BRIDGING INTERPRETABILITY AND OPTIMIZATION: PROVABLY ATTRIBUTION-WEIGHTED ACTOR–CRITIC IN REPRODUCING-KERNEL HILBERT SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Actor-critic (AC) methods are a cornerstone of reinforcement learning (RL) but offer limited interpretability. Current explainable RL methods seldom use *state attributions* to assist training. Rather, they treat all state features equally, thereby neglecting the heterogeneous impacts of individual state dimensions on the reward. We propose *RKHS–SHAP-based Advanced Actor–Critic (RSA2C)*, an attribution-aware, kernelized, two–timescale AC algorithm, including Actor, Value Critic, and Advantage Critic. The Actor is instantiated in a vector-valued reproducing kernel Hilbert space (RKHS) with a Mahalanobis-weighted operator-valued kernel, while the Value Critic and Advantage Critic reside in scalar RKHSs. These RKHS-enhanced components use sparsified dictionaries: the Value Critic maintains its own dictionary, while the Actor and Advantage Critic share one. State attributions, computed from the Value Critic via RKHS–SHAP (kernel mean embedding for on-manifold expectations and conditional mean embedding for off-manifold expectations), are converted into Mahalanobis-gated weights that modulate Actor gradients and Advantage Critic targets. Theoretically, we derive a global, non-asymptotic convergence bound under *state perturbations*, showing stability through the perturbation-error term and efficiency through the convergence-error term. Empirical results on three standard continuous-control environments show that our algorithm achieves efficiency, stability, and interpretability.

1 INTRODUCTION

Policy-gradient (PG) methods (Williams, 1992; Sutton et al., 1999) are a cornerstone of reinforcement learning (RL) (Sutton et al., 1998), directly optimizing the policy to maximize the expected discounted return. Actor–Critic (AC) algorithms (Konda & Tsitsiklis, 1999; Bhatnagar et al., 2009) reduce the variance of PG by pairing a policy (Actor) with a learned value estimator (Critic), thereby stabilizing on-policy training. However, standard AC (Romero et al., 2024) architectures remain *opaque*. The policy updates are driven by the advantage estimates of multi-dimensional state, yet the influence of each dimension on the return is not explicitly revealed. This has fueled a growing interest in explainable RL (XRL) mechanisms that can make optimization dynamics of AC more transparent.

Existing XRL methods can be broadly classified into *post-hoc* and *intrinsic* approaches. Post-hoc methods interpret a trained policy without altering its learning process, using tools such as saliency maps (Rosynski et al., 2020), counterfactual analyses (Gajcin & Dusparic, 2024), and Shapley-value (SHAP)–based explanations (Shapley, 1953; Lundberg & Lee, 2017; Li et al., 2021). Intrinsic methods build interpretability directly into the model, as in programmatic or symbolic policies (Verma et al., 2018), decision-tree policies (Li et al., 2024), or fuzzy-logic rule abstractions (Fu et al., 2024). However, they predominantly explain decisions at the *task*, *trajectory*, or *policy-structure* level. Very few methods provide the *dimension–level attributions*.

The structural coupling between state features and the PG update implies that AC requires *dimension-level attributions* that align with the additive property, which cannot be satisfied by methods such as LIME (Ribeiro et al., 2018) or saliency map (Rosynski et al., 2020). Shapley values (Shapley, 1953; Lundberg & Lee, 2017), as the general metric to provide principled and model-agnostic attributions, uniquely satisfy axiomatic properties including efficiency, linearity, symmetry, and the dummy

property, making them well-suited for AC methods. These properties align naturally with the additive structure of value and advantage functions, where the overall value contribution is built from the joint influence of individual state dimensions. However, classical SHAP-based methods rely on Monte Carlo sampling over coalitions, leading to high computational cost.

A scalable alternative is provided by RKHS–SHAP (Chau et al., 2022), which computes SHAP analytically using kernel mean embeddings (KMEs) and conditional mean embeddings (CMEs) in reproducing-kernel Hilbert spaces (RKHS). Instead of sampling coalitions explicitly, RKHS–SHAP represents coalition expectations as inner products between kernel embeddings, converting the exponential computation cost into tractable linear operations in a feature space and greatly reducing variance. In addition, the reproducing property of RKHS, as a non-parametric method, enables these expectations to be computed in closed form through inner products with KME/CME, yielding stable and low-variance attributions. Specifically, KME yields *observational* (on-manifold) SHAP by averaging over the empirical state distribution, ensuring that attributions reflect environment-consistent configurations. CME further enables *interventional* (off-manifold) SHAP by conditioning on subsets of dimensions while respecting the underlying data geometry, thereby probing the effect of feature coalitions without generating invalid states. These properties address the core limitations like computational scalability, robustness under correlation, and manifold consistency in classical SHAP, providing a principled and efficient mechanism for dimension-level attribution in AC.

