nu^t$ and $\nu' = d^{t+1} \times \pi^t(\cdot \mid s)$ gives

$$\left|\mathbb{E}_{d^{t+1}\times\pi^t}[e_t]\right| \;\le\; \sqrt{C_+}\,\sqrt{\mathbb{E}_{\nu^t}[e_t^2]}.$$

Therefore,

$$A_t \le 2\sqrt{C_+}\,\sqrt{\mathbb{E}_{\nu^t}[e_t^2]}. \tag{32}$$

An analogous argument using Assumption 5.4 (with $\nu^\star$ in place of ν^{t+1}) yields

$$B_t \le 2\sqrt{C_\star}\,\sqrt{\mathbb{E}_{\nu^t}[e_t^2]}. \tag{33}$$

Finally, decompose

$$e_t(s,a) = \phi_{s,a}^\top(w_t - w_t^\star) + \big(\phi_{s,a}^\top w_t^\star - Q^{\pi^t}(s,a)\big).$$

By the triangle inequality in $L_2(\nu^t)$,

$$\sqrt{\mathbb{E}_{\nu^t}[e_t^2]} \;\le\; \sqrt{\mathbb{E}_{\nu^t}\Big[\big(\phi_{s,a}^\top(w_t - w_t^\star)\big)^2\Big]} + \sqrt{\mathbb{E}_{\nu^t}\Big[\big(\phi_{s,a}^\top w_t^\star - Q^{\pi^t}(s,a)\big)^2\Big]} \tag{34}$$

$$\le \|w_t - w_t^\star\|_{\Sigma_t} + \varepsilon_{\mathrm{app}} \;\le\; \varepsilon_{\mathrm{stat}} + \varepsilon_{\mathrm{app}}, \tag{35}$$

where the last two inequalities follow from Assumptions 5.2 and 5.3, respectively.

Applying the master lemma. Combining (32), (33), and (35) yields, for all $t \ge 0$,

$$A_t \le \bar{A} := 2\sqrt{C_+}\big(\varepsilon_{\mathrm{stat}} + \varepsilon_{\mathrm{app}}\big), \qquad B_t \le \bar{B} := 2\sqrt{C_\star}\big(\varepsilon_{\mathrm{stat}} + \varepsilon_{\mathrm{app}}\big).$$

Lemma 3.7(i) then gives the claimed non-asymptotic bound with the error floor $\bar{\chi}\bar{A} + \bar{B} = \varepsilon_{\mathrm{floor}}(\bar{\chi})$, and Lemma 3.7(ii) gives the claimed asymptotic bound with the floor $(\chi + \zeta)\bar{A} + \bar{B} = \varepsilon_{\mathrm{floor}}(\chi + \zeta)$. □

D.2 Proof of a helper lemma (tail-averaged SGD)

Lemma D.1 (Estimating $w_t^\star$ with an unbiased label oracle)**.** *Fix t and consider the regression target*

$$w_t^\star \in \arg\min_{w\in\mathbb{R}^d} \; \tfrac{1}{2}\,\mathbb{E}_{(s,a)\sim\nu^t}\big[(w^\top\phi_{s,a} - Q^t(s,a))^2\big]$$

with $|Q^t(s,a)| \le G$ and approximation error $\mathbb{E}_{(s,a)\sim\nu^t}[(w_t^{\star\top}\phi_{s,a} - Q^t(s,a))^2] \le \varepsilon_{\mathrm{app}}^2$. Additionally suppose that:

- *(i) (Bounded features) $\|\phi_{s,a}\|_2 \le L$ for all (s,a).*
- *(ii) (Uniform covariance lower bound) For all t, $\Sigma_t := \mathbb{E}_{(s,a)\sim\nu^t}\big[\phi_{s,a}\phi_{s,a}^\top\big] \succeq \mu I_d$ for some $\mu > 0$.*
- *(iii) (Unbiased label oracle) The label oracle returns $\widehat{Q}^t(s,a)$ with*

$$\mathbb{E}\Big[\widehat{Q}^t(s,a) \mid s,a\Big] = Q^t(s,a), \qquad \mathbb{E}\Big[(\widehat{Q}^t(s,a))^2 \mid s,a\Big] \le G^2.$$

At each outer iteration t, compute w_t by running K steps of SGD on the quadratic loss

$$f_t(w) := \frac{1}{2}\,\mathbb{E}_{(s,a)\sim\nu^t}\big[(w^\top\phi_{s,a} - Q^t(s,a))^2\big],$$

Algorithm 1 Regression for $w_t^\star$

Require: Policy π^t; features $\phi_{s,a} \in \mathbb{R}^d$; current state of the trajectory; burn-in length m; rollout horizon H; tail index $\tau \in \{0, \dots, H-1\}$; number of rollouts $B = 1$; number of SGD steps K; step size α.
Ensure: Parameter $\widehat{w} \in \mathbb{R}^d$.

Initialize $w_{t,0} \leftarrow 0$
for $k = 1, 2, \dots, K$ **do**
  Set $s_{k,0}$ to the current state of the trajectory
  for $j = 0, 1, \dots, m-1$ **do**
    Sample $a_{k,j} \sim \pi^t(\cdot \mid s_{k,j})$
    Sample $s_{k,j+1} \sim P(\cdot \mid s_{k,j}, a_{k,j})$
  end for
  Sample $a_{k,m} \sim \pi^t(\cdot \mid s_{k,m})$
  Set $(s_k, a_k) \leftarrow (s_{k,m}, a_{k,m})$
  Continue the trajectory for a rollout of H transitions from (s_k, a_k) and compute $\widehat{Q}_k \leftarrow \widehat{Q}^{\pi^t}_{H,\tau}(s_k, a_k)$ from it by the formula of Lemma C.2 with $B = 1$
  $g_k \leftarrow (w_{t,k-1}^\top \phi_{s_k,a_k} - \widehat{Q}_k)\, \phi_{s_k,a_k}$
  $w_{t,k} \leftarrow w_{t,k-1} - \alpha g_k$
end for
Return $\widehat{w} \leftarrow \frac{1}{\lceil K/2 \rceil} \sum_{k=\lfloor K/2 \rfloor + 1}^{K} w_{t,k}$

Algorithm 2 Confidence boosting for w_t

Require: Policy π^t; current state of the trajectory; number of runs R; inputs m, H, τ, $B = 1$, K, α of Algorithm 1.
Ensure: Parameter $w_t \in \mathbb{R}^d$.

for $i = 1, 2, \dots, R$ **do**
  Run Algorithm 1, continuing the trajectory, and let $w^{(i)}$ be its output
end for
Let $\widehat{\Sigma}_t$ be the average of $\phi_{s,a}\phi_{s,a}^\top$ over the RK pairs (s,a) of the R runs
for $i = 1, 2, \dots, R$ **do**
  $r_i \leftarrow \min\{r \geq 0 : |\{j : \|w^{(j)} - w^{(i)}\|_{\widehat{\Sigma}_t} \leq r\}| > R/2\}$
end for
Return $w_t \leftarrow w^{(i^\star)}$ for the smallest $i^\star \in \arg\min_i r_i$

with independent and identically distributed labeled pairs $(s, a, \widehat{Q}^t(s,a))$, where $(s,a) \sim \nu^t$ and the label satisfies (iii), constant step size $\alpha = \frac{1}{4L^2}$, initialization $w_{t,0} = 0$, and tail-averaged output $w_t := \frac{1}{\lceil K/2 \rceil} \sum_{k=\lfloor K/2 \rfloor + 1}^{K} w_{t,k}$.

Then, the SGD output satisfies the statistical-error bound

$$\mathbb{E}\big[\|w_t - w_t^\star\|_{\Sigma_t}^2\big] \;\leq\; \frac{256\, L^4 G^2}{\mu^2 K^2}\, \exp\!\Big(-\frac{\mu\, \lfloor K/2 \rfloor}{4L^2}\Big) + \frac{8}{K}\Big(dG^2 + \frac{L^2}{\mu}\, \varepsilon_{\mathrm{app}}^2\Big).$$

Consequently, for any desired level $\varepsilon_{\mathrm{stat}} \in (0, G]$, it suffices to choose

$$K \;\geq\; \frac{16}{\varepsilon_{\mathrm{stat}}^2}\Big(dG^2 + \frac{L^2}{\mu}\, \varepsilon_{\mathrm{app}}^2\Big) + \frac{8L^2}{\mu} \log \frac{16 G^2}{\varepsilon_{\mathrm{stat}}^2}$$

to guarantee $\mathbb{E}\big[\|w_t - w_t^\star\|_{\Sigma_t}^2\big] \leq \varepsilon_{\mathrm{stat}}^2$.

Proof. With a constant step size, averaged SGD [4] forgets its initialization only at rate $1/K$; averaging only the second half of the iterates [15] makes this error decay geometrically. We prove the bound directly by splitting the error of SGD into an initialization part and a noise part; bounding the noise separately for the label noise and for the approximation error keeps L^2/μ off the label-noise term.

Write $w^\star = w_t^\star$. For a labeled pair (s, a, y) with $y = \widehat{Q}^t(s, a)$, let $x := \phi_{s,a}$ and let $\xi := (y - w^{\star\top} x)\, x$ be the residual; (x_k, y_k, ξ_k) denotes the k-th pair. Since $\mathbb{E}[y \mid s, a] = Q^t(s, a)$, first-order optimality of $w^\star$ gives $\mathbb{E}[\xi] = \mathbb{E}[(Q^t(s, a) - w^{\star\top} x)\, x] = 0$.

Projection bound. The minimizer satisfies the normal equation $\Sigma_t w^\star = \mathbb{E}[Q^t(s, a)\, x]$, so by the Cauchy–Schwarz inequality and $|Q^t(s, a)| \le G$,

$$\|w^\star\|_{\Sigma_t}^2 = \mathbb{E}\big[Q^t(s, a)\, x^\top w^\star\big] \le G\, \|w^\star\|_{\Sigma_t}, \qquad \text{hence} \qquad \|w^\star\| \le \frac{\|w^\star\|_{\Sigma_t}}{\sqrt{\mu}} \le \frac{G}{\sqrt{\mu}}.$$

Initialization and noise parts. The update $w_{t,k} = w_{t,k-1} - \alpha(w_{t,k-1}^\top x_k - y_k)\, x_k$ gives $w_{t,k} - w^\star = (I_d - \alpha x_k x_k^\top)(w_{t,k-1} - w^\star) + \alpha \xi_k$. Hence $w_{t,k} - w^\star = b_k + v_k$, where

$$b_k = (I_d - \alpha x_k x_k^\top)\, b_{k-1}, \quad b_0 = -w^\star, \qquad\qquad v_k = (I_d - \alpha x_k x_k^\top)\, v_{k-1} + \alpha \xi_k, \quad v_0 = 0.$$

Let $\mathcal{T}(A) := \mathbb{E}[(I_d - \alpha x x^\top) A (I_d - \alpha x x^\top)]$. This linear map preserves the positive semidefinite order and is self-adjoint for the trace inner product. Since $\|x\| \le L$ and $\alpha L^2 \le 1$,

$$\mathcal{T}(I_d) = I_d - 2\alpha \Sigma_t + \alpha^2\, \mathbb{E}[\|x\|^2 x x^\top] \preceq I_d - \alpha \Sigma_t \preceq (1 - \alpha\mu) I_d,$$

so $\operatorname{Tr} \mathcal{T}(A) = \operatorname{Tr}(A\, \mathcal{T}(I_d)) \le (1 - \alpha\mu) \operatorname{Tr}(A)$ for every $A \succeq 0$. Since b_{k-1} is independent of the k-th pair, $\mathbb{E}[b_k b_k^\top] = \mathcal{T}(\mathbb{E}[b_{k-1} b_{k-1}^\top])$, and therefore

$$\mathbb{E}\|b_k\|^2 \le (1 - \alpha\mu)^k\, \|w^\star\|^2 \le (1 - \alpha\mu)^k\, G^2/\mu.$$

For the noise part, $\mathbb{E}[v_k] = 0$ by induction, so the cross terms vanish and $C_k := \mathbb{E}[v_k v_k^\top] = \mathcal{T}(C_{k-1}) + \alpha^2\, \mathbb{E}[\xi \xi^\top]$. Conditioning on (s, a) separates the label noise from the approximation error:

$$\mathbb{E}[\xi \xi^\top] = \underbrace{\mathbb{E}\big[\operatorname{Var}(y \mid s, a)\, x x^\top\big]}_{=: S_1} + \underbrace{\mathbb{E}\big[(Q^t(s, a) - w^{\star\top} x)^2\, x x^\top\big]}_{=: S_2},$$

where $S_1 \preceq G^2 \Sigma_t$ because $\operatorname{Var}(y \mid s, a) \le \mathbb{E}[y^2 \mid s, a] \le G^2$, and $\operatorname{Tr}(S_2) \le L^2 \varepsilon_{\text{app}}^2$ because $\|x\| \le L$. By linearity, $C_k = C_k^{(1)} + C_k^{(2)}$ with $C_k^{(i)} = \mathcal{T}(C_{k-1}^{(i)}) + \alpha^2 S_i$ and $C_0^{(i)} = 0$. Induction gives $C_k^{(1)} \preceq \alpha G^2 I_d$, because $\mathcal{T}(\alpha G^2 I_d) + \alpha^2 G^2 \Sigma_t \preceq \alpha G^2 I_d$. The trace contraction gives $\operatorname{Tr} C_k^{(2)} \le (1 - \alpha\mu) \operatorname{Tr} C_{k-1}^{(2)} + \alpha^2 L^2 \varepsilon_{\text{app}}^2$, hence $\operatorname{Tr} C_k^{(2)} \le \alpha (L^2/\mu)\, \varepsilon_{\text{app}}^2$. Together,

$$\mathbb{E}\|v_k\|^2 = \operatorname{Tr}(C_k) \le \alpha \Big(d G^2 + \frac{L^2}{\mu}\, \varepsilon_{\text{app}}^2 \Big).$$

