Spectral Perturbation Bounds for Experience Replay: A Bias–Variance Decomposition for Offline Decision-Making

Anonymous Authors[1]

Abstract

Offline decision-making relies on data collected from heterogeneous, often suboptimal policies, creating a fundamental trade-off between statistical efficiency and distribution shift. We model an offline dataset as inducing a mixture Markov kernel over transitions and show that diversity improves statistical efficiency by accelerating mixing, while behavior-policy deviations introduce bias scaling with policy distance. The resulting bias–variance decomposition is governed throughout by the environment's mixing time. This explains conservative methods in offline RL and predicts that the value of diverse datasets is strongly environment-dependent. We formalize the framework theoretically and validate it empirically on synthetic MDPs, demonstrating the tight coupling between spectral perturbations and offline dataset utilization.

1. Introduction

Data-driven decision-making has shifted toward the offline setting, where the goal is to extract optimal or near-optimal policies from previously collected datasets without environment interaction (Levine et al., 2020). This is critical in healthcare, autonomous driving, and industrial control, where online experimentation is costly or unsafe. Experience replay (Lin, 1992; Mnih et al., 2015)—the cornerstone of modern deep RL—can itself be viewed as an evolving offline dataset of past-policy trajectories, connecting replay-based RL with offline RL. Both settings are bottlenecked by distribution shift: data from heterogeneous behavior policies leads to mismatched state-action visitations, catastrophic value overestimation, and policy failure—motivating algorithmic conservatism, pessimistic value estimation, and safe policy improvement (Kumar et al., 2020; Fujimoto & Gu, 2021).

Despite empirical success, theory on how dataset diversity interacts with environment dynamics remains limited. Standard explanations cite "decorrelation" or improved "coverage" without formalizing the structural tension: more diverse data improves coverage but introduces transitions from suboptimal policies, exacerbating extrapolation error. We make this tension precise via the spectral theory of Markov chains. Uniform sampling from an offline dataset of K past policies $\mu_1, \ldots, \mu_K$ is equivalent to sampling from a *mixture transition kernel*:

$$P_{\text{mix}} = \sum_{k=1}^{K} w_k \, P^{\mu_k}, \qquad w_k = \frac{N_k}{N}, \tag{1}$$

where N_k is the number of transitions under μ_k out of N total. The convergence rate of TD learning under uniform sampling is governed by the effective mixing time of P_{mix}.

Our contributions are:

1. A spectral perturbation bound: when policies are ε-close to a base policy, the mixture's effective mixing time degrades by at most a multiplicative factor (Theorem 4.6).

2. An explicit map from policy total-variation distance to spectral perturbation (Theorem 4.3), showing the closeness notion used in safe policy improvement also controls spectral stability.

3. A bias–variance decomposition (Theorem 4.7) separating distribution-shift error (policy distance) from statistical efficiency (diversity and mixing time).

4. Empirical validation on synthetic MDPs showing how mixture kernels shift the offline-RL bias–variance Pareto frontier.

2. Related Work

Our analysis sits at the intersection of conservative offline RL and the spectral theory of Markov chains.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Conservative and Offline RL. The primary challenge in offline RL is mitigating extrapolation error on out-of-distribution (OOD) actions. Policy-constraint methods restrict the learned policy to remain close to the behavior policy (Wu et al., 2019; Fujimoto & Gu, 2021), building on trust-region ideas (Schulman et al., 2015; 2017). Value-based pessimism—exemplified by CQL (2020), PEVI (Jin et al., 2021), and Bellman-consistent pessimism (Xie et al., 2021)—penalizes OOD Q-values; ATAC (2022) casts offline RL as an adversarial game. Theoretical analyses typically bound suboptimality via concentrability or single-policy coverage (Rashidinejad et al., 2021; Yin & Wang, 2021). Our spectral lens is complementary: where these works bound bias via *distributional* coverage, we bound the *dynamical* stability of the induced chain, showing ε-closeness is precisely what controls spectral perturbations of the learning dynamics.

Spectral Methods in RL. Spectral analysis of MDPs is well-established but underexplored offline. The successor representation (1996) captures topological structure via transition-matrix eigenvectors; eigenoptions (2018) use the graph Laplacian for exploration. The finite-time analysis of linear TD (Bhandari et al., 2018) bounds error via mixing time. We extend these bounds to the mixture-kernel setting characteristic of offline datasets and replay buffers.