Integrating RKHS–SHAP into AC algorithms introduces two fundamental challenges: *attribution reliability* and *optimization stability*. On the one hand, using RKHS–based attributions modifies the functional form of the Critic, potentially violating compatible function approximation and disrupting the contraction properties of the (soft) Bellman operator, especially when the kernel dictionary evolves over time. Such shifts in the underlying RKHS can introduce bias or drift into value estimation, compromising convergence guarantees. On the other hand, RL states are rarely clean or stationary in practice. They are perturbed by stochastic transition dynamics, sensing and communication noise, representation drift, and distributional shifts across training. These perturbations propagate into advantage estimation, making attribution-sensitive methods highly vulnerable to spurious correlations unless carefully regularized. As the attribution directly shapes the policy-gradient update, instability in SHAP values can amplify approximation errors, leading to oscillatory or brittle learning dynamics. Hence, the key research question is *how to design an efficient and provable attribution-aware RKHS-based Actor–Critic that computes state attributions online while maintaining stability.*

1.1 Contribution

This paper proposes *RKHS-SHAP-based Advanced Actor-Critic (RSA2C)*, an attribution-aware two–timescale AC that transforms state attributions into a Mahalanobis distance of the policy. By grounding attribution computation and policy updates in RKHS, RSA2C provides a principled mechanism to highlight influential state dimensions, stabilize gradient updates. This explainable algorithm delivers stable and efficient training for continuous control with theoretical guarantees. Our contributions are summarized as follows:

(i) Kernelized two–timescale Actor–Critic. To jointly improve decision efficiency and interpretability, we propose *RSA2C*, an RKHS-enhanced AC framework that embeds an Actor and two Critics in an RKHS and injects adaptive state feature importance via RKHS–SHAP directly into a Mahalanobis-weighted operator-valued kernel (OVK), which naturally models vector-valued policies and encodes correlations among state dimensions. The *Actor* is instantiated in an OVK RKHS with Mahalanobis weights, while the *Value Critic* and *Advantage Critic* are scalar RKHSs to approximate value and advantage functions, respectively. Computational complexity is controlled by a sparse dictionary maintained via approximate linear dependence (ALD). Crucially, its computation is closed-form and scales linearly with the controllable dictionary size, making it lightweight and suitable for online RL.

(ii) State attribution to learning signal via RKHS–SHAP. We compute SHAP *from Value Critic* using two RKHS–SHAP routes, i.e., KME for on-manifold expectations (RSA2C-KME) and CME for off-manifold expectations (RSA2C-CME). These signals are gated by Mahalanobis weights, and injected into the *Actor* and the *Advantage Critic* targets to modulate updates with budgeted online cost. Under this state-level auxiliary signal, RSA2C achieves both efficiency and intrinsic interpretability.

(iii) Global non-asymptotic convergence under state perturbations. We establish a global, non-asymptotic convergence bound for RSA2C under *state perturbations*. This is achieved by

decomposing the learning gap into a perturbation error and a convergence error, which together quantify stability and efficiency. Moreover, the perturbation error is divided into an attribution-induced term and a policy-learning term, demonstrating stability. The convergence error is divided into a refined tracking term and a two–timescale approximation term, demonstrating efficiency.

(iv) Empirics and intrinsic interpretability. We conduct simulations on three standard continuous-control environments with a focus on effectiveness, stability, and interpretability. Results show that RSA2C improves returns while preserving intrinsic interpretability, incurring only a modest runtime overhead. Besides, RSA2C–CME maintains stable performance for various state perturbations, showing strong robustness to stochastic disturbances.

1.2 Related Works

Explainable RL. XRL generally includes post-hoc and intrinsic explainability (Milani et al., 2024). Post-hoc attribution methods from supervised learning have inspired XRL tools, such as saliency maps (Rosynski et al., 2020), RKHS–SHAP (Chau et al., 2022), LIME (Ribeiro et al., 2016; 2018), Integrated Gradients (Sundararajan et al., 2017), and SmoothGrad (Smilkov et al., 2017). However, they typically describe learned behavior, not utilizing it during training. Intrinsic interpretability lines constrain the policy class, e.g., rule-based policies (Bastani et al., 2018; Li et al., 2024) or decision-tree (Custode & Iacca, 2023), programmatic policies (Verma et al., 2018), and logic controller (Hein et al., 2018). These approaches improve transparency but ignore the ability of the learning loop.

RKHS-based RL. Kernel methods yield nonparametric function approximation with explicit control of geometry and capacity. Recent kernel RL provides finite-time guarantees and provably efficient algorithms in RKHS (Domingues et al., 2021). Sample-efficient GP-based Critics continue to advance via sparse or ensemble designs (Polyzos et al., 2021; Zhang et al., 2020). Vector-valued kernels further enable multi-output policies and critics, with recent theory and online identification methods (Alvarez et al., 2012). However, these kernel-based approaches often lack transparency in decision-making.

Two-timescale AC. Early AC was formalized by Konda & Tsitsiklis (1999) and later extended to natural AC (NAC) using the natural policy gradient (Bhatnagar et al., 2009). A rich body of work has established asymptotic convergence for two–timescale AC/NAC under both independent and identically distributed and Markovian sampling (Agarwal et al., 2020; Hu et al., 2022; Cayci et al., 2024) or non-asymptotic convergence (Xu et al., 2020), yet non-asymptotic convergence guarantees and sample-complexity bounds for two–timescale RKHS-enhanced AC remain largely open gaps.