Tail averaging. Let u_k be either b_k or v_k, and let $\mathcal{K} := \{\lfloor K/2 \rfloor + 1, \dots, K\}$, with $q := |\mathcal{K}| = \lceil K/2 \rceil$. For $j \ge i$, the pairs after the i-th are independent of u_i and their residuals have mean zero, so $\mathbb{E}[u_j u_i^\top] = (I_d - \alpha \Sigma_t)^{j-i}\, \mathbb{E}[u_i u_i^\top]$. Expanding the square, the indices $i \le j$ therefore contribute at most $2 \operatorname{Tr}\big(\Sigma_t (I_d - \alpha \Sigma_t)^{j-i}\, \mathbb{E}[u_i u_i^\top]\big)$, which is nonnegative since $\Sigma_t (I_d - \alpha \Sigma_t)^l \succeq 0$ for every $l \ge 0$ (as $\alpha \Sigma_t \preceq I_d$). Summing over $j \ge i$ and using $\sum_{l \ge 0} \Sigma_t (I_d - \alpha \Sigma_t)^l = I_d/\alpha$,

$$\mathbb{E}\Big\| \frac{1}{q} \sum_{k \in \mathcal{K}} u_k \Big\|_{\Sigma_t}^2 \le \frac{2}{q^2} \sum_{i \in \mathcal{K}} \sum_{l \ge 0} \operatorname{Tr}\big(\Sigma_t (I_d - \alpha \Sigma_t)^l\, \mathbb{E}[u_i u_i^\top]\big) = \frac{2}{\alpha q^2} \sum_{i \in \mathcal{K}} \mathbb{E}\|u_i\|^2.$$

Conclusion. Summing the geometric series over the tail gives $\sum_{k \in \mathcal{K}} \mathbb{E}\|b_k\|^2 \le (1 - \alpha\mu)^{\lfloor K/2 \rfloor} G^2/(\alpha\mu^2)$, and $\sum_{k \in \mathcal{K}} \mathbb{E}\|v_k\|^2 \le q\, \alpha\, (d G^2 + (L^2/\mu) \varepsilon_{\text{app}}^2)$. Since $w_t - w^\star$ is the tail average of $b_k + v_k$ and $\|u + v\|_{\Sigma_t}^2 \le 2\|u\|_{\Sigma_t}^2 + 2\|v\|_{\Sigma_t}^2$,

$$\mathbb{E}\|w_t - w^\star\|_{\Sigma_t}^2 \le \frac{4 G^2 (1 - \alpha\mu)^{\lfloor K/2 \rfloor}}{\alpha^2 \mu^2 q^2} + \frac{4}{q} \Big(d G^2 + \frac{L^2}{\mu}\, \varepsilon_{\text{app}}^2 \Big),$$

and $\alpha = 1/(4L^2)$, $q \ge K/2$ and $1 - \alpha\mu \le e^{-\alpha\mu}$ give the stated bound. For the choice of K, the first summand of the lower bound on K makes the noise term at most $\varepsilon_{\text{stat}}^2/2$. For the initialization term, let $\ell := \log(16 G^2/\varepsilon_{\text{stat}}^2) \ge 1$. The second summand gives $K \ge 8 (L^2/\mu)\, \ell$, so $256 L^4/(\mu^2 K^2) \le 4$, and $\mu \lfloor K/2 \rfloor / (4L^2) \ge \ell - \frac{1}{8}$ by $\lfloor K/2 \rfloor \ge (K - 1)/2$ and $\mu \le L^2$. The initialization term is therefore at most $4 e^{1/8} G^2 e^{-\ell} < \varepsilon_{\text{stat}}^2/2$. □

D.3 Proof of Lemma 5.7 (LFA critic upper bound)

Lemma D.2 (LFA critic upper bound, detailed form)**.** *Suppose Assumption 5.2 holds, $\|\phi_{s,a}\|_2 \le L$ for all (s,a), and $\Sigma_t \succeq \mu I_d$ for all t, for some $\mu > 0$. Fix an outer iteration t, let $\mathcal{F}_{t-1}$ denote the σ-algebra generated by the trajectory up to the start of this critic call, and let w_t be the output of Algorithm 2, whose runs of Algorithm 1 use step size $\alpha = \frac{1}{4L^2}$. For any target critic accuracy $\varepsilon_{\mathrm{stat}} \in (0,1)$ and confidence $\delta \in (0,1)$, choose H, τ, B, m and K as in Lemma D.3 at accuracy $\varepsilon_{\mathrm{stat}}/20$, and $R = \lceil 2\log(2/\delta)\rceil$. Then*

$$\Pr\big(\|w_t - w_t^\star\|_{\Sigma_t} \le \varepsilon_{\mathrm{stat}} \,\big|\, \mathcal{F}_{t-1}\big) \;\ge\; 1-\delta \quad \textit{almost surely.}$$

Moreover, Algorithm 2 uses at most $RK(m+BH)$ transitions, so under the above choices the total number of transitions is

$$RK(m+BH) \;=\; \widetilde{O}\left(\Big(\frac{d\,t_{\mathrm{mix}}^3}{\varepsilon_{\mathrm{stat}}^2} + \frac{L^2}{\mu}\,t_{\mathrm{mix}}\Big(1+\frac{\varepsilon_{\mathrm{app}}^2}{\varepsilon_{\mathrm{stat}}^2}\Big)\Big)\log\frac{2}{\delta}\right),$$

where $\widetilde{O}(\cdot)$ hides absolute constants and polylogarithmic factors in $t_{\mathrm{mix}}, L, \mu, d, 1/\varepsilon_{\mathrm{stat}}$.

The proof combines a bound in expectation for a single run of Algorithm 1 with the selection step of Algorithm 2. We first prove the former.

Lemma D.3 (One run of Algorithm 1)**.** *Suppose Assumption 5.2 holds, $\|\phi_{s,a}\|_2 \le L$ for all (s,a), and $\Sigma_t \succeq \mu I_d$ for all t, for some $\mu > 0$. Fix an outer iteration t and let $\widehat{w}$ be the output of a run of Algorithm 1 with policy π^t and step size $\alpha = \frac{1}{4L^2}$. Let $\mathcal{G}$ denote the σ-algebra generated by the trajectory up to the start of this run; the policy π^t and the current state of the trajectory are $\mathcal{G}$-measurable.*

Then for any target critic accuracy $\varepsilon_{\mathrm{stat}} \in (0,1)$, the integer parameter choices, with $H \ge 4$,

$$H \;=\; \Theta\left(t_{\mathrm{mix}}\,\log\frac{2t_{\mathrm{mix}}}{\varepsilon_{\mathrm{stat}}}\right), \qquad \tau \;=\; \left\lceil\frac{H}{2}\right\rceil, \qquad B \;=\; 1,$$

$$m \;=\; \Theta\left(t_{\mathrm{mix}}\,\log\frac{t_{\mathrm{mix}}\,(1+L^2/\mu)}{\varepsilon_{\mathrm{stat}}}\right),$$

$$K \;=\; \Theta\left(\frac{d\,t_{\mathrm{mix}}^2\,\log(2t_{\mathrm{mix}}/\varepsilon_{\mathrm{stat}})}{\varepsilon_{\mathrm{stat}}^2} + \frac{L^2}{\mu}\Big(\log\frac{2d\,t_{\mathrm{mix}}}{\varepsilon_{\mathrm{stat}}} + \frac{\min\{\varepsilon_{\mathrm{app}}^2, (1+2t_{\mathrm{mix}})^2\}}{\varepsilon_{\mathrm{stat}}^2}\Big)\right)$$

suffice to guarantee

$$\mathbb{E}\big[\|\widehat{w} - w_t^\star\|_{\Sigma_t}^2 \,\big|\, \mathcal{G}\big] \;\le\; \varepsilon_{\mathrm{stat}}^2 \quad \textit{almost surely.}$$

Moreover, Algorithm 1 uses at most $K(m+BH)$ transitions, so under the above choices (with $BH = H = \widetilde{O}(t_{\mathrm{mix}})$) the total number of transitions is

$$K(m+BH) \;=\; \widetilde{O}\left(\frac{d\,t_{\mathrm{mix}}^3}{\varepsilon_{\mathrm{stat}}^2} + \frac{L^2}{\mu}\,t_{\mathrm{mix}}\Big(1+\frac{\varepsilon_{\mathrm{app}}^2}{\varepsilon_{\mathrm{stat}}^2}\Big)\right),$$

where $\widetilde{O}(\cdot)$ hides absolute constants and polylogarithmic factors in $t_{\mathrm{mix}}, L, \mu, 1/\varepsilon_{\mathrm{stat}}$.

Proof. Fix t.

Coupling with independent pairs. Throughout, we condition on $\mathcal{G}$, which fixes π^t and the current state of the trajectory. Write $Z_k := (s_k, a_k, \widehat{Q}_k)$ for the k-th pair of Algorithm 1 and its label, so that $\widehat{w}$ is a function of $Z_1, \dots, Z_K$. Let $Z_1', \dots, Z_K'$ be independent, where $Z_k' := (s_k', a_k', \widehat{Q}_k')$ has $(s_k', a_k') \sim \nu^t$ and a label $\widehat{Q}_k'$ computed in the same way from a rollout that starts at (s_k', a_k'); this is the sampling of Lemma D.1. Conditional on the trajectory up to the start of the k-th burn-in, the law of s_k is $P_{\pi^t}^m(\cdot \mid s_{k,0})$, which by (11) is within $\delta_m := 2^{-\lfloor m/t_{\mathrm{mix}}\rfloor}$ of d^{π^t} in total variation. Since $a_k \mid s_k \sim \pi^t(\cdot \mid s_k)$ and $\widehat{Q}_k$ is computed from the rollout that starts at (s_k, a_k), the law of Z_k satisfies the same TV bound relative to the law of Z_k', and, by convexity of the total variation

distance, so does its conditional law given only $Z_1, \dots, Z_{k-1}$, which averages these laws. Now replace $Z_K, Z_{K-1}, \dots, Z_1$ in turn by independent draws distributed as $Z'_K, Z'_{K-1}, \dots, Z'_1$. The step that replaces Z_k changes only the conditional law of the k-th coordinate given the first $k-1$, from that of Z_k to the law of Z'_k, so the joint law moves in total variation by the average distance between these two laws, which is at most δ_m. Hence the joint laws $\mathbb{P}$ of $(Z_1, \dots, Z_K)$ and $\mathbb{P}'$ of $(Z'_1, \dots, Z'_K)$ satisfy $\|\mathbb{P} - \mathbb{P}'\|_{\mathrm{TV}} \le K\delta_m$. Since the Z_k take finitely many values, an optimal coupling of $\mathbb{P}$ and $\mathbb{P}'$ [24, Proposition 4.7 and Remark 4.8] lets us realize $Z'_1, \dots, Z'_K$ on an extended probability space so that $Z'_k \neq Z_k$ for some k with probability at most $K\delta_m$. Let $\widehat{w}'$ be the output of the same SGD run on $Z'_1, \dots, Z'_K$; it equals $\widehat{w}$ whenever $Z'_k = Z_k$ for all k.

