FUNCTIONAL CRITICS FOR OFF-POLICY ACTOR-CRITIC: FROM STABILITY TO EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Actor-critic methods are central to modern reinforcement learning, yet off-policy actor-critic theory remains far removed from practical algorithms, due in part to the interaction between the deadly triad and the continually changing target policy. We revisit functional critics, i.e., value functions that explicitly condition on the policy, and show that this representation provides a shared foundation for off-policy training stability and policy-space exploration. We first develop a convergent off-policy actor-critic framework under linear function approximation that accommodates target-based TD, partial coverage, and evolving behavior policies. Our analysis combines a policy-evaluation guarantee under partial coverage with policy-gradient generalization of functional critics to yield a dual trust-coverage framework that controls policy-gradient error under evolving behavior policies while enabling greater reuse of off-policy data. We further develop a replay-buffer analysis under i.i.d. sampling, yielding a principled replay-sampling design for minimizing the replay-sample complexity required to reach a prescribed evaluation accuracy. We next show that functional critics provide a natural representation for model-free posterior-sampling-style exploration. Standard critic ensembles do not naturally represent joint uncertainty over return functions across policy space, whereas functional-critic ensembles represent policy-conditioned return functions directly. This motivates a particle-based, model-free analogue of posterior-sampling-style exploration without explicit model learning. Finally, we instantiate these ideas in a deep functional actor-critic algorithm that, despite omitting several standard off-policy AC heuristics, achieves strong performance across the full Dog and Humanoid task series of the DeepMind Control Suite.

1 INTRODUCTION

Off-policy actor-critic (AC) methods combine policy optimization with the ability to reuse experience collected by behavior policies, making them particularly attractive for continuous and high-dimensional control (Degris et al., 2012; Silver et al., 2014; Lillicrap et al., 2015; Haarnoja et al., 2018). Despite their empirical success, however, the theory of off-policy AC remains far removed from practical algorithms. Two central challenges are **training stability** and **exploration**: the former concerns how to reliably improve the policy from off-policy experience, while the latter concerns how to collect informative experience in the first place.

A **functional critic**, or policy-conditioned value function (Harb et al., 2020; Tang et al., 2022; Faccio et al., 2020), models a value functional $Q(s, a, \pi)$ by explicitly taking the policy as input. Functional critics were originally motivated by the "moving target" problem in AC: as the actor changes, a conventional critic must repeatedly relearn a new value function. Yet their broader theoretical role remains unclear, particularly for exploration. In this work, we show that functional critics provide a shared foundation for off-policy training stability and exploration.

Training stability. Off-policy AC combines two sources of difficulty. Even for a fixed target policy, temporal-difference learning with function approximation and off-policy data can diverge due to the "deadly triad" (Baird et al., 1995; Tsitsiklis & Van Roy, 1996; Sutton & Barto, 2018). At the same time, actor updates continuously change the policy being evaluated. Existing convergence theory therefore relies on algorithms and assumptions that differ substantially from common practice: gradient or emphatic TD (Sutton et al., 2009; 2016; Zhang et al., 2020) rather than target-based TD;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

fixed behavior policies with strong coverage assumptions rather than behavior policies that track the actor; and samples from a stationary behavior distribution rather than a replay buffer.

We formalize functional critics under linear function approximation and establish a convergent off-policy AC framework that accommodates three practical ingredients: **target-based TD, partial coverage, and evolving behavior policies**. We separately develop an i.i.d. replay-buffer analysis that connects replay sampling to policy-evaluation accuracy and replay-sample complexity. Several technical innovations are central to the analysis. First, we establish a policy-evaluation guarantee under partial coverage, requiring only coverage of feature directions relevant to the target policy rather than full state-action coverage. Second, functional critics provide a policy-gradient generalization property that controls gradient error as the actor moves away from the reference policy at which the critic was evaluated. Together, they yield a dual trust-coverage framework that controls evaluation error arising from off-policy experience and gradient-generalization error arising from target-policy evolution, establishing convergence in policy-gradient norm while permitting greater reuse of off-policy data. Third, under i.i.d. replay sampling, we decompose the policy-evaluation error into finite-buffer statistical and TD-optimization components, exposing a trade-off between *sample diversity* and *feature coverage*. This yields a principled replay-sampling and evaluation-budget design for minimizing the replay-sample complexity required to reach a prescribed evaluation accuracy.

Exploration through the functional lens. Functional critics also reveal a connection to efficient exploration. Posterior Sampling Reinforcement Learning (PSRL) provides a principled exploration framework, but explicitly maintaining a posterior over environment models and repeatedly solving sampled MDPs is computationally prohibitive in large-scale problems. Existing model-free approximations instead represent uncertainty through ensembles or randomized priors (Osband et al., 2016; 2018; 2019). We illustrate that ordinary critic ensembles do not naturally represent joint uncertainty over return functions across policy space. Functional-critic ensembles instead represent policy-conditioned return functions directly, motivating a particle-based, model-free analogue of posterior-sampling-style exploration.

From theory to practice. We instantiate these ideas in an AC algorithm using a neural functional-critic architecture. Our implementation deliberately omits several standard off-policy AC heuristics, including minimum twin-target values, exploration-noise schedules, and entropy regularization. It achieves strong performance across the full Dog and Humanoid task series of the DeepMind Control Suite (Tassa et al., 2018).

Additional related work on policy-conditioned value functions, off-policy actor-critic theory, modern off-policy practice, and posterior-sampling-style exploration is provided in Appendix A.

Contributions. Our contributions are threefold:

1. **Bridging off-policy AC theory and practice.** We establish an off-policy AC convergence guarantee under linear function approximation that jointly accommodates target-based TD, partial coverage, and evolving behavior policies. Our analysis combines a partial-coverage policy-evaluation guarantee with policy-gradient generalization of functional critics to obtain the dual trust-coverage framework. We additionally develop an i.i.d. replay-buffer analysis that yields a principled replay-sampling and evaluation-budget design.

2. **Policy-space exploration through functional critics.** We reveal a structural connection between functional critics and posterior-sampling-style exploration. Standard critic ensembles do not naturally represent joint uncertainty over return functions across policy space, whereas functional-critic ensembles represent policy-conditioned return functions directly. This motivates a particle-based, model-free analogue of posterior-sampling exploration without explicit model learning.

3. **A neural functional actor-critic implementation.** We translate the functional-critic representation into a practical actor-critic algorithm with a tailored neural architecture. The implementation deliberately omits several heuristics commonly used by modern off-policy AC methods, yet achieves strong performance across the full Dog and Humanoid task series of the DeepMind Control Suite using a single hyperparameter configuration. These results demonstrate the practical viability of neural functional critics beyond the linear theoretical setting.

2 PRELIMINARIES

MDP and Off-Policy Actor-Critic. We consider an infinite-horizon discounted Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$ with finite state and action spaces, transition kernel p, reward $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, and discount $\gamma \in [0, 1)$. For a policy π, its value and action-value functions are

$$V_\pi(s) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right],$$
$$Q_\pi(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)}[V_\pi(s')].$$

We parameterize the policy by $\theta \in \Theta \subseteq \mathbb{R}^d$ and write π_θ, V_θ, and Q_θ accordingly. Given an initial-state distribution ρ, the objective is $J_\rho(\theta) := \mathbb{E}_{s\sim\rho}[V_\theta(s)]$.

In off-policy RL, experience from a *behavior policy* μ is reused to optimize π_θ. Actor-critic (AC) alternates critic evaluation and actor updates, while off-policy critic learning with function approximation can suffer from deadly-triad instability (Sutton et al., 2009; 2016; Zhang et al., 2021).

Off-Policy Gradients with Functional Critics. Because the target-policy state distribution cannot be sampled directly, the standard policy gradient cannot be directly estimated from behavior data. Existing off-policy formulations instead optimize a surrogate J_{d^μ} with a policy-dependent emphasis correction (Imani et al., 2018). We instead differentiate $J_\rho(\theta)$ directly:

$$\nabla_\theta J_\rho(\theta) = \sum_s \rho(s) \sum_a \left[\nabla_\theta \pi_\theta(a|s) Q_\theta(s, a) + \pi_\theta(a|s) \nabla_\theta^{\mathrm{pol}} Q_\theta(s, a) \right]. \tag{1}$$

The second term in equation 1 is difficult to evaluate with a conventional critic because Q_θ depends implicitly on the policy parameters. A functional critic instead models this dependence explicitly as $\hat{Q}(\pi_\theta, s, a; \xi)$. Following Faccio et al. (2020), this yields the direct gradient estimate

$$\nabla_\theta J_\rho(\pi_\theta; \xi) = \mathbb{E}_{s\sim\rho} \left[\sum_a \left(\hat{Q}(\pi_\theta, s, a; \xi) \nabla_\theta \pi_\theta(a|s) + \pi_\theta(a|s) \nabla_\theta^{\mathrm{pol}} \hat{Q}(\pi_\theta, s, a; \xi) \right) \right]. \tag{2}$$

Here, $\nabla_\theta^{\mathrm{pol}}$ denotes differentiation with respect to the policy parameters through the policy dependence of the value function, while holding the state–action arguments fixed. For a functional critic, this corresponds to differentiation through its explicit policy input.

Posterior Sampling for Reinforcement Learning (PSRL). PSRL extends Thompson sampling to MDPs. It maintains a posterior over the environment, samples an MDP at the beginning of each episode, computes its optimal policy, executes that policy, and updates the posterior with the collected experience. Although PSRL provides a principled exploration mechanism (Osband et al., 2013), maintaining a posterior over transition and reward models and repeatedly solving sampled MDPs is generally computationally prohibitive in high-dimensional continuous spaces.

3 METHODOLOGY

We now develop the functional actor-critic framework and its theoretical foundations. Section 3.1 introduces the general meta-algorithm. Section 3.2 establishes convergence under linear function approximation, including target-based evaluation with partial coverage and the dual trust-coverage mechanism for evolving behavior policies. Section 3.3 separately studies replay-buffer sampling and critic-budget allocation under an i.i.d. sampling model.

3.1 FUNCTIONAL ACTOR-CRITIC META-ALGORITHM

Our framework maintains an actor π_θ and a functional critic $\hat{Q}(\pi_\theta, s, a; \xi)$ that explicitly takes the policy as input. At each iteration, off-policy experience is used to update the critic toward the Bellman equation of the current actor, while the actor is updated using the direct functional gradient in equation 2. Because the policy is an explicit input, a single critic can generalize across evolving actors rather than relearning an independent value function after each policy update.

Algorithm 1 summarizes this template; subsequent sections instantiate the functional policy evaluator and its update schedule.

Algorithm 1 Functional Actor-Critic Meta-Algorithm

1: **Input:** $\theta_0, \xi_0, T, \{\eta_t\}_{t=1}^T$
2: **for** $t = 1, \ldots, T$ **do**
3: Collect/sample off-policy data $\mathcal{D}_t$
4: $\xi_t \leftarrow \text{FUNCTIONALPOLICYEVALUATOR}(\xi_{t-1}, \pi_{\theta_{t-1}}, \mathcal{D}_t)$
5: $\theta_t \leftarrow \theta_{t-1} + \eta_t \nabla_\theta J_\rho(\pi_\theta; \xi_t)|_{\theta=\theta_{t-1}}$
6: **end for**
7: **Return** θ_T

3.2 Convergence Analysis under Linear Function Approximation

We now instantiate the functional actor-critic framework under linear function approximation and establish convergence for target-based off-policy AC with partial coverage and evolving behavior policies. We present the main results and defer complete proofs to Appendix F.

3.2.1 Linear Function Approximation Setup

We analyze our framework under the following structural and regularity assumptions.

Assumption 1 (Strong Mixing)**.** *For every $\pi \in \Pi$, the induced Markov chain is irreducible and aperiodic, with unique stationary distribution d^π.*

Assumption 2 (Linear Functional Representation)**.** *For the policy class Π, there exist a feature map $\phi : \Pi \times \mathcal{S} \times \mathcal{A} \to \mathbb{B}^d$ and $\xi^\star \in \mathbb{R}^d$, $\|\xi^\star\|_2 \le L$, such that*

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

Assumption 3 (Bellman Completeness)**.** *For any $\xi \in \mathbb{R}^d$, there exists $\xi' \in \mathbb{R}^d$ such that*

$$\langle \phi^\pi(s,a), \xi' \rangle = \mathcal{T}^\pi \left(\lceil \langle \phi^\pi(\cdot,\cdot), \xi \rangle \rceil \right)(s,a) + \mathcal{E}_{\text{approx}}, \quad \forall (\pi, s, a) \in \Pi \times \mathcal{S} \times \mathcal{A},$$

where $\lceil \cdot \rceil$ projects onto $[0, (1-\gamma)^{-1}]$, and $\mathcal{E}_{approx}$ is the inherent approximation error.

Assumption 1 is a standard ergodicity condition for off-policy stochastic approximation (Zhang et al., 2020; 2021; Chen et al., 2023). Assumptions 2–3 impose functional realizability and approximate Bellman completeness across policies; further discussion appears in Appendix F.1.

3.2.2 Target-Based Evaluation with Partial Coverage

We evaluate a target policy π from data collected by a behavior policy μ using target-based TD, with the complete procedure given in Algorithm 5. Unlike existing guarantees that typically require full coverage by μ, we require coverage only of feature directions relevant to the target policy.

Define the on-policy and off-policy feature covariance as follows,

$$\mathbf{A}_\pi := \mathbb{E}_{d^\pi} \left[\phi^\pi(s,a) \phi^\pi(s,a)^\top \right], \qquad \widetilde{\mathbf{A}}_\pi := \mathbb{E}_{d^\mu} \left[\phi^\pi(s,a) \phi^\pi(s,a)^\top \right]. \tag{3}$$

Definition 1 (Evaluation Trust Metric)**.** For target policy π and behavior policy μ, define

$$\mathcal{C}(\pi, \mu) := \mathbb{E}_{d^\pi} \left[\|\phi^\pi(s,a)\|^2_{\widetilde{\mathbf{A}}_\pi^{-1}} \right], \tag{4}$$

where $\|\mathbf{u}\|_{\mathbf{A}^{-1}} := \lim_{\lambda \to 0_+} \|\mathbf{u}\|_{(\mathbf{A}+\lambda \mathbf{I})^{-1}}$.

Finite $\mathcal{C}(\pi, \mu)$ requires coverage only of feature directions relevant under π, rather than every state–action pair; The key technical step is to establish Bellman contraction in the $\mathbf{A}_\pi$-weighted L^2 norm.

Theorem 1 (Evaluation Guarantee with Partial Coverage)**.** *Let Assumptions 1–3 hold. With suitable α and $K \asymp \epsilon^{-2}$, the target-based evaluator satisfies*

$$\mathbb{E} \left[\|\hat{Q}_T - Q^\pi\|^2_{d^\pi} \right] \lesssim \gamma^T + \mathcal{C}(\pi, \mu) \left(\epsilon^2 + \mathcal{E}^2_{\text{approx}} \right) + \mathcal{E}^2_{\text{approx}}. \tag{5}$$

To our knowledge, Theorem 1 is the first target-based TD policy-evaluation guarantee that replaces pointwise full coverage with a feature-based coverage condition, allowing convergence whenever $\mathcal{C}(\pi, \mu) < \infty$. The corresponding parameter-space guarantee is

$$\mathbb{E} \left[\|\hat{\xi}_T - \xi^\star\|^2_{\mathbf{A}_\pi} \right] \lesssim \gamma^T + \mathcal{C}(\pi, \mu) \left(\epsilon^2 + \mathcal{E}^2_{\text{approx}} \right) + \mathcal{E}^2_{\text{approx}}, \tag{6}$$

which will serve as the evaluation guarantee for the full actor-critic analysis.

Algorithm 2 Functional Actor–Critic with Dual Trust-Coverage

Input: $\phi, \epsilon, \eta, C_{\max}$
Initialize: $\bar{\pi}^{(0)}$, $\mu \leftarrow \bar{\pi}^{(0)}$
for $k = 0, 1, \ldots$ **do**
  if $\mathcal{C}(\bar{\pi}^{(k)}, \mu) > C_{\max}$ **then**
    $\mu \leftarrow \bar{\pi}^{(k)}$
  end if
  $\xi^{(k)} \leftarrow \textsc{FunctionalPolicyEvaluator}(\mu, \bar{\pi}^{(k)}, \phi, \epsilon)$
  $\pi_0^{(k)} \leftarrow \bar{\pi}^{(k)}$
  for $t = 1, 2, \ldots$ **do**
    $\pi_t^{(k)} \leftarrow \pi_{t-1}^{(k)} + \eta \hat{\nabla} J(\pi_{t-1}^{(k)}, \xi^{(k)})$
    if $\Delta_{k,t} > 2$ **then**
      $\bar{\pi}^{(k+1)} \leftarrow \pi_t^{(k)}$; **break**
    end if
  end for
end for

3.2.3 Off-Policy AC with Dual Trust-Coverage: Algorithm and Convergence

We now extend the partial-coverage evaluation guarantee to the full actor-critic setting. At epoch k, a reference policy $\bar{\pi}^{(k)}$ anchors the functional critic, with $\mathbf{A}_k := \mathbf{A}_{\bar{\pi}^{(k)}}$. Two sources of error must be controlled: (i) the error in evaluating $\bar{\pi}^{(k)}$ from off-policy experience, and (ii) the gradient-generalization error as the current actor $\pi_t^{(k)}$ evolves away from $\bar{\pi}^{(k)}$.

The first is controlled by the evaluation trust metric $\mathcal{C}(\bar{\pi}^{(k)}, \mu)$. Whenever $\mathcal{C}(\bar{\pi}^{(k)}, \mu) > C_{\max}$, the behavior policy is refreshed as $\mu \leftarrow \bar{\pi}^{(k)}$. After evaluating the reference policy to obtain $\xi^{(k)}$, gradient compatibility controls the actor update at the reference policy, while a second, *gradient-generalization trust metric* $\Delta_{k,t}$ governs subsequent reuse of the critic as the actor evolves away from the reference. Algorithm 2 summarizes this dual trust-coverage mechanism.

By Theorem 1, whenever $\mathcal{C}(\bar{\pi}^{(k)}, \mu) \le C_{\max}$, the anchored critic satisfies

$$\sqrt{\mathbb{E}\left[\|\xi^{(k)} - \xi^\star\|_{\mathbf{A}_k}^2\right]} \lesssim \bar{\epsilon} := \sqrt{C_{\max}}\,(\epsilon + \mathcal{E}_{\text{approx}})\,. \tag{7}$$

To transfer equation 7 to the gradient evaluation error, we introduce the functional-gradient operator,

$$\mathbf{G}_{\pi_\theta} := \mathbb{E}_{s\sim\rho}\left[\sum_a \left(\nabla_\theta \pi_\theta(a \mid s)\phi^{\pi_\theta}(s,a)^\top + \pi_\theta(a \mid s) D\phi^{\pi_\theta}(s,a)\right)\right], \tag{8}$$

under which the gradient error can be represented as $\hat{\nabla} J_\rho(\pi;\xi) - \nabla J_\rho(\pi) = \mathbf{G}_\pi(\xi - \xi^\star)$. We impose the following compatibility condition,

Assumption 4 (Gradient Compatibility)**.** *There exists a finite constant $\kappa_{\mathrm{g}} > 0$ such that for every reference policy π,*

$$\mathbf{G}_\pi^\top \mathbf{G}_\pi \preceq \kappa_{\mathrm{g}}^2 \mathbf{A}_\pi. \tag{9}$$

Assumption 4 ensures that critic error controlled in the $\mathbf{A}_\pi$ seminorm also controls the corresponding policy-gradient error, up to a factor of κ_{g}. Together with equation 7, it therefore yields gradient error $\mathcal{O}(\kappa_{\mathrm{g}}\bar{\epsilon})$ at the reference policy. Away from the reference policy, critic estimation error can be amplified through both the functional features ϕ and their policy derivatives $D\phi$. We quantify this amplification by the *gradient-generalization trust metric*

$$\Delta_{k,t} := \max\{\mathcal{C}_1^{(k,t)}, \mathcal{C}_2^{(k,t)}\}, \tag{10}$$

where $\mathcal{C}_1^{(k,t)}$ and $\mathcal{C}_2^{(k,t)}$ measure, respectively, amplification through ϕ and $D\phi$; their explicit definitions are given in Appendix F.3. The resulting gradient error satisfies

$$\mathbb{E}\left[\|\nabla J_\rho(\pi_t^{(k)}) - \hat{\nabla} J_\rho(\pi_t^{(k)}, \xi^{(k)})\|_\infty\right] \lesssim \Delta_{k,t}\bar{\epsilon}. \tag{11}$$

The evaluation criterion $\mathcal{C}$ controls reference-policy evaluation. The first actor update is controlled by Assumption 4, while $\Delta_{k,t}$ controls subsequent reuse of the anchored critic as the actor evolves.

Under the standard smoothness assumption that ϕ^π and $D\phi^\pi$ are Lipschitz in the policy parameters, we obtain the following convergence guarantee.

Theorem 2 (Convergence of Functional Actor-Critic)**.** *Let Assumptions 1–3, Assumption 4, and the smoothness condition above hold. Let $c_{\mathrm{g}} := \max\{\kappa_{\mathrm{g}}^2, 16d\}$. For a suitable stepsize η, the policies generated by Algorithm 2 over M actor updates satisfy*

$$\frac{1}{M}\mathbb{E}\left[\sum_{k}\sum_{t=0}^{T_k-1}\|\nabla J_\rho(\pi_t^{(k)})\|^2\right] \le \frac{2(J_\rho^\star - J_\rho(\pi^{(0)}))}{\eta M} + \tilde{\mathcal{O}}(c_{\mathrm{g}}\bar{\epsilon}^2). \tag{12}$$

The dual trust-coverage mechanism therefore exposes two complementary reuse horizons: the first criterion governs how long behavior data can be reused, while the second governs how many actor updates can reuse the same functional critic.

3.3 Replay Sampling and Critic Budgets

The analysis above, following recent off-policy AC theory, already allows Markovian samples generated by a behavior policy. Yet it still shares a major abstraction with existing theory: policy evaluation draws samples directly from the behavior process, whereas replay-buffer training is a standard component of practical off-policy AC. Bridging this gap is more than an implementation detail. While replay sampling is often handled heuristically in practice, a principled treatment should connect the sampling design directly to convergence and the replay-sample complexity required to reach a prescribed evaluation accuracy. In this section, we propose a optimization problem for buffer sampling distribution design. For simplicity, throughout this section, we assume that samples in the buffer are independent, $\mathcal{E}_{\mathrm{approx}} = 0$, and $\mathbf{A}_\pi \succ 0$.

There are two competing considerations. On the one hand, broadly reusing both old and new experience increases *sample diversity* and reduces finite-data statistical error. On the other hand, when the behavior policy tracks the target policy, newer experience tends to provide better *feature coverage* of the current target and can therefore accelerate TD evaluation. We show that these two effects appear as distinct terms in the policy evaluation error, yielding a principled diversity–coverage trade-off.