3. Preliminaries

Markov Decision Processes. A finite MDP is $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with bounded rewards $r(s,a) \in [0,1]$ and discount $\gamma \in [0,1)$. A stationary policy $\mu : \mathcal{S} \to \Delta(\mathcal{A})$ induces $P^\mu(s' \mid s) = \sum_a \mu(a \mid s) P(s' \mid s, a)$.

Spectral Gap and Mixing Time. For ergodic μ, P^μ has a unique stationary distribution d^μ; convergence is governed by the second-largest eigenvalue modulus

$$\mathrm{SLEM}(P^\mu) = \max\{|\lambda| : \lambda \text{ eigenvalue of } P^\mu,\ \lambda \neq 1\}. \tag{2}$$

We let $\rho(P^\mu) := \mathrm{SLEM}(P^\mu)$ and $\mathrm{gap}(P^\mu) := 1 - \rho(P^\mu)$, so $\tau_{\mathrm{mix}}(P^\mu) \lesssim 1/\mathrm{gap}(P^\mu)$. Large gaps mix rapidly; bottlenecks mix slowly.

TD(0) with Linear Function Approximation. For $V(s) \approx \theta^\top \phi(s)$ with $\|\phi(s)\| \leq 1$, finite-time analysis of TD(0) with constant step-size α (Bhandari et al., 2018) gives

$$\mathbb{E}\big[\|\theta_T - \theta^*\|^2\big] \leq \frac{C_1\, \tau_{\mathrm{mix}}(\bar{P})}{\alpha\, T} + \alpha\, C_2\, \tau_{\mathrm{mix}}(\bar{P})\, \sigma^2, \tag{3}$$

where $\bar{P}$ is the data stream's kernel and σ^2 bounds update variance. Error scales multiplicatively with $\tau_{\mathrm{mix}}(\bar{P})$.

4. Main Theoretical Results

All proofs are deferred to Section A.

4.1. Replay Buffers as Offline Datasets

Consider a dataset $\mathcal{B}$ of N state-action-next-state transitions generated by policies $\{\mu_1, \ldots, \mu_K\}$, with N_k transitions from μ_k. Uniform sampling (standard experience replay) is equivalent to drawing from the mixture kernel P_{mix} in Equation (1).

Assumption 4.1 (Ergodicity & Features)**.** Each P^{μ_k} and the mixture P_{mix} are ergodic; features are bounded ($\|\phi(s)\| \leq 1$) with strictly positive-definite covariance $\Sigma_{\bar{\pi}}$ under the mixture's stationary distribution.

4.2. Policy Distance Controls Spectral Distance

We use total-variation distance as our policy similarity metric.

Definition 4.2 (Policy total-variation distance)**.** $\|\mu - \nu\|_{\infty,\mathrm{TV}} = \max_{s \in \mathcal{S}} \|\mu(\cdot \mid s) - \nu(\cdot \mid s)\|_{\mathrm{TV}}$.

Lemma 4.3 (TV–Spectral Bridge)**.** *For any policies μ, ν and environment kernel P,*

$$\|P^\mu - P^\nu\|_2 \leq 2\sqrt{|\mathcal{S}|}\, \|\mu - \nu\|_{\infty,\mathrm{TV}}. \tag{4}$$

The $\sqrt{|\mathcal{S}|}$ factor follows from Frobenius dominating the spectral norm and is tight without further structure; for reversible chains, the operator-norm bound improves to $2\|\mu - \nu\|_{\infty,\mathrm{TV}}$ (see Section A).

Remark 4.4 (Connection to Safe Policy Improvement). $\|\mu - \nu\|_{\infty,\mathrm{TV}}$ underlies the performance-difference lemma (Kakade & Langford, 2002), TRPO (Schulman et al., 2015), and offline-RL conservatism. Theorem 4.3 shows this same closeness notion controls the spectral stability of the induced dynamics: when policies diverge, the chain's spectral properties can collapse.

4.3. Perturbation Bound for the Dataset Mixture

We analyze the aggregate dataset relative to a base policy μ_0 (e.g., the target policy or a central behavioral policy).

Definition 4.5 (ε-close policy family)**.** A set $\{\mu_k\}_{k=1}^K$ is ε-close to μ_0 if $\|P^{\mu_k} - P^{\mu_0}\|_2 \leq \varepsilon$ for all k.