2 Preliminaries

2.1 Markov Decision Process

Consider a Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma)$ with discount factor $\gamma \in (0,1)$, state space $\mathcal{S} \subseteq \mathbb{R}^d$, and action space $\mathcal{A} \subseteq \mathbb{R}^m$. At each time $t \geq 0$, the agent observes $\mathbf{s}_t \in \mathcal{S}$, selects $\mathbf{a}_t \sim \pi(\cdot \mid \mathbf{s}_t)$ with policy $\pi(\cdot \mid \mathbf{s}_t) \in \Delta(\mathcal{A})$, receives reward $r(\mathbf{s}_t, \mathbf{a}_t) \in [0,1]$, and transitions to $\mathbf{s}_{t+1} \sim \mathcal{T}(\cdot \mid \mathbf{s}_t, \mathbf{a}_t)$, where $\mathcal{T}(\cdot \mid \mathbf{s}_t, \mathbf{a}_t)$ is a Markov kernel on $\mathcal{S}$. Let $\mathbf{s}_0 \sim \rho_0$ denote the initial-state distribution. For conciseness, we write $r_{t+1} = r(\mathbf{s}_t, \mathbf{a}_t)$. The value and action-value functions are $V^\pi(\mathbf{s}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} \middle| \mathbf{s}_0 = \mathbf{s}\right]$ and $Q^\pi(\mathbf{s}, \mathbf{a}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} \middle| \mathbf{s}_0 = \mathbf{s},\ \mathbf{a}_0 = \mathbf{a}\right]$, respectively, so that both V^π and Q^π quantify the expected accumulated reward over the entire horizon. The advantage function is defined as $A^\pi(\mathbf{s}, \mathbf{a}) \coloneqq Q^\pi(\mathbf{s}, \mathbf{a}) - V^\pi(\mathbf{s})$.

Define the discounted visitation distribution as $d_\gamma^\pi(\mathbf{s}) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr(\mathbf{s}_t = \mathbf{s} \mid \pi, \rho_0)$ with $d_\gamma^\pi(\mathbf{s}, \mathbf{a}) = d_\gamma^\pi(\mathbf{s})\pi(\mathbf{a} \mid \mathbf{s})$. The objective of RL can be defined as $J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim d_\gamma^\pi}[r(\mathbf{s}, \mathbf{a})] = \mathbb{E}_{\mathbf{s}_0 \sim \rho_0}[V^\pi(\mathbf{s}_0)]$. The discounted visitation rewrites time-averaged expectations as expectations w.r.t. single distribution d_γ^π, which can be estimated using on-policy rollouts when the policy drifts slowly.

2.2 RKHS-SHAP

Shapley value. Let $\mathcal{X} = \{1, \ldots, d\}$ index input features and $\mathcal{C} \subseteq \mathcal{X}$ be a coalition. For a fixed input $\mathbf{x} \in \mathbb{R}^d$, a cooperative game is specified by a characteristic value function $v_{\mathbf{x}} : 2^{\mathcal{X}} \to \mathbb{R}$. The

Shapley value for feature $i \in \mathcal{X}$ is $\phi_i(v_{\mathbf{x}}) = \sum_{\mathcal{C} \subseteq \mathcal{X} \setminus \{i\}} \frac{|\mathcal{C}|!\left(d-|\mathcal{C}|-1\right)!}{d!} \Big(v_{\mathbf{x}}(\mathcal{C} \cup \{i\}) - v_{\mathbf{x}}(\mathcal{C}) \Big)$, which is exponential to compute naively. Shapley values average a feature's marginal contribution over all coalitions and satisfy efficiency, symmetry, null-player, and additivity.

KernelSHAP. Given a predictive model $f : \mathbb{R}^d \to \mathbb{R}$ and input $\mathbf{x}$, KernelSHAP (Lundberg & Lee, 2017) estimates $\{\phi_i(v_{\mathbf{x}})\}_{i=1}^d$ by fitting a locally linear surrogate to sampled coalitions with a Shapley-inspired kernel, avoiding retraining f on all subsets. To define $v_{\mathbf{x}}(\mathcal{C})$, one must complete the unobserved coordinates $\bar{\mathcal{C}} = \mathcal{X} \setminus \mathcal{C}$:

$$\text{Off-manifold (interventional):} \quad v_{\mathbf{x}}(\mathcal{C}) = \mathbb{E}_{\mathbf{x}' \sim p(\mathbf{x}')} \left[f\left(\mathrm{concat}(\mathbf{x}^{\mathcal{C}}, \mathbf{x}'^{\bar{\mathcal{C}}}) \right) \right], \tag{1a}$$

$$\text{On-manifold (observational):} \quad v_{\mathbf{x}}(\mathcal{C}) = \mathbb{E}_{\mathbf{x}' \sim p(\mathbf{x}' | \mathbf{x}^{\mathcal{C}})} \left[f(\mathbf{x}') \right], \tag{1b}$$

where the latter preserves feature correlations (Štrumbelj & Kononenko, 2014). Here, $\mathrm{concat}(\cdot, \cdot)$ denotes the vector obtained by keeping the coordinates in $\mathcal{C}$ from $\mathbf{x}$ and imputing the remaining coordinates from $\mathbf{x}'$. In practice, off-manifold imputations are simple but may break natural dependencies (e.g., kinematics or physical constraints). On-manifold imputations respect correlations by conditioning on observed coordinates, but require conditional distributions, which are hard to model in high dimensions and expensive to estimate online.

RKHS–SHAP. By embedding marginal and conditional distributions into an RKHS, RKHS–SHAP (Chau et al., 2022) circumvents explicit density models and enables analytic evaluation of the expectations in Eq. (1). Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a bounded positive-definite kernel with feature map $\psi : \mathbb{R}^d \to \mathcal{H}_k$. The KME of $P_{\mathbf{x}}$ is $\mu_{\mathbf{x}} := \mathbb{E}[\psi(\mathbf{x})] \in \mathcal{H}_k$, with empirical estimate $\widehat{\mu}_{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \psi(\mathbf{x}_i)$ (Muandet et al., 2017). For CME, we require the distribution of the missing block given the observed block; let $\mathbf{x}_{\bar{\mathcal{C}}}$ be the unobserved coordinates and $\mathbf{x}_{\mathcal{C}}$ the observed ones. With a kernel on $\mathbb{R}^{|\bar{\mathcal{C}}|}$ and feature map φ, the CME is defined as $\mu_{\mathbf{x}_{\bar{\mathcal{C}}} | \mathbf{x}_{\mathcal{C}}} = \mathbb{E}[\varphi(\mathbf{x}_{\bar{\mathcal{C}}}) \mid \mathbf{x}_{\mathcal{C}}]$.