Almost-sure bound on $\|\widehat{w} - w_t^\star\|_{\Sigma_t}$. With $B = 1$, each label equals the quantity Y_1 in the proof of Lemma C.2, so $|\widehat{Q}_k| \le 2H$ by the range bound there. Each SGD step is $w_{t,k} = (I_d - \alpha\phi\phi^\top)w_{t,k-1} + \alpha\widehat{Q}_k\phi$ with $\phi = \phi_{s_k,a_k}$, and $\|I_d - \alpha\phi\phi^\top\|_{\mathrm{op}} \le 1$ since $\alpha\|\phi\|_2^2 \le 1/4$, so $\|w_{t,k}\|_2 \le \|w_{t,k-1}\|_2 + 2\alpha LH = \|w_{t,k-1}\|_2 + H/(2L)$. Hence $\|w_{t,k}\|_2 \le kH/(2L)$, and the tail average satisfies $\|\widehat{w}\|_2 \le KH/(2L)$. Moreover, $\|w_t^\star\|_2 \le \|Q^{\pi^t}\|_\infty/\sqrt{\mu} \le L\,\|Q^{\pi^t}\|_\infty/\mu$ by the bound on $\|w^\star\|$ in the proof of Lemma D.1 and $\mu \le L^2$. For the Poisson bound $\|Q^{\pi^t}\|_\infty \le 1 + 2\,t_{\mathrm{mix}}$, the series for Q^{π^t} in the proof of Lemma C.2 has a first term of absolute value at most 1, and by (28) the absolute values of its remaining terms sum to at most $\sum_{\ell \ge 0} 2^{-\lfloor \ell/t_{\mathrm{mix}} \rfloor} = 2\,t_{\mathrm{mix}}$. Since $\Sigma_t \preceq L^2 I_d$, almost surely

$$\|\widehat{w} - w_t^\star\|_{\Sigma_t} \;\le\; L\big(\|\widehat{w}\|_2 + \|w_t^\star\|_2\big) \;\le\; \frac{KH}{2} + \frac{(1 + 2\,t_{\mathrm{mix}})L^2}{\mu}.$$

Decomposition. Define the conditional-mean label

$$\widehat{Q}(s,a) \;:=\; \mathbb{E}\Big[\widehat{Q}^{\pi^t}_{H,\tau}(s,a) \mid s,a\Big],$$

and the population minimizer

$$\tilde{w}_t \in \arg\min_w \mathbb{E}_{(s,a)\sim\nu^t}\Big[\big(w^\top\phi_{s,a} - \widehat{Q}(s,a)\big)^2\Big].$$

Split the expectation according to whether $Z'_k = Z_k$ for all k. If so, $\widehat{w} = \widehat{w}'$, and we write $\widehat{w}' - w_t^\star = (\widehat{w}' - \tilde{w}_t) + (\tilde{w}_t - w_t^\star)$ and use $\|u+v\|_{\Sigma_t}^2 \le 2\|u\|_{\Sigma_t}^2 + 2\|v\|_{\Sigma_t}^2$; if not, which has probability at most $K\delta_m$, we use the almost-sure bound. This gives

$$\mathbb{E}\big[\|\widehat{w} - w_t^\star\|_{\Sigma_t}^2\big] \le 2\,\mathbb{E}\big[\|\widehat{w}' - \tilde{w}_t\|_{\Sigma_t}^2\big] + 2\,\|\tilde{w}_t - w_t^\star\|_{\Sigma_t}^2 + K\delta_m\Big(\frac{KH}{2} + \frac{(1+2\,t_{\mathrm{mix}})L^2}{\mu}\Big)^2. \quad (36)$$

Bias term $\|\tilde{w}_t - w_t^\star\|_{\Sigma_t}^2$. Since $\phi^\top w_t^\star$ and $\phi^\top\tilde{w}_t$ are the $L^2(\nu^t)$ projections of Q^{π^t} and $\widehat{Q}$ onto the span of the features, and projections do not increase distances,

$$\|\tilde{w}_t - w_t^\star\|_{\Sigma_t} \le \sqrt{\mathbb{E}_{(s,a)\sim\nu^t}\Big[\big(\widehat{Q}(s,a) - Q^{\pi^t}(s,a)\big)^2\Big]} \le \|\widehat{Q} - Q^{\pi^t}\|_\infty.$$

By the proof of Lemma C.2 (bias bound (29)),

$$\|\widehat{Q} - Q^{\pi^t}\|_\infty \;\le\; 12\,t_{\mathrm{mix}}\,2^{-\lfloor(\tau-1)/t_{\mathrm{mix}}\rfloor}.$$

With the choice $\tau = \lceil H/2 \rceil$ and $H = \Theta(t_{\mathrm{mix}}\,\log(2t_{\mathrm{mix}}/\varepsilon_{\mathrm{stat}}))$, the right-hand side is $O(\varepsilon_{\mathrm{stat}})$, hence $\|\widehat{Q} - Q^{\pi^t}\|_\infty \le \varepsilon_{\mathrm{stat}}/3$ and $\|\tilde{w}_t - w_t^\star\|_{\Sigma_t}^2 \le \varepsilon_{\mathrm{stat}}^2/9$ for a suitable constant in the $\Theta(\cdot)$ choice of H.

SGD term $\mathbb{E}[\|\widehat{w}' - \tilde{w}_t\|_{\Sigma_t}^2]$. By definition of $\widehat{Q}$, the labels of Algorithm 1 ($B = 1$) satisfy $\mathbb{E}[\widehat{Q}^{\pi^t}_{H,\tau}(s,a) \mid s,a] = \widehat{Q}(s,a)$. The SGD rate of Lemma D.1 depends on the label second moment, which we bound directly (rather than via the high-probability guarantee of Lemma C.2):

$$\begin{aligned}\mathbb{E}[(\widehat{Q}^{\pi^t}_{H,\tau}(s,a))^2 \mid s,a] \;&\le\; \|\widehat{Q}\|_\infty^2 + \mathrm{Var}(\widehat{Q}^{\pi^t}_{H,\tau}(s,a) \mid s,a) \;=\; O(t_{\mathrm{mix}}^2) + O(H\,t_{\mathrm{mix}}) \\ &=\; O\big(t_{\mathrm{mix}}^2\log(2t_{\mathrm{mix}}/\varepsilon_{\mathrm{stat}})\big),\end{aligned}$$

where the last step uses $H = \Theta(t_{\mathrm{mix}} \log(2t_{\mathrm{mix}}/\varepsilon_{\mathrm{stat}}))$, the mean bound uses the bias and Poisson bounds above, and the variance bound follows from $\mathbb{E}[Y_i^2] \le 2(1 + 4\,t_{\mathrm{mix}})(\tau + \tau^2/(H - \tau))$ in the proof of Lemma C.2, with $\tau = \lceil H/2 \rceil$. Therefore, Lemma D.1 applies to $Z'_1, \dots, Z'_K$, with target $\widehat{Q}$ in place of Q^t (so that $w_t^\star$ there is $\tilde{w}_t$, and $|\widehat{Q}| \le G$ by Jensen's inequality), label second moment $G^2 = O(t_{\mathrm{mix}}^2 \log(2t_{\mathrm{mix}}/\varepsilon_{\mathrm{stat}}))$ with $G \ge 1$, and $2\bar{\varepsilon}_{\mathrm{app}}^2 + 2\varepsilon_{\mathrm{stat}}^2/9$ in place of $\varepsilon_{\mathrm{app}}^2$. Here $\bar{\varepsilon}_{\mathrm{app}} := \min\{\varepsilon_{\mathrm{app}}, 1 + 2\,t_{\mathrm{mix}}\}$ satisfies Assumption 5.2 as well, since $w = 0$ gives $\mathbb{E}_{\nu^t}[(Q^{\pi^t})^2] \le (1 + 2\,t_{\mathrm{mix}})^2$ by the Poisson bound above, and the minimality of $\tilde{w}_t$ together with $\|\widehat{Q} - Q^{\pi^t}\|_\infty \le \varepsilon_{\mathrm{stat}}/3$ gives

$$\begin{aligned}\mathbb{E}_{\nu^t}\big[(\tilde{w}_t^\top \phi - \widehat{Q})^2\big] &\le \mathbb{E}_{\nu^t}\big[(w_t^{\star\top} \phi - \widehat{Q})^2\big] \\ &\le 2\,\mathbb{E}_{\nu^t}\big[(w_t^{\star\top} \phi - Q^{\pi^t})^2\big] + 2\|\widehat{Q} - Q^{\pi^t}\|_\infty^2 \le 2\bar{\varepsilon}_{\mathrm{app}}^2 + \tfrac{2}{9}\varepsilon_{\mathrm{stat}}^2.\end{aligned}$$

At accuracy $\varepsilon_{\mathrm{stat}}/3 \le G$, the choice of K in Lemma D.1 is then of the order stated, which ensures $\mathbb{E}[\|\widehat{w}' - \tilde{w}_t\|_{\Sigma_t}^2] \le \varepsilon_{\mathrm{stat}}^2/9$.

Burn-in term. Since K and H are bounded by polynomials in t_{mix}, $1 + L^2/\mu$ and $1/\varepsilon_{\mathrm{stat}}$ (using $d \le L^2/\mu$, from $d\mu \le \mathrm{Tr}(\Sigma_t) \le L^2$, and $\bar{\varepsilon}_{\mathrm{app}} \le 1 + 2\,t_{\mathrm{mix}}$), a suitable constant in the $\Theta(\cdot)$ choice of m gives $m \ge t_{\mathrm{mix}} \lceil \log_2\!\big(2K(KH/2 + (1 + 2\,t_{\mathrm{mix}})L^2/\mu)^2/\varepsilon_{\mathrm{stat}}^2\big) \rceil$. Then $\delta_m = 2^{-\lfloor m/t_{\mathrm{mix}} \rfloor}$ makes the last term of (36) at most $\varepsilon_{\mathrm{stat}}^2/2$. In particular, $\delta_m \le 1/(2K)$, since $\varepsilon_{\mathrm{stat}} < 1$.

Sample complexity. Plugging the three bounds into (36) gives $\mathbb{E}[\|\widehat{w} - w_t^\star\|_{\Sigma_t}^2] \le 4\varepsilon_{\mathrm{stat}}^2/9 + \varepsilon_{\mathrm{stat}}^2/2 < \varepsilon_{\mathrm{stat}}^2$, conditionally on $\mathcal{G}$. Finally, each SGD step uses m transitions for its burn-in and H transitions for its rollout ($B = 1$), so the per-step cost is $m + H = \widetilde{O}(t_{\mathrm{mix}})$ and the total number of transitions is $K(m + BH) = \widetilde{O}\big(d\,t_{\mathrm{mix}}^3/\varepsilon_{\mathrm{stat}}^2 + (L^2/\mu)\,t_{\mathrm{mix}}(1 + \varepsilon_{\mathrm{app}}^2/\varepsilon_{\mathrm{stat}}^2)\big)$. □

Proof of Lemma D.2. Let $w^{(1)}, \dots, w^{(R)}$ be the outputs of the R runs of Algorithm 1 in Algorithm 2, and let $\mathcal{G}_{i-1}$ be the σ-algebra generated by the trajectory up to the start of run i, so that $\mathcal{G}_0 = \mathcal{F}_{t-1}$. All runs follow π^t. Call run i *accurate* if $\|w^{(i)} - w_t^\star\|_{\Sigma_t} \le \varepsilon_{\mathrm{stat}}/5$, and write $\|u\|_{\widehat{\Sigma}_t} := \sqrt{u^\top \widehat{\Sigma}_t u}$.

Most runs are accurate. Let $\mathbf{1}_i$ indicate that run i is not accurate; it is $\mathcal{G}_i$-measurable. Lemma D.3 at accuracy $\varepsilon_{\mathrm{stat}}/20$, applied with $\mathcal{G} = \mathcal{G}_{i-1}$, and Markov's inequality give $\mathbb{E}[\mathbf{1}_i \mid \mathcal{G}_{i-1}] \le (\varepsilon_{\mathrm{stat}}/20)^2/(\varepsilon_{\mathrm{stat}}/5)^2 = 1/16$. Hence $\mathbb{E}[4^{\mathbf{1}_i} \mid \mathcal{G}_{i-1}] \le 1 + 3/16 \le e^{3/16}$, and iterating conditional expectations gives $\mathbb{E}[4^{\sum_i \mathbf{1}_i} \mid \mathcal{F}_{t-1}] \le e^{3R/16}$. By Markov's inequality,

$$\Pr\Big(\textstyle\sum_{i=1}^R \mathbf{1}_i \ge R/2 \,\Big|\, \mathcal{F}_{t-1}\Big) \;\le\; e^{3R/16}\, 4^{-R/2} \;=\; e^{-R(\log 2 - 3/16)} \;\le\; e^{-R/2} \;\le\; \delta/2.$$

So, outside an event of conditional probability at most $\delta/2$, more than half of the runs are accurate. The runs need not be independent.

The empirical covariance is accurate. List the RK pairs of all runs in the order in which they are visited, with features $x_1, \dots, x_{RK}$, and let $\mathcal{H}_{l-1}$ be the σ-algebra generated by the trajectory up to the start of the burn-in that precedes the l-th pair. Set $\Phi_l := (\mu/L^2)\,\Sigma_t^{-1/2} x_l x_l^\top \Sigma_t^{-1/2}$ and $\Phi_{\le l} := \Phi_1 + \cdots + \Phi_l$. Then $0 \preceq \Phi_l \preceq I_d$, because $\|\Sigma_t^{-1/2} x_l\|_2^2 \le L^2/\mu$, and $\Phi_{\le RK} = (\mu RK/L^2)\,\Sigma_t^{-1/2} \widehat{\Sigma}_t \Sigma_t^{-1/2}$. As in the proof of Lemma D.3, the conditional law of the l-th pair given $\mathcal{H}_{l-1}$ is within $2^{-\lfloor m/t_{\mathrm{mix}} \rfloor}$ of ν^t in total variation. For every unit vector u, the function $(s, a) \mapsto (u^\top \Sigma_t^{-1/2} \phi_{s,a})^2$ takes values in $[0, L^2/\mu]$ and has mean 1 under ν^t, so

$$\Big\| \frac{L^2}{\mu}\, \mathbb{E}[\Phi_l \mid \mathcal{H}_{l-1}] - I_d \Big\|_{\mathrm{op}} \;\le\; \frac{L^2}{\mu}\, 2^{-\lfloor m/t_{\mathrm{mix}} \rfloor} \;\le\; \frac{1}{4},$$

where the last step uses $2^{-\lfloor m/t_{\mathrm{mix}} \rfloor} \le 1/(2K)$, from the choice of m in that proof, and $K \ge 32(L^2/\mu)\log(4d) \ge 2L^2/\mu$, which a suitable constant in the choice of K there guarantees. Hence $\frac{3\mu}{4L^2} I_d \preceq \mathbb{E}[\Phi_l \mid \mathcal{H}_{l-1}] \preceq \frac{5\mu}{4L^2} I_d$.