Replay distributions and evaluation error. Fix a target policy π and a replay buffer $\mathcal{B} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$. A replay distribution is a vector $\mathbf{w} = (w_1, \dots, w_n) \in \Delta^n$ specifying the probability of sampling each buffer entry. Define

$$\mathbf{\Sigma}_{\mathbf{w}} := \sum\nolimits_{i=1}^n w_i \phi^\pi(s_i,a_i)\phi^\pi(s_i,a_i)^\top, \qquad \mathbf{M}_{\mathbf{w}} := \sum\nolimits_{i=1}^n w_i^2 \phi^\pi(s_i,a_i)\phi^\pi(s_i,a_i)^\top, \tag{13}$$

and the two replay-design criteria

$$\mathcal{V}_{\mathcal{B}}(\pi,\mathbf{w}) := \mathrm{tr}\big(\mathbf{A}_\pi \mathbf{\Sigma}_{\mathbf{w}}^{-1}\mathbf{M}_{\mathbf{w}}\mathbf{\Sigma}_{\mathbf{w}}^{-1}\big), \quad \mathcal{C}_{\mathcal{B}}(\pi,\mathbf{w}) := \mathrm{tr}\big(\mathbf{A}_\pi \mathbf{\Sigma}_{\mathbf{w}}^{-1}\big). \tag{14}$$

Here $\mathcal{V}_{\mathcal{B}}$ captures the finite-buffer statistical cost of the sampling design, whereas $\mathcal{C}_{\mathcal{B}}$ measures how well the replayed features cover directions relevant to the target policy. The squared weights in $\mathbf{M}_{\mathbf{w}}$ arise because the statistical variance of a weighted observation scales with w_i^2.

Recall that Algorithm 5 performs K inner TD updates for each of T target-network iterations. Under replay sampling, each inner TD update draws one transition from $\mathcal{B}$ according to $\mathbf{w}$; hence a complete policy-evaluation call uses TK replay samples in total.

We can then establish the following replay-based policy-evaluation guarantee.

Lemma 1 (Replay evaluation guarantee)**.** *Suppose Assumptions 1–3 hold. There exist constants $\alpha > 0$, $c_{\mathrm{stat}} > 0$, and $c_{\mathrm{TD}} > 0$ such that, for any integers $K, T \ge 1$, the output of Algorithm 5 using replay distribution $\mathbf{w}$ and step size α satisfies, up to logarithm factors,*

$$\mathbb{E}\left[\|\hat{\xi}_T - \xi^\star\|_{\mathbf{A}_\pi}^2\right] \le \frac{c_{\mathrm{stat}}\mathcal{V}_{\mathcal{B}}(\pi,\mathbf{w})}{(1-\gamma)^3} + \frac{c_{\mathrm{TD}}\mathcal{C}_{\mathcal{B}}(\pi,\mathbf{w})}{K(1-\gamma)^4} + \frac{3\gamma^{2T}}{(1-\gamma)^2}. \tag{15}$$

Statistical and optimization errors. For a fixed replay distribution $\mathbf{w}$, let $\xi^{\mathrm{LS}}(\mathbf{w})$ denote the corresponding weighted least-squared solution. The policy-evaluation error admits the decomposition

$$\mathbb{E}\big[\|\hat{\xi}_T - \xi^\star\|_{\mathbf{A}_\pi}^2\big] \le 2\mathbb{E}\big[\|\xi^{\mathrm{LS}}(\mathbf{w}) - \xi^\star\|_{\mathbf{A}_\pi}^2\big] + 2\mathbb{E}\big[\|\hat{\xi}_T - \xi^{\mathrm{LS}}(\mathbf{w})\|_{\mathbf{A}_\pi}^2\big]. \tag{16}$$

The first term is the finite-buffer statistical error under $\mathbf{w}$ and cannot be reduced by additional TD updates; the second is the TD-to-LS optimization gap. Lemma 1 controls these terms through $\mathcal{V}_{\mathcal{B}}(\pi, \mathbf{w})$ and $\mathcal{C}_{\mathcal{B}}(\pi, \mathbf{w})/K$, respectively, with γ^{2T} accounting for the residual target-network error. Thus, $\mathcal{V}_{\mathcal{B}}$ sets the accuracy floor, while $\mathcal{C}_{\mathcal{B}}/K$ governs the cost of approaching it.

Diversity–coverage trade-off. These two error components favor different replay distributions. Let $\mathbf{u} = (1/n, \ldots, 1/n)$ denote uniform replay. The statistical criterion satisfies $\mathcal{V}_{\mathcal{B}}(\pi, \mathbf{w}) \geq \mathcal{V}_{\mathcal{B}}(\pi, \mathbf{u})$, so uniform replay minimizes the statistical error by making the broadest use of the finite buffer. This favors *sample diversity*: experience collected under both older and newer behavior policies contributes to reducing the finite-data statistical error.

In contrast, the TD optimization term favors replay distributions with small $\mathcal{C}_{\mathcal{B}}(\pi, \mathbf{w})$, i.e., distributions whose sampled features provide strong coverage of the current target policy. When the behavior policy tracks the target policy, this can favor more recent experience. Hence replay sampling faces a diversity–coverage trade-off: uniform sampling provides the best statistical limit, whereas a more coverage-focused distribution can require fewer TD updates to approach its corresponding limit. A concrete numerical example illustrating this trade-off is shown in Fig. 5 in the Appendix G.2.

Replay design and sample complexity. The decomposition above suggests choosing the replay distribution together with the evaluation budget. For a prescribed evaluation accuracy ϵ_{c} and fixed number T of target-network iterations, we consider

$$\min_{\mathbf{w} \in \Delta^n,\ K \in \mathbb{N}} K \qquad \text{s.t.} \quad \frac{c_{\mathrm{stat}} \mathcal{V}_{\mathcal{B}}(\pi, \mathbf{w})}{(1-\gamma)^3} + \frac{c_{\mathrm{TD}} \mathcal{C}_{\mathcal{B}}(\pi, \mathbf{w})}{K(1-\gamma)^4} + \frac{3\gamma^{2T}}{(1-\gamma)^2} \leq \epsilon_{\mathrm{c}}^2. \tag{17}$$

Since each TD update consumes one replay sample, minimizing K for fixed T is equivalent to minimizing the total replay-sample complexity TK. For a fixed replay budget K, the corresponding sampling-design problem and its interpretation are provided in Appendix G.2.

Remark 1 (From policy evaluation to policy optimization). Lemma 1 and equation 17 optimize replay under the $\mathbf{A}_\pi$-weighted policy-evaluation metric. Such guarantee can be transferred to gradient estimation as well. More precisely, for each reference policy π, by equation 8 and Assumption 4, we have $\mathbb{E}\big[\|\hat{\nabla} J_\rho(\pi, \hat{\xi}_T) - \nabla J_\rho(\pi)\|_2^2\big] \leq \kappa_{\mathrm{g}}^2 \mathbb{E}\big[\|\hat{\xi}_T - \xi^\star\|_{\mathbf{A}_\pi}^2\big]$. Thus the same replay program controls gradient accuracy at the reference policy.

Empirical results for the resulting replay-sampling strategy are reported in Appendix D.

4 Functional Critics for Posterior-Sampling-Style Exploration

As reviewed in Section 2, PSRL maintains a posterior over MDPs and, for each rollout, executes the optimal policy of a sampled MDP. Because return depends jointly on the MDP and the policy, each sampled MDP defines a return functional over the entire policy space (Fig. 1 (a)). This observation motivates a model-free posterior-sampling-style construction that represents uncertainty jointly across policy space, rather than only at individual policies.

Model-free exploration methods often use ensembles of standard critics (Osband et al., 2016; 2018; 2019). Fig. 1 (b)–(c) illustrates why such ensembles do not naturally represent uncertainty jointly over return functions across policy space. With a shared target critic, all critics estimate the value of the same policy and therefore characterize uncertainty only at that policy (Fig. 1 (b)). With independent target critics, ensemble members instead estimate different policy-dependent quantities, which are not samples of a common return functional over policy space (Fig. 1 (c)).

Functional critics provide a direct representation of the relevant object. Each functional critic models an entire policy-conditioned return $\hat{Q}(\pi, s, a)$, so an ensemble represents a collection of candidate return functionals over policy space (Fig. 1 (d)). Sampling one such functional and improving an actor against it therefore motivates a model-free analogue of posterior-sampling-style exploration, without explicitly maintaining or solving sampled MDP models.

We instantiate this idea with ensembles of actors and functional critics. During actor optimization, critics are sampled from the ensemble and actors are updated using the corresponding functional gradients. The complete procedure is given in Algorithm 3 in Appendix B.5.

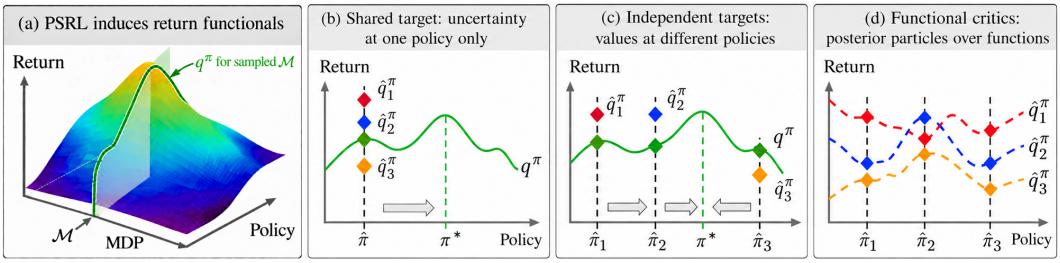


Figure 1: From PSRL to functional-critic ensembles. (a) An MDP sample induces a return functional over policy space. (b) With a shared target policy, a standard critic ensemble represents uncertainty only at that policy. (c) With independent target policies, ensemble members estimate different policy-dependent quantities rather than samples from a common posterior over return functions. (d) Functional-critic ensembles represent particles over entire policy-conditioned return functions.

5 Practical Implementation

We instantiate the functional actor-critic framework with neural networks and the ensemble construction of Section 4. The main experiments evaluate the resulting neural functional actor-critic system; heuristic neural instantiations of evaluation-trust behavior reuse and replay-sampling mechanisms, together with targeted empirical evaluations, are reported in the Appendix E and D.

Neural Functional Critics. We maintain an ensemble of functional critics $\{\hat{Q}^{(i)}\}_{i=1}^n$. Each critic consists of an actor encoder $E_{\text{act}}^{(i)}$, a state-action encoder $E_{\text{sa}}^{(i)}$, and a joint encoder $E_{\text{joint}}^{(i)}$:

$$\hat{Q}^{(i)}(\pi_\theta, s, a) = E_{\text{joint}}^{(i)}\left(E_{\text{act}}^{(i)}(\pi_\theta), E_{\text{sa}}^{(i)}(s,a)\right). \tag{18}$$

Target critics use delayed copies of the state-action and joint encoders, while sharing the actor encoder. Each critic is trained by target-based TD on actor-transition pairs drawn from the product of the current actor ensemble and replay data.

To represent a policy efficiently, E_{act} evaluates the actor at a small set of trainable states $\{\zeta_j\}$ and processes the resulting actions and hidden activations with a transformer encoder. The evaluation states are learned end-to-end together with the critic.

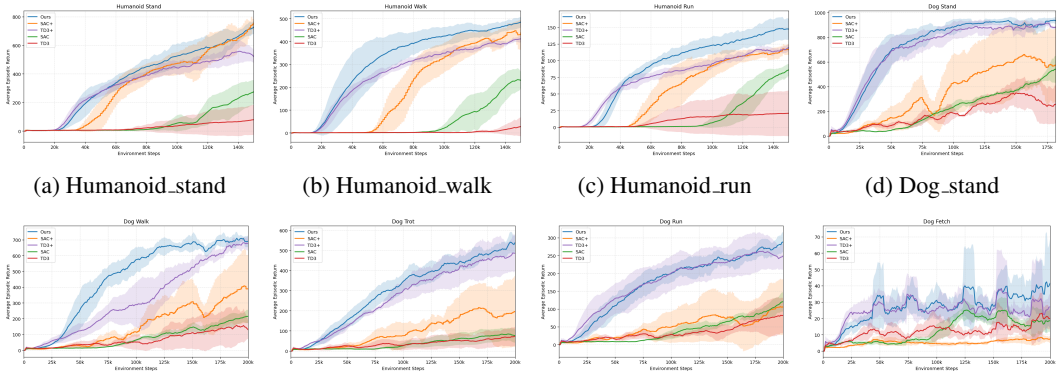
Actor Ensemble and Exploration. We use an ensemble of deterministic actors $\{\pi_{\theta^{(i)}}\}_{i=1}^n$, avoiding additional hyperparameters associated with stochastic actors. Each functional critic is trained across the actor ensemble, while every actor update independently samples a critic from which to obtain the functional policy gradient. For deterministic actors, gradients propagate through both the action and explicit policy inputs of the functional critic; the resulting gradient is given in Appendix B.2.

Omitted Standard Heuristics. Our implementation intentionally omits several heuristics widely used in modern off-policy AC practice: clipped/minimum double-Q targets, exploration noise, target-policy smoothing, and entropy regularization (Fujimoto et al., 2018; Haarnoja et al., 2018). Further architectural details and the complete training procedure are provided in Appendix B.

5.1 Empirical Results

Baselines and Setup. We evaluate on the Dog and Humanoid task series from the DeepMind Control Suite (Tunyasuvunakool et al., 2020). These two series feature relatively high-dimensional state and action spaces and distinct morphologies, providing challenging and complementary testbeds; we evaluate the full series rather than selecting individual tasks.

We compare against SAC and TD3 (Haarnoja et al., 2018; Fujimoto et al., 2018), together with strengthened variants SAC+ and TD3+. Recent high-performance RL systems increasingly rely on aggressive critic updating, often expressed through a high update-to-data (UTD) ratio, together with critic ensembles and layer normalization (Chen et al., 2021; Li et al., 2023; Ball et al., 2023), a recipe underlying several recent strong online and offline RL methods (Zhou et al., 2024; Xiao et al., 2025; Nauman et al., 2024; Seo et al., 2025). We incorporate this recipe into SAC and TD3 using the extensively tested ALF codebase (Xu et al., 2021) and carefully tune the resulting SAC+ and TD3+



(a) Humanoid_stand (b) Humanoid_walk (c) Humanoid_run (d) Dog_stand

(e) Dog_walk (f) Dog_trot (g) Dog_run (h) Dog_fetch

Figure 2: Average episodic return on Dog and Humanoid tasks from DM Control. Curves show the mean over four seeds and shaded regions indicate one standard deviation; all experiments use four parallel environments.

baselines. As shown in Fig. 2, both substantially outperform their vanilla counterparts, confirming that the modern training recipe itself materially strengthens the baselines and providing substantially stronger comparators for our method.

After independent tuning, both SAC+ and TD3+ use 10 critic updates per actor update. We use the same 10:1 critic-to-actor update ratio for our method in all main experiments. Our method, SAC+, and TD3+ use 10 critics, while the vanilla SAC and TD3 retain their original two-critic architectures. Functional-critic updates additionally reuse each sampled transition across the actor ensemble; the comparisons therefore match environment interaction, critic ensemble size, and nominal update frequency, but not total critic-side computation. We match network backbones across methods where their architectures permit, and our method uses a single hyperparameter configuration across all eight tasks. Full experimental protocols, baseline configurations, architectures, and hyperparameters are provided in Appendix C.

Prior policy-conditioned value-function methods are discussed in Appendix A.1. Their published implementations and evaluations operate in substantially different training regimes, and a subsequent benchmark found the major variants noncompetitive with TD3 (Griesbach, 2023). We therefore focus the main comparison on carefully tuned modern SAC+ and TD3+ baselines.

Results. Figure 2 shows that our method matches or exceeds the strongest baseline in mean return across all evaluated Dog and Humanoid tasks. It exhibits stable learning across independent runs and favorable environment-sample efficiency, despite omitting the standard off-policy AC heuristics described above. These results demonstrate the practical viability of the proposed neural functional actor-critic system on challenging continuous-control tasks. Targeted experiments in Appendices D and E further illustrate how the theoretically motivated replay-sampling and trust-coverage mechanisms can be instantiated with learned neural features.

6 Conclusion

We studied functional critics as a shared framework for off-policy actor-critic learning. Under linear functional approximation, we establish convergence with target-based TD, partial feature coverage, and evolving behavior policies, and develop a replay-sampling analysis that exposes a diversity–coverage trade-off. Functional-critic ensembles further motivate model-free posterior-sampling-style exploration over policy-conditioned return functions. Our neural functional actor-critic system performs strongly on challenging continuous-control tasks while omitting several standard off-policy AC heuristics; targeted Appendix experiments further illustrate neural instantiations of the evaluation-trust and replay-sampling mechanisms. Our theory remains limited to linear functional approximation, and the exploration construction is posterior-sampling-inspired rather than an exact Bayesian posterior approximation. Extending the guarantees to nonlinear critics and developing calibrated functional uncertainty are important directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and editing the manuscript, including improving clarity, conciseness, grammar, and presentation. We also used Codex and Claude to assist with the implementation of extension methods reported in the Appendix, and used ChatGPT to cross-check portions of the replay-sampling derivations and theoretical results.

Generative AI tools were not used to formulate or develop the theoretical results, design the research methodology or experiments, or implement the algorithms used in the main-text experiments. All AI-assisted code and technical checks were independently reviewed and verified by the authors, and all AI-assisted text was reviewed and revised by the authors. We take full responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Complete proofs and additional derivations for our theoretical results are provided in the Appendix. Details of the deep functional actor–critic implementation, experimental protocols, baseline configurations, and hyperparameters are also provided in the Appendix. All methods compared in the main experiments were implemented and evaluated within the same ALF reinforcement-learning codebase (Xu et al., 2021). We plan to release our implementation, including the appendix extensions, together with the experiment configurations needed to reproduce the reported results.

REFERENCES

Anthony Atkinson, Alexander Donev, and Randall Tobias. *Optimum experimental designs, with SAS*, volume 34. OUP Oxford, 2007.

Leemon Baird et al. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the twelfth international conference on machine learning*, pp. 30–37, 1995.

Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pp. 1577–1594. PMLR, 2023.

Anas Barakat, Pascal Bianchi, and Julien Lehmann. Analysis of a target-based actor-critic algorithm with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 991–1040. PMLR, 2022.

Diogo Carvalho, Francisco S Melo, and Pedro Santos. A new convergent variant of q-learning with linear function approximation. *Advances in Neural Information Processing Systems*, 33:19412–19421, 2020.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith Ross. Randomized ensembled double q-learning: Learning fast without a model. *arXiv preprint arXiv:2101.05982*, 2021.

Zaiwei Chen, Sheng Zhang, Thinh T Doan, John-Paul Clarke, and Siva Theja Maguluri. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, 146:110623, 2022.

Zaiwei Chen, John-Paul Clarke, and Siva Theja Maguluri. Target network and truncation overcome the deadly triad in-learning. *SIAM Journal on Mathematics of Data Science*, 5(4):1078–1101, 2023.

Thomas Degris, Martha White, and Richard S Sutton. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pp. 2701–2709. PMLR, 2020.

Francesco Faccio, Louis Kirsch, and Jürgen Schmidhuber. Parameter-based value functions. *arXiv preprint arXiv:2006.09226*, 2020.

Francesco Faccio, Aditya Ramesh, Vincent Herrmann, Jean Harb, and Jürgen Schmidhuber. General policy evaluation and improvement by learning to identify few but crucial states. *arXiv preprint arXiv:2207.01566*, 2022.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Sebastian Griesbach. Improving policy-conditioned value functions. Master's thesis, University of Tübingen, March 2023.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Jean Harb, Tom Schaul, Doina Precup, and Pierre-Luc Bacon. Policy evaluation networks. *arXiv preprint arXiv:2002.11833*, 2020.

Ehsan Imani, Eric Graves, and Martha White. An off-policy policy gradient theorem using emphatic weightings. *Advances in neural information processing systems*, 31, 2018.

Qiyang Li, Aviral Kumar, Ilya Kostrikov, and Sergey Levine. Efficient deep reinforcement learning requires regulating overfitting. *arXiv preprint arXiv:2304.10466*, 2023.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Yifei Min, Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Variance-aware off-policy evaluation with linear function approximation. *Advances in neural information processing systems*, 34:7598–7610, 2021.

Michal Nauman, Mateusz Ostaszewski, Krzysztof Jankowski, Piotr Miłoś, and Marek Cygan. Bigger, regularized, optimistic: scaling for compute and sample efficient continuous control. *Advances in neural information processing systems*, 37:113038–113071, 2024.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Ian Osband, Benjamin Van Roy, Daniel J Russo, and Zheng Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019.

Younggyo Seo, Carmelo Sferrazza, Haoran Geng, Michal Nauman, Zhao-Heng Yin, and Pieter Abbeel. Fasttd3: Simple, fast, and capable reinforcement learning for humanoid control. *arXiv preprint arXiv:2505.22642*, 2025.

David Silver, Gavin Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 387–395, 2014.

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

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th annual international conference on machine learning*, pp. 993–1000, 2009.

Richard S Sutton, A Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*, 17(73):1–29, 2016.

Hongyao Tang, Zhaopeng Meng, Jianye Hao, Chen Chen, Daniel Graves, Dong Li, Changmin Yu, Hangyu Mao, Wulong Liu, Yaodong Yang, et al. What about inputting policy in value function: Policy representation and policy-extended value function approximator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8441–8449, 2022.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020. ISSN 2665-9638. doi: https://doi.org/10.1016/j.simpa.2020.100022. URL https://www.sciencedirect.com/science/article/pii/S2665963820300099.

Wei Xiao, Jiacheng Liu, Zifeng Zhuang, Runze Suo, Shangke Lyu, and Donglin Wang. Efficient online rl fine tuning with offline pre-trained policy only. *arXiv preprint arXiv:2505.16856*, 2025.

Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*, 2022.

Wei Xu, Haonan Yu, Haichao Zhang, Break Yang, Yingxiang Hong, Le Zhao, Qinxun Bai, Andrew Choi, and ALF contributors. ALF: Agent Learning Framework, 2021. URL https://github.com/HorizonRobotics/alf.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *International Conference on Machine Learning*, pp. 11204–11213. PMLR, 2020.

Shangtong Zhang, Hengshuai Yao, and Shimon Whiteson. Breaking the deadly triad with a target network. In *International Conference on Machine Learning*, pp. 12621–12631. PMLR, 2021.

Zhiyuan Zhou, Andy Peng, Qiyang Li, Sergey Levine, and Aviral Kumar. Efficient online reinforcement learning fine-tuning need not retain offline data. *arXiv preprint arXiv:2412.07762*, 2024.

Table 1: Comparison with prior policy-conditioned value-function methods. The listed design differences should not all be attributed to policy conditioning itself; see the discussion below.