Assuming a product kernel $k = \prod_{i=1}^d k^{(i)}$ so that $\mathcal{H}_k = \bigotimes_{i=1}^d \mathcal{H}_{k^{(i)}}$, and assuming $f \in \mathcal{H}_k$ (or is approximated therein), the coalition value functionals admit RKHS inner-product forms:

$$v_{\mathbf{x}}(\mathcal{C}) = \left\langle f,\ v^{\mathcal{C}}(\mathbf{x}) \right\rangle_{\mathcal{H}_k}, \qquad v^{\mathcal{C}}(\mathbf{x}) = \begin{cases} \left(\bigotimes_{i \in \mathcal{C}} \psi^{(i)}(x_i) \right) \ \otimes \ \mu_{\mathbf{x}_{\bar{\mathcal{C}}}}, & \text{off-manifold,} \\ \left(\bigotimes_{i \in \mathcal{C}} \psi^{(i)}(x_i) \right) \ \otimes \ \mu_{\mathbf{x}_{\bar{\mathcal{C}}} | \mathbf{x}_{\mathcal{C}}}, & \text{on-manifold.} \end{cases} \tag{2}$$

Here, $\otimes$ denotes the tensor (outer) product of feature maps. Formally, for any coordinates $i \in \mathcal{C}$ and $i \in \bar{\mathcal{C}}$, the tensor (outer) product feature map is defined as $\otimes_{i \in \mathcal{C}} \psi^{(i)}(x_i) \in \otimes_{i \in \mathcal{C}} \mathcal{H}_{k^{(i)}}$, with $\left\langle \bigotimes_{i \in C} \psi^{(i)}(x_i), \bigotimes_{i \in C} \psi^{(i)}(x_i') \right\rangle = \prod_{i \in C} \left\langle \psi^{(i)}(x_i), \psi^{(i)}(x_i') \right\rangle_{\mathcal{H}_{k^{(i)}}}$. The expectations in Eq. (1) reduce to inner products in $\mathcal{H}_k$ rather than explicit (conditional) densities.

Computational cost. Exact CMEs require inverses of Gram operators and cost $\mathcal{O}(n^3)$ for n samples. With a Random Fourier Feature (RFF) (Rahimi & Recht, 2007) or Nyström method (Yang et al., 2012), one can reduce the complexity of evaluating empirical CME from $\mathcal{O}(n^3)$ to $\mathcal{O}(q^2 n + q^3)$ (Muandet et al., 2017), where q is the dimension of the feature map and typically, $q \ll n$ (Li et al., 2019).

3 Algorithm Design of RSA2C

To jointly improve decision efficiency and interpretability, we propose RSA2C, a kernel-based algorithm that embeds an Actor and two Critics in RKHS and injects adaptive state feature importance via RKHS–SHAP directly into a Mahalanobis-weighted OVK. We utilize the ALD-based sparse dictionary to control time complexity. Here, RKHS–SHAP is not merely used for explanation. Its attributions, derived from the value approximation, emphasize stable and influential state dimensions, producing smoother advantages and more stable policy-gradient updates. An overview is given in Fig. 1 and described in Algorithm 1.

Architecture. RSA2C comprises three RKHS-enhanced components with two-timescale mechanisms: (i) an *Actor* whose stochastic Gaussian policy has its mean represented in a *vector-valued RKHS* $\mathcal{H}_K$; (ii) an *Advantage Critic* that estimates $A(\mathbf{s}, \mathbf{a})$ in a *scalar-valued RKHS* with *compatible features* (aligned with $\nabla \log \pi$) to yield low-variance policy gradients; (iii) a *Value Critic* that estimates $V(\mathbf{s})$

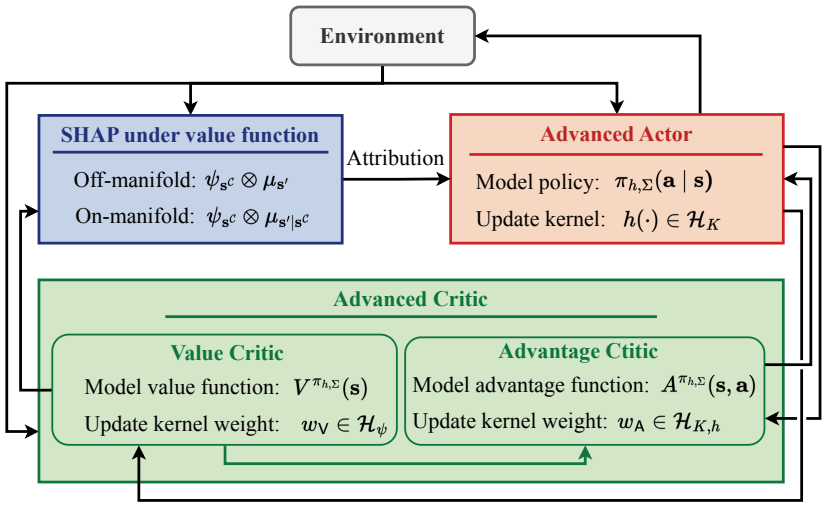


Figure 1: Overview diagram of RSA2C consisting of Actor, Value Critic and Advantage Critic.