We now follow the iterative Golden–Thompson argument of Ahlswede and Winter [43, Section 1.3.1], which applies to our dependent sequence because each conditional bound below is a deterministic

multiple of I_d. Convexity of $z \mapsto e^{\pm z/2}$ on $[0,1]$ gives $e^{\pm\Phi_l/2} \preceq I_d + (e^{\pm 1/2} - 1)\,\Phi_l$. With the bounds on $\mathbb{E}[\Phi_l \mid \mathcal{H}_{l-1}]$ above, $1 + z \le e^z$, $e^{1/2} \le 17/10$ and $e^{-1/2} \le 5/8$, this gives

$$\mathbb{E}\big[e^{\Phi_l/2} \,\big|\, \mathcal{H}_{l-1}\big] \preceq e^{7\mu/(8L^2)}\, I_d, \qquad \mathbb{E}\big[e^{-\Phi_l/2} \,\big|\, \mathcal{H}_{l-1}\big] \preceq e^{-9\mu/(32L^2)}\, I_d.$$

The Golden–Thompson inequality gives $\operatorname{Tr} e^{\pm\Phi_{\le l}/2} \le \operatorname{Tr}\big(e^{\pm\Phi_{\le l-1}/2}\, e^{\pm\Phi_l/2}\big)$, and $e^{\pm\Phi_{\le l-1}/2} \succeq 0$ is $\mathcal{H}_{l-1}$-measurable, so taking conditional expectations and iterating over $l = RK, \dots, 1$ yields

$$\mathbb{E}\big[\operatorname{Tr} e^{\Phi_{\le RK}/2} \,\big|\, \mathcal{F}_{t-1}\big] \le d\, e^{7\mu RK/(8L^2)}, \qquad \mathbb{E}\big[\operatorname{Tr} e^{-\Phi_{\le RK}/2} \,\big|\, \mathcal{F}_{t-1}\big] \le d\, e^{-9\mu RK/(32L^2)}.$$

By Markov's inequality, the event $\lambda_{\max}(\Phi_{\le RK}) \ge 2\mu RK/L^2$ has conditional probability at most $d\, e^{-\mu RK/(8L^2)}$, and the event $\lambda_{\min}(\Phi_{\le RK}) \le \mu RK/(2L^2)$ has conditional probability at most $d\, e^{-\mu RK/(32L^2)}$. Hence $\frac{1}{2}\Sigma_t \preceq \widehat{\Sigma}_t \preceq 2\Sigma_t$ fails with conditional probability at most $2d\, e^{-\mu RK/(32L^2)} \le \delta/2$. Here we used $K \ge 32(L^2/\mu)\log(4d)$, together with $R \ge 2\log(2/\delta) \ge 1 + \log(1/\delta)$ and $\log(4d) \ge 1$, which give $R\log(4d) \ge \log(4d/\delta)$.

The selected run has error at most $\varepsilon_{\mathrm{stat}}$. On the intersection of the two events above, which has conditional probability at least $1-\delta$, more than half of the runs are accurate and $\frac{1}{2}\|u\|^2_{\Sigma_t} \le \|u\|^2_{\widehat{\Sigma}_t} \le 2\|u\|^2_{\Sigma_t}$ for every $u \in \mathbb{R}^d$. The last bound holds for all u at once, so it applies to the data-dependent differences $w^{(i)} - w^{(j)}$. Any two accurate runs satisfy $\|w^{(i)} - w^{(j)}\|_{\widehat{\Sigma}_t} \le \sqrt{2} \cdot 2\varepsilon_{\mathrm{stat}}/5$, so every accurate run has $r_i \le 2\sqrt{2}\,\varepsilon_{\mathrm{stat}}/5$, and hence $r_{i^\star} \le 2\sqrt{2}\,\varepsilon_{\mathrm{stat}}/5$. The ball of radius $r_{i^\star}$ around $w^{(i^\star)}$ in $\|\cdot\|_{\widehat{\Sigma}_t}$ contains more than half of the runs, so it contains an accurate run $w^{(j)}$, and

$$\|w_t - w_t^\star\|_{\Sigma_t} \le \|w^{(i^\star)} - w^{(j)}\|_{\Sigma_t} + \|w^{(j)} - w_t^\star\|_{\Sigma_t} \le \sqrt{2} \cdot \frac{2\sqrt{2}\,\varepsilon_{\mathrm{stat}}}{5} + \frac{\varepsilon_{\mathrm{stat}}}{5} = \varepsilon_{\mathrm{stat}}.$$

This is the deterministic selection argument of Hsu and Sabato [14, Proposition 8]; the two previous steps remove the independence that their probability bounds require, through conditional arguments along a single trajectory.

Sample complexity. Each run uses $K(m + BH)$ transitions, so Algorithm 2 uses $RK(m + BH)$. At accuracy $\varepsilon_{\mathrm{stat}}/20$, the parameters of Lemma D.3 change only by constant factors, so with $R = \lceil 2\log(2/\delta)\rceil$ the total is as stated. □

D.4 Proof of Theorem 5.10 (LFA sample complexity)

Theorem D.4 (LFA sample complexity, detailed form)**.** *Suppose Assumption 5.4 and the hypotheses of Lemma 5.7 hold, and $D_0^\star < \infty$. Recall the initial Lyapunov value $\Psi_0 = \Delta_0 + D_0^\star/(\eta_{-1}\bar{\chi})$. Fix a target suboptimality $\varepsilon \in (0,1)$ with $\varepsilon \ge 8(\bar{\chi}\sqrt{C_+} + \sqrt{C_\star})\varepsilon_{\mathrm{app}}$ and a confidence $\delta \in (0,1)$, and set $\varepsilon_{\mathrm{stat}} := \varepsilon/(8(\bar{\chi}\sqrt{C_+} + \sqrt{C_\star}))$ and $T := \lceil \bar{\chi}\log\max\{e, 2\Psi_0/\varepsilon\}\rceil$. Run the update (2) with $\tilde{Q}^t(s,a) = \phi_{s,a}^\top w_t$ for T steps under the hypotheses of Lemma 3.7(i), where w_t is the output of Algorithm 2 with the parameters of Lemma D.2 at accuracy $\varepsilon_{\mathrm{stat}}$ and confidence δ/T. Then, with probability at least $1-\delta$,*

$$\Delta_T \;\le\; \varepsilon.$$

Moreover, each outer iteration uses at most $RK(m+BH)$ transitions, with $R = \lceil 2\log(2T/\delta)\rceil$, hence the total number of transitions is at most $T\,RK(m+BH)$, which under these choices is

$$T\,RK(m+BH) \;=\; \widetilde{O}\left(\left(\frac{d\,t_{\mathrm{mix}}^3\,(\bar{\chi}\sqrt{C_+} + \sqrt{C_\star})^2\,\bar{\chi}}{\varepsilon^2} + \frac{L^2}{\mu}\,t_{\mathrm{mix}}\,\bar{\chi}\right)\log\frac{2T}{\delta}\right),$$

where $\widetilde{O}(\cdot)$ hides polylogarithmic factors in $t_{\mathrm{mix}}, L, \mu, d, \bar{\chi}, C_+, C_\star, 1/\varepsilon, \Psi_0$.

Proof. As for Theorem C.4, the proof has two ingredients: a union bound converting the per-call guarantee of Lemma D.2 into critic accuracy along the PMD trajectory, and a deterministic composition with Theorem 5.6.

Critic accuracy along the trajectory. For each $t = 0, 1, \ldots, T-1$, the policy π^t and the state of the trajectory at the start of critic call t are $\mathcal{F}_{t-1}$-measurable, so Lemma D.2 at confidence δ/T gives

$$\Pr\big(\|w_t - w_t^\star\|_{\Sigma_t} > \varepsilon_{\text{stat}} \,\big|\, \mathcal{F}_{t-1}\big) \;\le\; \delta/T \quad \text{almost surely.}$$

Taking expectations and a union bound over $t \in \{0, \ldots, T-1\}$ then gives

$$\Pr\big(\exists\, t < T : \|w_t - w_t^\star\|_{\Sigma_t} > \varepsilon_{\text{stat}}\big) \;\le\; \delta.$$

The critic calls share one trajectory, but the union bound needs no independence between them. On the complementary good event, Assumption 5.3 holds along the realized PMD trajectory up to time T, which is all that the bound of Theorem 5.6 at $t = T$ uses.

Composition with Theorem 5.6. Condition on the good event, intersected with the probability-one event of Assumption 5.4. Theorem 5.6 gives $(\Delta_T - \varepsilon_{\text{floor}}(\bar\chi))_+ \le (1 - 1/\bar\chi)^T \Psi_0$. Since $\varepsilon \ge 8(\bar\chi\sqrt{C_+} + \sqrt{C_\star})\varepsilon_{\text{app}}$ gives $\varepsilon_{\text{app}} \le \varepsilon_{\text{stat}}$, the error floor satisfies

$$\varepsilon_{\text{floor}}(\bar\chi) = 2(\bar\chi\sqrt{C_+} + \sqrt{C_\star})\big(\varepsilon_{\text{stat}} + \varepsilon_{\text{app}}\big) \;\le\; 4(\bar\chi\sqrt{C_+} + \sqrt{C_\star})\,\varepsilon_{\text{stat}} \;=\; \frac{\varepsilon}{2}.$$

Using $(1 - 1/\bar\chi)^T \le \exp(-T/\bar\chi)$, the choice of T ensures $(1 - 1/\bar\chi)^T \Psi_0 \le \varepsilon/2$. Therefore $\Delta_T \le \varepsilon$ with probability at least $1 - \delta$.

Sample complexity. Each outer iteration runs Algorithm 2 once, at accuracy $\varepsilon_{\text{stat}}$ and confidence δ/T. By Lemma D.2, this uses $RK(m + BH) = \widetilde{O}\big((d\, t_{\text{mix}}^3/\varepsilon_{\text{stat}}^2 + (L^2/\mu)\, t_{\text{mix}}) \log(2T/\delta)\big)$ transitions, since $\varepsilon_{\text{app}} \le \varepsilon_{\text{stat}}$ absorbs the approximation term. Substituting $\varepsilon_{\text{stat}}$ and multiplying by $T = \widetilde{\Theta}(\bar\chi)$ gives the stated total. □

D.5 Proof of Proposition 5.9 (LFA critic lower bound)

We first restate the proposition with the explicit construction and constants used in the proof.

Proposition D.5 (LFA critic lower bound, detailed form)**.** *For every $d \ge 1$, $p \in (0, 1/4)$, and $\varepsilon_{\text{stat}} \in (0, 1/(256\, p)]$, there exists a family of one-action average-reward MDPs on $2d$ states with costs in $[0, 1]$ satisfying Assumption 3.1, whose $1/4$-mixing time satisfies $t_{\text{mix}} \in [1/(16p),\, 3/p]$, and admitting an exactly realizable linear differential Q-function $Q^\theta(s, a) = w_\theta^{\star\top} \phi(s, a)$ with $\|\phi(s, a)\|_2 \le 1$ and stationary covariance $\Sigma = \mathbb{E}_{\nu^\theta}[\phi\phi^\top] = (1/d)\, I_d$, such that any (possibly adaptive) generative-model algorithm that outputs $\widehat{w} \in \mathbb{R}^d$ with*

$$\Pr_\theta(\|\widehat{w} - w_\theta^\star\|_\Sigma \le \varepsilon_{\text{stat}}) \;\ge\; 2/3 \qquad \textit{for every } \theta \in \{0, 1\}^d$$

must use

$$N \;\ge\; \frac{1}{2^{21} \cdot 27}\, \frac{d\, t_{\text{mix}}^3}{\varepsilon_{\text{stat}}^2} \;=\; \Omega\!\left(\frac{d\, t_{\text{mix}}^3}{\varepsilon_{\text{stat}}^2}\right)$$

generative-model transitions.

Here, an algorithm is adaptive if it chooses each query based on the samples seen so far.

Proof. The proof has six steps: construct the direct-sum reset family, verify its mixing properties, solve the centered Poisson equation, compute the per-coordinate gap and per-query KL, then assemble Assouad's inequality.

Construction. Fix $d \ge 1$ and $p \in (0, 1/4)$. Let $\mathcal{S} = [d] \times \{-1, +1\}$ with one action $\mathcal{A} = \{a\}$, and parameterize the family by $\theta \in \{0, 1\}^d$ via local transition rates $q_j(\theta) := p + \theta_j\, \Delta$, where $\Delta \in (0, p/2]$ will be fixed in Step 6. The transition kernel is

$$P_\theta\big((j, \sigma), \cdot\big) \;:=\; (1 - p)\, R_\theta\big((j, \sigma), \cdot\big) \;+\; p\, u(\cdot),$$

where u is the uniform distribution on $\mathcal{S}$ (so $u(j, \sigma) = 1/(2d)$) and R_θ is the local sign-flip kernel

$$R_\theta\big((j, \sigma), (j, \sigma)\big) \;:=\; 1 - q_j, \qquad R_\theta\big((j, \sigma), (j, -\sigma)\big) \;:=\; q_j,$$

with $R_\theta((j, \sigma), \cdot) = 0$ on all other states. Cost $c(j, \sigma) := (1 + \sigma)/2 \in \{0, 1\}$ and features $\phi(j, \sigma) := \sigma\, e_j \in \mathbb{R}^d$.