	Policy representation	**Critic training**	**Policy distribution**	**Actor update**
PSSVF / PSVF	raw parameters	MC / TD	parameter perturbations	policy-input derivative only
PVN	probe-state actions	MC on fixed policy set	pre-collected policies	policy-input derivative only
PAVF	raw parameters	TD	parameter perturbations	DPG + policy term
PeVFA	learned OPR / SPR	MC, on-policy	single improvement path	standard PPO
Ours	probe-state actions + hidden activations; end-to-end transformer	off-policy target-based TD + replay	concurrent actor ensemble	DPG + policy term

A Additional Related Work

A.1 Policy-Conditioned and Parameter-Based Value Functions

Making the value function an explicit function of the policy has been explored in several forms, differing both in how the policy is represented and in how the policy-conditioned value function is used for actor optimization.

Policy-conditioned value functions. Policy Evaluation Networks (PVNs) (Harb et al., 2020) represent a policy through its behavior on a set of probing states and learn a value function over a pre-collected set of policies. Once trained, the differentiable policy input can be used to optimize a new policy by backpropagating through the learned value function. The original experiments focused on relatively small policy networks and simple control problems rather than online off-policy actor-critic learning.

Parameter-Based Value Functions (Faccio et al., 2020) instead use the policy parameters themselves as inputs to the value function. The state-value variants PSSVF and PSVF update the policy directly through the policy dependence of the learned value function, without the standard action-gradient pathway used by deterministic actor-critic methods. Importantly, the Q-function variant PAVF is different: its actor update contains both a DPG-style action-gradient term and a derivative through the explicit policy input. PAVF is therefore the closest formal antecedent of the two-path functional gradient used in our deterministic implementation; our distinction from PAVF is not the presence of this second gradient term. A subsequent work (Faccio et al., 2022) replaces raw policy parameters with behavior on a small set of learned probing states, further reducing the dependence of the representation on the particular policy-network parameterization.

Policy embeddings for value estimation. PeVFA (Tang et al., 2022) uses policy conditioning primarily to improve value estimation. It learns policy representations from policy-network parameters or state–action behavior using auxiliary representation-learning objectives. Its practical PPO-PeVFA instantiation retains the standard PPO actor update: the policy-conditioned value function is used for value estimation rather than introducing the additional policy-input derivative used by PAVF and our method. PPO-PeVFA is evaluated in an on-policy setting, with historical policies and their associated trajectories used to train the value function, and studies generalization primarily along a policy-improvement trajectory. Its ablations also illustrate an important practical challenge for policy-conditioned value functions: high-dimensional raw parameter representations can degrade performance, motivating more structured policy representations.

Relation to our method. Our implementation differs from prior policy-conditioned value-function methods along several distinct design axes. Table 1 summarizes the main differences.

First, our policy representation is behaviorally grounded: the actor encoder evaluates the policy on learned probing states and processes the resulting actions and hidden activations with a transformer. The representation is trained end-to-end through the critic TD objective rather than through a separate representation-learning loss.

Second, the functional critic is trained by off-policy target-based TD on replay data and across the current actor ensemble. Consequently, a replayed transition can provide Bellman supervision for multiple policies without requiring a separately collected trajectory for each policy.

Third, the critic is trained on a concurrently evolving ensemble of actors, rather than only on parameter perturbations or policies lying along a single policy-improvement trajectory.

Finally, our deterministic actor update retains the standard DPG action-gradient pathway while also differentiating through the explicit policy input of the functional critic, as given in Eq. equation 25. This gradient structure itself is not unique to our work: PAVF also contains both pathways. Our contribution instead lies in combining this gradient structure with a scalable neural policy representation, modern off-policy actor-critic training, and a theoretical framework that analyzes functional critics under partial coverage and evolving behavior policies.

More broadly, our work connects the policy-conditioned value-function literature to modern off-policy actor-critic design and analysis. Prior policy-conditioned approaches established that value functions can generalize across policies, but were largely developed in substantially different training regimes, often using Monte Carlo evaluation, parameter perturbations, or comparatively small-scale actor-critic implementations. Our method instead embeds policy conditioning into a modern off-policy actor-critic system based on replay, target-based TD, high-rate critic optimization, and critic ensembles.

At the same time, our theoretical analysis studies the functional-critic representation under practical off-policy conditions that are largely absent from the earlier policy-conditioned literature, including target-based TD, partial feature coverage, and evolving behavior policies. Conversely, existing off-policy actor-critic theory typically analyzes ordinary state-action critics rather than value functions that explicitly generalize across policies. Thus, an important distinction of our work is to bring these two lines together: policy-conditioned value-function modeling on the one hand, and modern off-policy actor-critic training and convergence analysis on the other.

These distinctions should not be interpreted as an attribution of the empirical gains to policy conditioning alone. Off-policy TD and replay are standard actor-critic ingredients, while the concurrent policy distribution is a property of the actor ensemble. Likewise, the standard action-gradient pathway is inherited from conventional deterministic actor-critic methods. The explicit policy conditioning and its learned representation are the functional-critic-specific ingredients. Our main experiments evaluate the resulting integrated system and do not isolate the contribution of each of these design choices individually.

Prior empirical comparisons. A subsequent benchmark (Griesbach, 2023) found several earlier policy-conditioned value-function methods noncompetitive with standard algorithms such as TD3 (Fujimoto et al., 2018). That benchmark predates the critic-ensemble, high-update-ratio, and layer-normalization recipe used by our modern baselines, so we do not interpret it as evidence that policy conditioning is intrinsically unsuitable for modern off-policy learning. For this reason, the main experiments compare against carefully tuned SAC+ and TD3+ baselines, while the methodological relationship to earlier policy-conditioned approaches is documented here.

A.2 Off-Policy Actor-Critic Theory and Coverage Conditions

Convergence analyses for off-policy actor-critic have largely relied on gradient or emphatic TD (Sutton et al., 2009; 2016; Imani et al., 2018; Zhang et al., 2020), or on target networks and truncation to stabilize TD with function approximation (Carvalho et al., 2020; Zhang et al., 2021; Chen et al., 2023); target-based actor-critic under linear approximation was further analyzed by Barakat et al. (2022). These results typically assume a fixed behavior policy together with strong coverage conditions. Feature-based coverage quantities are also standard in offline RL with linear structure (Duan et al., 2020; Min et al., 2021; Xiong et al., 2022; Yin et al., 2022), primarily in least-squares or pessimism-based analyses.

These analyses generally study ordinary state–action value approximators, rather than a single critic that explicitly conditions on and generalizes across evolving policies. Our analysis brings functional critics into this off-policy setting, establishing a target-based TD guarantee under target-relevant feature coverage and incorporating it into an actor-critic scheme with evolving behavior policies.

A.3 Modern Off-Policy Actor-Critic Practice

Modern continuous-control systems often combine large critic update-to-data ratios, critic ensembles, and normalization techniques (Chen et al., 2021; Li et al., 2023; Ball et al., 2023), a recipe used by several recent high-performing RL methods (Nauman et al., 2024; Zhou et al., 2024; Xiao et al., 2025; Seo et al., 2025). The earlier policy-conditioned value-function methods discussed in Section A.1 were developed in substantially different training regimes and do not establish how policy conditioning behaves within this modern off-policy recipe.

Our implementation brings functional critics into this regime, adopting these ingredients where applicable, while our SAC+ and TD3+ baselines apply the same modern training recipe to SAC (Haarnoja et al., 2018) and TD3 (Fujimoto et al., 2018) within the same ALF codebase (Xu et al., 2021). As reflected by the large gap between these strengthened baselines and their vanilla counterparts in Section 5.1, this training recipe itself accounts for a substantial part of absolute performance and should not be attributed to the functional-critic representation.

A.4 Posterior Sampling and Model-Free Exploration

Posterior Sampling for Reinforcement Learning (PSRL) (Osband et al., 2013) samples an environment hypothesis and acts according to the corresponding policy, but explicitly maintaining and solving sampled models is difficult in high-dimensional continuous-control problems. Model-free approximations therefore often use bootstrapped critic ensembles or randomized priors to induce temporally coherent exploration (Osband et al., 2016; 2018; 2019). Our construction differs in representing entire policy-conditioned return functions, rather than critic uncertainty only at individual policies, motivating the posterior-sampling-style construction of Section 4. We do not interpret the learned critic ensemble as an exact Bayesian posterior: in particular, our critics do not use bootstrap masks or randomized priors, and critic samples are used during actor optimization rather than as episode-level sampled environment hypotheses.

B Deep Functional Actor–Critic Implementation

This section provides additional architectural and algorithmic details for the neural functional actor–critic implementation used in Section 5.

B.1 Neural Functional Critic Architecture

We maintain an ensemble of n functional critics $\{\hat{Q}^{(i)}\}_{i=1}^{n}$. Each critic contains three components: an actor encoder $E^{(i)}_{\mathrm{act}}$, a state–action encoder $E^{(i)}_{\mathrm{sa}}$, and a joint encoder $E^{(i)}_{\mathrm{joint}}$. Writing $\Xi^{(i)} = \{\xi^{(i)}_{\mathrm{act}}, \xi^{(i)}_{\mathrm{sa}}, \xi^{(i)}_{\mathrm{joint}}\}$, the critic takes the form

$$\hat{Q}^{(i)}(\pi_\theta, s, a; \Xi^{(i)}) = E^{(i)}_{\mathrm{joint}}\left(E^{(i)}_{\mathrm{act}}(\pi_\theta; \xi^{(i)}_{\mathrm{act}}), E^{(i)}_{\mathrm{sa}}(s, a; \xi^{(i)}_{\mathrm{sa}}); \xi^{(i)}_{\mathrm{joint}}\right). \tag{19}$$

For each functional critic, we maintain a target critic with delayed copies of the state–action and joint encoders, while sharing the actor encoder:

$$\hat{Q}^{(i)}_{\mathrm{tar}}(\pi_\theta, s, a) = E^{(i)}_{\mathrm{joint}}\left(E^{(i)}_{\mathrm{act}}(\pi_\theta; \xi^{(i)}_{\mathrm{act}}), E^{(i)}_{\mathrm{sa}}(s, a; \xi^{\prime(i)}_{\mathrm{sa}}); \xi^{\prime(i)}_{\mathrm{joint}}\right). \tag{20}$$

Thus, policy representations evolve jointly with the primary critic, whereas the state–action and joint components of the target critic are updated slowly.

Given a policy π, transition (s, a, r, s'), and target critic i, the TD target is

$$y^{(i)}(\pi; s, a, r, s') = r + \gamma \hat{Q}^{(i)}_{\mathrm{tar}}\left(\pi, s', \pi(s')\right). \tag{21}$$

A key distinction from a standard critic ensemble is that each functional critic is trained across the *entire* current actor ensemble

$$\mathcal{A}_t = \{\pi_{\theta_t^{(j)}}\}_{j=1}^{n}. \tag{22}$$

For a transition minibatch B_C, critic i is therefore trained on the product set $\mathcal{A}_t \times B_C$ using

$$\mathcal{L}_Q^{(i)} = \frac{1}{|\mathcal{A}_t||B_C|} \sum_{\pi \in \mathcal{A}_t} \sum_{\tau \in B_C} \left[y^{(i)}(\pi; \tau) - \hat{Q}^{(i)}(\pi, s_\tau, a_\tau) \right]^2. \tag{23}$$

This cross-policy training is what allows every critic particle to provide value and gradient information for every actor in the ensemble.

B.2 Deterministic Functional Policy Gradient

The functional policy gradient in Eq. equation 2 is written for stochastic policies. Our continuous-control experiments instead use deterministic actors $a = \pi_\theta(s)$. For a functional critic $\hat{Q}(\pi_\theta, s, a; \xi)$, we define the corresponding deterministic surrogate objective as

$$J_\rho(\pi_\theta; \xi) := \mathbb{E}_{s \sim \rho} \left[\hat{Q}(\pi_\theta, s, \pi_\theta(s); \xi) \right]. \tag{24}$$

The policy parameters therefore affect the objective through two distinct computational paths: through the action $a = \pi_\theta(s)$ and through the explicit policy input of the functional critic. Applying the chain rule gives

$$\nabla_\theta J_\rho(\pi_\theta; \xi) = \mathbb{E}_{s \sim \rho} \left[\nabla_\theta \pi_\theta(s) \nabla_a \hat{Q}(\pi_\theta, s, a; \xi) \Big|_{a=\pi_\theta(s)} + \nabla_\theta^{\mathrm{pol}} \hat{Q}(\pi_\theta, s, a; \xi) \Big|_{a=\pi_\theta(s)} \right]. \tag{25}$$

Here, $\nabla_\theta^{\mathrm{pol}} \hat{Q}(\pi_\theta, s, a; \xi)$ denotes differentiation through the explicit policy input of the functional critic while holding its state and action inputs fixed. Thus, the first term in Eq. equation 25 is the standard deterministic policy-gradient term, whereas the second term captures the additional dependence of the predicted return on the policy represented by the functional critic.

In our implementation, both terms in Eq. equation 25 are backpropagated to the actor. Specifically, the first term propagates through the action input $a = \pi_\theta(s)$, while the second propagates through the actor encoder $E_{\mathrm{act}}(\pi_\theta)$ used as the explicit policy input to the critic. For the policy representation described in Section 5, this second pathway differentiates through the actor outputs and hidden activations evaluated at the trainable evaluation states $\{\zeta_j\}$. During the actor update, the critic parameters, evaluation states, and the state–action inputs of this second pathway are held fixed.

For an actor update using critic $\hat{Q}^{(i)}$, we therefore optimize Eq. equation 24 with $\hat{Q} = \hat{Q}^{(i)}$ and backpropagate the full gradient in Eq. equation 25, rather than detaching the explicit policy input.

B.3 Actor Encoder

The actor encoder must represent a high-dimensional policy compactly while remaining differentiable with respect to the actor parameters. Motivated by evaluation functionals, we introduce a set of trainable evaluation states

$$\mathcal{Z} = \{\zeta_k\}_{k=1}^{N_\zeta}, \tag{26}$$

which serve as learned probes of the input policy. For each ζ_k, the actor is evaluated and the resulting action together with selected hidden activations is used to construct a policy token. The resulting sequence of policy tokens is processed by a transformer encoder to obtain $E_{\mathrm{act}}(\pi_\theta)$.

In our current implementation, the policy tokens use activations from the final two layers of the actor network, providing a lightweight representation of the policy while avoiding direct encoding of all actor parameters. The evaluation states $\{\zeta_k\}$ are themselves trainable and are optimized end-to-end through the functional-critic TD loss in Eq. equation 23.

This construction may be viewed as learning a set of evaluation functionals δ_{ζ_k} over the policy space, where $\delta_{\zeta_k}(\pi) = \pi(\zeta_k)$, augmented with internal actor representations. It provides the functional critic with a policy-dependent representation without requiring an explicit vectorization of the full actor network.

B.4 Ensemble Functional Actor-Critic for Posterior-Sampling Exploration

We use an ensemble of deterministic actors $\mathcal{A}_t = \{\pi_{\theta_t^{(i)}}\}_{i=1}^n$. The actor ensemble plays two roles. First, it supplies the policy training set on which each functional critic is fit. Second, it provides multiple candidate policies that can be improved under different sampled critic particles.

Importantly, actors and functional critics are *not* permanently paired. At every actor update, actor i independently samples an index

$$c_t^{(i)} \sim \mathrm{Unif}(\{1, \dots, n\}), \tag{27}$$

and is updated using the functional policy gradient induced by $\hat{Q}^{(c_t^{(i)})}$:

$$\theta_{t+1}^{(i)} = \theta_t^{(i)} + \eta_t \nabla_\theta J_\rho \left(\pi_\theta; \xi^{(c_t^{(i)})}\right) \Big|_{\theta=\theta_t^{(i)}}. \tag{28}$$

Because every functional critic is defined over the policy space, any critic particle can be used to improve any actor. Resampling a critic for every actor update implements the particle-based posterior-sampling mechanism described in Section 4.

B.5 Ensemble Functional Actor–Critic for Posterior-Sampling-Style Exploration

Algorithm 3 instantiates the posterior-sampling-style construction described in Section 4. We maintain n actors and n functional critics. Each critic is trained across the current actor ensemble, while actor updates independently sample critics from the ensemble and use their corresponding functional policy gradients.

Algorithm 3 Ensemble Functional Actor–Critic

Input: $\{\theta_0^{(i)}\}_{i=1}^n$, $\{\xi_0^{(j)}\}_{j=1}^n$, T, $\{\eta_t\}_{t=1}^T$
for $t = 1, \dots, T$ **do**
  Collect/sample off-policy data $\mathcal{D}_t$
  $\mathcal{A}_{t-1} \leftarrow \{\pi_{\theta_{t-1}^{(i)}}\}_{i=1}^n$
  {Update functional-critic ensemble}
  for $j = 1, \dots, n$ **do**
    $\xi_t^{(j)} \leftarrow$ FunctionalPolicyEvaluator$(\xi_{t-1}^{(j)}, \mathcal{A}_{t-1}, \mathcal{D}_t)$
  end for
  {Update actors with independently sampled critics}
  for $i = 1, \dots, n$ **do**
    Sample $c_t^{(i)} \sim \mathrm{Unif}(\{1, \dots, n\})$
    $\theta_t^{(i)} \leftarrow \theta_{t-1}^{(i)} + \eta_t \nabla_\theta J_\rho(\pi_\theta; \xi_t^{(c_t^{(i)})})\big|_{\theta=\theta_{t-1}^{(i)}}$
  end for
end for
Return $\{\theta_T^{(i)}\}_{i=1}^n$

For the deterministic actors used in our experiments, the actor update in Algorithm 3 uses the full deterministic functional policy gradient in Eq. equation 25, including both the action-gradient pathway and the derivative through the explicit policy input.

B.6 Training Algorithm

Algorithm 4 summarizes the complete neural functional actor–critic training procedure. Each functional critic is trained across the current actor ensemble, while each actor independently samples a functional critic at every actor update.

In all main experiments, we use $n = 10$ functional critics and a critic-to-actor update ratio of $G = 10$. Additional hyperparameters are reported in Table 2.

Algorithm 4 Neural Ensemble Functional Actor–Critic

Input: ensemble size n, critic-to-actor update ratio G, actor learning rate η, target-update coefficient ρ
Initialize actor parameters $\{\theta_0^{(i)}\}_{i=1}^n$
Initialize functional critics $\{\Xi_0^{(i)}\}_{i=1}^n$
Initialize target critic parameters $\{(\xi_{\mathrm{sa}}^{'(i)}, \xi_{\mathrm{joint}}^{'(i)})\}_{i=1}^n$
Initialize replay buffer $\mathcal{R}$
for $t = 0, 1, \ldots$ **do**
  if starting a new episode **then**
    Sample rollout actor $r \sim \mathrm{Unif}(\{1, \ldots, n\})$
  end if
  Execute $a_t = \pi_{\theta_t^{(r)}}(s_t)$ and store (s_t, a_t, r_t, s_{t+1}) in $\mathcal{R}$
  $\mathcal{A}_t \leftarrow \{\pi_{\theta_t^{(i)}}\}_{i=1}^n$
  {Update functional-critic ensemble}
  for $g = 1, \ldots, G$ **do**
    Sample transition batch $B_C \sim \mathcal{R}$
    for $j = 1, \ldots, n$ **do**
      Compute TD targets Eq. 21 on $\mathcal{A}_t \times B_C$
      Update $\Xi^{(j)}$ by minimizing Eq. 23
      Soft-update target critic:

$$\xi_{\mathrm{sa}}^{'(j)} \leftarrow \rho \xi_{\mathrm{sa}}^{'(j)} + (1-\rho)\xi_{\mathrm{sa}}^{(j)}$$
$$\xi_{\mathrm{joint}}^{'(j)} \leftarrow \rho \xi_{\mathrm{joint}}^{'(j)} + (1-\rho)\xi_{\mathrm{joint}}^{(j)}$$

    end for
  end for
  Sample actor-update batch $B_A \sim \mathcal{R}$
  {Update actors with independently sampled critics}
  for $i = 1, \ldots, n$ **do**
    Sample $c^{(i)} \sim \mathrm{Unif}(\{1, \ldots, n\})$
    Evaluate the deterministic functional policy gradient in Eq. 25 using $\hat{Q}^{(c^{(i)})}$ and B_A
    $\theta_{t+1}^{(i)} \leftarrow \theta_t^{(i)} + \eta \hat{\nabla}_\theta J_\rho(\pi_{\theta_t^{(i)}}; \hat{Q}^{(c^{(i)})})$
  end for
end for

B.7 Omitted Off-Policy AC Heuristics

The minimalist implementation deliberately excludes several techniques widely used in modern off-policy actor–critic implementations.

For critic learning, we do not use clipped double-Q or minimum-target value estimation (Fujimoto et al., 2018; Haarnoja et al., 2018). For actor learning and data collection, the actors are deterministic and we do not inject exploration noise. We further omit target-policy smoothing (Fujimoto et al., 2018) and entropy regularization or automatic entropy tuning (Haarnoja et al., 2018).

These choices are not intended to argue that the corresponding heuristics are generally unnecessary. Rather, they isolate a deliberately simple implementation that remains close to the structural ingredients analyzed in the paper: policy-conditioned critic generalization and functional-ensemble exploration.

C Experimental Details

C.1 Benchmarks and Evaluation Protocol

We evaluate the complete Dog and Humanoid task series used in our benchmark from the DeepMind Control Suite (Tunyasuvunakool et al., 2020): Humanoid Stand, Humanoid Walk, Humanoid Run,

Dog Stand, Dog Walk, Dog Trot, Dog Run, and Dog Fetch. These task families provide two distinct morphologies with relatively high-dimensional state and action spaces.

For each task, experiments are run with four parallel environments and four independent random seeds. Figure 2 reports mean episodic return as a function of environment steps, with shaded regions denoting one standard deviation. We evaluate the full set of tasks in these two series rather than selecting individual tasks.

C.2 Baseline Construction

We compare against the original SAC and TD3 algorithms (Haarnoja et al., 2018; Fujimoto et al., 2018), as well as strengthened variants SAC+ and TD3+ implemented using the extensively tested ALF codebase (Xu et al., 2021).

The vanilla SAC and TD3 baselines retain their original two-critic architectures and critic-to-actor update ratios of 1:1 and 2:1, respectively. SAC+ and TD3+ use 10 critics and, after independent tuning, a 10:1 critic-to-actor update ratio. Our method likewise uses 10 functional critics and a 10:1 ratio in all main experiments.

Our method also uses 10 functional critics and the same 10:1 critic-to-actor update ratio in all main experiments. Thus, the main comparison against SAC+ and TD3+ matches both the critic ensemble size and the frequency of critic updates, while retaining method-specific architectural components. SAC+ retains the stochastic Gaussian actor and entropy regularization of SAC, without introducing additional exploration mechanisms beyond those in the original algorithm. TD3+ retains the deterministic actor of TD3, together with its additive rollout exploration noise and target-policy smoothing noise. Our method uses the deterministic functional-actor ensemble described in Appendix B.