Algorithm 1 RSA2C

input Discount factor γ, kernels for Actor and Value Critic, stepsizes α_t^{h} and α_t^{v};
1: **initialize**: Initial state $\mathbf{s}_0$, empty dictionaries $\mathcal{D}_{\mathsf{A}}$ and $\mathcal{D}_{\mathsf{V}}$, covariance matrix Σ_0, Actor's feature mapping $h(\mathbf{s}_0) = 0$, Value Critic's feature mapping w_{V}, and weight of Advantage Critic w_{A};
2: **for** Epoch $t = 1, \ldots, T$ **do**
3: Collect data from the system using policy $\pi_{h,\Sigma}$ with $\mathbf{s}_t \sim \mathcal{T}_\gamma(\cdot \mid \mathbf{s}_{t-1}, \mathbf{a}_{t-1})$;
4: Compute SHAP under value function using Eq. (10) and obtain the weighted kernel by Eq. (3);
5: Optimize the mean of Actor by $h = h + \alpha_t^{\mathsf{h}} \nabla_h J(h, \Sigma)$ with $\nabla_h J(h, \Sigma)$ defined in Eq. (4);
6: (Optional) Reduce the covariance Σ of Actor;
7: Optimize Value Critic by $\psi = \psi + \alpha_t^{\mathsf{v}} \nabla_\psi J(\psi)$ with $\nabla_{w_{\mathsf{V}}} J(w_{\mathsf{V}})$ defined in Eq. (8);
8: Sparsify the dictionary sets $\mathcal{D}_{\mathsf{A}}$ and $\mathcal{D}_{\mathsf{V}}$;
9: Update the weight of Advantage Critic;
10: **end for**
output policy $\pi_{h,\Sigma}$.

in a scalar-valued RKHS via the temporal difference (TD) method. A restart-type transition kernel is incorporated to improve mixing and stabilize targets, i.e., $\mathcal{T}_\gamma(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) := (1-\gamma)\rho_0(\mathbf{s}') + \gamma \mathcal{T}(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$. For any policy π, the stationary state distribution of the Markov chain induced by $\mathcal{T}_\gamma$ coincides with the discounted visitation distribution d_γ^π (i.e., expectations under d_γ^π equal stationary expectations under $\mathcal{T}_\gamma$). This equivalence facilitates writing policy gradients and TD objectives w.r.t. d_γ^π while estimating the expectations online using streaming samples from the on-policy chain.

On-policy two–timescale mechanism. RSA2C is an online, on-policy method, which means rollouts at each iteration are sampled from the current Gaussian policy $\pi_{h,\Sigma}$ until episode termination. Parameters are updated on mini-batches drawn from the most recent on-policy trajectories. Learning proceeds on two–timescale schedules: a *fast* timescale for the Value Critic and a *slow* one for the Actor. Let α_t^{h} and α_t^{v} be the Actor and Value Critic stepsizes, respectively, satisfying

$$\sum_t \alpha_t^{\mathsf{h}} = \infty, \quad \sum_t (\alpha_t^{\mathsf{h}})^2 < \infty, \qquad \sum_t \alpha_t^{\mathsf{v}} = \infty, \quad \sum_t (\alpha_t^{\mathsf{v}})^2 < \infty, \qquad \alpha_t^{\mathsf{v}} / \alpha_t^{\mathsf{h}} \to 0.$$

SHAP-aware kernelization. SHAP scores rescale state features inside the Mahalanobis weights, so importance weighting influences similarities, gradients, and parameter updates, not only providing a post-hoc explanation in prior works. Expectations needed for on-manifold or off-manifold imputations are computed via KME and CME as in RKHS–SHAP, avoiding explicit density models.

3.1 ADVANCED ACTOR–CRITIC FRAMEWORK

We adopt three RKHS-enhanced components on *two time-scales*: a kernelized Gaussian *Actor* in a vector-valued RKHS, an *Advantage Critic* with compatible features sharing the Actor dictionary, and a *Value Critic* in a scalar RKHS with its *own* dictionary. Online sparsification for both dictionaries is performed via the ALD method. First, we define a new kernel as follows.

Definition 3.1 (Adaptive Mahalanobis-weighted OVK)**.** *Let $\mathcal{S} \subseteq \mathbb{R}^d$ denote the state space, and $\Sigma_{\mathsf{K}} \succeq 0$ be a positive semidefinite operator. Define the scalar kernel*

$$\kappa_{\boldsymbol{\phi}}(\mathbf{s}, \mathbf{s}_j) = \exp\left(-\tfrac{1}{2}(\mathbf{s} - \mathbf{s}_j)^\top \mathbf{W}(\mathbf{s} - \mathbf{s}_j)\right), \tag{3}$$

where $\mathbf{W} = \mathrm{diag}(\tilde{\boldsymbol{\phi}})$ with $\tilde{\phi}_i = \max(\phi_i, \varepsilon_0)$ for SHAP-based importances $\phi_i \geq 0$ and a small floor $\varepsilon_0 > 0$ ensuring $\mathbf{W} \succ 0$. The corresponding OVK is $K(\mathbf{s}, \mathbf{s}_j) = \kappa_{\boldsymbol{\phi}}(\mathbf{s}, \mathbf{s}_j)\Sigma_{\mathsf{K}}$. Since $\kappa_{\boldsymbol{\phi}}$ is positive definite (PD) and $\Sigma_{\mathsf{K}} \succeq 0$, the kernel K is operator-valued PD (Micchelli & Pontil, 2005).