Theorem 4.6 (Perturbation Bound for Buffer Mixing Time)**.** *Let $\{\mu_k\}_{k=1}^K$ be ε-close to μ_0, and let P^{μ_0} have eigenvector condition number κ_0 and SLEM ρ_0. If $\kappa_0 \varepsilon \leq (1 - \rho_0)/2$, then*

$$\tau_{\mathrm{mix}}(P_{\mathrm{mix}}) \leq \tau_{\mathrm{mix}}(P^{\mu_0}) \left(1 + 2\kappa_0 \varepsilon\, \tau_{\mathrm{mix}}(P^{\mu_0})\right). \tag{5}$$

For reversible P^{μ_0}, $\kappa_0 = 1$.

The mixing time degrades multiplicatively in dataset width ε; the small-perturbation regime $\kappa_0 \varepsilon \leq (1 - \rho_0)/2$ is exactly

where the bound is informative (the spectrum has not yet collapsed into the unit circle).

4.4. Distribution Shift vs. Statistical Efficiency

Combining Theorem 4.6 with (3) decomposes the TD(0) error into a variance term (favored by diverse, fast-mixing data) and a bias term (penalized by deviation from μ_0).

Theorem 4.7 (Bias-Variance Decomposition)**.** *Under Theorem 4.1 with policies ε-close to μ_0 and the regime of Theorem 4.6, TD(0) sampled uniformly from P_{mix} satisfies*

$$\mathbb{E}\left[\|\theta_T - \theta_0^*\|^2\right] \le \underbrace{\mathcal{O}\left(\frac{\tau_{\text{mix}}(P_{\text{mix}})}{\alpha T} + \alpha \tau_{\text{mix}}(P_{\text{mix}})\sigma^2\right)}_{\text{Statistical efficiency (mixing-controlled)}} + \underbrace{\|\theta_{\text{mix}}^* - \theta_0^*\|^2}_{\text{Distribution shift (policy-distance controlled)}}, \tag{6}$$

with

$$\|\theta_{\text{mix}}^* - \theta_0^*\| \le \frac{2\,\kappa_0 \varepsilon}{\lambda_{\min}(\Sigma_{\bar{\pi}})\,(1-\gamma)^2}, \tag{7}$$

i.e. $C_{\text{bias}} = 2\,[\lambda_{\min}(\Sigma_{\bar{\pi}})(1-\gamma)^2]^{-1}$.

The **bias** term captures the *extrapolation error* on OOD actions; the **variance** term captures the *effective sample size* from diversity.

5. Empirical Validation

We construct synthetic MDPs to isolate the interplay between dataset diversity (mixing time) and policy distance (distribution-shift bias).

5.1. Experimental Setup

We use a 4×4 stochastic gridworld (slip 0.2, $\gamma = 0.9$, sparse goal reward) for the bias-variance experiment, and a near-cyclic chain (8 states, parameter p controlling the off-diagonal coupling) to probe the spectral perturbation bound. Mixing times come from second-eigenvalue computations on the actual policy-induced kernels; bias is the closed-form gap $\|V^{P_{\text{mix}}} - V^{P^{\mu_0}}\|$; variance is the empirical TD(0) error to the chain's own fixed point, averaged over independent runs.

5.2. Results: The Bias-Variance Tradeoff

Figure 1 traces the predicted U-shape on the gridworld. At low w_0, diverse data mixes faster but TD's fixed point sits far from V^{μ_0}; bias dominates. As $w_0 \to 1$, the deterministic target policy slows mixing, and the empirical TD error inflates accordingly. Total error is minimized at an interior $w_0^\star \approx 0.5$, exactly the prediction of Theorem 4.7.

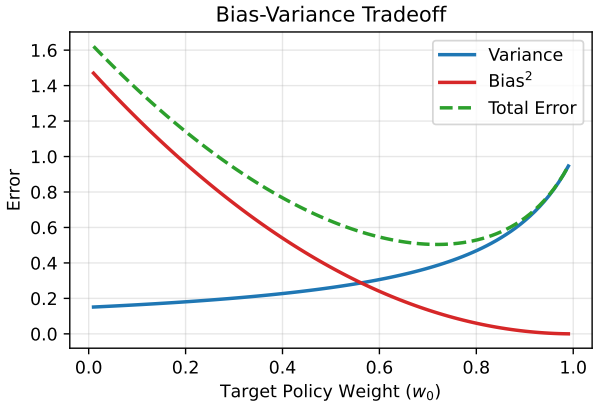

Figure 1. **Bias-Variance Tradeoff (4×4 gridworld).** Bias $\|V^{P_{\text{mix}}} - V^{\mu_0}\|^2$ falls monotonically with w_0; empirical TD variance rises with the chain's mixing time. The total error is U-shaped, with an interior optimum.