C.3 Prior Policy-Conditioned Value Functions

We did not include earlier policy-conditioned value-function (PCVF) methods as primary baselines. A comprehensive benchmark by Griesbach (2023) found the major PCVF variants to be noncompetitive with standard RL algorithms such as TD3, consistent with our own preliminary attempts. Our experiments therefore focus on the substantially stronger SAC+ and TD3+ baselines; we refer readers to Griesbach (2023) for a detailed comparison among earlier PCVF approaches.

Table 2: Hyperparameters of our functional actor–critic implementation, shared across all reported tasks.

Hyperparameter	**Value**
optimizer	Adam
learning rate	3×10^{-4}
batch size	256
discount factor (γ)	0.99
target update rate (τ)	0.005
critic ensemble size	10
critic-to-actor update ratio	10
actor network hidden layers	$(256, 256)$
critic state–action encoder hidden layers	$(256, 256)$
critic joint encoder hidden layers	$(256, 256)$
actor transformer encoder layers	4
actor transformer encoder heads	1
trainable evaluation states	512
actor layers exposed to actor encoder	last two
actor encoding dimension	128
state–action encoding dimension	64
activation function	ReLU

Table 3: Key algorithmic differences among the compared methods.

	Ours	SAC+	TD3+	SAC	TD3
Number of critics	10	10	10	2	2
Critic-to-actor update ratio	10:1	10:1	10:1	1:1	2:1
Layer normalization	Yes	Yes	Yes	No	No
Actor policy	Det.	Stoch.	Det.	Stoch.	Det.
Additional rollout action noise	No	No	Yes	No	Yes
Target-policy smoothing	No	No	Yes	No	Yes
Entropy regularization	No	Yes	No	Yes	No

C.4 Hyperparameters and Baseline Configurations

Hyperparameters of our method used across the experiments reported in Fig. 2 are summarized in Table 2. Key algorithmic differences among the compared methods are summarized in Table 3.

D Heuristic Solution to Replay Sampling

In this section, we develop a heuristic neural instantiation of the replay sampling mechanism proposed in Section 3.3. Starting from the base neural ensemble functional actor–critic in Algorithm 4, we modify only the critic replay update: instead of using uniform replay weights, we compute a feature-based sampling distribution over the replay minibatch and use it to form a weighted TD objective. Because the linear features required by the theory are not directly available in the neural implementation, we approximate them using frozen critic representations and evaluate the resulting modification on Dog_run and Humanoid_run.

D.1 Feature Covariance with Frozen Neural Features

While our convergence guarantees for Algorithm 2 are based on the linear functional representation **Assumption 2**, practical implementations typically adopt a nonlinear neural parameterization of the value function, such as Section 5. Note that this does not contradict with **Assumption 2**, because the latter only assumes the existence of a linear representation in hindsight, which needs not be the neural value function we use in practice. Nonetheless, this does bring difficulty as we do not know the linear feature representation ϕ^π and thereby cannot directly approximate the feature covariance with Monte Carlo methods.

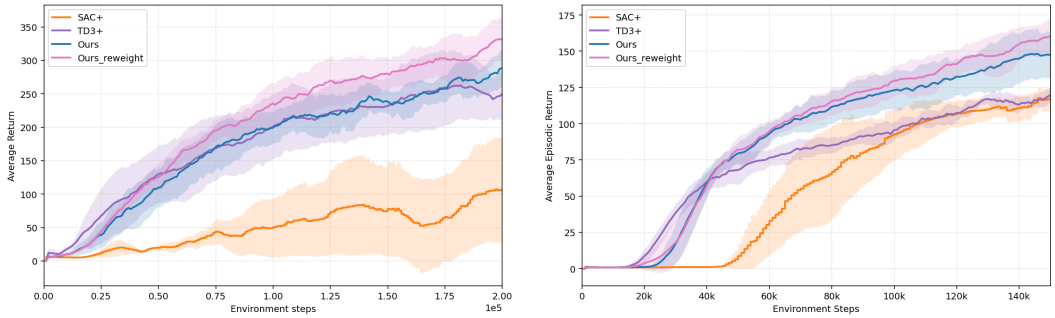
To address this issue, we propose freezing the reference features at each update of the critic network $\hat{Q}^{(i)}(\pi, s, a)$ in the critic ensemble. To retain the linear structure in **Assumption 2**, we take the penultimate activation as the linear feature $\hat{\phi}^{(i),\pi}$ and use them to approximate the population covariance $\hat{\mathbf{A}}_\pi$ by a sample average, where the expectation in equation 3 over d^π is replaced by state-action samples from a small recent buffer. The weighted covariance matrices over the buffer in equation 13 also use the frozen reference features above; but to reduce the computational burden, we first uniformly draw a minibatch $n_{\text{mini}} < n$ from the entire size-n buffer $\mathcal{B}$ and then assemble the matrices in equation 13 over the minibatch and solve the distribution design $\hat{\mathbf{w}}$ in equation 17 over the minibatch.

D.2 Approximately Solving the Nonconvex Program

It remains to solve the generally nonconvex optimization program in equation 17 for the distribution design $\hat{\mathbf{w}}$ over the drawn minibatch $\mathcal{B}_{\text{mini}}$. For an approximate solution, we parameterize $\hat{\mathbf{w}}$ with a softmax representation

$$\hat{\mathbf{w}}_i = \frac{\min\{p_i, p_{\max}\}}{\sum_{j \in \mathcal{B}_{\text{mini}}} \min\{p_j, p_{\max}\}}, \qquad p_i = \mathrm{softmax}(\mathbf{z})_i$$

where $p_{\max} \in (0, 1]$ is a hyperparameter ceiling and $\mathbf{z} \in \mathbb{R}^{n_{\text{mini}}}$ is the trainable parameter. We initialize $\mathbf{z}$ to the zero vector so $\hat{\mathbf{w}}$ is initially uniform, and then we apply K_{lbfgs} steps of the L-



(a) Dog_run (b) Humanoid_run

Figure 3: Targeted evaluation of the replay-reweighted variant `Ours_reweight` on Dog_run and Humanoid_run. Starting from Algorithm 4, we modify only the critic replay update: for each candidate replay minibatch, we compute the heuristic distribution $\hat{\mathbf{w}}$ described above and use it to weight the squared TD loss in Eq. equation 23. We use $p_{\max} = 0.1$ and $K_{\text{lbfgs}} = 1$. On Dog_run, replay reweighting is enabled from the beginning for all four runs and improves environment-sample efficiency over the original algorithm. On Humanoid_run, early activation was not beneficial, motivating delayed activation after the initial feature-learning phase; because activation timing is not identical across all runs, this panel is intended as a diagnostic of feature-dependent activation rather than a fixed-protocol four-seed comparison.

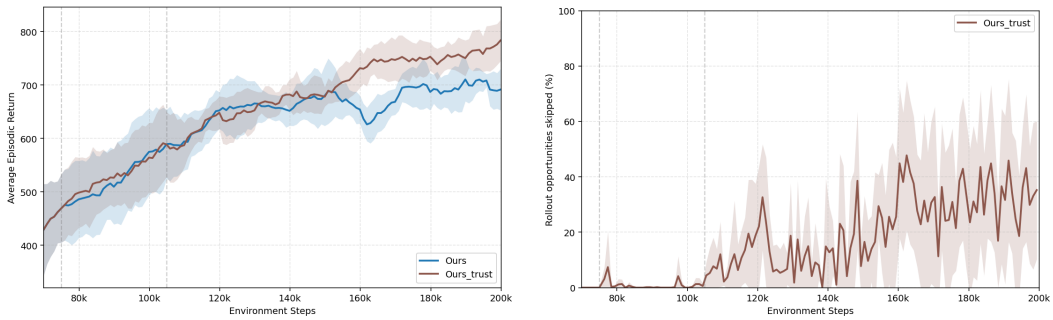
BFGS optimizer to optimize the objective in equation 17 over $\mathbf{z}$ and arrive at the final distribution design $\hat{\mathbf{w}}$.

Modification to Algorithm 4. All components of Algorithm 4 are kept unchanged except for the critic replay update. For each critic update, we first draw a candidate minibatch $\mathcal{B}_{\text{mini}}$ from the replay buffer, compute the heuristic replay weights $\hat{\mathbf{w}}$ described above, and replace the uniformly averaged critic TD loss by its $\hat{\mathbf{w}}$-weighted counterpart. We refer to this variant as `Ours_reweight`.

Delayed Replay Sampling. Another practical distinction from the convergence analysis in Section 3.2 is that the feature representation ϕ^π is not known and needs to be learned over time. Recall that we use the penultimate activation of our neural functional critics as the approximate feature representation $\hat{\phi}^{(i),\pi}$. It is therefore possible that a feature-based module like Replay Sampling can be unhelpful or even harmful in the early stage, when the functional critics have not trained sufficiently and the learned features are misrepresented. Indeed, in the Humanoid_run task in Figure 3, we found that enabling Replay Sampling from the beginning is not helpful compared to `Ours`. Thus in Figure 3b, we enable Replay Sampling only after 90k environment steps for three of the four independent runs. Figure 3a enables Replay Sampling from the beginning for all four independent runs. These observations suggest that feature-based replay control becomes useful only after the learned neural representation is sufficiently informative. We therefore view the Humanoid_run experiment as a diagnostic of this feature-maturation effect rather than as a controlled comparison under a single activation schedule. Automatically determining when to activate the feature-based replay mechanism remains an open practical question.

E Heuristic Neural Instantiation of Trust Coverage

This section develops a heuristic neural instantiation of the trust-coverage-based behavior-policy refresh mechanism in Algorithm 2. Specifically, we integrate the evaluation-trust criterion $\mathcal{C}(\pi, \mu)$ from Algorithm 2 into the neural ensemble functional actor–critic of Algorithm 4. This experiment instantiates only the evaluation-trust component of Algorithm 2; the gradient-generalization criterion $\Delta_{k,t}$ governing reuse of the evaluated critic across actor updates is not implemented here. The actor and critic updates otherwise follow Algorithm 4; the modification concerns only whether a new behavior-policy rollout is collected or the existing behavior data are reused.



(a) Dog_walk return curve

(b) Skip rate of behavior policy refresh based on trust coverage

Figure 4: Targeted evaluation of the trust-guided behavior-reuse variant `Ours_trust` on Dog_walk. The variant combines the neural ensemble functional actor–critic in Algorithm 4 with the evaluation-trust behavior-refresh criterion from Algorithm 2. Before collecting a fresh rollout, we compute the neural approximation of $\mathcal{C}(\pi, \mu)$ described above and skip the rollout when the estimate remains below the chosen threshold. To study delayed activation of the learned feature-based criterion, two runs enable refresh skipping at 75k environment steps and two at 105k environment steps (dashed lines). The threshold for each run is chosen as the 33-percentile of empirically estimated $\mathcal{C}(\pi, \mu)$ values over random subsets of the replay buffer. The trust-based variant improves environment-sample efficiency in this targeted study while skipping a substantial fraction of behavior-policy refreshes. The improvement coincides with periods of frequent refresh skipping, providing initial evidence that the evaluation-trust criterion can serve as a practical signal for data reuse.

Because Algorithm 2 is formulated under linear functional approximation whereas Algorithm 4 uses neural functional critics, implementing the trust criterion requires neural approximations of the population feature covariances $\mathbf{A}_\pi$ and $\widetilde{\mathbf{A}}_\pi$, and hence of $\mathcal{C}(\pi, \mu)$ in Eq. equation 4.

As in Appendix D.1, we use the penultimate activations of frozen functional critics as approximations to the linear feature representation in Assumption 2. For the expectation over the behavior-policy stationary distribution d^μ in $\widetilde{\mathbf{A}}_\pi$, we use the sample average over the large replay buffer $\mathcal{B}$. The expectations over the target-policy stationary distribution d^π in $\mathbf{A}_\pi$ and Eq. equation 4 are approximated using a much smaller recent buffer $\mathcal{B}_{\text{recent}}$.

Our goal here is not a full benchmark of the trust-coverage mechanism, but a targeted study of whether the learned trust signal can identify periods in which behavior-policy refreshes may be safely skipped and data reused.

Trust-Guided Behavior Reuse. Algorithm 2 refreshes the behavior policy when the evaluation-trust criterion indicates insufficient coverage of the current reference policy. We transfer this mechanism to Algorithm 4 by evaluating a neural approximation of $\mathcal{C}(\pi, \mu)$ before collecting a new rollout. When the estimated trust metric remains below the prescribed threshold, the new rollout is skipped and the existing behavior data are reused; otherwise, a fresh rollout is collected as in the base neural algorithm. We refer to this variant as `Ours_trust`.

Because the trust criterion is computed from learned neural features rather than the unknown linear features in Assumption 2, its reliability can depend on the quality of the learned representation. In particular, activating trust-based refresh skipping too early, when the critic features are still poorly learned, can be unhelpful or even harmful. We therefore study delayed activation of this mechanism in the experiment below.

F Proof for Section 3.2

F.1 Discussion of the Linear Functional Approximation Assumptions

Assumption 1 is a standard ergodicity condition for stochastic approximation with Markovian noise (Zhang et al., 2020; 2021; Chen et al., 2023). Some prior works impose slightly weaker conditions;

we assume irreducibility and aperiodicity for simplicity and to ensure a unique, well-behaved stationary distribution.

Assumption 2 generalizes the standard linear realizability assumption by allowing the feature representation $\phi^\pi(s,a)$ to depend explicitly on the policy. Thus, rather than representing the value function of a single fixed policy, the functional features jointly represent the mapping from the policy class Π to the corresponding value functions. Geometrically, this representation can be interpreted locally through a first-order expansion over the policy manifold, with policy-dependent features capturing variations of the value function along its tangent directions.

We consider the policy space Π as a smooth Riemannian manifold. For any fixed state-action pair (s,a), the value function can be understood as a functional map $\mathcal{Q}_{s,a} : \Pi \to \mathbb{R}$. While standard linear function approximation focuses on the representability of a single policy, Assumption 2 leverages the intrinsic local continuity and smoothness of the policy manifold Π.

For any anchored reference policy $\pi_0 \in \Pi$, let $T_{\pi_0}\Pi$ denote the tangent space at π_0, with the associated exponential map $\exp_{\pi_0}(\cdot) : T_{\pi_0}\Pi \to \Pi$ and its inverse, the logarithmic map $\log_{\pi_0}(\cdot) : \Pi \to T_{\pi_0}\Pi$. For any policy π in a sufficiently small neighborhood of π_0, a first-order Taylor expansion on the manifold yields:

$$Q^\pi(s,a) = \mathcal{Q}_{s,a}(\exp_{\pi_0}(\log_{\pi_0}(\pi))) \approx Q^{\pi_0}(s,a) + \langle \nabla_\pi \mathcal{Q}_{s,a}(\pi_0), \log_{\pi_0}(\pi)\rangle_{T_{\pi_0}\Pi},$$

where $\nabla_\pi \mathcal{Q}_{s,a}(\pi_0)$ represents the functional gradient of the value with respect to the anchor policy.

By identifying a set of basis functions $\Phi(s,a)$ that span the directional derivatives of the value functional, the displacement vector $\log_{\pi_0}(\pi) \in T_{\pi_0}\Pi$ interacts linearly with the state-action features. Specifically, we can define the **functional feature map** as the tensor product $\phi^\pi(s,a) := \mathrm{vec}(\Phi(s,a) \otimes \log_{\pi_0}(\pi))$. Under this coordinate-free construction, the value function locally satisfies:

$$Q^\pi(s,a) \approx \phi^\pi(s,a)^\top w_{\pi_0}. \tag{29}$$

In this light, Assumption 2 holds locally for any π close to π_0, with the underlying representation depending on the choice of the anchor policy. Our dual-layer mechanism is precisely designed to respect the boundaries of this local validity: while the first layer ($\mathcal{C}(\pi,\mu)$) ensures the data is sufficient to build the anchored critic, the second layer ($\Delta_{k,t}$) monitors the displacement on the tangent space. Crucially, the **Gradient Generalization Trust Metric** $\Delta_{k,t}$ serves as a data-aware proxy for the Riemannian distance $d(\pi,\pi_0)$ on the manifold. By enforcing $\Delta_{k,t} \le 2$, Algorithm 2 refreshes the anchor policy exactly when the policy π drifts beyond the "trust region" where the manifold's first-order approximation—and thus the functional critic's gradient—remains reliable.

Assumption 3 similarly extends approximate Bellman completeness to the functional setting. Unlike the Linear MDP assumption used in some prior work (Xiong et al., 2022; Yin et al., 2022; Min et al., 2021), it does not require the transition dynamics themselves to admit a linear representation in the features.

F.2 Additional Details for Target-Based Evaluation with Partial Coverage

We provide additional details for the target-based policy evaluator analyzed in Section 3.2, together with further discussion of the partial-coverage guarantee.

Target-based policy evaluator. For a target policy π and behavior policy μ, we use the following target-network TD procedure.

The auxiliary target weights $\hat{\xi}_t$ are held fixed within each inner loop and refreshed after K TD updates. Such target-based procedures are well studied under linear function approximation (Carvalho et al., 2020; Chen et al., 2023; Zhang et al., 2021). Our contribution is to establish convergence without requiring pointwise full coverage by the behavior policy.

Interpretation of feature-based coverage. Recall the on-policy and off-policy feature covariances

$$\mathbf{A}_\pi = \mathbb{E}_{d^\pi}\left[\phi^\pi(s,a)\phi^\pi(s,a)^\top\right], \qquad \widetilde{\mathbf{A}}_\pi = \mathbb{E}_{d^\mu}\left[\phi^\pi(s,a)\phi^\pi(s,a)^\top\right].$$

Algorithm 5 Target-Network Based Policy Evaluation

Input: behavior policy μ; target policy π; feature map ϕ^π; accuracy level ϵ; inner iterations K; outer iterations T
Initialize: $\hat{\xi}_0 \leftarrow \mathbf{0}$
for $t = 0, 1, \ldots, T-1$ **do**
  Initialize $\xi_{t,0} \leftarrow \mathbf{0}$
  for $k = 0, 1, \ldots, K-1$ **do**
    Sample (s_k, a_k, r_k, s'_k) from μ
    $y_k \leftarrow r_k + \gamma \sum_a \pi(a|s'_k) \left[\phi^\pi(s'_k, a)^\top \hat{\xi}_t\right]$
    $\xi_{t,k+1} \leftarrow \xi_{t,k} + \alpha \phi^\pi(s_k, a_k) \left(y_k - \phi^\pi(s_k, a_k)^\top \xi_{t,k}\right)$
  end for
  $\hat{\xi}_{t+1} \leftarrow \xi_{t,K}$
end for
Output: $\hat{\xi}_T$

The evaluation trust metric $\mathcal{C}(\pi, \mu)$ in equation 4 measures whether data collected under μ cover the feature directions that matter under the target policy π. Unlike pointwise full coverage, this condition allows $\widetilde{\mathbf{A}}_\pi$ to be singular, provided the feature directions required by π lie in the span supported by the behavior data. The inverse norm in equation 4 is therefore interpreted through the limiting regularization, equivalently the Moore–Penrose pseudoinverse on the supported subspace.

The key analytical step behind Theorem 1 is to establish Bellman contraction in the L^2 norm weighted by the on-policy covariance $\mathbf{A}_\pi$. This choice of metric allows the analysis to ignore uncovered directions that are irrelevant to the target policy and is what permits the weaker feature-based coverage requirement.

Specialization to standard linear function approximation. Theorem 1 only requires Assumptions 2 and 3 for the particular target policy being evaluated. Consequently, it also applies directly to the standard linear function approximation setting (Carvalho et al., 2020; Chen et al., 2023; Zhang et al., 2021). When $\mathcal{C}(\pi, \mu) < \infty$, the resulting bound has the same ϵ-dependence as least-squares policy-evaluation guarantees (Min et al., 2021; Duan et al., 2020).

Weight-space consequence. The value-space guarantee in Theorem 1 immediately implies a corresponding bound on the critic parameters:

Corollary 1 (Convergence in Weight Space)**.** *Under the conditions of Theorem 1,*

$$\mathbb{E}\left[\|\hat{\xi}_T - \xi^\star\|^2_{\mathbf{A}_\pi}\right] \lesssim \gamma^T + \mathcal{C}(\pi, \mu)\left(\epsilon^2 + \mathcal{E}^2_{\text{approx}}\right) + \mathcal{E}^2_{\text{approx}}.$$

This parameter-space form is used in the analysis of the full functional actor-critic algorithm.

F.3 Additional Details for Dual Trust-Coverage

We provide additional details for the dual trust-coverage mechanism analyzed in Section 3.2, including the derivation of the gradient-generalization trust metric and several useful special cases.

Notation. Within epoch k, let $\pi_t^{(k)} := \pi_{\theta_t^{(k)}}$ denote the current actor and $\bar{\pi}^{(k)}$ the reference policy at which the functional critic is evaluated. We write

$$\phi_t^{(k)}(s, a) := \phi^{\pi_t^{(k)}}(s, a), \qquad \mathbf{A}_k := \mathbf{A}_{\bar{\pi}^{(k)}}.$$

The covariance $\mathbf{A}_k$ remains fixed within epoch k and serves as the metric anchor for both trust criteria.

Reference-policy evaluation. By Theorem 1, whenever

$$\mathcal{C}(\bar{\pi}^{(k)}, \mu) \leq C_{\max},$$

the critic parameter $\xi^{(k)}$ obtained by evaluating $\bar{\pi}^{(k)}$ satisfies

$$\sqrt{\mathbb{E}\big[\|\xi^{(k)} - \xi^\star\|^2_{\mathbf{A}_k}\big]} \lesssim \bar{\epsilon} := \sqrt{C_{\max}}\,(\epsilon + \mathcal{E}_{\text{approx}})\,. \tag{30}$$

Thus, the first trust criterion controls the accuracy of the functional critic at the reference policy using off-policy experience.

Gradient-error decomposition. The functional policy-gradient estimate at the current actor $\pi_t^{(k)}$ reuses the anchored critic $\xi^{(k)}$. Under linear functional representation, the resulting gradient error decomposes as

$$\begin{aligned}\hat{\nabla} J_\rho(\pi_t^{(k)}, \xi^{(k)}) - \nabla J_\rho(\pi_t^{(k)}) = &\underbrace{\mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}} \left\langle \phi_t^{(k)}(s,a), \xi^{(k)} - \xi^\star \right\rangle \nabla \pi_t^{(k)}(a|s)\right]}_{:=\mathcal{G}_1^{(k,t)}} \\ &+ \underbrace{\mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}} \pi_t^{(k)}(a|s) D\phi_t^{(k)}(s,a)(\xi^{(k)} - \xi^\star)\right]}_{:=\mathcal{G}_2^{(k,t)}}.\end{aligned} \tag{31}$$

The first term captures amplification of critic error through the functional features themselves, while the second captures amplification through their policy derivatives.