Unlike Lever & Stafford (2015), which treats all state features as equally important, $\mathbf{W}$ adapts the base kernel to heterogeneous feature relevance via RKHS-SHAP-derived importances. Notably, $\mathbf{W}$ is diagonal only in its matrix representation, but the quantities placed on the diagonal originate from interaction-aware SHAP features, and thus do not discard feature correlations. If cross-feature interactions are desired, $\mathbf{W}$ may be generalized from diagonal to a full SPD matrix; we adopt a diagonal $\mathbf{W}$ for robustness and efficiency. Also, we use two dictionaries for RSA2C: $\mathcal{D}_{\mathsf{A}} = \{\mathbf{s}_j\}_{j=1}^{q_{\mathsf{A}}}$ shared by the *Actor* and the *Advantage Critic*, and $\mathcal{D}_{\mathsf{V}} = \{\tilde{\mathbf{s}}_j\}_{j=1}^{q_{\mathsf{V}}}$ used solely by the *Value Critic*. Both are maintained online by ALD with standard residual result; see Appendix A.

Advanced Actor. With $\mathcal{D}_{\mathsf{A}} = \{\mathbf{s}_j\}_{j=1}^{q}$ and stack the coefficients $\mathbf{C} = [\mathbf{c}_1, \cdots, \mathbf{c}_q]^\top \in \mathbb{R}^{q \times m}$ with $\mathbf{c}_j \in \mathbb{R}^m$. Under Definition 3.1, we represent the policy mean $h : \mathcal{S} \to \mathcal{A}$ in the vector-valued RKHS $\mathcal{H}_K$ as $h(\mathbf{s}) = \sum_{j=1}^{q} K(\mathbf{s}, \mathbf{s}_j)\mathbf{c}_j$. The Gaussian policy is

$$\pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s}) = \mathcal{N}\left(h(\mathbf{s}), \Sigma\right), \tag{4}$$

where $\Sigma \in \mathbb{R}^{m \times m}$ is a positive-definite covariance matrix. The policy gradient with an advantage baseline w.r.t. the discounted visitation distribution is

$$\nabla_h J(h, \Sigma) = \tfrac{1}{1-\gamma}\mathbb{E}_{(\mathbf{s},\mathbf{a}) \sim d_\gamma^{\pi_{h,\Sigma}}}\left[A_{w_{\mathsf{A}}}^{\pi_{h,\Sigma}}(\mathbf{s}, \mathbf{a})\nabla_h \log \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s})\right],$$

where $A_{w_{\mathsf{A}}}^{\pi_{h,\Sigma}}$ is the approximate advantage function; see Eq. (5). Notably, we update RKHS policy with Fréchet gradient $\nabla_h J(h, \Sigma)$ and perform online sparsification; see Appendices B.1 and A.

Advanced Critic. We estimate advantage and value with kernel approximators designed for policy-gradient compatibility, with the proof in Appendix B.2. The Advantage Critic is established under compatible features on dictionary $\mathcal{D}_{\mathsf{A}}$ and OVK K, giving scalar-valued kernel $K_h\left((\mathbf{s}, \mathbf{a}), (\mathbf{s}_j, \mathbf{a}_j)\right) = \big(\mathbf{a} - h(\mathbf{s})\big)^\top \Sigma^{-1/2} K(\mathbf{s}, \mathbf{s}_j)\Sigma^{-1/2}\big(\mathbf{a}_j - h(\mathbf{s}_j)\big)$, which is positive semidefinite if K is PD. With the feature map $\nu(\mathbf{s}, \mathbf{a}) = K(\mathbf{s}, \cdot)\Sigma^{-1/2}\big(\mathbf{a} - h(\mathbf{s})\big) \in \mathcal{H}_K$, we approximate

$$A_{w_{\mathsf{A}}}^{\pi_{h,\Sigma}}(\mathbf{s}, \mathbf{a}) = \left\langle w_{\mathsf{A}}, \nu(\mathbf{s}, \mathbf{a})\right\rangle_{\mathcal{H}_K}, \tag{5}$$

and estimate w_{A} in closed-form with the ALD method.

For the Value Critic, we deliberately avoid applying RKHS–SHAP reweighting to prevent circular dependence between the value estimates and their own attribution scores. Keeping the Value Critic unweighted avoids error amplification caused by biased SHAP values and provides a more stable baseline for policy evaluation. Specifically, using a scalar RKHS $\mathcal{H}_k$ with kernel $k(\mathbf{s}, \mathbf{s}_j) = \langle \psi(\mathbf{s}), \psi(\mathbf{s}_j)\rangle$ and dictionary $\mathcal{D}_{\mathsf{V}}$, we set

$$V_{w_{\mathsf{V}}}^{\pi_{h,\Sigma}}(\mathbf{s}) = \langle w_{\mathsf{V}}, \psi(\mathbf{s})\rangle_{\mathcal{H}_k} = \sum\nolimits_{j \in \mathcal{D}_{\mathsf{V}}} \eta_j k(\mathbf{s}, \mathbf{s}_j), \qquad w_{\mathsf{V}} = \sum\nolimits_{j \in \mathcal{D}_{\mathsf{V}}} \eta_j \psi(\mathbf{s}_j). \tag{6}$$

with w_{V} fitted by minimizing the regularized TD loss

$$J(w_{\mathsf{V}}) = \mathbb{E}\left[\tfrac{1}{2}\Big(V_{w_{\mathsf{V}}}(\mathbf{s}) - \big(r(\mathbf{s}, \mathbf{a}) + \gamma V_{w_{\mathsf{V}}}(\mathbf{s}')\big)\Big)^2\right] + \tfrac{\lambda}{2}\|w_{\mathsf{V}}\|_{\mathcal{H}_k}^2, \tag{7}$$

with the gradient

$$\nabla_{w_{\mathsf{V}}} J(w_{\mathsf{V}}) = \mathbb{E}\left[\Big(V_{w_{\mathsf{V}}}(\mathbf{s}) - r(\mathbf{s}, \mathbf{a}) - \gamma V_{w_{\mathsf{V}}}(\mathbf{s}')\Big)\psi(\mathbf{s})\right] + \lambda w_{\mathsf{V}}. \tag{8}$$