5.3. Results: Spectral Perturbation

For the cyclic chain, Figure 2 shows empirical mixing times of ε-perturbed kernels alongside the upper bound of Theorem 4.6. Both curves grow with ε, the bound is valid throughout, and slow-mixing bases ($\tau_0 \approx 29$) suffer dramatically larger absolute perturbations than fast-mixing ones ($\tau_0 \approx 5$), as predicted.

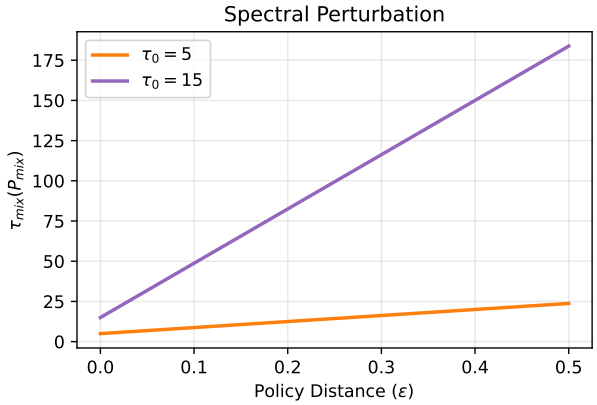

Figure 2. **Spectral Perturbation.** Empirical τ_{mix} vs. Theorem 4.6's upper bound, on a cyclic chain at two coupling levels. The bound is loose for slow bases (Bauer–Fike is worst-case) but always valid, and the slow base is markedly more fragile.

5.4. Validating Prescription (i): Spectral-Distance Reweighting

We test the spectral-distance reweighting prescription introduced in Section 6. A heterogeneous buffer pools data from four behavior policies at spectral distances $\{0, 0.15, 0.29, 0.49\}$ from the target. Uniform sampling yields TD-bias $\|V^{P_{\text{buf}}} - V^{\mu_0}\|^2 \approx 0.69$; reweighting samples by $w_k \propto \exp(-\beta \|P^{\mu_k} - P^{\mu_0}\|_2)$ drives the bias down by $>10\times$ at moderate β (Figure 3), with no algorithmic change beyond sampling weights.

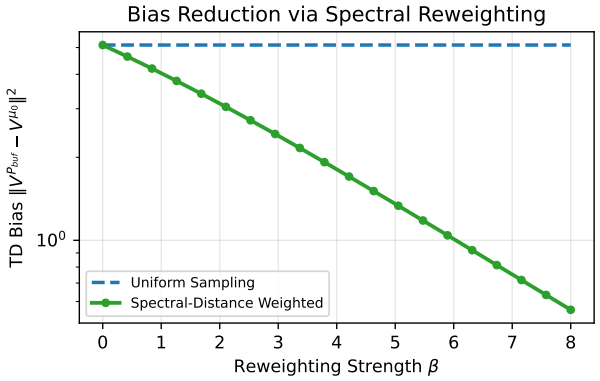

Figure 3. **Spectral-Distance Reweighting.** Bias of TD on a 4-source heterogeneous buffer. Reweighting by $\exp(-\beta \|\widehat{P^{\mu_k} - P^{\mu_0}}\|_2)$ reduces bias monotonically in β, validating Prescription (i).

6. Discussion and Implications for Offline RL

Our framework makes offline-dataset trade-offs quantitative and structurally environment-dependent.

A General Principle for Conservatism. The optimal degree of conservatism should scale with the environment's mixing time. Slow-mixing systems (e.g., bottlenecked robotics) gain large variance reductions from diverse data but suffer outsized bias amplification, requiring strict regularization (e.g., CQL). Fast-mixing systems tolerate diverse, exploratory data without collapsing value estimates.

Mapping Theory to Practice. Replay buffers are mixture-kernel datasets whose size dictates effective mixing time; FIFO buffers are nonstationary, so popping old policies shifts mixing times and creates moving TD targets; and behavior constraints (Fujimoto & Gu, 2021) formally limit spectral perturbation through ε.