Gradient estimation guarantee. Let $e_k := \xi^{(k)} - \xi^\star$ and write $\delta_{k,t} := \hat{\nabla} J_\rho(\pi_t^{(k)}, \xi^{(k)}) - \nabla J_\rho(\pi_t^{(k)})$. With the feature defined in equation 8, we have $\delta_{k,t} = \mathbf{G}_{\pi_t^{(k)}} e_k$. Since $\pi_0^{(k)} = \bar{\pi}^{(k)}$, Assumption 4 gives $\|\delta_{k,0}\|_2^2 = e_k^\top \mathbf{G}_{\bar{\pi}^{(k)}}^\top \mathbf{G}_{\bar{\pi}^{(k)}} e_k \le \kappa_{\mathrm{g}}^2 \|e_k\|^2_{\mathbf{A}_k}$. Thus the evaluation guarantee at the start of a fixed epoch transfers to its first gradient estimate with amplification at most κ_{g}^2.

Gradient-generalization trust metric. To control these two terms using equation 30, define

$$\Delta_{k,t} := \max\{\mathcal{C}_1^{(k,t)}, \mathcal{C}_2^{(k,t)}\}, \tag{32}$$

where

$$\mathcal{C}_1^{(k,t)} := \left\| \mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}} |\nabla \pi_t^{(k)}(a|s)|\, \|\phi_t^{(k)}(s,a)\|_{\mathbf{A}_k^{-1}}\right]\right\|_\infty, \tag{33}$$

$$\mathcal{C}_2^{(k,t)} := \mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}} \pi_t^{(k)}(a|s)\, \|D\phi_t^{(k)}(s,a)\|_{\mathbf{A}_k^{-1},\infty}\right]. \tag{34}$$

By Cauchy–Schwarz,

$$\|\mathcal{G}_1^{(k,t)}\|_\infty \le \mathcal{C}_1^{(k,t)} \|\xi^{(k)} - \xi^\star\|_{\mathbf{A}_k}, \qquad \|\mathcal{G}_2^{(k,t)}\|_\infty \le \mathcal{C}_2^{(k,t)} \|\xi^{(k)} - \xi^\star\|_{\mathbf{A}_k}.$$

Combining these inequalities with equation 30 gives

$$\mathbb{E}\Big[\|\nabla J_\rho(\pi_t^{(k)}) - \hat{\nabla} J_\rho(\pi_t^{(k)}, \xi^{(k)})\|_\infty\Big] \lesssim \Delta_{k,t}\bar{\epsilon}. \tag{35}$$

Hence, $\Delta_{k,t}$ quantifies how reliably the anchored functional critic generalizes from $\bar{\pi}^{(k)}$ to the evolving target policy $\pi_t^{(k)}$.

Smoothness. We use the following standard regularity condition to guarantee that the trust metric remains controlled for sufficiently small policy updates.

Assumption 5 (Smoothness)**.** *There exists $L > 0$ such that for any $\theta, \theta' \in \Theta$ and all $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$\begin{aligned}\|\phi^{\pi_\theta}(s,a) - \phi^{\pi_{\theta'}}(s,a)\|_2 \le L\|\theta - \theta'\|_2, \quad \|D\phi^{\pi_\theta}(s,a) - D\phi^{\pi_{\theta'}}(s,a)\|_2 \le L\|\theta - \theta'\|_2, \\ |\pi_\theta(a|s) - \pi_{\theta'}(a|s)| \le L\|\theta - \theta'\|_2, \quad \|\nabla_\theta \pi_\theta(a|s) - \nabla_\theta \pi_{\theta'}(a|s)\|_2 \le L\|\theta - \theta'\|_2.\end{aligned}$$

This is the standard continuity requirement used to control policy drift in actor-critic convergence analyses (Zhang et al., 2020; 2021; Barakat et al., 2022).

Interpretation of the two trust layers. The two criteria control distinct sources of policy-gradient error. The threshold $C_{\max}$ controls the reference-policy evaluation error by requiring sufficient feature coverage from the behavior data. The threshold on $\Delta_{k,t}$ controls the additional gradient-generalization error incurred as the actor evolves away from the reference policy. Thus, the algorithm refreshes the behavior policy only when off-policy experience no longer provides sufficient coverage, and refreshes the functional critic only when its gradient estimate no longer generalizes reliably to the current actor.

This separation is also what enables sample-efficient reuse: behavior-policy data may be reused across multiple reference-policy evaluations as long as the first trust criterion holds, while a single functional critic may be reused across multiple actor updates as long as the second criterion holds.

Remark 2 (Similarity-Based Generalization). The benefit of critic reuse depends on the functional critic generalizing across nearby policies. In the limiting case where the effective epoch length is one, so that the critic is refreshed after every actor update, the framework reduces to a standard two-timescale actor-critic scheme. Functional generalization therefore extends the classical setting by permitting multiple policy-improvement steps per critic evaluation whenever policy similarity can be exploited.

Remark 3 (Behavior-Policy Special Cases). Several common behavior-policy choices arise as special cases. Setting $\mu = \bar{\pi}^{(k)}$ ensures coverage by construction and recovers an on-policy actor-critic regime, while a fixed behavior policy satisfying conventional full-coverage conditions yields a standard off-policy regime. The dual trust-coverage mechanism additionally permits intermediate settings in which μ is reused until its feature coverage becomes insufficient.

F.4 A Finite-time Bound for Markovian SAA under (Possibly) Degenerate Norm

In this section, we provide a detailed proof of a finite-time contraction bound for Markovian stochastic approximation under a possibly degenerate design, which serves as a foundation for our proof of Theorem 1 in the next section. The proof follows a Lyapunov drift argument as in Chen et al. (2022) while carefully tracking the semi-norm induced by a positive semi-definite matrix.

Notations. For any positive semi-definite matrix $\mathbf{A} \succeq \mathbf{0}$ and vectors $\mathbf{u}, \mathbf{v}$ with the same dimension, we define $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{A}} := \mathbf{u}^\top \mathbf{A} \mathbf{v}$ and $\|\mathbf{v}\|_{\mathbf{A}}^2 := \mathbf{v}^\top \mathbf{A} \mathbf{v}$. Let $\lambda_{\max}(\mathbf{A})$ denote the largest eigenvalue of $\mathbf{A}$. When $\mathbf{A}$ is singular, let $\lambda_{\min}^+(\mathbf{A})$ denote the smallest *strictly positive* eigenvalue of $\mathbf{A}$ (and W.L.O.G. $\lambda_{\min}^+(\mathbf{A}) = 0$ when $\mathbf{A} = \mathbf{0}$). Let $\mathbf{\Pi}_{\mathbf{A}}$ be the orthogonal projection operator onto $\mathrm{Range}(\mathbf{A})$ and $\mathbf{\Pi}_{\mathbf{A}}^{\perp}$ the orthogonal projection operator onto $\ker(\mathbf{A})$, so that $\mathbf{\Pi}_{\mathbf{A}} + \mathbf{\Pi}_{\mathbf{A}}^{\perp} = \mathbf{I}$. For a finite-state Markov chain $\{\mathbf{y}_k\}_{k \geq 0}$ over $\mathcal{Y}$ with transition kernel P and stationary distribution ν, we define the β-mixing time as

$$\tau_\beta := \min\left\{k \geq 0 : \max_{y \in \mathcal{Y}} \|P^k(y, \cdot) - \nu(\cdot)\|_{\mathrm{TV}} \leq \beta\right\}.$$

Theorem 3. *Consider the Markovian stochastic approximation recursion with arbitrary initialization $\mathbf{x}_0 \in \mathbb{R}^d$ and*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha\, \mathbf{F}(\mathbf{x}_k, \mathbf{y}_k). \tag{36}$$

Let $\bar{\mathbf{F}}(\mathbf{x}) := \mathbb{E}_{\mathbf{y} \sim \nu}[\mathbf{F}(\mathbf{x}, \mathbf{y})]$. Suppose that for some $\mathbf{A} \succeq \mathbf{0}, \mathbf{A} \neq \mathbf{0}$ satisfying $\lambda_{\max}(\mathbf{A}) \leq 1$:

1. ***(Geometric mixing).*** *The chain $\{\mathbf{y}_k\}$ has a unique stationary distribution ν, and $\max_{\mathbf{y} \in \mathcal{Y}} \|P^k(\mathbf{y}, \cdot) - \nu(\cdot)\|_{\mathrm{TV}} \leq C\rho^k$ for some $C > 0$ and $\rho \in (0, 1)$.*
2. ***(Euclidean Lipschitz + boundedness).*** *There exist $L_1, L_2 > 0$ such that for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ and $\mathbf{y} \in \mathcal{Y}$,*
$$\|\mathbf{F}(\mathbf{x}_1, \mathbf{y}) - \mathbf{F}(\mathbf{x}_2, \mathbf{y})\|_2 \leq L_1 \|\mathbf{x}_1 - \mathbf{x}_2\|_2, \qquad \|\mathbf{F}(\mathbf{0}, \mathbf{y})\|_2 \leq L_2.$$
3. ***(Semi-norm drift).*** *The equation $\bar{\mathbf{F}}(\mathbf{x}) = 0$ has at least one solution (not necessarily unique), and fix any such $\mathbf{x}^\star$. Moreover, for all $\mathbf{x} \in \mathbb{R}^d$,*
$$\left\langle \mathbf{x} - \mathbf{x}^\star, \bar{\mathbf{F}}(\mathbf{x}) \right\rangle_{\mathbf{A}} \leq -\kappa \|\mathbf{x} - \mathbf{x}^\star\|_{\mathbf{A}}^2 \tag{37}$$
for some constant $\kappa > 0$.

4. ***(Range-compatibility for degenerate* $\mathbf{A}$*).*** *For all* $\mathbf{x} \in \mathbb{R}^d$ *and* $\mathbf{y} \in \mathcal{Y}$*, the increment satisfies*

$$\mathbf{F}(\mathbf{x}, \mathbf{y}) \in \mathrm{Range}(\mathbf{A}), \qquad \mathbf{\Pi}_{\mathbf{A}}^{\perp} \mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}. \tag{38}$$

Define $L := \max\{L_1, L_2\}$*. Assume* $\mathbf{A} \neq \mathbf{0}$ *so that* $\lambda_{\min}^{+}(\mathbf{A})$ *is well-defined, and choose* α *such that*

$$\alpha\tau_\alpha \leq \min\Big\{\frac{1}{4L},\ \frac{\kappa\,\lambda_{\min}^{+}(\mathbf{A})}{260\,\lambda_{\max}(\mathbf{A})\,L^2}\Big\}. \tag{39}$$

Then for all $k \geq \tau_\alpha$*,*

$$\mathbb{E}\big[\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|_{\mathbf{A}}^2\big] \leq (1 - \kappa\alpha)\,\mathbb{E}\big[\|\mathbf{x}_k - \mathbf{x}^\star\|_{\mathbf{A}}^2\big] + \frac{260\,\lambda_{\max}(\mathbf{A})\,L^2}{\lambda_{\min}^{+}(\mathbf{A})}\,\alpha^2\tau_\alpha \cdot \mathcal{C}_\star^2, \tag{40}$$

where

$$\mathcal{C}_\star^2 := \|\mathbf{\Pi}_{\mathbf{A}}^{\perp}(\mathbf{x}_0 - \mathbf{x}^\star)\|_2^2 + (\|\mathbf{x}^\star\|_2 + 1)^2. \tag{41}$$

Proof of Theorem 3. Fix $t = \tau_\alpha$ and define the conditional expectation

$$\mathbb{E}_k[\cdot] := \mathbb{E}\big[\cdot \mid \mathbf{x}_{k-t}, \mathbf{y}_{k-t}\big], \qquad k \geq t.$$

We begin with a one-step expansion of the $\mathbf{A}$-semi-norm:

$$\mathbb{E}_k\big[\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|_{\mathbf{A}}^2\big] - \mathbb{E}_k\big[\|\mathbf{x}_k - \mathbf{x}^\star\|_{\mathbf{A}}^2\big] = 2\mathbb{E}_k\big[\langle \mathbf{x}_k - \mathbf{x}^\star, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle_{\mathbf{A}}\big] + \mathbb{E}_k\big[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_{\mathbf{A}}^2\big]$$
$$= \underbrace{2\alpha\,\mathbb{E}_k\big[\langle \mathbf{x}_k - \mathbf{x}^\star, \bar{\mathbf{F}}(\mathbf{x}_k)\rangle_{\mathbf{A}}\big]}_{:=\text{(a)}} + \underbrace{2\alpha\,\mathbb{E}_k\big[\langle \mathbf{x}_k - \mathbf{x}^\star, \mathbf{F}(\mathbf{x}_k, \mathbf{y}_k) - \bar{\mathbf{F}}(\mathbf{x}_k)\rangle_{\mathbf{A}}\big]}_{:=\text{(b)}} + \underbrace{\alpha^2\,\mathbb{E}_k\big[\|\mathbf{F}(\mathbf{x}_k, \mathbf{y}_k)\|_{\mathbf{A}}^2\big]}_{:=\text{(c)}}.$$

Now it remains to bound three terms (a)–(c) separately.

Step 1: negative drift term (a). We have it holds by equation 37 that

$$2\alpha\,\mathbb{E}_k\big[\langle \mathbf{x}_k - \mathbf{x}^\star, \bar{\mathbf{F}}(\mathbf{x}_k)\rangle_{\mathbf{A}}\big] \leq -2\kappa\alpha\,\mathbb{E}_k\big[\|\mathbf{x}_k - \mathbf{x}^\star\|_{\mathbf{A}}^2\big]. \tag{42}$$

To control terms (b), (c), we first introduce several preliminary lemmas as the following:

Lemma 2 (Global growth bound)**.** *Under Assumption 2, for all* $\mathbf{x} \in \mathbb{R}^d$ *and* $\mathbf{y} \in \mathcal{Y}$*,*

$$\|\mathbf{F}(\mathbf{x}, \mathbf{y})\|_2 \leq L(\|\mathbf{x}\|_2 + 1), \qquad \|\mathbf{F}(\mathbf{x}, \mathbf{y})\|_{\mathbf{A}} \leq \sqrt{\lambda_{\max}(\mathbf{A})}\,L(\|\mathbf{x}\|_2 + 1).$$

The same bounds hold for $\|\bar{\mathbf{F}}(\mathbf{x})\|_2$ *and* $\|\bar{\mathbf{F}}(\mathbf{x})\|_{\mathbf{A}}$*.*

Proof. By Lipschitzness and the bound at the origin,

$$\|\mathbf{F}(\mathbf{x}, \mathbf{y})\|_2 \leq \|\mathbf{F}(\mathbf{x}, \mathbf{y}) - \mathbf{F}(\mathbf{0}, \mathbf{y})\|_2 + \|\mathbf{F}(\mathbf{0}, \mathbf{y})\|_2 \leq L_1\|\mathbf{x}\|_2 + L_2 \leq L(\|\mathbf{x}\|_2 + 1).$$

For the $\mathbf{A}$-semi-norm, use $\|\mathbf{v}\|_{\mathbf{A}}^2 = \mathbf{v}^\top\mathbf{A}\mathbf{v} \leq \lambda_{\max}(\mathbf{A})\|\mathbf{v}\|_2^2$. Finally, $\bar{\mathbf{F}}(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim\nu}[\mathbf{F}(\mathbf{x}, \mathbf{y})]$ and Jensen's inequality give the same bounds for $\bar{\mathbf{F}}$. □

Lemma 3 (Range invariance and change-of-norm bounds)**.** *Assume equation 38. Then for all* $k \geq 0$*,*

$$\mathbf{\Pi}_{\mathbf{A}}^{\perp}\mathbf{x}_k = \mathbf{\Pi}_{\mathbf{A}}^{\perp}\mathbf{x}_0, \qquad \mathbf{\Pi}_{\mathbf{A}}^{\perp}(\mathbf{x}_k - \mathbf{x}^\star) = \mathbf{\Pi}_{\mathbf{A}}^{\perp}(\mathbf{x}_0 - \mathbf{x}^\star).$$

Moreover, for any $\mathbf{v} \in \mathbb{R}^d$*,*

$$\|\mathbf{\Pi}_{\mathbf{A}}\mathbf{v}\|_2^2 \leq \frac{1}{\lambda_{\min}^{+}(\mathbf{A})}\|\mathbf{v}\|_{\mathbf{A}}^2, \qquad \|\mathbf{v}\|_2^2 \leq \frac{1}{\lambda_{\min}^{+}(\mathbf{A})}\|\mathbf{v}\|_{\mathbf{A}}^2 + \|\mathbf{\Pi}_{\mathbf{A}}^{\perp}\mathbf{v}\|_2^2. \tag{43}$$

Proof. By equation 38, we have

$$\mathbf{\Pi}_{\mathbf{A}}^{\perp}\mathbf{x}_{k+1} = \mathbf{\Pi}_{\mathbf{A}}^{\perp}\mathbf{x}_k + \alpha\,\mathbf{\Pi}_{\mathbf{A}}^{\perp}\mathbf{F}(\mathbf{x}_k, \mathbf{y}_k) = \mathbf{\Pi}_{\mathbf{A}}^{\perp}\mathbf{x}_k,$$

hence $\mathbf{\Pi}_{\mathbf{A}}^{\perp}\mathbf{x}_k = \mathbf{\Pi}_{\mathbf{A}}^{\perp}\mathbf{x}_0$ for all k. The identity for $\mathbf{x}_k - \mathbf{x}^\star$ follows immediately.

For equation 43, write the eigendecomposition $\mathbf{A} = \mathbf{U}\mathrm{diag}(\lambda_1, \ldots, \lambda_d)\mathbf{U}^\top$ with $\lambda_i \geq 0$. Let $\mathbf{U}_+$ collect eigenvectors with $\lambda_i > 0$. Then $\mathbf{\Pi}_{\mathbf{A}} = \mathbf{U}_+\mathbf{U}_+^\top$ and

$$\|\mathbf{v}\|_{\mathbf{A}}^2 = \sum_{\lambda_i > 0} \lambda_i\,(\mathbf{u}_i^\top\mathbf{v})^2 \geq \lambda_{\min}^{+}(\mathbf{A}) \sum_{\lambda_i > 0} (\mathbf{u}_i^\top\mathbf{v})^2 = \lambda_{\min}^{+}(\mathbf{A})\|\mathbf{\Pi}_{\mathbf{A}}\mathbf{v}\|_2^2,$$

which gives the first inequality. The second follows from $\|\mathbf{v}\|_2^2 = \|\mathbf{\Pi}_{\mathbf{A}}\mathbf{v}\|_2^2 + \|\mathbf{\Pi}_{\mathbf{A}}^{\perp}\mathbf{v}\|_2^2$. □

Lemma 4 (Mixing bias for Markovian sampling)**.** *For any $k \geq t = \tau_\alpha$ and any deterministic $\mathbf{x} \in \mathbb{R}^d$,*

$$\Big\|\mathbb{E}\big[\mathbf{F}(\mathbf{x}, \mathbf{y}_k) \mid \mathbf{y}_{k-t} = y\big] - \bar{\mathbf{F}}(\mathbf{x})\Big\|_2 \leq 2L\alpha\,(\|\mathbf{x}\|_2 + 1) \quad \textit{for all } y \in \mathcal{Y}. \tag{44}$$

Consequently,

$$\Big\|\mathbb{E}\big[\mathbf{F}(\mathbf{x}, \mathbf{y}_k) \mid \mathbf{y}_{k-t} = y\big] - \bar{\mathbf{F}}(\mathbf{x})\Big\|_{\mathbf{A}} \leq 2\sqrt{\lambda_{\max}(\mathbf{A})}\, L\alpha\,(\|\mathbf{x}\|_2 + 1).$$

Proof. Since $\mathcal{Y}$ is finite, we can write

$$\mathbb{E}[\mathbf{F}(\mathbf{x}, \mathbf{y}_k) \mid \mathbf{y}_{k-t} = y] = \sum_{y' \in \mathcal{Y}} P^t(y, y')\, \mathbf{F}(\mathbf{x}, y'), \qquad \bar{\mathbf{F}}(\mathbf{x}) = \sum_{y' \in \mathcal{Y}} \nu(y')\, \mathbf{F}(\mathbf{x}, y').$$

Hence,

$$\begin{aligned}\Big\|\mathbb{E}[\mathbf{F}(\mathbf{x}, \mathbf{y}_k) \mid \mathbf{y}_{k-t} = y] - \bar{\mathbf{F}}(\mathbf{x})\Big\|_2 &\leq \sum_{y' \in \mathcal{Y}} \big|P^t(y, y') - \nu(y')\big|\, \|\mathbf{F}(\mathbf{x}, y')\|_2 \\ &\leq 2\|P^t(y, \cdot) - \nu(\cdot)\|_{\mathrm{TV}} \cdot \max_{y' \in \mathcal{Y}} \|\mathbf{F}(\mathbf{x}, y')\|_2 \\ &\leq 2\alpha \cdot L(\|\mathbf{x}\|_2 + 1),\end{aligned}$$

where we used the definition of total variation and Lemma 2 for the last step. The $\mathbf{A}$-bound follows from $\|\cdot\|_{\mathbf{A}} \leq \sqrt{\lambda_{\max}(\mathbf{A})}\|\cdot\|_2$. □

Finally, we show the following lemma, which mirrors Chen et al. (2022) (Lemma 2.3 therein) for constant stepsize.