Step 1: Stationary distribution. The uniform distribution u is invariant under P_θ. Indeed, the reset part moves every state to u, and for the local part,

$$(uR_\theta)(k,\sigma) \;=\; u(k,\sigma)(1-q_k) + u(k,-\sigma)\,q_k \;=\; \tfrac{1}{2d}(1-q_k) + \tfrac{1}{2d}\,q_k \;=\; \tfrac{1}{2d} \;=\; u(k,\sigma).$$

Hence $uP_\theta = (1-p)\,u + p\,u = u$. Since $P_\theta(s,s') \geq p\,u(s') = p/(2d) > 0$ for all s, s', the chain is irreducible and aperiodic, so $d^\theta = u$ is the unique stationary distribution.

Step 2: Mixing time. Decomposing $P_\theta = (1-p)\,R_\theta + p\,\mathbf{1}u$, where the kernel $\mathbf{1}u$ moves every state to u, and using $u\,R_\theta = u$,

$$P_\theta^k \;=\; (1-p)^k\,R_\theta^k \;+\; \big(1-(1-p)^k\big)\,\mathbf{1}u,$$

so $\mu_0 P_\theta^k - u = (1-p)^k(\mu_0 R_\theta^k - u)$ and $\sup_{\mu_0}\|\mu_0 P_\theta^k - u\|_{\mathrm{TV}} \leq (1-p)^k$. Setting this $\leq 1/4$ and using $\log(1/(1-p)) \geq p$ gives the upper bound

$$t_{\mathrm{mix}} \;\leq\; \lceil \log 4/p \rceil \;\leq\; (\log 4 + 1)/p \;<\; 3/p. \tag{37}$$

For the lower bound, define the coordinate-local function $f_j(k,\sigma) := \mathbf{1}\{k = j\}\,\sigma$, which has $\|f_j\|_\infty \leq 1$ and $\mathbb{E}_u[f_j] = 0$. Direct computation shows f_j is an eigenfunction of P_θ with eigenvalue $\lambda_j := (1-p)(1-2q_j)$. Starting from $(j,+1)$,

$$\big\|\delta_{(j,+1)}\,P_\theta^k - u\big\|_{\mathrm{TV}} \;\geq\; \tfrac{1}{2}\,\mathbb{E}_{(j,+1)}[f_j(s_k)] \;=\; \tfrac{1}{2}\,\lambda_j^k.$$

With $q_j \leq p + \Delta \leq 3p/2$, $\lambda_j \geq (1-p)(1-3p) \geq 1 - 4p$ for $p \leq 1/16$. For $k \leq 1/(16p)$, $(1-4p)^k \geq \exp(-4pk/(1-4p)) \geq \exp(-1/3) > 1/2$. Hence for $p \leq 1/16$ and $k \leq 1/(16p)$,

$$\big\|\delta_{(j,+1)}\,P_\theta^k - u\big\|_{\mathrm{TV}} \;>\; 1/4.$$

For $p \in [1/16, 1/4)$, the trivial bound $t_{\mathrm{mix}} \geq 1$ implies $t_{\mathrm{mix}} \geq 1/(16p)$. Combining the upper and lower bounds,

$$1/(16p) \;\leq\; t_{\mathrm{mix}} \;\leq\; 3/p.$$

Step 3: Poisson equation and linear realizability. Since $d^\theta = u$, the average cost is $\rho(\theta) = \mathbb{E}_u[c] = 1/2$. The centered Poisson equation $\rho + h^\theta = c + P_\theta h^\theta$ with $\sum_s u(s)h^\theta(s) = 0$ admits the ansatz $h^\theta(j,\sigma) = \beta_j(\theta)\,\sigma$ (the centering is automatic since $\sum_\sigma \sigma = 0$). Computing $(P_\theta h^\theta)(j,\sigma) = (1-p)(1-2q_j)\,\beta_j\,\sigma$ (the reset term vanishes since $\mathbb{E}_u[h^\theta] = 0$) and substituting,

$$\tfrac{1}{2} + \beta_j\,\sigma \;=\; \tfrac{1+\sigma}{2} \;+\; (1-p)(1-2q_j)\,\beta_j\,\sigma \quad\Longrightarrow\quad \beta_j(\theta) \;=\; \frac{1}{2\big(p + 2q_j(1-p)\big)}.$$

With $q_j \in [p, 3p/2]$, the denominator lies in $[5p, 8p]$, so $\beta_j(\theta) \in [1/(8p),\, 1/(5p)] = \Theta(1/p)$.

Since there is one action, $Q^\theta(j,\sigma,a) \;=\; h^\theta(j,\sigma) \;=\; \beta_j(\theta)\,\sigma \;=\; w_\theta^{\star\top}\phi(j,\sigma)$ with $w_\theta^\star := (\beta_1(\theta), \ldots, \beta_d(\theta))^\top$. The approximation error is exactly zero. The covariance matrix is

$$\Sigma \;=\; \mathbb{E}_{\nu^\theta}[\phi\phi^\top] \;=\; \frac{1}{2d}\sum_{j=1}^{d}\sum_{\sigma}(\sigma e_j)(\sigma e_j)^\top \;=\; \frac{1}{d}\,I_d,$$

giving $\|\phi\|_2 \leq 1$ ($L = 1$) and $\mu = 1/d$.

Step 4: Per-coordinate gap. Write $\beta(q) := 1/(2p + 4q(1-p))$. Since $1/\beta(p) = p(6-4p) \in [5p, 6p]$ and $1/\beta(p+\Delta) \in [5p, 8p]$ for $\Delta \leq p/2$,

$$|\beta(p+\Delta) - \beta(p)| \;=\; 4(1-p)\,\Delta\,\beta(p)\,\beta(p+\Delta) \;\geq\; \frac{3\,\Delta}{(6p)(8p)} \;=\; \frac{1}{16}\,\frac{\Delta}{p^2},$$

where we used $4(1-p) \geq 3$ for $p \leq 1/4$. Define $\delta_w := |\beta(p+\Delta) - \beta(p)| \geq \Delta/(16p^2)$.

Step 5: Per-query KL. Fix θ, θ' differing only in coordinate j, with $q_j = p$ and $q'_j = p + \Delta$. A query from any state (j', σ') with $j' \neq j$ has identical next-state law under both hypotheses (only q_j changes), so the per-query KL is zero.

For a query from (j, σ), the only changing atoms are (j, σ) and $(j, -\sigma)$:

$$\begin{aligned} P_q\big((j,\sigma),\,(j,\sigma)\big) &= P_{\text{stay}}(q) := (1-p)(1-q) + p/(2d), \\ P_q\big((j,\sigma),\,(j,-\sigma)\big) &= P_{\text{flip}}(q) := (1-p)\,q + p/(2d), \end{aligned}$$

By the standard bound of the KL divergence by the chi-square divergence,

$$\mathrm{KL}(P_q \,\|\, P_{q'}) \;\le\; (1-p)^2\,\Delta^2 \left(\frac{1}{P_{\text{stay}}(q')} + \frac{1}{P_{\text{flip}}(q')} \right).$$

For $p \le 1/4$ and $\Delta \le p/2$, $P_{\text{flip}}(q') \ge (1-p)p \ge 3p/4$ (so $1/P_{\text{flip}}(q') \le 4/(3p)$) and $P_{\text{stay}}(q') \ge (1-p)(1-p-\Delta) \ge (3/4)(5/8) = 15/32$, hence $1/P_{\text{stay}}(q') \le 32/15 \le 8/(15p)$ (using $p \le 1/4$, so $1/p \ge 4$). Combining,

$$\frac{1}{P_{\text{stay}}(q')} + \frac{1}{P_{\text{flip}}(q')} \;\le\; \left(\frac{8}{15} + \frac{4}{3} \right) \frac{1}{p} \;=\; \frac{28}{15\,p} \;<\; \frac{2}{p}.$$

Since $(1-p)^2 \le 1$,

$$\mathrm{KL}\big(P_\theta(\cdot \mid j,\sigma) \,\big\|\, P_{\theta'}(\cdot \mid j,\sigma)\big) \;\le\; 2\,\frac{\Delta^2}{p}. \tag{38}$$

Step 6: Assouad assembly. Suppose, for contradiction, an algorithm using at most N generative-model queries satisfies the success condition. Let $\mathbb{P}_\theta$ be the law under θ of the transcript $(U, S_1, Y_1, \ldots, S_N, Y_N, \widehat{w})$, where U is the algorithm's internal randomness, S_n is the queried state at round n, and Y_n is the observed next state. The query rule S_n is measurable with respect to $(U, S_1, Y_1, \ldots, S_{n-1}, Y_{n-1})$ and has the same conditional law under every θ; only the next-state law $P_\theta(\cdot \mid S_n)$ depends on θ. Hence the KL chain rule gives

$$\mathrm{KL}\big(\mathbb{P}_\theta \,\big\|\, \mathbb{P}_{\theta'}\big) \;=\; \sum_{n=1}^{N} \mathbb{E}_\theta\big[\mathrm{KL}\big(P_\theta(\cdot \mid S_n) \,\big\|\, P_{\theta'}(\cdot \mid S_n)\big)\big].$$

Letting $N_j := \#\{n : S_n \in \{j\} \times \{-1, +1\}\}$ be the number of queries to coordinate j (so $\sum_j N_j \le N$) and using (38) with zero cross-coordinate KL,

$$\mathrm{KL}(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\theta^{(j)}}) \;\le\; 2\,\frac{\Delta^2}{p}\,\mathbb{E}_\theta[N_j],$$

where $\theta^{(j)}$ is θ with bit j flipped. Since $\beta(q) = 1/(2p + 4q(1-p))$ is *decreasing* in q, we have $w^\star_{\theta,j} = \beta(p) > \beta(p+\Delta) = w^\star_{\theta^{(j)},j}$ when $\theta_j = 0$. Accordingly, define the bit decoder

$$\widehat{\theta}_j \;:=\; \mathbf{1}\{\widehat{w}_j < (\beta(p) + \beta(p+\Delta))/2\}.$$

If $\widehat{\theta}_j \neq \theta_j$, then $|\widehat{w}_j - w^\star_{\theta,j}| \ge \delta_w/2$, so each Hamming error contributes $\delta_w^2/(4d)$ to $\|\widehat{w} - w^\star_\theta\|_\Sigma^2 = (1/d)\|\widehat{w} - w^\star_\theta\|_2^2$:

$$d_H(\widehat{\theta}, \theta) \;\le\; \frac{4d}{\delta_w^2}\,\|\widehat{w} - w^\star_\theta\|_\Sigma^2. \tag{39}$$

Choose $\Delta := 128\,p^2\varepsilon_{\text{stat}}$. Then $\delta_w \ge \Delta/(16p^2) = 8\varepsilon_{\text{stat}}$, so $\delta_w^2 \ge 64\,\varepsilon_{\text{stat}}^2$. The constraint $\Delta \le p/2$ becomes $\varepsilon_{\text{stat}} \le 1/(256\,p)$, which is the stated regime. Under the success hypothesis $\Pr_\theta(\|\widehat{w} - w^\star_\theta\|_\Sigma \le \varepsilon_{\text{stat}}) \ge 2/3$ and using (39),

$$\mathbb{E}_\theta[d_H(\widehat{\theta}, \theta)] \;\le\; \tfrac{2}{3} \cdot \tfrac{4d\,\varepsilon_{\text{stat}}^2}{64\,\varepsilon_{\text{stat}}^2} + \tfrac{1}{3} \cdot d \;=\; \tfrac{2}{3} \cdot \tfrac{d}{16} + \tfrac{d}{3} \;=\; \tfrac{3d}{8}.$$

Averaging over $\theta \in \{0,1\}^d$, the average Hamming risk satisfies $R_H(\widehat{\theta}) \le 3d/8$.

We use the standard Hamming-risk form of Assouad's inequality [46]: pairing each θ with $\theta^{(j)}$ for each j and applying Pinsker plus Jensen,

$$R_H(\widehat{\theta}) \;\geq\; \frac{d}{2}\left[1 - \sqrt{\frac{1}{2d\,2^{d-1}}\sum_{j=1}^{d}\sum_{\theta:\theta_j=0}\mathrm{KL}(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\theta^{(j)}})}\right].$$

By the per-query KL bound and $\sum_j N_j \leq N$, $\sum_j \sum_{\theta:\theta_j=0} \mathbb{E}_\theta[N_j] \leq 2^d\, N$, so the expression under the square root is bounded by $2\,N\,\Delta^2/(d\,p)$, giving

$$R_H(\widehat{\theta}) \;\geq\; \frac{d}{2}\left[1 - \sqrt{2\,\frac{N\,\Delta^2}{d\,p}}\right].$$

If $N \leq d\,p/(128\,\Delta^2)$, the radicand is $\leq 1/64$, hence $R_H(\widehat{\theta}) \geq (d/2)(1-1/8) = 7d/16$. But we showed $R_H(\widehat{\theta}) \leq 3d/8 < 7d/16$, a contradiction. Therefore

$$N \;>\; \frac{d\,p}{128\,\Delta^2} \;=\; \frac{d\,p}{128\cdot 128^2\,p^4\,\varepsilon_{\mathrm{stat}}^2} \;=\; \frac{1}{2^{21}}\,\frac{d}{p^3\,\varepsilon_{\mathrm{stat}}^2}.$$

Using $1/p^3 \geq t_{\mathrm{mix}}^3/27$ from (37),

$$N \;\geq\; \frac{1}{2^{21}\cdot 27}\,\frac{d\,t_{\mathrm{mix}}^3}{\varepsilon_{\mathrm{stat}}^2}. \qquad \square$$

D.6 Sufficient conditions for the covariance condition of Lemma 5.7

For $\pi \in \Pi$, let $\Sigma_\pi := \mathbb{E}_{(s,a)\sim\nu^\pi}[\phi_{s,a}\phi_{s,a}^\top]$, so that $\Sigma_t = \Sigma_{\pi^t}$.