3.2 SHAP COMPUTING.

We compute state feature attributions from the Value Critic via RKHS–SHAP. Let $\mathcal{X} = \{1, \dots, d\}$ index state features and $\mathcal{C} \subseteq \mathcal{X}$ be a coalition. From Eq. (6), define the coalition value as

$$v_{\mathbf{s}}(\mathcal{C}) = \mathbb{E}\left[V_{w_{\mathsf{V}}}^{\pi_h,\Sigma}\left(\mathrm{concat}(\mathbf{s}^{\mathcal{C}}, \mathbf{s}'^{\bar{\mathcal{C}}})\right)\right] = \langle w_{\mathsf{V}}, \mu_{\mathcal{C}}(\mathbf{s})\rangle_{\mathcal{H}_k}, \tag{9}$$

where the expectation follows either the off-manifold or on-manifold imputations, along with

$$\mu_{\mathcal{C}}^{(\mathrm{off})}(\mathbf{s}) = \left(\otimes_{i\in\mathcal{C}}\, \psi^{(i)}(\mathbf{s}_i)\right) \otimes \mu_{\mathbf{s}_{\bar{\mathcal{C}}}}; \quad \mu_{\mathcal{C}}^{(\mathrm{on})}(\mathbf{s}) = \left(\otimes_{i\in\mathcal{C}}\, \psi^{(i)}(\mathbf{s}_i)\right) \otimes \mu_{\mathbf{s}_{\bar{\mathcal{C}}}|\mathbf{s}_{\mathcal{C}}}, \tag{10}$$

with $\mu_{\mathbf{s}_{\bar{\mathcal{C}}}}$ and $\mu_{\mathbf{s}_{\bar{\mathcal{C}}}|\mathbf{s}_{\mathcal{C}}}$ the (conditional) mean embeddings estimated via KME/CME, leading to RSA2C-KME/RSA2C-CME. Finally, the SHAP attribution for feature i is given by

$$\phi_i(v_{\mathbf{s}}) = \sum\nolimits_{\mathcal{C}\subseteq\mathcal{X}\setminus\{i\}} \frac{|\mathcal{C}|!\big(d-|\mathcal{C}|-1\big)!}{d!}\Big(v_{\mathbf{s}}(\mathcal{C}\cup\{i\}) - v_{\mathbf{s}}(\mathcal{C})\Big). \tag{11}$$

4 THEORETICAL GUARANTEES

In this section, we show the stability and efficiency of RSA2C; see Theorems 4.8 and 4.10. Notably, Lemma C.3 in Appendix C.2 shows the stability of the RKHS–SHAP score produced by the Value Critic. For conciseness, let π_h denote $\pi_{h,\Sigma}$ herein. We first specify the assumptions considered.

Assumption 4.1 (Stationary, deterministic and Markovian adversary)**.** $b(\mathbf{s})$ is a deterministic function $b : \mathcal{S} \to \mathcal{S}$, which only depends on the current state $\mathbf{s}$, and b does not change over time.

Given the same $\mathbf{s}$, the adversary generates the same (stationary) perturbation. If the adversary can perturb a state $\mathbf{s}$ arbitrarily without bounds, the problem can be trivial. To fit our analysis to the most realistic settings, we restrict the power of an adversary. We define perturbation set $B(\mathbf{s})$, to restrict the adversary to perturb a state $\mathbf{s}$ only to a predefined set of states:

Definition 4.2 (Adversary perturbation set)**.** *We define a set $B(\mathbf{s})$ which contains all allowed perturbations of the adversary. Formally, $b(\mathbf{s}) \in B(\mathbf{s})$ where $B(\mathbf{s})$ is a set of states and $\mathbf{s} \in \mathcal{S}$.*

Here, $B(\mathbf{s})$ is usually a set of task-specific "neighboring" states of $\mathbf{s}$ (e.g., bounded sensor measurement errors), which makes the observation still meaningful (yet not accurate) even with perturbations.

We first introduce Proposition 4.3 as follows. See the proof in Appendix C.5.1.

Proposition 4.3. *For all $h \in \mathcal{H}$, the Fisher information matrix induced by policy π_h and initial state distribution ρ_0 satisfies: For some constant $\lambda_F > 0$,*

$$\mathbf{F}(h) = \mathbb{E}_{\nu_{\pi_h}}[\nabla_h \log \pi_h(\mathbf{a}|\mathbf{s}) \nabla_h \log \pi_h(\mathbf{a}|\mathbf{s})^\top] \succeq \lambda_F \cdot \mathbf{I}_m.$$

Assumption 4.4 (Feature boundedness and Lipschitzness)**.** The Value Critic feature map is bounded: $\|\psi(s)\|^2_{\mathcal{H}_\psi} \le M_k$. Whenever perturbation bounds are invoked, there exist finite constants L_k, L_ψ such that k and ψ are Lipschitz in their arguments for Value Critic.