Lemma 5. *Assume $\alpha t \leq \frac{1}{4L}$. Then for any $k \geq t$,*

$$\|\mathbf{x}_k - \mathbf{x}_{k-t}\|_2 \leq 2L\,\alpha t\,(\|\mathbf{x}_{k-t}\|_2 + 1), \tag{45}$$
$$\|\mathbf{x}_k - \mathbf{x}_{k-t}\|_2 \leq 4L\,\alpha t\,(\|\mathbf{x}_k\|_2 + 1). \tag{46}$$

Consequently,

$$\|\mathbf{x}_k - \mathbf{x}_{k-t}\|_{\mathbf{A}} \leq \sqrt{\lambda_{\max}(\mathbf{A})}\,\|\mathbf{x}_k - \mathbf{x}_{k-t}\|_2 \leq 2\sqrt{\lambda_{\max}(\mathbf{A})}\, L\,\alpha t\,(\|\mathbf{x}_{k-t}\|_2 + 1).$$

Proof. From the recursion,

$$\|\mathbf{x}_{j+1}\|_2 - \|\mathbf{x}_j\|_2 \leq \|\mathbf{x}_{j+1} - \mathbf{x}_j\|_2 = \alpha\|\mathbf{F}(\mathbf{x}_j, \mathbf{y}_j)\|_2 \leq \alpha L(\|\mathbf{x}_j\|_2 + 1),$$

so $(\|\mathbf{x}_{j+1}\|_2 + 1) \leq (1 + L\alpha)(\|\mathbf{x}_j\|_2 + 1)$. Iterating from $j = k - t$ to $k - 1$ yields

$$\|\mathbf{x}_j\|_2 + 1 \leq (1 + L\alpha)^{j-(k-t)}(\|\mathbf{x}_{k-t}\|_2 + 1) \leq e^{L\alpha t}(\|\mathbf{x}_{k-t}\|_2 + 1).$$

Since $\alpha t \leq 1/(4L)$, we have $L\alpha t \leq 1/4$ and thus $e^{L\alpha t} \leq 1 + 2L\alpha t \leq 2$. Therefore $\|\mathbf{x}_j\|_2 + 1 \leq 2(\|\mathbf{x}_{k-t}\|_2 + 1)$ for all $j \in [k - t, k]$. Now, to show equation 45, noticing that

$$\begin{aligned}\|\mathbf{x}_k - \mathbf{x}_{k-t}\|_2 &\leq \sum_{j=k-t}^{k-1} \|\mathbf{x}_{j+1} - \mathbf{x}_j\|_2 = \alpha \sum_{j=k-t}^{k-1} \|\mathbf{F}(\mathbf{x}_j, \mathbf{y}_j)\|_2 \\ &\leq \alpha \sum_{j=k-t}^{k-1} L(\|\mathbf{x}_j\|_2 + 1) \leq \alpha t \cdot 2L(\|\mathbf{x}_{k-t}\|_2 + 1).\end{aligned}$$

equation 46 follows similarly by running the same argument backward (or using $\|\mathbf{x}_{k-t}\|_2 + 1 \leq 2(\|\mathbf{x}_k\|_2 + 1)$, which follows from the same estimate with indices swapped). □

Step 2: Bounding term (b) Inspired by Lemma 2.4 of Chen et al. (2022) we establish the following result based on above lemmas regarding term (b)

Lemma 6 (Markovian noise inner product bound)**.** *Assume $\alpha t \le \frac{1}{4L}$. Then for any $k \ge t$,*

$$\mathbb{E}_k\big[\langle \mathbf{x}_k - \mathbf{x}^\star, \mathbf{F}(\mathbf{x}_k, \mathbf{y}_k) - \bar{\mathbf{F}}(\mathbf{x}_k)\rangle_{\mathbf{A}}\big] \le \frac{64\,\lambda_{\max}(\mathbf{A})\,L^2}{\lambda^+_{\min}(\mathbf{A})}\,\alpha t\,\Big(\mathbb{E}_k[\|\mathbf{x}_k - \mathbf{x}^\star\|^2_{\mathbf{A}}] + \lambda_{\max}(\mathbf{A})\mathcal{C}^2_\star\Big), \quad (47)$$

where $\mathcal{C}_\star$ is defined in equation 41.

Proof. Write $t = \tau_\alpha$ and decompose, for $k \ge t$:

$$\begin{aligned}
&\mathbb{E}_k\big[\langle \mathbf{x}_k - \mathbf{x}^\star, \mathbf{F}(\mathbf{x}_k, \mathbf{y}_k) - \bar{\mathbf{F}}(\mathbf{x}_k)\rangle_{\mathbf{A}}\big] \\
&= \mathbb{E}_k\big[\langle \mathbf{x}_k - \mathbf{x}_{k-t}, \mathbf{F}(\mathbf{x}_k, \mathbf{y}_k) - \bar{\mathbf{F}}(\mathbf{x}_k)\rangle_{\mathbf{A}}\big] && \text{(T1)}\\
&\quad + \big\langle \mathbf{x}_{k-t} - \mathbf{x}^\star,\ \mathbb{E}_k[\mathbf{F}(\mathbf{x}_{k-t}, \mathbf{y}_k)] - \bar{\mathbf{F}}(\mathbf{x}_{k-t})\big\rangle_{\mathbf{A}} && \text{(T2)}\\
&\quad + \mathbb{E}_k\big[\langle \mathbf{x}_{k-t} - \mathbf{x}^\star, \mathbf{F}(\mathbf{x}_k, \mathbf{y}_k) - \mathbf{F}(\mathbf{x}_{k-t}, \mathbf{y}_k)\rangle_{\mathbf{A}}\big] && \text{(T3)}\\
&\quad + \mathbb{E}_k\big[\langle \mathbf{x}_{k-t} - \mathbf{x}^\star, \bar{\mathbf{F}}(\mathbf{x}_{k-t}) - \bar{\mathbf{F}}(\mathbf{x}_k)\rangle_{\mathbf{A}}\big]. && \text{(T4)}
\end{aligned}$$

We bound the four terms.

Bound for (T1). By Cauchy–Schwarz in the $\mathbf{A}$-inner product and Lemma 2,

$$\|\mathbf{F}(\mathbf{x}_k, \mathbf{y}_k) - \bar{\mathbf{F}}(\mathbf{x}_k)\|_{\mathbf{A}} \le \|\mathbf{F}(\mathbf{x}_k, \mathbf{y}_k)\|_{\mathbf{A}} + \|\bar{\mathbf{F}}(\mathbf{x}_k)\|_{\mathbf{A}} \le 2\sqrt{\lambda_{\max}(\mathbf{A})}\,L(\|\mathbf{x}_k\|_2 + 1).$$

Also, Lemma 5 implies

$$\|\mathbf{x}_k - \mathbf{x}_{k-t}\|_{\mathbf{A}} \le \sqrt{\lambda_{\max}(\mathbf{A})}\|\mathbf{x}_k - \mathbf{x}_{k-t}\|_2 \le 4\sqrt{\lambda_{\max}(\mathbf{A})}\,L\,\alpha t\,(\|\mathbf{x}_k\|_2 + 1),$$

where we used equation 46. Therefore,

$$\text{(T1)} \le 8\,\lambda_{\max}(\mathbf{A})\,L^2\,\alpha t\,\mathbb{E}_k\big[(\|\mathbf{x}_k\|_2 + 1)^2\big]. \quad (48)$$

Bound for (T2). By Cauchy–Schwarz and Lemma 4,

$$\text{(T2)} \le \mathbb{E}_k\big[\|\mathbf{x}_{k-t} - \mathbf{x}^\star\|_{\mathbf{A}}\big] \cdot 2\sqrt{\lambda_{\max}(\mathbf{A})}\,L\alpha\,(\|\mathbf{x}_{k-t}\|_2 + 1).$$

Using $ab \le \frac{1}{2}(a^2 + b^2)$ and then Lemma 5 (which implies $\|\mathbf{x}_{k-t}\|_2 + 1 \le 2(\|\mathbf{x}_k\|_2 + 1)$), we obtain

$$\text{(T2)} \le \sqrt{\lambda_{\max}(\mathbf{A})}\,L\alpha\,\mathbb{E}_k\big[\|\mathbf{x}_{k-t} - \mathbf{x}^\star\|^2_{\mathbf{A}}\big] + 4\sqrt{\lambda_{\max}(\mathbf{A})}\,L\alpha\,(\|\mathbf{x}_k\|_2 + 1)^2. \quad (49)$$

Since $t \ge 1$, we may upper bound α by αt in the final aggregation.

Bound for (T3)+(T4). First note that Assumption 2 implies, for any $\mathbf{y}$,

$$\|\mathbf{F}(\mathbf{x}_k, \mathbf{y}) - \mathbf{F}(\mathbf{x}_{k-t}, \mathbf{y})\|_2 \le L_1\|\mathbf{x}_k - \mathbf{x}_{k-t}\|_2 \le L\|\mathbf{x}_k - \mathbf{x}_{k-t}\|_2,$$

and similarly

$$\|\bar{\mathbf{F}}(\mathbf{x}_k) - \bar{\mathbf{F}}(\mathbf{x}_{k-t})\|_2 \le L\|\mathbf{x}_k - \mathbf{x}_{k-t}\|_2.$$

Thus

$$\|\mathbf{F}(\mathbf{x}_k, \mathbf{y}_k) - \mathbf{F}(\mathbf{x}_{k-t}, \mathbf{y}_k)\|_{\mathbf{A}} + \|\bar{\mathbf{F}}(\mathbf{x}_k) - \bar{\mathbf{F}}(\mathbf{x}_{k-t})\|_{\mathbf{A}} \le 2\sqrt{\lambda_{\max}(\mathbf{A})}\,L\|\mathbf{x}_k - \mathbf{x}_{k-t}\|_2.$$

Then by Cauchy–Schwarz and Lemma 5 (using equation 45),

$$\begin{aligned}
\text{(T3)} + \text{(T4)} &\le \mathbb{E}_k\big[\|\mathbf{x}_{k-t} - \mathbf{x}^\star\|_{\mathbf{A}}\big] \cdot 2\sqrt{\lambda_{\max}(\mathbf{A})}\,L\,\|\mathbf{x}_k - \mathbf{x}_{k-t}\|_2 \\
&\le \mathbb{E}_k\big[\|\mathbf{x}_{k-t} - \mathbf{x}^\star\|_{\mathbf{A}}\big] \cdot 4\sqrt{\lambda_{\max}(\mathbf{A})}\,L^2\,\alpha t\,(\|\mathbf{x}_{k-t}\|_2 + 1). \quad (50)
\end{aligned}$$

Using $\|\mathbf{x}_{k-t}\|_2 + 1 \le 2(\|\mathbf{x}_k\|_2 + 1)$ and then $ab \le \frac{1}{2}(a^2 + b^2)$, we get

$$\text{(T3)} + \text{(T4)} \le 2\,\lambda_{\max}(\mathbf{A})\,L^2\,\alpha t\,\mathbb{E}_k\big[\|\mathbf{x}_{k-t} - \mathbf{x}^\star\|^2_{\mathbf{A}}\big] + 8\,\lambda_{\max}(\mathbf{A})\,L^2\,\alpha t\,(\|\mathbf{x}_k\|_2 + 1)^2. \quad (51)$$

Then we have by Lemma 3,

$$\|\mathbf{x}_k\|_2 \le \|\mathbf{x}_k - \mathbf{x}^\star\|_2 + \|\mathbf{x}^\star\|_2 \le \frac{1}{\sqrt{\lambda^+_{\min}(\mathbf{A})}}\|\mathbf{x}_k - \mathbf{x}^\star\|_{\mathbf{A}} + \|\mathbf{\Pi}^\perp_{\mathbf{A}}(\mathbf{x}_0 - \mathbf{x}^\star)\|_2 + \|\mathbf{x}^\star\|_2.$$

Thus, by $(a + b)^2 \le 2a^2 + 2b^2$,

$$\begin{aligned}
(\|\mathbf{x}_k\|_2 + 1)^2 &\le \frac{2}{\lambda^+_{\min}(\mathbf{A})}\|\mathbf{x}_k - \mathbf{x}^\star\|^2_{\mathbf{A}} + 4\Big(\|\mathbf{\Pi}^\perp_{\mathbf{A}}(\mathbf{x}_0 - \mathbf{x}^\star)\|_2^2 + (\|\mathbf{x}^\star\|_2 + 1)^2\Big) \\
&= \frac{2}{\lambda^+_{\min}(\mathbf{A})}\|\mathbf{x}_k - \mathbf{x}^\star\|^2_{\mathbf{A}} + 4\mathcal{C}^2_\star.
\end{aligned} \quad (52)$$

Putting All Together. Plugging equation 52 into equation 48 and equation 51, and also upper bounding the α factors in equation 49 by αt (since $t \geq 1$), we obtain that (T1)–(T4) are each bounded by a constant multiple of $\frac{\lambda_{\max}(\mathbf{A})L^2}{\lambda^+_{\min}(\mathbf{A})}\alpha t\big(\mathbb{E}_k\|\mathbf{x}_k - \mathbf{x}^\star\|^2_{\mathbf{A}} + \lambda_{\max}(\mathbf{A})\mathcal{C}_\star^2\big)$. This then gives equation 47. □

Step 3: bounding term (c).

Lemma 7 (Squared increment bound)**.** *For any $k \geq 0$,*

$$\mathbb{E}\big[\|\mathbf{F}(\mathbf{x}_k, \mathbf{y}_k)\|^2_{\mathbf{A}}\big] \leq \frac{4\,\lambda_{\max}(\mathbf{A})\,L^2}{\lambda^+_{\min}(\mathbf{A})}\,\mathbb{E}\big[\|\mathbf{x}_k - \mathbf{x}^\star\|^2_{\mathbf{A}}\big] + 8\,\lambda_{\max}(\mathbf{A})\,L^2\,\mathcal{C}_\star^2. \tag{53}$$

Proof. By Lemma 2,

$$\|\mathbf{F}(\mathbf{x}_k, \mathbf{y}_k)\|^2_{\mathbf{A}} \leq \lambda_{\max}(\mathbf{A})\,L^2(\|\mathbf{x}_k\|_2 + 1)^2.$$

Use equation 52 to bound $(\|\mathbf{x}_k\|_2 + 1)^2$ by $\frac{2}{\lambda^+_{\min}(\mathbf{A})}\|\mathbf{x}_k - \mathbf{x}^\star\|^2_{\mathbf{A}} + 4\mathcal{C}_\star^2$, then take expectations and simplify constants. □

Step 4: Putting All Together. Substitute equation 42, Lemma 6 and Lemma 7 into terms (a)–(c). For $k \geq t$,

$$\begin{aligned}\mathbb{E}_k\big[\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|^2_{\mathbf{A}}\big] \leq \Big(1 - 2\kappa\alpha\Big)\,\mathbb{E}_k\big[\|\mathbf{x}_k - \mathbf{x}^\star\|^2_{\mathbf{A}}\big] \\ + 2\alpha \cdot \frac{64\,\lambda_{\max}(\mathbf{A})\,L^2}{\lambda^+_{\min}(\mathbf{A})}\,\alpha t\Big(\mathbb{E}_k[\|\mathbf{x}_k - \mathbf{x}^\star\|^2_{\mathbf{A}}] + \lambda_{\max}(\mathbf{A})\mathcal{C}_\star^2\Big) \\ + \alpha^2\Big(\frac{4\,\lambda_{\max}(\mathbf{A})\,L^2}{\lambda^+_{\min}(\mathbf{A})}\,\mathbb{E}_k[\|\mathbf{x}_k - \mathbf{x}^\star\|^2_{\mathbf{A}}] + 8\,\lambda_{\max}(\mathbf{A})\,L^2\,\mathcal{C}_\star^2\Big).\end{aligned}$$

Since $t \geq 1$, we have $\alpha^2 \leq \alpha^2 t$ and thus we may upper bound the last line by an $\alpha^2 t$-term. Collecting the coefficients (and noting $2 \cdot 64 = 128$) yields

$$\begin{aligned}\mathbb{E}_k\big[\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|^2_{\mathbf{A}}\big] \leq \Big(1 - 2\kappa\alpha + \frac{130\,\lambda_{\max}(\mathbf{A})\,L^2}{\lambda^+_{\min}(\mathbf{A})}\,\alpha^2 t\Big)\,\mathbb{E}_k\big[\|\mathbf{x}_k - \mathbf{x}^\star\|^2_{\mathbf{A}}\big] \\ + \frac{130\,\lambda_{\max}(\mathbf{A})\,L^2}{\lambda^+_{\min}(\mathbf{A})}\,\alpha^2 t \cdot \lambda_{\max}(\mathbf{A})\mathcal{C}_\star^2.\end{aligned} \tag{54}$$

Under the stepsize choice equation 39, we have

$$\frac{130\,\lambda_{\max}(\mathbf{A})\,L^2}{\lambda^+_{\min}(\mathbf{A})}\,\alpha t \leq \kappa,$$

hence the coefficient in equation 54 satisfies

$$1 - 2\kappa\alpha + \frac{130\,\lambda_{\max}(\mathbf{A})\,L^2}{\lambda^+_{\min}(\mathbf{A})}\,\alpha^2 t \leq 1 - \kappa\alpha.$$

Therefore,

$$\mathbb{E}_k\big[\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|^2_{\mathbf{A}}\big] \leq (1 - \kappa\alpha)\,\mathbb{E}_k\big[\|\mathbf{x}_k - \mathbf{x}^\star\|^2_{\mathbf{A}}\big] + \frac{260\,\lambda_{\max}(\mathbf{A})\,L^2}{\lambda^+_{\min}(\mathbf{A})}\,\alpha^2 t \cdot \mathcal{C}_\star^2,$$

where we used $\lambda_{\max}(\mathbf{A})\mathcal{C}_\star^2 \leq \frac{2\lambda_{\max}(\mathbf{A})}{\lambda^+_{\min}(\mathbf{A})}\mathcal{C}_\star^2$ up to constants and absorbed into 260. Finally, take total expectation of both sides to obtain equation 40. □

F.5 Proof of Theorem 1

Now we are ready to prove the evaluation result with behaviour policy μ and target policy π.

Throughout this section, we simply write ϕ^π by ϕ and index the state and action spaces as $\mathcal{S} = \{s_1, \dots, s_n\}$ and $\mathcal{A} = \{a_1, \dots, a_m\}$. Let $d_\mu \in \mathbb{R}^{mn}$ denote the stationary distribution of (S, A) under μ (allowing zero components), and define

$$\mathbf{D} := \mathit{diag}\big(\bar{d}_\mu(s_1, a_1), \dots, \bar{d}_\mu(s_n, a_m)\big) \in \mathbb{R}^{mn \times mn},$$
$$\mathbf{\Phi} := \big[\phi(s_1, a_1), \dots, \phi(s_n, a_m)\big]^\top \in \mathbb{R}^{mn \times d}.$$

Note that $\mathbf{D}$ (and hence $\mathbf{\Phi}^\top \mathbf{D} \mathbf{\Phi}$) is allowed to be singular. We also use $\mathcal{H} := \mathcal{H}^\pi$ to denote the Bellman evaluation operator:

$$[\mathcal{H}(Q)](s, a) := r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) \sum_{a' \in \mathcal{A}} \pi(a'|s') Q(s', a').$$

Equivalently, for $(\mathbf{P}_\pi Q)(s, a) := \sum_{s'} P(s'|s, a) \sum_{a'} \pi(a'|s') Q(s', a')$ we have

$$\mathcal{H}(Q) = \mathbf{r} + \gamma \mathbf{P}_\pi Q.$$

Our first step is a convergence analysis of the inner-loop TD update:

Step 1: Local Convergence Inner-loop TD with fixed target. Fix a target network parameter $w \in \mathbb{R}^d$ and consider the TD-type update driven by a Markov trajectory (S_k, A_k, S_{k+1}) under μ:

$$\xi_{k+1} = \xi_k + \alpha\, \phi(S_k, A_k)\Big(r(S_k, A_k) + \gamma \sum_{a' \in \mathcal{A}} \pi(a'|S_{k+1}) \lceil \phi(S_{k+1}, a')^\top w \rceil - \phi(S_k, A_k)^\top \xi_k\Big).$$

This is a Markovian SA instance of equation 36 with $\mathbf{y}_k = (S_k, A_k, S_{k+1})$ and

$$\begin{aligned} \mathbf{F}\big(\xi, (s, a, s')\big) &:= \phi(s, a)\Big(- \phi(s, a)^\top \xi + r(s, a) + \gamma \sum_{a' \in \mathcal{A}} \pi(a'|s') \lceil \phi(s', a')^\top w \rceil\Big), \\ \bar{\mathbf{F}}(\xi) &:= \mathbb{E}_{\substack{(S,A) \sim d_\mu \\ S' \sim P(\cdot|S,A)}} \Big[\mathbf{F}\big(\xi, (S, A, S')\big)\Big] = \mathbf{\Phi}^\top \mathbf{D}\Big(- \mathbf{\Phi}\xi + \mathcal{H}\big(\lceil \mathbf{\Phi} w \rceil\big)\Big). \end{aligned} \tag{55}$$

Now it remains to verify Theorem 3.

Verification of Theorem 3. With $\mathbf{A} := \mathbf{\Phi}^\top \mathbf{D} \mathbf{\Phi} \succeq \mathbf{0}$, we have:

Condition 1. By the aperiodic assumption, we have the Markov chain $\{(S_k, A_k, S_{k+1})\}$ induced by μ is geometrically mixing on its support:

$$\max_{y \in \mathcal{Y}} \|P^k(y, \cdot) - \nu(\cdot)\|_{\mathrm{TV}} \le C \rho^k,$$

where ν is the stationary distribution over $\mathcal{Y} := \{(s, a, s') : d_\mu(s, a) P(s'|s, a) > 0\}$. Consequently, the mixing time in Theorem 3 satisfies the explicit bound

$$\tau_\alpha := \min\Big\{k \ge 0 : \max_{y \in \mathcal{Y}} \|P^k(y, \cdot) - \nu(\cdot)\|_{\mathrm{TV}} \le \alpha\Big\} \le \left\lceil \frac{\log(C/\alpha)}{\log(1/\rho)} \right\rceil. \tag{56}$$

Condition 2. By our Assumption, we have $\|\phi(s, a)\|_2 \le 1$ for all (s, a). Then for all (s, a, s'),

$$\|\mathbf{F}(\xi_1, (s, a, s')) - \mathbf{F}(\xi_2, (s, a, s'))\|_2 = \|\phi(s, a)\phi(s, a)^\top (\xi_2 - \xi_1)\|_2 \le \|\phi(s, a)\|_2^2 \|\xi_1 - \xi_2\|_2 \le \|\xi_1 - \xi_2\|_2.$$

Moreover, using $r \in [0, 1]$ and $|\lceil \phi^\top w \rceil| \le 1/(1 - \gamma)$,

$$\|\mathbf{F}(\mathbf{0}, (s, a, s'))\|_2 \le \|\phi(s, a)\|_2 \Big(1 + \gamma B\Big) \le 1 + 1/(1 - \gamma).$$

Thus we may take $L_1 = 1$ and $L_2 = 2/(1 - \gamma)$.

Condition 3. Let $\xi^\star(w)$ be any solution to $\bar{\mathbf{F}}(\xi) = \mathbf{0}$:

$$\mathbf{\Phi}^\top \mathbf{D} \mathbf{\Phi}\, \xi^\star(w) = \mathbf{\Phi}^\top \mathbf{D}\, \mathcal{H}(\lceil \mathbf{\Phi} w \rceil).$$

Then for any $\xi \in \mathbb{R}^d$,

$$\begin{aligned}\langle \xi - \xi^\star, \bar{\mathbf{F}}(\xi)\rangle_{\mathbf{A}} &= \langle \xi - \xi^\star, \bar{\mathbf{F}}(\xi) - \bar{\mathbf{F}}(\xi^\star)\rangle_{\mathbf{A}} = \langle \xi - \xi^\star, \mathbf{A}(\xi^\star - \xi)\rangle_{\mathbf{A}} \\ &= -(\xi - \xi^\star)^\top \mathbf{A}^2 (\xi - \xi^\star) \le -\lambda^+_{\min}(\mathbf{A})\, (\xi - \xi^\star)^\top \mathbf{A} (\xi - \xi^\star) \\ &= -\lambda^+_{\min}(\mathbf{A})\, \|\xi - \xi^\star\|^2_{\mathbf{A}}.\end{aligned}$$

Hence Condition 3 holds with $\kappa = \lambda^+_{\min}(\mathbf{A})$.