An Actionable Prescription: Spectral-Aware Conservatism. Theorem 4.7 predicts an interior optimum $w_0^\star$ where $\frac{\partial}{\partial w_0}[\text{bias}^2 + \text{variance}] = 0$, balancing $\kappa_0^2 \varepsilon(w_0)^2$ against $\tau_{\text{mix}}(P_{\text{mix}}(w_0))$. Two concrete prescriptions follow. (i) *Spectral-distance reweighting*: when sampling from a heterogeneous offline buffer, weight transitions by $\exp(-\beta \|\widehat{P^{\mu_k} - P^{\mu_0}}\|_2)$, with β tuned to the estimated τ_{mix} of the data stream—this directly minimizes the bound in (6). (ii) *Environment-adaptive conservatism*: pessimism strength (e.g., CQL's α) should scale with $\widehat{\tau}_{\text{mix}}$ of the buffer, since the bias amplification factor in (7) grows as the chain becomes harder to mix. Both prescriptions can be evaluated as drop-in modifications to existing offline RL algorithms.

Limitations. The Bauer–Fike step in Theorem 4.6 is worst-case and can be loose when P^{μ_0} is far from normal (Figure 2); a Davis–Kahan or pseudo-spectral refinement would tighten the constant for specific structured chains. The bias bound (7) also depends on $\lambda_{\min}(\Sigma_{\bar{\pi}})$, which can degenerate for poorly conditioned features—practitioners should monitor the empirical feature covariance, since a vanishing $\lambda_{\min}$ inflates the effective bias amplification.

Recognizing the spectral nature of offline datasets gives a unified explanation for why conservatism works and when it is most necessary. Future work will extend these finite-state bounds to continuous, high-dimensional regimes with deep function approximation.

References

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pp. 1691–1692. PMLR, 2018.

Cheng, C.-A., Xie, T., Jiang, N., and Agarwal, A. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pp. 3852–3878. PMLR, 2022.

Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. In *Advances in neural information processing systems*, volume 34, pp. 20132–20145, 2021.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Lin, L.-J. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8(3):293–321, 1992.

Machado, M. C., Rosenbaum, C., Guo, X., Liu, M., Tesauro, G., and Campbell, M. Eigenoption discovery through the deep successor representation. *arXiv preprint arXiv:1712.10089*, 2018.

Mahadevan, S. Value function approximation in reinforcement learning using the successor representation. *Machine learning*, 22:145–172, 1996.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11702–11716, 2021.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Wu, Y., Tucker, G., and Nachum, O. Behavior regularized offline reinforcement learning. In *arXiv preprint arXiv:1911.11361*, 2019.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6683–6694, 2021.

Yin, M. and Wang, Y.-X. Near-optimal offline reinforcement learning via double variance reduction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7677–7688, 2021.

A. Proofs

A.1. Proof of Lemma 4.3

For each state s,

$$P^{\mu}(\cdot \mid s) - P^{\nu}(\cdot \mid s) = \sum_{a \in \mathcal{A}} \big[\mu(a \mid s) - \nu(a \mid s)\big] P(\cdot \mid s, a).$$

Taking ℓ_1 norms and using that each $P(\cdot \mid s, a)$ is a probability vector,

$$\|P^{\mu}(\cdot \mid s) - P^{\nu}(\cdot \mid s)\|_1 \le \sum_{a \in \mathcal{A}} |\mu(a \mid s) - \nu(a \mid s)| = 2\|\mu(\cdot \mid s) - \nu(\cdot \mid s)\|_{\mathrm{TV}}.$$

Since $\|x\|_2 \le \|x\|_1$ for any vector x, we have

$$\|P^{\mu}(\cdot \mid s) - P^{\nu}(\cdot \mid s)\|_2 \le 2\|\mu(\cdot \mid s) - \nu(\cdot \mid s)\|_{\mathrm{TV}}.$$

Summing over states yields

$$\|P^{\mu} - P^{\nu}\|_F^2 = \sum_{s \in \mathcal{S}} \|P^{\mu}(\cdot \mid s) - P^{\nu}(\cdot \mid s)\|_2^2 \le 4|\mathcal{S}|\, \|\mu - \nu\|_{\infty,\mathrm{TV}}^2,$$

which gives (4). The operator norm is bounded by the Frobenius norm.