Spanning features. Suppose $\mathrm{span}\{\phi_{s,a_s} : s \in \mathcal{S}\} = \mathbb{R}^d$ for every $(a_s)_{s\in\mathcal{S}} \in \mathcal{A}^{\mathcal{S}}$. If $u^\top \Sigma_\pi u = \sum_s d^\pi(s)\sum_a \pi(a \mid s)\,(u^\top\phi_{s,a})^2 = 0$, then, since $d^\pi > 0$ (Lemma 3.2), $u^\top\phi_{s,a} = 0$ whenever $\pi(a \mid s) > 0$; choosing one such a_s in every state gives $u \perp \mathrm{span}\{\phi_{s,a_s}\}_s = \mathbb{R}^d$, so $u = 0$. Hence $\Sigma_\pi \succ 0$ for every $\pi \in \Pi$. The map $\pi \mapsto \Sigma_\pi$ is continuous, as $\pi \mapsto d^\pi$ is (see the proof of Lemma 3.2), and Π is compact, so $\mu := \min_{\pi\in\Pi}\lambda_{\min}(\Sigma_\pi) > 0$ and $\Sigma_\pi \succeq \mu I_d$ for all π. When $d \leq |\mathcal{S}|$, the spanning condition holds whenever the features are in general position, that is, whenever any d of them are linearly independent, and hence for almost every choice of features.

ε-exploration. Without the spanning condition, enforce $\pi^t(a \mid s) \geq \varepsilon/|\mathcal{A}|$ as in Remark 5.5. Then $\nu^t(s,a) \geq \underline{d}\,\varepsilon/|\mathcal{A}|$, so $\Sigma_t \succeq (\underline{d}\,\varepsilon/|\mathcal{A}|)\sum_{s,a}\phi_{s,a}\phi_{s,a}^\top$, and the condition holds with $\mu = (\underline{d}\,\varepsilon/|\mathcal{A}|)\,\lambda_{\min}\big(\sum_{s,a}\phi_{s,a}\phi_{s,a}^\top\big)$, which is positive whenever $\mathrm{span}\{\phi_{s,a} : (s,a) \in \mathcal{S}\times\mathcal{A}\} = \mathbb{R}^d$.

Hard instances. The family of Proposition 5.9 has features $\phi(j,\sigma) = \sigma e_j$ of norm 1 and covariance $\Sigma = I_d/d$ (see the proof of Proposition D.5). Lemma 5.7 assumes $\|\phi_{s,a}\|_2 \leq L$ and $\Sigma_t \succeq \mu I_d$, which force $d\mu \leq \mathrm{Tr}(\Sigma_t) \leq L^2$. Multiplying the features of the family by $\sqrt{d\mu}$ and dividing $w_\theta^\star$ by $\sqrt{d\mu}$ gives features of norm $\sqrt{d\mu} \leq L$ and covariance μI_d, and leaves the differential Q-function unchanged; an estimate $\widehat{w}$ for the rescaled family corresponds to $\sqrt{d\mu}\,\widehat{w}$ for the original one, with the same error in the covariance norm. The lower bound therefore holds for every L and μ allowed by Lemma 5.7, so the condition does not make the estimation problem easier. Finally, a generative model simulates a single trajectory with one query per transition, so the lower bound also applies to the single-trajectory oracle of Lemma 5.7.

E Mixing time details

Under Assumption 3.1, for each fixed policy π the induced Markov chain with kernel P_π is finite-state, irreducible, and aperiodic, and therefore admits a unique stationary distribution d^π and converges to it at a geometric rate in total variation. In particular, every such π has a finite mixing time

$$t_{\mathrm{mix}}(\pi) \;:=\; \min\left\{k \geq 1 : \sup_{\mu_0\in\mathcal{P}(\mathcal{S})} \left\|\mu_0\,P_\pi^k - d^\pi\right\|_{\mathrm{TV}} \leq \tfrac{1}{4}\right\} \;<\; \infty. \tag{40}$$

A self-contained proof is given in Lemma E.1 below.

Lemma 4.5 upgrades (40) to a single constant t_{mix} that works *uniformly* over all policies $\pi \in \Pi$; its proof is given at the end of this appendix. The argument is non-constructive, so we first record a standard condition under which t_{mix} is explicit.

This condition is uniform minorization (see, e.g., Murthy et al. [30, Assumption 3.1]). One convenient MDP-level formulation is: there exist $\alpha \in (0,1)$ and a reference distribution $\nu \in \mathcal{P}(\mathcal{S})$ such that

$$P(s' \mid s,a) \;\geq\; \alpha\,\nu(s') \qquad \text{for all } s,s' \in \mathcal{S},\ a \in \mathcal{A}. \tag{41}$$

Since $P_\pi(\cdot \mid s) = \sum_a \pi(a \mid s)\,P(\cdot \mid s,a)$, (41) implies $P_\pi(\cdot \mid s) \geq \alpha\,\nu$ for every π and s, and consequently yields uniform mixing with an explicit bound on t_{mix}. Define the residual kernel $R_\pi(\cdot \mid s) := (P_\pi(\cdot \mid s) - \alpha\,\nu)/(1-\alpha)$; then $P_\pi = \alpha\,\mathbf{1}\,\nu + (1-\alpha)\,R_\pi$, and for any distributions μ_0, μ_0',

$$\|\mu_0 P_\pi - \mu_0' P_\pi\|_{\mathrm{TV}} \;=\; (1-\alpha)\,\|(\mu_0 - \mu_0')\,R_\pi\|_{\mathrm{TV}} \;\leq\; (1-\alpha)\,\|\mu_0 - \mu_0'\|_{\mathrm{TV}}.$$

Taking $\mu_0' = d^\pi$ and iterating yields $\|\mu_0\,P_\pi^k - d^\pi\|_{\mathrm{TV}} \leq (1-\alpha)^k$. Requiring $(1-\alpha)^k \leq 1/4$ gives $t_{\mathrm{mix}} \leq \lceil \log 4 / \log(1/(1-\alpha)) \rceil = O(1/\alpha)$ as $\alpha \to 0$.

The same condition also makes $\bar{\chi}$ explicit. Stationarity gives $d^\pi(s') = \sum_s d^\pi(s)\,P_\pi(s' \mid s) \geq \alpha\,\nu(s')$ for every π and s'. If ν has full support, the constant of Lemma 3.2 therefore satisfies $\underline{d} \geq \alpha \min_s \nu(s)$, and since $d^\star \leq 1$,

$$\sup_{t \geq 0} \chi_t \;\leq\; \frac{1}{\underline{d}} \;\leq\; \frac{1}{\alpha\,\min_s \nu(s)},$$

so $\bar{\chi} = 1/(\alpha \min_s \nu(s))$ is a valid choice in Lemma 3.7(i). Under (41), both t_{mix} and $\bar{\chi}$ are thus explicit.

Lemma E.1 (Finite mixing time for a fixed policy)**.** *Let π be a policy such that P_π is irreducible and aperiodic. Then $t_{\mathrm{mix}}(\pi)$ defined by (40) is finite.*

Proof. Fix π. Since P_π is finite-state, irreducible, and aperiodic, it is *primitive*; hence there exists an integer $m_\pi \geq 1$ such that $P_\pi^{m_\pi}(s' \mid s) > 0$ for all $s, s' \in \mathcal{S}$ (see Levin and Peres [24, Proposition 1.7]). Let

$$\varepsilon_\pi \;:=\; \min_{s,s' \in \mathcal{S}} P_\pi^{m_\pi}(s' \mid s) \;>\; 0, \qquad \alpha_\pi \;:=\; \min\{|\mathcal{S}|\,\varepsilon_\pi,\, 1/2\} \in (0, 1/2],$$

and let ν be the uniform distribution on $\mathcal{S}$. Then $P_\pi^{m_\pi}(\cdot \mid s) \geq \alpha_\pi\,\nu$ for every s. Defining the residual kernel $R_\pi(\cdot \mid s) := (P_\pi^{m_\pi}(\cdot \mid s) - \alpha_\pi\,\nu)/(1-\alpha_\pi)$, we obtain $P_\pi^{m_\pi} = \alpha_\pi\,\mathbf{1}\,\nu + (1-\alpha_\pi)\,R_\pi$, and by the same non-expansive argument as above,

$$\sup_{\mu_0} \left\|\mu_0\,P_\pi^{q\,m_\pi} - d^\pi\right\|_{\mathrm{TV}} \;\leq\; (1-\alpha_\pi)^q \qquad \forall\, q \geq 0.$$

Choosing q large enough that $(1-\alpha_\pi)^q \leq 1/4$, namely $q \geq \lceil \log 4 / \log(1/(1-\alpha_\pi)) \rceil$, yields $t_{\mathrm{mix}}(\pi) \leq q\,m_\pi < \infty$. □

Proof of Lemma 4.5, first part. For $\pi \in \Pi$ and $k \geq 1$, write $\mathrm{TV}_\pi(k) := \sup_{\mu_0 \in \mathcal{P}(\mathcal{S})} \|\mu_0\,P_\pi^k - d^\pi\|_{\mathrm{TV}}$. We use three facts. *(i) Monotonicity.* Since d^π is stationary and P_π is non-expansive in total variation, $\mathrm{TV}_\pi(k+1) \leq \mathrm{TV}_\pi(k)$. *(ii) Continuity.* For fixed k, the map $\mu_0 \mapsto \|\mu_0\,P_\pi^k - d^\pi\|_{\mathrm{TV}}$ is convex, so its supremum over the simplex is attained at a vertex, $\mathrm{TV}_\pi(k) = \max_{s \in \mathcal{S}} \|P_\pi^k(\cdot \mid s) - d^\pi\|_{\mathrm{TV}}$. Both $\pi \mapsto P_\pi^k$ and $\pi \mapsto d^\pi$ are continuous on Π (the latter is shown in the proof of Lemma 3.2), so $\pi \mapsto \mathrm{TV}_\pi(k)$ is continuous as a finite maximum of continuous functions. *(iii) Pointwise decay.* For each π, the proof of Lemma E.1 gives $\mathrm{TV}_\pi(q\,m_\pi) \leq (1-\alpha_\pi)^q \to 0$, so there is $k_\pi \geq 1$ with $\mathrm{TV}_\pi(k_\pi) < 1/4$.

By (ii), $U_\pi := \{\pi' \in \Pi : \mathrm{TV}_{\pi'}(k_\pi) < 1/4\}$ is an open neighborhood of π. Since $\Pi = \mathcal{P}(\mathcal{A})^{\mathcal{S}}$ is compact, finitely many $U_{\pi_1}, \ldots, U_{\pi_n}$ cover it. Let $t_{\mathrm{mix}} := \max_i k_{\pi_i}$. Any $\pi' \in \Pi$ lies in some U_{π_i}, and (i) gives $\mathrm{TV}_{\pi'}(t_{\mathrm{mix}}) \leq \mathrm{TV}_{\pi'}(k_{\pi_i}) < 1/4$. □

Proof of (11). With $\mathrm{TV}_\pi(k)$ as in the proof above, let $\overline{\mathrm{TV}}_\pi(k) := \sup_{\mu_0,\mu_0' \in \mathcal{P}(\mathcal{S})} \|\mu_0 P_\pi^k - \mu_0' P_\pi^k\|_{\mathrm{TV}}$. Then $\mathrm{TV}_\pi \le \overline{\mathrm{TV}}_\pi \le 2\,\mathrm{TV}_\pi$, and $\overline{\mathrm{TV}}_\pi$ is submultiplicative, $\overline{\mathrm{TV}}_\pi(k+l) \le \overline{\mathrm{TV}}_\pi(k)\,\overline{\mathrm{TV}}_\pi(l)$ [24, Lemmas 4.10–4.11]. The first part of Lemma 4.5 gives $\overline{\mathrm{TV}}_\pi(t_{\mathrm{mix}}) \le 2\,\mathrm{TV}_\pi(t_{\mathrm{mix}}) \le 1/2$, so $\mathrm{TV}_\pi(\ell\, t_{\mathrm{mix}}) \le \overline{\mathrm{TV}}_\pi(t_{\mathrm{mix}})^\ell \le 2^{-\ell}$ for every integer $\ell \ge 0$. Since TV_π is nonincreasing by (i), $\mathrm{TV}_\pi(k) \le 2^{-\lfloor k/t_{\mathrm{mix}} \rfloor}$ for all $k \ge 0$, which is (11). The factor $1/2$ per block, rather than the $1/4$ at $k = t_{\mathrm{mix}}$, is the price of a single bound valid for every k.