Condition 4. Noticing that for any $(s, a, s') \in \mathcal{Y}$, we have $\mathbf{F}(\xi, (s, a, s')) \propto \phi(s, a)$ and moreover, since (s, a) belongs to the support of d_μ on $\mathcal{Y}$, we have $d_\mu(s, a) > 0$. Noticing that Then by $\mathrm{Range}(\mathbf{A}) = \mathrm{span}\{\phi(s, a) : d_\mu(s, a) > 0\}$,we have $\phi(s, a) \in \mathrm{Range}(\mathbf{A})$ for all (s, a) appearing in $\mathcal{Y}$. Therefore $\mathbf{F}(\xi, (s, a, s')) \in \mathrm{Range}(\mathbf{A})$ for all $(s, a, s') \in \mathcal{Y}$. □

Now we can apply Theorem 3 for all $k \ge \tau_\alpha$ to arrive at for $c_0 = \lambda^+_{\min}(\mathbf{A}), c_1 \asymp \frac{1}{(1-\gamma)^2 \lambda^+_{\min}(\mathbf{A})}$

$$\mathbb{E}\big[\|\xi_{k+1} - \xi^\star(w)\|^2_{\mathbf{A}}\big] \le (1 - c_0\alpha)\, \mathbb{E}\big[\|\xi_k - \xi^\star(w)\|^2_{\mathbf{A}}\big] + c_1 \alpha^2 \tau_\alpha. \tag{57}$$

By recursion, for any $K \ge \tau_\alpha$, we have then

$$\mathbb{E}\big[\|\xi_K - \xi^\star(w)\|^2_{\mathbf{A}}\big] \le (1 - c_0\alpha)^K\, \mathbb{E}\big[\|\xi_0 - \xi^\star(w)\|^2_{\mathbf{A}}\big] + \underbrace{c_1/c_0}_{\asymp \left((1-\gamma)\lambda^+_{\min}(\mathbf{A})\right)^{-2}} \alpha\tau_\alpha. \tag{58}$$

Remark 4 (Choice of (α, K) for inner-loop accuracy ϵ.). Fix any $\epsilon \in (0, 1)$ and define $V_0(w) := \mathbb{E}\big[\|\xi_0 - \xi^\star(w)\|^2_{\mathbf{A}}\big]$. Using $(1 - c_0\alpha)^K \le \exp(-c_0\alpha K)$, it suffices to ensure

$$\frac{3e^{-2c_0\alpha K}}{(1-\gamma)^2} \le \frac{\epsilon^2}{2}, \quad \frac{256(1 + \alpha K)}{(1-\gamma)^2}\left[\frac{\alpha\tau_\alpha}{c_0}\log(2e\tau_\alpha T) + \left(\frac{\alpha\tau_\alpha}{c_0}\right)^2\right] \le \frac{\epsilon^2}{2}.$$

Under the geometric mixing bound equation 56, a convenient choice is

$$\alpha \propto \frac{\epsilon^2 (1-\gamma)^2 \log(1/\rho)}{\log^2\left(\frac{1}{(1-\gamma)^2\epsilon^2}\right) \log T}, \tag{59}$$

and

$$K := \left\lceil \frac{1}{2c_0\alpha} \log\left(\frac{6}{(1-\gamma)^2\epsilon^2}\right)\right\rceil + \tau_\alpha. \tag{60}$$

With equation 59 and equation 60, we have $\mathbb{E}\big[\|\xi_K - \xi^\star(w)\|^2_{\mathbf{A}}\big] \le \epsilon$. In particular, since $\tau_\alpha = O(\log(1/\alpha))$ by equation 56, this gives an inner-loop complexity of order $K = \tilde{O}\big(\epsilon^{-2}\big)$ up to problem-dependent constants.

Convergence of the target variables. Now we are ready to complete the proof of Theorem 1 based on the above inner loop bound and contraction of Bellman operator. The following proof is inspired by the arguments of Chen et al. (2023) under ℓ_∞ norm, but here we make a careful modification change of it to the on-policy L^2 norm, which is the key step to tackle the partial coverage of μ and utilize previous convergence result under $\mathbf{A}$.

For clarity, we denote $\{w_k\}_{k\ge 0}$ be the target-network sequence generated by the outer loop update, and at outer iteration $k \ge 1$ define the returned estimate

$$\hat{\mathbf{Q}}_k := \mathbf{\Phi}\xi_K(w_{k-1}) \in \mathbb{R}^{mn}.$$

Let $\mathbf{Q}^\star := \mathbf{Q}^\pi$ denote the true evaluation Q-function under π. We denote the joint (s, a) stationary distribution of π by $\bar{d}_\pi$, which satisfies $\bar{d}_\pi^\top \mathbf{P}_\pi = \bar{d}_\pi^\top$.

Using $\mathcal{H}(Q) = \mathbf{r} + \gamma \mathbf{P}_\pi Q$ and the entrywise inequality $|\mathbf{P}_\pi \mathbf{v}| \le \mathbf{P}_\pi |\mathbf{v}|$, we have

$$|\mathcal{H}(\lceil \hat{\mathbf{Q}}_{k-1} \rceil) - \mathcal{H}(\mathbf{Q}^\star)| = \gamma |\mathbf{P}_\pi(\lceil \hat{\mathbf{Q}}_{k-1} \rceil - \mathbf{Q}^\star)| \le \gamma \mathbf{P}_\pi |\hat{\mathbf{Q}}_{k-1} - \mathbf{Q}^\star|.$$

Therefore, for each $k \geq 1$, we have pointwisely,

$$|\hat{\mathbf{Q}}_k - \mathbf{Q}^\star| = \left|\mathcal{H}(\lceil \hat{\mathbf{Q}}_{k-1} \rceil) - \mathcal{H}(\mathbf{Q}^\star) + \hat{\mathbf{Q}}_k - \mathbf{\Phi}\xi^\star(w_{k-1}) - \mathcal{H}(\lceil \hat{\mathbf{Q}}_{k-1} \rceil) + \mathbf{\Phi}\xi^\star(w_{k-1})\right|$$
$$\leq \gamma \mathbf{P}_\pi \left|\hat{\mathbf{Q}}_{k-1} - \mathbf{Q}^\star\right| + |\hat{\mathbf{Q}}_k - \mathbf{\Phi}\xi^\star(w_{K-1})| + \underbrace{|\mathcal{H}(\lceil \hat{\mathbf{Q}}_{k-1} \rceil) - \mathbf{\Phi}\xi^\star(w_{K-1})|}_{\leq \mathcal{E}_{\text{approx}} + |\mathbf{\Phi}(\xi' - \xi^\star(w_{k-1}))|}. \quad (61)$$

Where ξ' is the best-approximation representation in Assumption 3.

Taking squared in both sides and using weighted geometric inequality, we can have pointwisely

$$|\hat{\mathbf{Q}}_k - \mathbf{Q}^\star|^2 \leq \gamma \underbrace{\left(\mathbf{P}_\pi |\hat{\mathbf{Q}}_{k-1} - \mathbf{Q}^\star|\right)^2}_{\leq \mathbf{P}_\pi |\hat{\mathbf{Q}}_{k-1} - \mathbf{Q}^\star|^2 \text{ by Cauchy-Schwartz}} + \frac{16}{(1-\gamma)^2}|\hat{\mathbf{Q}}_k - \mathbf{\Phi}\xi^\star(w_{k-1})|^2$$
$$+ 4\mathcal{E}^2_{\text{approx}} + 4|\mathbf{\Phi}(\xi' - \xi^\star(w_{k-1}))|^2.$$

Left-multiplying by $\bar{d}_\pi^\top$ and using $\bar{d}_\pi^\top \mathbf{P}_\pi = d_\pi^\top$ then yields the L^2 recursion

$$\mathbb{E}_{(s,a)\sim d_\pi}\left[|\hat{Q}_k(s,a) - Q^\star(s,a)|^2\right] \leq \gamma\, \mathbb{E}_{\bar{d}_\pi}\left[|\hat{Q}_{k-1}(s,a) - Q^\star(s,a)|^2\right]$$
$$+ \frac{16}{(1-\gamma)^2}\bar{d}_\pi^\top[|\hat{\mathbf{Q}}_k - \mathbf{\Phi}\xi^\star(w_{k-1})|^2] \quad (62)$$
$$+ 4\mathcal{E}^2_{\text{approx}} + 4\bar{d}_\pi^\top|\mathbf{\Phi}(\xi' - \xi^\star(w_{k-1}))|^2.$$

Bounding the Second Term. For the second term, we have by

$$|\phi(s,a)^\top(\xi_K(w_{k-1}) - \xi^\star(w_{k-1})| \leq \|\phi(s,a)\|_{\mathbf{A}^{-1}} \|\xi_K(w_{k-1}) - \xi^\star(w_{k-1})\|_{\mathbf{A}},$$

for any $\mathbf{A}$, with

$$\|\phi(s,a)\|_{\mathbf{A}^{-1}} = \lim_{\lambda \to 0^+} \|\phi(s,a)\|_{(\mathbf{A}+\lambda\mathbf{I})^{-1}} = \begin{cases} \|\phi(s,a)\|_{\mathbf{A}^\dagger}, & \text{if } \phi(s,a) \in \text{Range}(\mathbf{A}), \\ +\infty, & \text{otherwise.} \end{cases}$$

Hence,

$$\mathbb{E}_{\bar{d}_\pi}\left|\mathbf{\Phi}(\xi_K(w_{k-1}) - \xi^\star(w_{k-1}))\right|^2 \leq \mathbb{E}_{\bar{d}_\pi}\left[\|\phi(s,a)\|^2_{\mathbf{A}^{-1}}\right] \cdot \|\xi_K(w_{k-1}) - \xi^\star(w_{k-1})\|^2_{\mathbf{A}} \quad (63)$$

Thus, if the inner-loop is run with accuracy $\mathbb{E}[\|\xi_K(w) - \xi^\star(w)\|^2_{\mathbf{A}}] \leq \epsilon^2$ (i.e. with the parameter setting as we discussed in Remark 4), then

$$\mathbb{E}\left[d_\pi^\top\left|\mathbf{\Phi}(\xi_K(w_{k-1}) - \xi^\star(w_{k-1}))\right|^2\right] \leq \epsilon^2 \cdot \mathbb{E}_{(s,a)\sim\bar{d}_\pi}\left[\|\phi(s,a)\|^2_{\mathbf{A}^{-1}}\right]. \quad (64)$$

Bounding the Last Term. Similarly, by definition of $\xi^\star(w_{k-1})$, we have

$$\mathbf{\Phi}^\top \mathbf{D}\mathbf{\Phi}\xi^\star(w_{k-1}) = \mathbf{\Phi}^\top \mathbf{D}\mathcal{H}(\lceil \hat{\mathbf{Q}}_{k-1} \rceil) = \mathbf{\Phi}^\top \mathbf{D}\mathbf{\Phi}\xi' + O(\mathcal{E}_{\text{approx}})$$
$$\implies \|\xi^\star(w_{k-1}) - \xi'\|^2_{\mathbf{A}} = (\xi^\star(w_{k-1}) - \xi')^\top \mathbf{A}\mathbf{A}^{-1}\mathbf{A}(\xi^\star(w_{k-1}) - \xi') = O(\mathcal{E}^2_{\text{approx}}/\lambda^+_{\min}(\mathbf{A}))$$

thus

$$\bar{d}_\pi^\top|\mathbf{\Phi}(\xi' - \xi^\star(w_{k-1})|^2 \leq \mathbb{E}_{\bar{d}_\pi}\left[\|\phi(s,a)\|^2_{\mathbf{A}^{-1}}\right] \cdot \underbrace{\|\xi' - \xi^\star(w_{k-1})\|^2_{\mathbf{A}}}_{=O(\mathcal{E}^2_{\text{approx}})}$$

Putting All Together. Combining all above bounds together and take recursion, we have any $T \geq 1$,

$$\mathbb{E}_{(s,a)\sim d_\pi}\left[|\hat{Q}_T(s,a) - Q^\star(s,a)|^2\right] \leq \gamma^T\, \mathbb{E}_{(s,a)\sim d_\pi}\left[|\hat{Q}_0(s,a) - Q^\star(s,a)|^2\right] +$$
$$\frac{\mathbb{E}_{(s,a)\sim d_\pi}\left[\|\phi(s,a)\|^2_{\mathbf{A}^{-1}}\right]}{1-\gamma}\left(\epsilon^2 + \mathcal{E}^2_{\text{approx}}\right) + \mathcal{E}^2_{\text{approx}}, \quad (65)$$

as desired.

F.6 Proof of Corollary 1

By our Assumption 2, we have

$$\mathbb{E}_{(s,a)\sim\bar{d}_\pi}|\hat{Q}_T(s,a)-Q^\pi(s,a)|^2=\sum_{s,a}\bar{d}_\pi(s,a)|\langle\phi^\pi(s,a),\hat{\xi}_T-\xi^\star\rangle|^2$$
$$=(\hat{\xi}_T-\xi^\star)^\top\underbrace{\big[\sum_{s,a}\bar{d}_\pi(s,a)\phi^\pi(s,a)\phi^\pi(s,a)^\top\big]}_{=\mathbf{A}_\pi}(\hat{\xi}_T-\xi^\star)=\|\hat{\xi}_T-\xi^\star\|^2_{\mathbf{A}_\pi}.$$

Then taking expectation in both sides and applying Theorem 1 leads to the desired result.

F.7 Proof of Theorem 2

To prove Theorem 2, we first develop the following high-probability version of the evaluation error bound, throughout this section, we use the same notation as in Appendix F.5.

Lemma 8. *At each inner loop of Algorithm 5 with target variable w, it holds with probability at least $1-\delta$,*

$$\|\xi_K-\xi^\star(w)\|^2_{\mathbf{A}}\le\frac{3e^{-2c_0\alpha K}}{(1-\gamma)^2}+\frac{256(1+\alpha K)}{(1-\gamma)^2}\bigg[\frac{\alpha\tau_\alpha}{c_0}\log\frac{2\tau_\alpha}{\delta}+\bigg(\frac{\alpha\tau_\alpha}{c_0}\bigg)^2\bigg].\tag{66}$$

Proof. For $\bar{\mathbf{F}},\mathbf{F}$ defined in equation 55, using $\bar{\mathbf{F}}(\xi)=-\mathbf{A}(\xi-\xi^\star(w))$ and the inner loop update formula, we have

$$\xi_K-\xi^\star(w)=-(\mathbf{I}-\alpha\mathbf{A})^K\xi^\star(w)+\alpha\sum_{i=0}^{K-1}(\mathbf{I}-\alpha\mathbf{A})^{K-1-i}\big[\mathbf{F}(\xi_i,\mathbf{y}_i)-\bar{\mathbf{F}}(\xi_i)\big].$$

Since $\lambda_{\max}(\mathbf{A})\le 1$ and $\xi^\star(w)\in\text{Range}(\mathbf{A})$, the first term can be bounded as follows:

$$(1-\alpha c_0)^K\|\xi^\star(w)\|_{\mathbf{A}}\le\frac{e^{-c_0\alpha K}}{1-\gamma}.$$

It remains to bound the second summation term. For this purpose, we first provide the following deterministic bounds. For the i-th update of the inner loop, we have

$$\begin{aligned}&\|\xi_{i+1}\|_2^2-\|\xi_i\|_2^2\\&\le\alpha\bigg[\bigg(r(S_i,A_i)+\gamma\sum_{a'\in\mathcal{A}}\pi(a'|S_{i+1})\lceil\phi(S_{i+1},a')^\top w\rceil\bigg)^2-(\phi(S_i,A_i)^\top\xi_i)^2\bigg]\\&\le\alpha\bigg(1+\frac{\gamma}{1-\gamma}\bigg)^2=\frac{\alpha}{(1-\gamma)^2}.\end{aligned}\tag{67}$$

Summing from $\xi_0=0$ gives $\|\xi_i\|_2\le\sqrt{\alpha K}/(1-\gamma)$ for $i\le K$. As a result,

$$\|\mathbf{F}(\xi_i,y)\|_2\le\|\xi_i\|_2+\frac{1}{1-\gamma}\le\frac{1+\sqrt{\alpha K}}{1-\gamma},\quad\forall y\in\mathcal{Y}.$$

Since $\lambda_{\max}(\mathbf{A})\le 1$, this gives

$$\big\|\alpha\sum_{i=0}^{\tau_\alpha-1}(\mathbf{I}-\alpha\mathbf{A})^{K-1-i}\big[\mathbf{F}(\xi_i,\mathbf{y}_i)-\bar{\mathbf{F}}(\xi_i)\big]\big\|_{\mathbf{A}}\le 2\alpha\frac{\tau_\alpha(1+\sqrt{\alpha K})}{1-\gamma}.$$

It remains to bound the terms with $i\ge\tau_\alpha$. Let $\mathcal{F}_i$ denote the historical information before the i-th update. Applying the argument in the proof of Lemma 4 gives

$$\big\|\mathbb{E}[\mathbf{F}(\xi_{i-\tau_\alpha},\mathbf{y}_i)\mid\mathcal{F}_{i-\tau_\alpha}]-\bar{\mathbf{F}}(\xi_{i-\tau_\alpha})\big\|_2\le 2\alpha\sup_{y\in\mathcal{Y}}\|\mathbf{F}(\xi_{i-\tau_\alpha},y)\|_2\le\frac{2\alpha(1+\sqrt{\alpha K})}{1-\gamma}.\tag{68}$$

Summing $\xi_{j+1} - \xi_j = \alpha \mathbf{F}(\xi_j, \mathbf{y}_j)$ gives

$$\|\xi_i - \xi_{i-\tau_\alpha}\|_2 \le \alpha \sum_{j=i-\tau_\alpha}^{i-1} \|\mathbf{F}(\xi_j, \mathbf{y}_j)\|_2 \le \frac{\alpha\tau_\alpha(1+\sqrt{\alpha K})}{1-\gamma}. \tag{69}$$

Then, for $D_i := \mathbf{F}(\xi_{i-\tau_\alpha}, \mathbf{y}_i) - \mathbb{E}[\mathbf{F}(\xi_{i-\tau_\alpha}, \mathbf{y}_i) \mid \mathcal{F}_{i-\tau_\alpha}]$, combining (68) and (69) with the Lipschitz bound for $\mathbf{F}$ in Appendix F.5 gives

$$\|\mathbf{F}(\xi_i, \mathbf{y}_i) - \bar{\mathbf{F}}(\xi_i) - D_i\|_2 \le 2\|\xi_i - \xi_{i-\tau_\alpha}\|_2 + \frac{2\alpha(1+\sqrt{\alpha K})}{1-\gamma} \le \frac{4\alpha\tau_\alpha(1+\sqrt{\alpha K})}{1-\gamma}.$$

Using $\alpha \sum_{j=0}^{K-1}(1-\alpha c_0)^j \le 1/c_0$, we obtain

$$\begin{aligned}
&\left\| \alpha \sum_{i=\tau_\alpha}^{K-1} (\mathbf{I} - \alpha\mathbf{A})^{K-1-i} \big[\mathbf{F}(\xi_i, \mathbf{y}_i) - \bar{\mathbf{F}}(\xi_i)\big] \right\|_{\mathbf{A}} \\
&\le \frac{4\alpha\tau_\alpha(1+\sqrt{\alpha K})}{c_0(1-\gamma)} + \left\| \alpha \sum_{i=\tau_\alpha}^{K-1} (\mathbf{I} - \alpha\mathbf{A})^{K-1-i} D_i \right\|_{\mathbf{A}}.
\end{aligned}$$

It remains to bound the last term. For this purpose, we split the sum into $\tau_\alpha + 1$ subsequences, each taking every $(\tau_\alpha + 1)$-st term, and apply the vector version of the Azuma–Hoeffding inequality to each subsequence with the boundedness condition

$$\|\alpha \mathbf{A}^{1/2}(\mathbf{I} - \alpha\mathbf{A})^{K-1-i} D_i\|_2 \le \frac{2\alpha(1+\sqrt{\alpha K})}{1-\gamma}(1-\alpha c_0)^{K-1-i},$$

and a failure probability of $\delta/(\tau_\alpha + 1)$. This gives, with probability at least $1 - \delta$,

$$\begin{aligned}
&\left\| \alpha \sum_{i=\tau_\alpha}^{K-1} (\mathbf{I} - \alpha\mathbf{A})^{K-1-i} D_i \right\|_{\mathbf{A}} \\
&\le \frac{2\sqrt{2}\alpha(1+\sqrt{\alpha K})}{1-\gamma} \sqrt{(\tau_\alpha+1)\log\frac{2(\tau_\alpha+1)}{\delta} \sum_{i=\tau_\alpha}^{K-1} (1-\alpha c_0)^{2(K-1-i)}} \\
&\le \frac{4\sqrt{2}(1+\sqrt{\alpha K})}{1-\gamma} \sqrt{\frac{\alpha\tau_\alpha}{c_0} \log\frac{2\tau_\alpha}{\delta}}.
\end{aligned}$$

In the last inequality, we use the geometric summation bound and $(\tau_\alpha + 1)\log(2(\tau_\alpha + 1)/\delta) \le 4\tau_\alpha \log(2\tau_\alpha/\delta)$. Combining these bounds with the bound for the first τ_α terms gives equation 66. □

Corollary 2. *With parameters specified in Remark 4. For every $u \ge 0$, the output of Algorithm 5 satisfies with probability at least $1 - e^{-u}$,*

$$\|\hat{\xi}_T - \xi^\star\|^2_{\mathbf{A}_\pi} \le \frac{3}{(1-\gamma)^2}\big[\gamma^{2T} + (1+\sqrt{\mathcal{C}(\pi,\mu)})^2 \mathcal{E}^2_{\text{approx}} + \mathcal{C}(\pi,\mu)\epsilon^2(1+u)\big]. \tag{70}$$

Proof. Bringing the parameter choices in equation 59 and equation 60 to Lemma 8, we have the initial term in equation 66 at most $\epsilon^2/2$, and its remaining term at $\delta = 1/(eT)$ at most $\epsilon^2/2$. For $u > 0$, applying Lemma 8 with $\delta = e^{-u}/T$ and taking union bound over $t \in [T]$ give, with probability at least $1 - e^{-u}$,

$$\|\xi_K(w_j) - \xi^\star(w_j)\|^2_{\mathbf{A}} \le \epsilon^2(1+u), \quad \forall 0 \le j \le T$$

Under such event, equation 70 follows by repeating the argument in Appendix F.5 and applying the identity in the proof of Corollary 1. □

Proof of Theorem 2. With $0 < \eta \le 1/L$ and e_k and $\delta_{k,t}$ defined in Appendix F.3, applying Assumption 5 leads to

$$J_\rho(\pi^{(k)}_{t+1}) - J_\rho(\pi^{(k)}_t) \ge \frac{\eta}{2}\|\nabla J_\rho(\pi^{(k)}_t)\|_2^2 - \frac{\eta}{2}\|\delta_{k,t}\|_2^2.$$

On the other hand, by Assumption 4 and the trust region design in equation 33–equation 34, it holds that $\|\delta_{k,t}\|_2^2 \le c_{\mathrm{g}}\|e_k\|_{\mathbf{A}_k}^2$. Summing over the M actor updates yields

$$\frac{1}{M}\sum_k \sum_{t=0}^{T_k-1} \|\nabla J_\rho(\pi_t^{(k)})\|_2^2 \le \frac{2(J_\rho^\star - J_\rho(\pi^{(0)}))}{\eta M} + \frac{1}{M}\sum_k \sum_{t=0}^{T_k-1} \|\delta_{k,t}\|_2^2$$
$$\le \frac{2(J_\rho^\star - J_\rho(\pi^{(0)}))}{\eta M} + c_{\mathrm{g}} \max_{k:T_k>0} \|e_k\|_{\mathbf{A}_k}^2$$

by $\sum_k T_k = M$.