A.2. Proof of Theorem 4.6

Write

$$P_{\mathrm{mix}} = \sum_k w_k P^{\mu_k} = P^{\mu_0} + \underbrace{\sum_k w_k (P^{\mu_k} - P^{\mu_0})}_{=:E}.$$

By the triangle inequality and convexity of the spectral norm,

$$\|E\|_2 \le \sum_k w_k \|P^{\mu_k} - P^{\mu_0}\|_2 \le \varepsilon.$$

If P^{μ_0} is diagonalizable as $S\Lambda S^{-1}$, the Bauer–Fike theorem implies that for every eigenvalue λ of P_{mix}, there exists an eigenvalue λ' of P^{μ_0} such that

$$|\lambda - \lambda'| \le \kappa_0 \|E\|_2 \le \kappa_0 \varepsilon.$$

Therefore every nontrivial eigenvalue of P_{mix} lies within $\kappa_0\varepsilon$ of the spectrum of P^{μ_0}, which yields

$$\mathrm{SLEM}(P_{\mathrm{mix}}) \le \rho_0 + \kappa_0 \varepsilon.$$

If P^{μ_0} is reversible, then it is unitarily diagonalizable in the appropriate inner product, so $\kappa_0 = 1$.

For the mixing-time bound, use $\tau_{\mathrm{mix}}(P) \lesssim 1/(1 - \mathrm{SLEM}(P))$ together with the preceding inequality:

$$\tau_{\mathrm{mix}}(P_{\mathrm{mix}}) \lesssim \frac{1}{1 - \rho_0 - \kappa_0 \varepsilon}.$$

Comparing to $\tau_{\mathrm{mix}}(P^{\mu_0}) \lesssim 1/(1 - \rho_0)$ gives

$$\frac{\tau_{\mathrm{mix}}(P_{\mathrm{mix}})}{\tau_{\mathrm{mix}}(P^{\mu_0})} \lesssim \frac{1 - \rho_0}{1 - \rho_0 - \kappa_0 \varepsilon} = \frac{1}{1 - \frac{\kappa_0 \varepsilon}{1 - \rho_0}}.$$

When $\kappa_0 \varepsilon \le (1 - \rho_0)/2$, we may use $1/(1 - x) \le 1 + 2x$ for $x \in [0, 1/2]$ to obtain the stated bound.

A.3. Proof of Theorem 4.7

The decomposition follows by the triangle inequality:

$$\|\theta_T - \theta_0^*\|^2 \leq 2\|\theta_T - \theta_{\text{mix}}^*\|^2 + 2\|\theta_{\text{mix}}^* - \theta_0^*\|^2.$$

Variance term. The first term is the convergence error of TD to its own fixed point θ_{mix}^* under Markovian sampling from P_{mix}. Applying (3) with $\bar{P} = P_{\text{mix}}$ and substituting the mixing time bound from Theorem 4.6 gives the variance terms in (6).

Bias term. We bound $\|\theta_{\text{mix}}^* - \theta_0^*\|$ by bounding the perturbation of $A^{-1}b$ when the stationary distribution shifts from d^{μ_0} to π_{mix}. Since $\theta^* = A^{-1}b$:

$$\begin{aligned}\theta_{\text{mix}}^* - \theta_0^* &= A_{\pi_{\text{mix}}}^{-1} b_{\pi_{\text{mix}}} - A_{d^{\mu_0}}^{-1} b_{d^{\mu_0}} \\ &= A_{\pi_{\text{mix}}}^{-1}\big(b_{\pi_{\text{mix}}} - b_{d^{\mu_0}}\big) + \big(A_{\pi_{\text{mix}}}^{-1} - A_{d^{\mu_0}}^{-1}\big) b_{d^{\mu_0}}.\end{aligned}$$

The first term is bounded by $\|A_{\pi_{\text{mix}}}^{-1}\| \cdot \|b_{\pi_{\text{mix}}} - b_{d^{\mu_0}}\|$. Since $b_{\bar{\pi}} = \mathbb{E}_{s\sim\bar{\pi}}[r(s)\phi(s)]$ and $\|\phi(s)\| \leq 1$, $|r(s)| \leq 1$, this shift is bounded proportionally to the stationary distribution difference, which itself scales with $\kappa_0\varepsilon$. The second term is controlled similarly using $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ and the Lipschitz dependence of $A_{\bar{\pi}}$ on $\bar{\pi}$. Combining these bounds gives the result.