Extension to unichain MDPs. Assumption 3.1 can be weakened to the unichain condition (each induced kernel P_π has a unique recurrent class) while retaining aperiodicity, at the cost of additional constants, and the mechanism is visible in our analysis. *For the convergence results*, whenever the iterates remain interior (automatic under KL, where the update is multiplicative), each path mismatch χ_t stays finite. What is lost is the a priori uniform bound $\chi_t \le \underline{d}^{-1}$ of Lemma 3.2, so a uniform path bound $\bar{\chi}$ must be imposed or verified rather than inherited; once it is supplied, the recursion of Lemma 3.7 applies verbatim. *For the sample bounds*, uniform mixing is the assumption of the mixing unichain setting of Li et al. [25, Assumption 1], who note that aperiodicity ensures it. The proof of Lemma 4.5 indeed carries over when every P_π, $\pi \in \Pi$, is unichain with an aperiodic recurrent class, with two changes. For continuity, every v with $vP_\pi = v$ satisfies $v = vP_\pi^k \to (v\mathbf{1})\, d^\pi$ by the pointwise decay, so the linear system in the proof of Lemma 3.2 still has the unique solution d^π and is therefore nonsingular, although d^π may now vanish on transient states. For the pointwise decay, every state reaches the recurrent class, which is aperiodic, so $P_\pi^{m_\pi}(s' \mid s) > 0$ for some $m_\pi \ge 1$, every $s \in \mathcal{S}$ and every s' in the recurrent class, and the proof of Lemma E.1 applies with ν uniform on that class. With t_{mix} also absorbing the time to reach the recurrent class, the sample bounds are stated exactly as before. In both halves the change is in constants, not in the form of the results. By contrast, the multichain case (some induced kernel P_π has more than one recurrent class) requires additional analysis, as in Lee and Ryu [22].

F Numerical experiments

Code availability. All code for reproducing the experiments in this appendix is available at

`https://github.com/enesarda22/avg-reward-pmd-experiments`

We report three experiments. The first two (Sections F.2 and F.3) are proof-of-concept diagnostics on a shared three-state MDP, targeting Corollary 4.3 and Theorem 4.2 respectively: they show that the asymptotic rate is governed by the tail mismatch χ rather than the worst-case $\bar{\chi}$, and that the error floor is independent of the step-size schedule, empirically realizing the decoupling argument of Section 3. The third (Section F.4) extends this story to four diverse MDPs, showing that the $\chi \ll \bar{\chi}$ separation is broadly observed in practice. All experiments run on a single CPU laptop in under five minutes total.

F.1 Setup

MDP family. We use a three-state ergodic average-*cost* MDP with $\mathcal{S} = \{0,1,2\}$ and $\mathcal{A} = \{0,1\}$. The transition kernel is

$$\begin{aligned} P(0,0) &= (1-p,\,0,\,p), & P(0,1) &= (0,\,1-p,\,p),\\ P(1,0) &= (0,\,1-p,\,p), & P(1,1) &= (1-p,\,0,\,p),\\ P(2,0) &= (0.19,\,0.80,\,0.01), & P(2,1) &= (0.95,\,0.04,\,0.01), \end{aligned}$$

with $p = 0.05$. The small self-loop at state 2 makes the chain aperiodic under all policies. Costs are $c(2,0) = 0$, $c(2,1) = 0.2$, and $c(s,a) = 0$ for $s \in \{0,1\}$ in the *multiple-minimizer* instance; we also consider a *unique-minimizer* instance obtained by adding a tie-breaker $\varepsilon > 0$ via $c_{\text{unique}}(0,0) = \varepsilon$, $c_{\text{unique}}(1,1) = \varepsilon$ with the other entries unchanged. The initial policy is nearly deterministic and placed at high probability on the costly action in state 2: $\pi^0(0 \mid 0) = \pi^0(1 \mid 1) = 0.999$, $\pi^0(1 \mid 0) = \pi^0(0 \mid 1) = 0.001$, $\pi^0(0 \mid 2) = 0.010$, $\pi^0(1 \mid 2) = 0.990$.

Algorithm. We run KL-PMD,

$$\pi^{t+1}(a \mid s) \;\propto\; \pi^t(a \mid s)\exp\big(-\eta_t\, \widehat{Q}^t(s,a)\big),$$

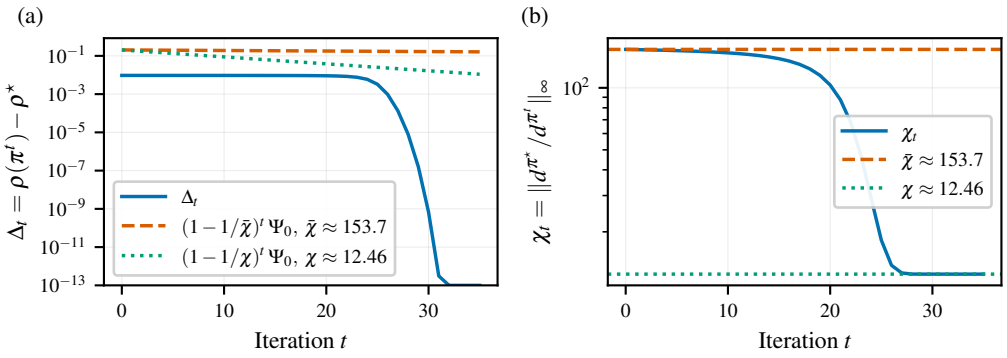

Figure 2: Exact KL-PMD on the multiple-minimizer instance. (a) Suboptimality Δ_t with the non-asymptotic upper bound $(1-1/\bar{\chi})^t\Psi_0$ (Lemma 3.7(i)) and asymptotic-rate reference $(1-1/\chi)^t\Psi_0$ (Lemma 3.7(ii)) overlaid. (b) Mismatch χ_t with horizontal references at $\bar{\chi} \approx 153.7$ and the tail estimate $\chi \approx 12.5$. The observed contraction is strictly tighter than both theoretical rates.

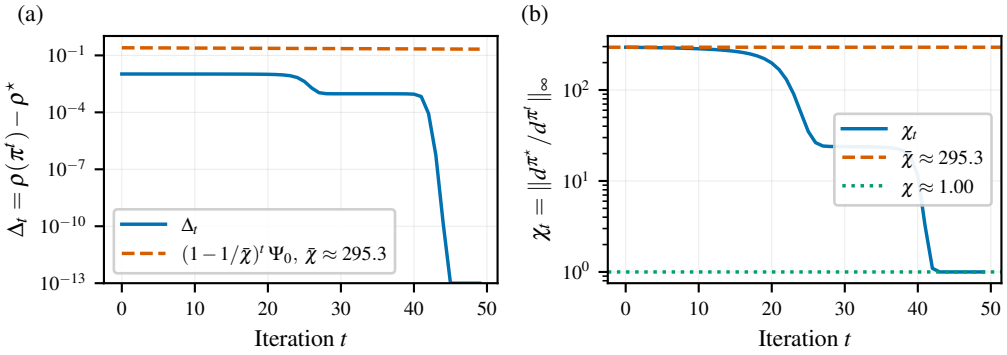

Figure 3: Exact KL-PMD on the unique-minimizer instance ($\varepsilon = 10^{-3}$). Same diagnostics as Figure 2. Here $\chi_t \to 1$ (tail $\chi \approx 1$) so the asymptotic reference collapses to zero and is omitted from panel (a); the observed trajectory exhibits the two-phase convergence predicted by Corollary 4.3, passing through an intermediate plateau at the tie-breaker scale before collapsing below 10^{-12}.

where $\widehat{Q}^t$ is either the exact differential action-value Q^{π^t} (Section F.2) or a noisy approximation (Section F.3). At each iteration we compute Q^{π^t} exactly by solving the Poisson equation for h^{π^t} in (16).

Diagnostics. We track the suboptimality gap $\Delta_t := \rho(\pi^t) - \rho^\star$ and the stationary-distribution mismatch $\chi_t := \|d^{\pi^\star}/d^{\pi^t}\|_\infty$ with $\pi^\star$ a fixed reference optimal policy; the "tail" estimate of χ is the median of χ_t over the last 20% of the run. On suboptimality plots we overlay the non-asymptotic upper bound $(1-1/\bar{\chi})^t\Psi_0$ from Lemma 3.7(i) and the asymptotic-rate reference $(1-1/\chi)^t\Psi_0$ from Lemma 3.7(ii); the prefactor $\Psi_0 = \Delta_0 + D_0^\star/(\eta_{-1}\bar{\chi})$ is the initial Lyapunov value, with $D_0^\star = D^{d^\star}(\pi^\star, \pi^0)$ the state-weighted KL divergence and $\eta_{-1} = \eta_0/b$ the largest step size consistent with the non-asymptotic hypothesis at $t = 0$.

F.2 Exact PMD: the asymptotic rate is governed by χ, not $\bar{\chi}$

The step-size schedule is the *mixed super-geometric*

$$\eta_t \;=\; \eta_0\, b^{\,t}\, \alpha^{\,t^2}, \qquad \eta_0 = 0.1, \quad b = 1.05, \quad \alpha = 1.005, \tag{42}$$

whose ratio $\eta_t/\eta_{t-1} = b\,\alpha^{2t-1}$ is bounded below by b for all $t \geq 1$, with the manual initialization $\eta_{-1} := \eta_0/b$ giving $\eta_0/\eta_{-1} = b$ at $t = 0$ (satisfying the floor condition of Lemma 3.7(i)) and

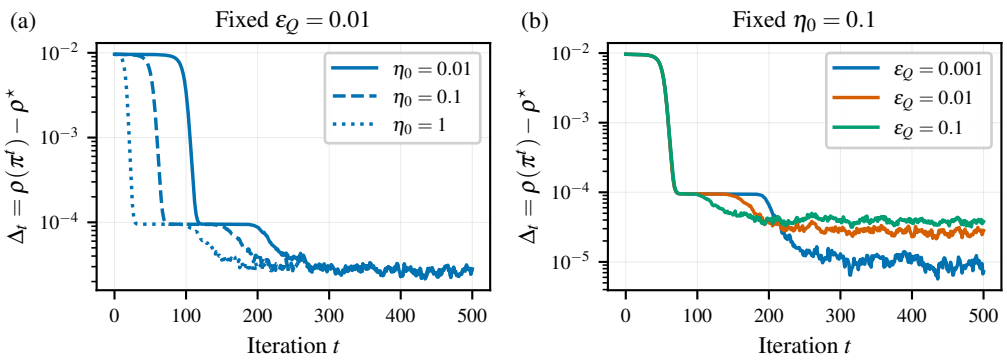

Figure 4: Inexact KL-PMD under uniform critic noise, on the unique-minimizer instance with tie-breaker $\varepsilon = 10^{-4}$; each curve is the mean of $N = 100$ seeds. (a) Step-size independence. At fixed $\varepsilon_Q = 10^{-2}$, three geometric schedules ($b = 1.05$, $\eta_0 \in \{10^{-2}, 10^{-1}, 1\}$) converge at very different rates but to essentially the same tail floor. (b) Floor depends on critic error. At fixed $\eta_0 = 10^{-1}$, sweeping $\varepsilon_Q \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ gives monotonically ordered tail floors.

diverges as $t \to \infty$ (satisfying Lemma 3.7(ii)). A plain geometric schedule $\eta_t = \eta_0 b^t$ has constant ratio and only satisfies the former.

Figures 2 and 3 illustrate the two regimes of Corollary 4.3. In the multiple-minimizer instance (Figure 2), PMD quickly collapses onto action 0 in state 2, but the action choices in states 0 and 1 are unconstrained at optimality, so the iterates settle at some optimum that need not match the reference $\pi^\star$ used to compute χ_t. Consequently χ_t stabilizes at a tail value $\chi \approx 12.5$, much smaller than the transient worst case $\bar{\chi} \approx 153.7$, and the observed Δ_t decays strictly below both the non-asymptotic and the asymptotic reference lines. In the unique-minimizer instance (Figure 3), the tie-breaker forces $\pi^t \to \pi^\star$ and hence $\chi_t \to 1$, driving the asymptotic rate $1 - 1/\chi \to 0$: Corollary 4.3's superlinear regime. Panel (a) exhibits the corresponding two-phase convergence, with Δ_t first settling at the tie-breaker scale and then collapsing superlinearly once χ_t approaches 1. In both instances the observed rate is at least as fast as the asymptotic rate predicts.

F.3 Inexact PMD: the error floor is independent of the step size

We now replace the exact differential value with $\widehat{Q}^t(s,a) = Q^{\pi^t}(s,a) + \xi^t(s,a)$, where $\xi^t(s,a) \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([-\varepsilon_Q, \varepsilon_Q])$ so that $\|\widehat{Q}^t - Q^{\pi^t}\|_\infty \le \varepsilon_Q$, exactly satisfying Assumption 4.1. The step-size schedule is the plain geometric $\eta_t = \eta_0 b^t$ with $b = 1.05$ (above the floor $\bar{\chi}/(\bar{\chi}-1) \approx 1.003$ required by Lemma 3.7(i)). We use the unique-minimizer instance with the tie-breaker reduced to $\varepsilon = 10^{-4}$ so that all three noise regimes considered below are non-trivial.

Figure 4(a) directly addresses the concern flagged in Section 3: aggressive step sizes do not amplify critic errors at the floor. Fixing $\varepsilon_Q = 10^{-2}$ and sweeping the initial step size η_0 across two orders of magnitude, the three trajectories enter the rate-dominated phase at very different iterations but settle to an indistinguishable tail value around 3×10^{-5}. Step size controls the rate, not the floor. Figure 4(b) fixes η_0 and sweeps ε_Q across the same range; the tail floors are monotonically ordered by the noise magnitude, consistent with Theorem 4.2. The separation is modest because the three-state MDP admits a deterministic optimum: once the policy is sharply concentrated, the only pathway for critic noise to contribute to Δ_t is by flipping the tie-broken actions in states 0 and 1, a discrete event whose frequency saturates once ε_Q exceeds the tie-breaker scale. The theoretical upper bound $2(\bar{\chi}+1)\varepsilon_Q$ is loose for this small MDP but its qualitative dependence on ε_Q is visible.