It remains to control $\max_{k:T_k>0} \|e_k\|_{\mathbf{A}_k}^2$. For this purpose, applying Corollary 2 with failure probability e^{-u}/M for each k and taking union bound leads to, with probability at least $1-e^{-u}$,

$$\max_{k:T_k>0} \|e_k\|_{\mathbf{A}_k}^2 \le \frac{12\max\{C_{\max}, d\}}{(1-\gamma)^2}\big[\mathcal{E}_{\mathrm{approx}}^2 + \epsilon^2(1+\log M + u)\big].$$

Taking integration with respect to u then gives

$$\mathbb{E}\left[\max_{k:T_k>0} \|e_k\|_{\mathbf{A}_k}^2\right] \le \frac{12\max\{C_{\max}, d\}}{(1-\gamma)^2}\big[\mathcal{E}_{\mathrm{approx}}^2 + \epsilon^2(2+\log M)\big], \tag{71}$$

as desired. □

G More Details and Proofs for Section 3.3

G.1 More explanation in diversity-coverage trade-off

In this section, we provide more details of the diversity-coverage trade-off explanation of Lemma 1.

The least-squares term is minimized under uniform design. We show that the weighted least-squared error term $\mathcal{V}_{\mathcal{B}}(\pi, \mathbf{w})$ is minimized under uniform sampling, $w_i = 1/n$. To see this, for the uniform weight vector $\mathbf{u}$, noticing the identity

$$\mathcal{V}_{\mathcal{B}}(\pi, \mathbf{w}) - \mathcal{V}_{\mathcal{B}}(\pi, \mathbf{w}_0) = \sum_{i=1}^n \|(w_i \mathbf{\Sigma}_{\mathbf{w}}^{-1} - \frac{1}{n}\mathbf{\Sigma}_{\mathbf{u}}^{-1})\phi^\pi(s_i, a_i)\|_{\mathbf{A}_\pi}^2 \ge 0$$

holds for all $\mathbf{w} \in \Delta^n$. This gives the desired claim. In particular, such minimizer is independent of the policy π to be evaluated, and just encourage the resulted distribution as diverse as possible over the buffer set.

The TD term is minimized under optimal L-design. Finding the weight minimizing the $\mathcal{C}_{\mathcal{B}}(\pi, \mathbf{w})$ term leads to the standard L-optimal design problem (Atkinson et al., 2007), which is convex and can be solved by Frank-Wolfe type algorithms. In contrast to the LS term, the TD term encourages the replay distribution to place more weight on samples that improve coverage of feature directions emphasized by the target-policy covariance $\mathbf{A}_\pi$.

Figure 5 illustrates the trade-off between the two terms in a toy environment. We linearly interpolate between the uniform design and the L-optimal design and plot how the objective in equation 17 varies along this interpolation.

G.2 Fixed-Budget Replay Design

The formulation in equation 17 jointly selects the replay distribution and the inner-loop TD budget to achieve a prescribed evaluation accuracy. A complementary view is to fix the replay budget K and optimize the sampling distribution to minimize the resulting policy-evaluation bound. Since the target-network residual is independent of $\mathbf{w}$, the fixed-budget replay design reduces to

$$\hat{\mathbf{w}}_K \in \arg\min_{\mathbf{w}\in\Delta^n} \frac{c_{\mathrm{stat}}\mathcal{V}_{\mathcal{B}}(\pi, \mathbf{w})}{(1-\gamma)^3} + \frac{c_{\mathrm{TD}}\mathcal{C}_{\mathcal{B}}(\pi, \mathbf{w})}{K(1-\gamma)^4}. \tag{72}$$

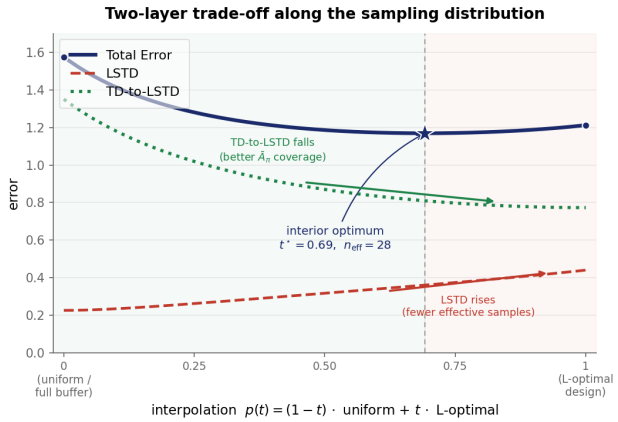


Figure 5: Illustration of the diversity-coverage trade-off in buffer sampling.

Equation equation 72 makes the diversity–coverage trade-off particularly transparent. For small K, the optimization term receives relatively greater weight, favoring replay distributions with stronger target-policy coverage and hence smaller $\mathcal{C}_{\mathcal{B}}$. As K increases, the TD-to-LS optimization gap becomes less important and the statistical term increasingly dominates; in the large-budget limit, the design therefore approaches the uniform distribution that minimizes $\mathcal{V}_{\mathcal{B}}$. Thus, the replay distribution naturally shifts from coverage-oriented sampling toward diversity-oriented sampling as the available evaluation budget increases.

The objective in equation 72 is generally non-convex in $\mathbf{w}$. In practice, we optimize it approximately using a small number of first-order updates.

G.3 Proof of Lemma 1

Proof. Throughout this proof, we use c_0 to denote an absolute constant that may vary between inequalities, and we adopt the notations used in the proof of Theorem 1 throughout this proof: for a target variable $u \in \mathbb{R}^d$, we denote $y_i(u) := r_i + \gamma \sum_a \pi(a|s_i')\lceil \phi^\pi(s_i', a)^\top u \rceil$ as the corresponding value estimation, under which the weighted least-squares solution is

$$\widehat{\xi}^\star_{\mathbf{w}}(u) = \mathbf{\Sigma}_{\mathbf{w}}^{-1} \sum_{i=1}^n w_i \phi^\pi(s_i, a_i) y_i(u).$$

As shown in Appendix F.5, we have $\mathbb{E}[\widehat{\xi}^\star_{\mathbf{w}}(u)] = \xi^\star(u)$, $\xi^\star(\xi^\star) = \xi^\star$, and

$$\|\xi^\star(u) - \xi^\star(v)\|_{\boldsymbol{A}_\pi} \le \gamma \|u - v\|_{\boldsymbol{A}_\pi}, \quad \|\xi^\star\|_{\boldsymbol{A}_\pi} \le \min\{L, (1-\gamma)^{-1}\}. \tag{73}$$

Bounding the least squared error. We define the residual matrix $\boldsymbol{R}_{\mathbf{w}} \in \mathbb{R}^{d \times |\mathcal{S}|}$ as

$$\big[\boldsymbol{R}_{\mathbf{w}}\big]_{j,s} := \left(\mathbf{\Sigma}_{\mathbf{w}}^{-1} \big[\sum_{i=1}^n w_i \phi^\pi(s_i, a_i) \mathbf{1}\{s_i' = s\} - \sum_i w_i \phi^\pi(s_i, a_i) p(s|s_i, a_i) \big] \right)_j .$$

and $\epsilon_{\mathcal{B}} = \gamma \big\| \boldsymbol{A}_\pi^{1/2} \boldsymbol{R}_{\mathbf{w}} \mathrm{diag}(d^\pi)^{-1/2} \big\|_{\mathrm{op}}$, then by Cauchy–Schwarz and Jensen's inequality,

$$\big\| \widehat{\xi}^\star_{\mathbf{w}}(u) - \xi^\star(u) \big\|_{\boldsymbol{A}_\pi} \le \epsilon_{\mathcal{B}} \|u\|_{\boldsymbol{A}_\pi}.$$

Combining this bound with equation 73 gives $\|\widehat{\xi}^\star_{\mathbf{w}}(\xi^\star) - \xi^\star\|_{\boldsymbol{A}_\pi} \le L\epsilon_{\mathcal{B}}$. Now it remains to control $\epsilon_{\mathcal{B}}$, for which we have the following upper bound results:

$$\mathbb{E}[\epsilon_{\mathcal{B}}^2] \le \frac{\mathcal{V}_{\mathcal{B}}(\pi, \mathbf{w})}{d^\pi_{\min}}, \quad \mathbb{E}[\epsilon_{\mathcal{B}}^4] \le \frac{c_0 \mathcal{V}_{\mathcal{B}}(\pi, \mathbf{w})^2}{(d^\pi_{\min})^2}. \tag{74}$$

To see the first inequality of equation 74, denote

$$\big[\boldsymbol{Z}_i\big]_{j,s} := \frac{\gamma w_i}{\sqrt{d^\pi(s)}} \left(\boldsymbol{A}_\pi^{1/2} \mathbf{\Sigma}_{\mathbf{w}}^{-1} \big[\phi^\pi(s_i, a_i) \mathbf{1}\{s_i' = s\} - \phi^\pi(s_i, a_i) p(s|s_i, a_i) \big] \right)_j ,$$

then it holds that $\epsilon_{\mathcal{B}} \le \big\| \sum_{i=1}^n \boldsymbol{Z}_i \big\|_F$ and by

$$\mathbb{E}\big[\|\boldsymbol{Z}_i\|_F^2\big] = \gamma^2 w_i^2 \|\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\phi^\pi(s_i,a_i)\|_{\boldsymbol{A}_\pi}^2 \sum_s \frac{p(s|s_i,a_i)(1-p(s|s_i,a_i))}{d^\pi(s)}, \tag{75}$$

it holds that

$$\mathbb{E}[\epsilon_{\mathcal{B}}^2] \le \frac{\gamma^2}{d_{\min}^\pi} \sum_{i=1}^n w_i^2 \|\Sigma_{\mathbf{w}}^{-1}\phi^\pi(s_i,a_i)\|_{\boldsymbol{A}_\pi}^2 = \frac{\gamma^2 \mathcal{V}_{\mathcal{B}}(\pi,\mathbf{w})}{d_{\min}^\pi},$$

as desired.

To see the second inequality of equation 74, noticing that by equation 75,

$$\mathbb{E}\big[\epsilon_{\mathcal{B}}^4\big] \le \mathbb{E}\big[\|\sum_{i=1}^n \boldsymbol{Z}_i\|_F^4\big] = \sum_{i=1}^n \mathbb{E}\big[\|\boldsymbol{Z}_i\|_F^4\big] + 2\sum_{i<j} \mathbb{E}\big[\|\boldsymbol{Z}_i\|_F^2 \|\boldsymbol{Z}_j\|_F^2\big] + 4\sum_{i<j} \mathbb{E}\big[\langle \boldsymbol{Z}_i, \boldsymbol{Z}_j\rangle^2\big]$$
$$\le 3\bigg(\sum_{i=1}^n \frac{4\gamma^2 w_i^2}{d_{\min}^\pi} \|\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\phi^\pi(s_i,a_i)\|_{\boldsymbol{A}_\pi}^2\bigg)^2 \le c_0 \frac{\gamma^4 \mathcal{V}_{\mathcal{B}}^2(\pi,\mathbf{w})}{(d_{\min}^\pi)^2}.$$

Bounding the TD error. For each outer loop step t and target variable u, consider the i-th step of the inner TD update, since $|y_i(u)| \le (1-\gamma)^{-1}$ and $\|\widehat{\xi}_{\mathbf{w}}^\star(u)\|_{\boldsymbol{A}_\pi} \le (1+\epsilon_{\mathcal{B}})/(1-\gamma)$,

$$\|\widehat{\xi}_{\mathbf{w}}^\star(u)\|_2 \le \frac{1+\epsilon_{\mathcal{B}}}{(1-\gamma)\sqrt{\lambda_{\min}(\boldsymbol{A}_\pi)}}, \quad |y_i(u) - \phi^\pi(s_i,a_i)^\top \widehat{\xi}_{\mathbf{w}}^\star(u)| \le \frac{2(1+\epsilon_{\mathcal{B}})}{(1-\gamma)\sqrt{\lambda_{\min}(\boldsymbol{A}_\pi)}}.$$

For $0 < \alpha \le 1/2$, write $\mathbf{H}_i = \mathbf{I} - \alpha\phi^\pi(s_i,a_i)\phi^\pi(s_i,a_i)^\top$. By the definition of $\widehat{\xi}_{\mathbf{w}}^\star(u)$,

$$\sum_{i=1}^n w_i \phi^\pi(s_i,a_i)\big[y_i(u) - \phi^\pi(s_i,a_i)^\top \widehat{\xi}_{\mathbf{w}}^\star(u)\big] = 0.$$

Thus $\mathbb{E}[\xi_{t,k} - \widehat{\xi}_{\mathbf{w}}^\star(u)] = -(\mathbf{I} - \alpha\boldsymbol{\Sigma}_{\mathbf{w}})^k \widehat{\xi}_{\mathbf{w}}^\star(u)$ and $\|\mathbb{E}[\xi_{t,k} - \widehat{\xi}_{\mathbf{w}}^\star(u)]\|_2 \le \|\widehat{\xi}_{\mathbf{w}}^\star(u)\|_2$. Expanding the TD update then gives

$$\mathbb{E}\big[(\xi_{t,k+1} - \widehat{\xi}_{\mathbf{w}}^\star(u))(\xi_{t,k+1} - \widehat{\xi}_{\mathbf{w}}^\star(u))^\top\big]$$
$$\preceq \sum_{i=1}^n w_i \mathbf{H}_i \mathbb{E}\big[(\xi_{t,k} - \widehat{\xi}_{\mathbf{w}}^\star(u))(\xi_{t,k} - \widehat{\xi}_{\mathbf{w}}^\star(u))^\top\big]\mathbf{H}_i + \frac{c_0\alpha^2(1+\epsilon_{\mathcal{B}})^2}{\lambda_{\min}(\boldsymbol{A}_\pi)(1-\gamma)^2}\boldsymbol{\Sigma}_{\mathbf{w}}.$$

Using $\sum_i w_i \mathbf{H}_i^2 \preceq \mathbf{I} - \alpha(2-\alpha)\boldsymbol{\Sigma}_{\mathbf{w}}$ and $\mathrm{tr}(\boldsymbol{A}_\pi) \le 1$ yield

$$\mathbb{E}[\|\xi_{t,K} - \widehat{\xi}_{\mathbf{w}}^\star(u)\|_{\boldsymbol{A}_\pi}^2] \le \frac{(1+\epsilon_{\mathcal{B}})^2}{\lambda_{\min}(\boldsymbol{A}_\pi)(1-\gamma)^2}\left[e^{-\alpha\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{w}})K} + \frac{c_0\alpha}{2-\alpha}\right]. \tag{76}$$

Putting all together. Now we are ready to combine equation 74 and equation 76 to prove equation 15. For a sufficiently large absolute constant c_1, we set

$$c_{\mathrm{stat}} = \frac{c_1(1+L)^2}{\lambda_{\min}(\boldsymbol{A}_\pi)^2 d_{\min}^\pi}, \quad c_{\mathrm{TD}} = \frac{c_1(1+L)^2}{\lambda_{\min}(\boldsymbol{A}_\pi)^3 d_{\min}^\pi}.$$

First, if the first two terms of equation 15 sum to at least $(1+L)^2$, applying the deterministic bound in equation 67 with the selection $\alpha = (1-\gamma)/K$ gives the desired result.

Otherwise, we have

$$\mathcal{V}_{\mathcal{B}}(\pi,\mathbf{w}) < \frac{(1+L)^2(1-\gamma)^3}{c_{\mathrm{stat}}}, \quad \mathcal{C}_{\mathcal{B}}(\pi,\mathbf{w}) \ge \frac{\lambda_{\min}(\boldsymbol{A}_\pi)}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{w}})}. \tag{77}$$

In this case, we set $\alpha = \min\{1/2, \log(2K)/[\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{w}})K]\}$, underwhich

$$e^{-\alpha\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{w}})K} + \alpha \le \frac{3\log(2K)}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{w}})K}. \tag{78}$$

Under the event $\mathcal{E} := \{\epsilon_{\mathcal{B}} \le (1-\gamma)/4\}$, using $(1+\epsilon_{\mathcal{B}})^2 \le 2$ in equation 76 gives,

$$\mathbb{E}\big[\|\xi_{t,K} - \widehat{\xi}^{\star}_{\mathbf{w}}(u)\|^2_{\boldsymbol{A}_\pi}\big] \le \frac{c_0 \log(2K)\mathcal{C}_{\mathcal{B}}(\pi,\mathbf{w})}{\lambda_{\min}(\boldsymbol{A}_\pi)^2 K(1-\gamma)^2}.$$

Combining equation 78 with equation 73 and $u = \hat{\xi}_t$ gives

$$\left(\mathbb{E}[\|\hat{\xi}_{t+1} - \xi^\star\|^2_{\boldsymbol{A}_\pi}]\right)^{1/2} \le (\gamma+\epsilon_{\mathcal{B}})\left(\mathbb{E}[\|\hat{\xi}_t - \xi^\star\|^2_{\boldsymbol{A}_\pi}]\right)^{1/2} + L\epsilon_{\mathcal{B}} + \sqrt{\frac{c_0 \log(2K)\mathcal{C}_{\mathcal{B}}(\pi,\mathbf{w})}{\lambda_{\min}(\boldsymbol{A}_\pi)^2 K(1-\gamma)^2}}.$$

Since $1-\gamma-\epsilon_{\mathcal{B}} \ge 3(1-\gamma)/4$ under $\mathcal{E}$ and $(\gamma+\epsilon_{\mathcal{B}})^T - \gamma^T = \epsilon_{\mathcal{B}} \sum_{j=0}^{T-1} (\gamma+\epsilon_{\mathcal{B}})^{T-1-j}\gamma^j \le \frac{\epsilon_{\mathcal{B}}}{1-\gamma-\epsilon_{\mathcal{B}}}$, summing from $\hat{\xi}_0 = 0$ and using equation 73 yields

$$\left(\mathbb{E}[\|\hat{\xi}_T - \xi^\star\|^2_{\boldsymbol{A}_\pi}]\right)^{1/2} \le \frac{\gamma^T}{1-\gamma} + \frac{c_0}{1-\gamma}\left[L\epsilon_{\mathcal{B}} + \sqrt{\frac{c_0 \log(2K)\mathcal{C}_{\mathcal{B}}(\pi,\mathbf{w})}{\lambda_{\min}(\boldsymbol{A}_\pi)^2 K(1-\gamma)^2}}\right].$$

Taking square in both sides and applying equation 74, we obtain

$$\mathbb{E}[\|\hat{\xi}_T - \xi^\star\|^2_{\boldsymbol{A}_\pi}\mathbf{1}_{\mathcal{E}}] \le \frac{3\gamma^{2T}}{(1-\gamma)^2} + \frac{c_0 L^2 \mathcal{V}_{\mathcal{B}}(\pi,\mathbf{w})}{d^\pi_{\min}(1-\gamma)^2} + \frac{c_0 \log(2K)\mathcal{C}_{\mathcal{B}}(\pi,\mathbf{w})}{\lambda_{\min}(\boldsymbol{A}_\pi)^2 K(1-\gamma)^4}. \tag{79}$$

On the other hand, under $\mathcal{E}^c$, using equation 76, equation 73 and the fact $\|\widehat{\xi}^{\star}_{\mathbf{w}}(u)\|_{\boldsymbol{A}_\pi} \le (1+\epsilon_{\mathcal{B}})/(1-\gamma)$gives

$$\mathbb{E}[\|\hat{\xi}_T - \xi^\star\|^2_{\boldsymbol{A}_\pi}] \le \frac{c_0(1+\epsilon_{\mathcal{B}})^2}{\lambda_{\min}(\boldsymbol{A}_\pi)(1-\gamma)^2}. \tag{80}$$

By $(1+\epsilon_{\mathcal{B}})^2 \le c_0\epsilon^4_{\mathcal{B}}/(1-\gamma)^4$ under $\mathcal{E}^c$ and equation 74, we obtain

$$\mathbb{E}[\|\hat{\xi}_T - \xi^\star\|^2_{\boldsymbol{A}_\pi}\mathbf{1}_{\mathcal{E}^c}] \le \frac{c_0\mathbb{E}[(1+\epsilon_{\mathcal{B}})^2\mathbf{1}_{\mathcal{E}^c}]}{\lambda_{\min}(\boldsymbol{A}_\pi)(1-\gamma)^2} \le \frac{c_0\mathcal{V}_{\mathcal{B}}(\pi,\mathbf{w})^2}{\lambda_{\min}(\boldsymbol{A}_\pi)(d^\pi_{\min})^2(1-\gamma)^6}. \tag{81}$$

Combining equation 79 and equation 81 together, we have

$$\begin{aligned}\mathbb{E}[\|\hat{\xi}_T - \xi^\star\|^2_{\boldsymbol{A}_\pi}] \le \frac{3\gamma^{2T}}{(1-\gamma)^2} + \frac{c_0 L^2 \mathcal{V}_{\mathcal{B}}(\pi,\mathbf{w})}{d^\pi_{\min}(1-\gamma)^2} + \frac{c_0 \log(2K)\mathcal{C}_{\mathcal{B}}(\pi,\mathbf{w})}{\lambda_{\min}(\boldsymbol{A}_\pi)^2 K(1-\gamma)^4} \\ + \frac{c_0\mathcal{V}_{\mathcal{B}}(\pi,\mathbf{w})^2}{\lambda_{\min}(\boldsymbol{A}_\pi)(d^\pi_{\min})^2(1-\gamma)^6}.\end{aligned}$$

Finally, by equation 77,

$$\frac{\mathcal{V}_{\mathcal{B}}(\pi,\mathbf{w})^2}{\lambda_{\min}(\boldsymbol{A}_\pi)(d^\pi_{\min})^2(1-\gamma)^6} \le \frac{\lambda_{\min}(\boldsymbol{A}_\pi)\mathcal{V}_{\mathcal{B}}(\pi,\mathbf{w})}{c_1 d^\pi_{\min}(1-\gamma)^3}.$$

Substituting this bound to the above inequality gives the desired result. □