F.4 Empirical $\chi \ll \bar{\chi}$ across diverse MDPs

The 3-state experiments of Sections F.2 and F.3 dissect the χ-vs-$\bar{\chi}$ separation in a controlled setting where the Poisson equation, the optimal policy, and both rate constants are computable in closed form. This subsection shows that the same separation arises across considerably more diverse problems. The figure referenced in the main body (Figure 1) is the result of the experiment described here.

MDP setups. We use four MDP families, chosen to span structural variety. Three are ergodic, so χ_t is finite by Lemma 3.2. Access-control is unichain and so also illustrates the extension of Appendix E: the PMD iterates keep every action probability positive, which keeps χ_t finite, and the path bound $\bar{\chi}$ is verified along each run, as that extension requires.

- *Random Dirichlet MDPs* ($|\mathcal{S}| = 20$, $|\mathcal{A}| = 4$): for each (s, a), the next-state distribution $P(\cdot \mid s, a)$ is sampled from Dirichlet(0.1) (yielding full-support kernels) and the cost $c(s, a)$ is sampled uniformly on $[0, 1]$. We sample 80 such MDPs and run PMD on each from the worst-cost policy, which plays the most costly action in every state, with zero probabilities floored at 10^{-300} so that the KL update and $D_0^\star$ are well defined. The variability across the ensemble then comes from the sampled MDPs themselves. This adversarial start gives these well-mixing MDPs a sizeable transient mismatch; it is also why their trajectories stay flat for the first ~ 20 iterations, the time the growing step sizes need to lift the floored probabilities.
- *Access-control queuing* [41, Example 10.2]: 10 identical servers, customers with priorities $\{1, 2, 4, 8\}$ arriving uniformly, each busy server completes with probability 0.06 per step; the agent admits or rejects each arrival. We use a single fixed instance and run PMD from 80 random Dirichlet(0.5) initial policies.
- 5×5 *GridWorld*: a slip-prone gridworld (slip probability 0.1) in cost-minimization form, with the goal-state transition acting as a teleport back to the start so the chain is ergodic. We use a single fixed instance and run PMD from 80 random Dirichlet(0.5) initial policies.
- *Jack's car rental* [41, Example 4.2]: the canonical two-location car-rental MDP with 441 states and 11 overnight-move actions, instantiated via the `gym-classics` `JacksCarRental-v0` environment [9]. The Poisson rentals/returns make every transition full-support, so the chain is ergodic under every policy without any smoothing. We run PMD from 80 random Dirichlet(0.5) initial policies.

Schedule and algorithm. All four ensembles use the mixed super-geometric schedule (42) with $\eta_0 = 10$, $b = 1.1$ and $\alpha = 1.001$, again with $\eta_{-1} := \eta_0/b$. Its ratio η_t/η_{t-1} diverges, as Lemma 3.7(ii) requires, and is at least b, so, as with doubling (Section 3.2), Lemma 3.7(i) holds with $\bar{\chi}$ replaced by $\max\{b/(b-1), \bar{\chi}\} = \max\{11, \bar{\chi}\}$; the ensembles below have median $\bar{\chi}$ between 14 and 372. The KL-PMD update is computed in log space, so the large late steps ($\eta_{60} \approx 1.1 \times 10^5$) cause no overflow. We use the exact differential value Q^{π^t} at every iteration via the Poisson solve described in Section F.1, so this is the exact-PMD setting of Corollary 4.3 (no critic noise).

Diagnostic. For each run we record the stationary-distribution mismatch $\chi_t = \|d^{\pi^\star}/d^{\pi^t}\|_\infty$ over $T = 60$ iterations and report the *normalized trajectory* $\chi_t/\bar{\chi}$, where $\bar{\chi} = \max_{0 \le t \le T} \chi_t$ is the per-run peak. The tail value χ of a run is the median of χ_t over its final 12 iterations. This normalization makes the separation between $\bar{\chi}$ (the constant driving the non-asymptotic rate) and the asymptotic constant χ visually comparable across MDPs of very different absolute scales. Figure 1 plots the per-ensemble median of $\chi_t/\bar{\chi}$ with the inter-quartile range shaded.

Result. Across all four families the median trajectory collapses by roughly 10–1000$\times$ within the first $\sim$ 20–40 iterations and remains essentially flat thereafter. The transient $\bar{\chi}$ regime that drives Theorem 4.2's non-asymptotic bound is therefore a brief warm-up across problem types of widely varying state-action sizes, transition structures, and optimal-policy geometry; the asymptotic regime of Corollary 4.3, where χ governs convergence, is the operationally relevant one.

Per-MDP convergence and mismatch. Figure 5 expands the body figure into the diagnostic format of Figures 2 and 3: for each MDP family we plot the median suboptimality Δ_t (panel (a)) and the median mismatch χ_t (panel (b)), with the inter-quartile range over the ensemble shaded in both. Panel (a) overlays the non-asymptotic rate reference $(1 - 1/\bar{\chi})^t \Psi_0$ (Lemma 3.7(i)) and the asymptotic-rate reference $(1 - 1/\chi)^t \Psi_0$ (Lemma 3.7(ii)), with $\Psi_0 := \Delta_0 + D_0^\star/(\eta_{-1}\bar{\chi})$ the initial Lyapunov value, evaluated at the ensemble-median $\bar{\chi}$ and tail χ (the asymptotic reference is omitted when $\chi \approx 1$ collapses it to zero, the superlinear regime of Corollary 4.3); panel (b) overlays horizontal references at the same $\bar{\chi}$ and tail χ values. The observed Δ_t is below the non-asymptotic reference in every panel. The asymptotic reference must be read differently: Corollary 4.3 is a limit statement, $\limsup_t \Delta_t^{1/t} \le 1 - 1/\chi$, which constrains the tail slope of Δ_t and asserts nothing at any finite t. In three of the four families $\chi \approx 1$, and the median Δ_t reaches the numerical floor of 10^{-15} within

40 iterations. In GridWorld, where $\chi \approx 7.5$, the median Δ_t stays below the asymptotic reference throughout and reaches the floor at $t = 58$, after which floating-point error dominates. In all four cases the empirical decay is significantly faster than the worst-case rate of Theorem 4.2.

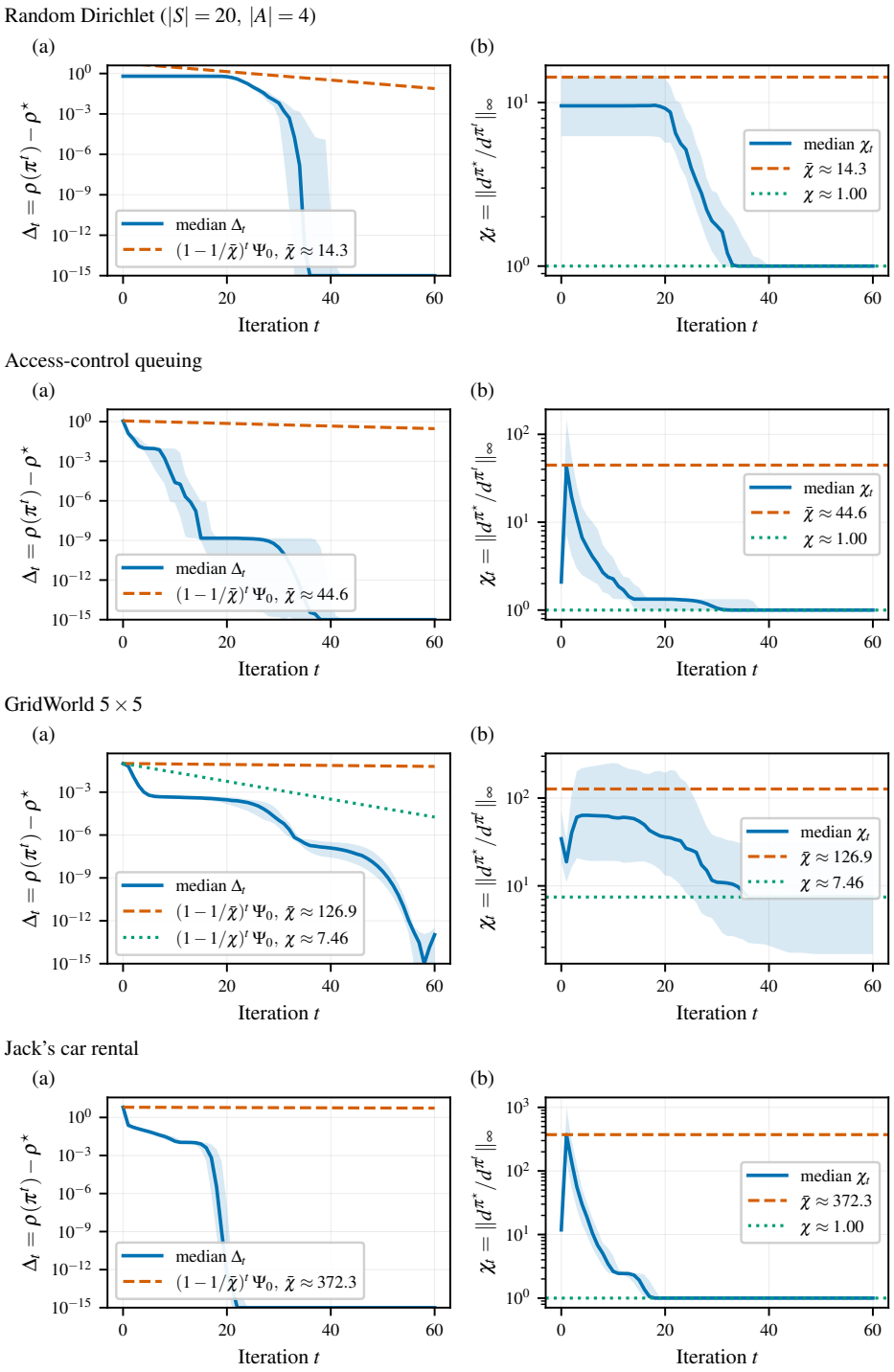

Figure 5: Per-MDP diagnostics for the rate-gap experiment (Section F.4): panel (a) median Δ_t with the non-asymptotic rate reference $(1 - 1/\bar{\chi})^t \Psi_0$ (Lemma 3.7(i)) and the asymptotic-rate reference $(1 - 1/\chi)^t \Psi_0$ (Lemma 3.7(ii)), both at ensemble-median values, where $\Psi_0 := \Delta_0 + D_0^\star/(\eta_{-1}\bar{\chi})$. The asymptotic reference is a limit statement plotted from $t = 0$, not a finite-time bound. Panel (b) median χ_t with horizontal references at $\bar{\chi}$ and tail χ. IQR shaded.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the bulleted Contributions paragraph in Section 1 name the master recursion (Lemma 3.7), linear/superlinear convergence in three settings (Theorems 4.2, 5.6 and Corollary 4.3), end-to-end sample complexities (Theorems 4.9, 5.10), and matching lower bounds (Propositions 4.8, 5.9); each is stated and proved in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 2 positions our analysis as ergodic and tabular/LFA, and Section 6 (Conclusion) explicitly names the scope limitations carried forward as future work: extending the ergodic analysis to the multichain setting and regret guarantees. The ergodicity requirement (Assumption 3.1) is stated where it is used.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumption 3.1 is a standing assumption, stated once in Section 3; every theorem, lemma, corollary, and proposition states all its other assumptions explicitly in the statement. Complete proofs appear in Appendices B (preliminaries), C (tabular case), D (LFA case), and E (mixing), with cross-references to the result being proved.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix F specifies every MDP family used (Sections F.1 and F.4), the step-size schedules and their parameters, the initial-policy distributions, ensemble sizes, and diagnostic definitions; the full code is also linked at the top of Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A repository link is given at the top of Appendix F. The repository contains a README with figure-to-script mappings.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section F.1 (3-state experiments) and Section F.4 (rate-gap experiments) specify the MDP construction, step-size schedule and parameters (η_0, b, α), initial-policy distribution, horizon, ensemble size, and diagnostic for every figure.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figures 1 (body) and 5 (appendix) plot per-MDP median trajectories together with the inter-quartile range computed empirically over the 80 runs of each ensemble (Section F.4).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As noted at the end of Appendix F's preamble, all experiments run on a single CPU laptop in under five minutes total; no GPU or cluster is required.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is purely theoretical with empirical illustrations on synthetic and standard textbook MDPs; no human subjects, scraped data, or potentially harmful artifacts are involved.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The contribution is foundational convergence theory for an existing optimization method (policy mirror descent) on average-reward MDPs; we see no direct path to societal impact, positive or negative, beyond improvements to the general theory of reinforcement learning.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: No data, models, or other artifacts with misuse risk are released; the only release is reproducibility code for the numerical experiments.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The MDP benchmarks are credited via the textbook [41] (Examples 4.2 and 10.2, cited in Section F.4); the Jack's car rental implementation uses the open-source `gym-classics` package, also referenced in Section F.4.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new datasets, models, or benchmarks. The accompanying code repository solely reproduces the figures reported in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used as writing and coding assistants, including drafting and checking proofs; the authors verified all arguments, code, and citations. LLMs are not part of the core methodology or experimental design of this paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.