Under the minimum separation rule, the off-diagonal entries of the RBF Gram matrix decay exponentially, yielding a strictly diagonally dominant and therefore PD matrix, summarized as follows.

Assumption 4.5 (Gram invertibility)**.** For the dictionaries $\mathcal{D}_{\mathsf{V}}$, $\mathcal{D}_{\mathsf{A}}$ of size q, the Gram matrix $\mathbf{K}_{\mathsf{V},\mathsf{V}}$ and $\mathbf{K}_{\mathsf{A},\mathsf{A}}$ are invertible. For RBF kernel with length-scale l, there are minimum separations $C_k, C_K > 0$ so that $\rho_{\mathsf{V}} := \exp(-\frac{C_k^2}{2l^2})$, $\rho_{\mathsf{A}} := \exp(-\frac{C_K^2}{2l^2})$, $\lambda_{\min}(\mathbf{K}_{\mathsf{V},\mathsf{V}}) \ge 1-(q-1)\rho_{\mathsf{V}}$, and $\lambda_{\min}(\mathbf{K}_{\mathsf{A},\mathsf{A}}) \ge 1-(q-1)\rho_{\mathsf{A}}$. Hence, $\|\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\| \le [1-(q-1)\rho_{\mathsf{V}}]^{-1}$ and $\|\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\| \le [1-(q-1)\rho_{\mathsf{A}}]^{-1}$.

When the environment transition dynamics are continuous and satisfy a mild drift condition, the resulting on-policy Markov chain is irreducible and uniformly ergodic, as described as follows.

Assumption 4.6 (On-policy geometric mixing)**.** The on-policy process is geometric mixing with effective sample size $n_{\mathrm{eff}} \asymp n/\tau_{\mathrm{mix}}$, where n is the total sample size and τ_{mix} is the mixing time. There exists a constant $C_b \ge 0$ such that $g(h_t) = \nabla \log \pi_h(\mathbf{a}|\mathbf{s})$ obeys $\mathbb{E}\|g(h_t) - \hat{g}(h_t)\|^2_{\mathcal{H}_K} \le C_b/n_{\mathrm{eff}}$.

Driven by the requirement to characterize the stability of the policy under perturbations and the induced variations in the stationary distribution, we introduce the following assumption.

Assumption 4.7. Let $\mathcal{H}_k$ be the RKHS for Value Critic and $\mathcal{H}_K$ be the RKHS for Actor. (i) The stationary state–action visitation distribution ν_{π_h} is C_ν-Lipschitz continuous w.r.t. h in the total variation norm $\left\|\nu_{\pi_h} - \nu_{\pi_{h'}}\right\|_{\mathrm{TV}} \leq C_\nu \|h - h'\|_{\mathcal{H}_K}$, $\forall h, h' \in \mathcal{H}_K$. (ii) The mapping $h \mapsto \pi_h(\cdot \mid \mathbf{s})$ is C_π-Lipschitz in $\mathcal{H}_K$; the underlying Markov chain induced by π_h is uniformly ergodic. Assumption 4.5 holds with $C_\nu = O\left(C_\pi\left(1 + \log \frac{1}{1-\rho_{\mathsf{A}}}\right)\right)$ for $\rho_{\mathsf{A}} < 1$, and analogously for $\rho_{\mathsf{V}} < 1$.

Next, we analyze the convergence under perturbations in three steps. First, we examine the impact of perturbations on the value function. Then, we investigate the convergence of the algorithm in the absence of perturbations. Finally, we derive the convergence result under perturbations, establishing the non-asymptotic convergence and sample complexity for RSA2C.

Step 1: Robustness with perturbation We first analyze the error under the non-perturbed condition as a baseline for evaluating the impact of perturbations. See the proof in Appendix C.3.

Theorem 4.8 (Performance gap under adversarial state perturbations)**.** *Let k be a product kernel with bounded components $|k^{(i)}(\mathbf{s}, \mathbf{s})| \leq M$, $\forall i \in \mathcal{X}$, and assume $\|w_{\mathsf{V}}\|^2_{\mathcal{H}_k} \leq M_V$, $M_S := \sup_{\mathbf{s}} \|\mathbf{s}\|_2$, and $M_\phi := \sup \|\phi\|$. Also, let $\varepsilon := \sup_{\widetilde{\mathbf{s}} \in B(\mathbf{s})} \|\mathbf{s} - \widetilde{\mathbf{s}}\|_2$, $\delta_\psi := \sup_{\mathcal{C} \subseteq \mathcal{X}} \|\psi(\mathbf{s}^{\mathcal{C}}) - \psi(\widetilde{\mathbf{s}}^{\mathcal{C}})\|^2_{\mathcal{H}_k}$, and $\delta_{\tilde{\mathsf{V}}}$ be the value–representation perturbation residual in Lemma C.3. Under Assumptions 4.2, and 4.4, there exists a finite constant $C_0 > 0$ such that:*

$$\mathbb{E}\left[|J(\pi) - J(\tilde{\pi})|\right] \leq C_0 \left(2\left(\varepsilon M_S + \varepsilon^2 M_\phi\right) + M_S^2 d \sqrt{\delta^2_{\tilde{\mathsf{V}}} M^d + M_V \delta_\psi M^d}\right), \tag{12a}$$

$$\mathbb{E}\left[J(\pi) - J(\tilde{\pi})\right] \leq C_0 \left(2\left(\varepsilon M_S + \varepsilon^2 M_\phi\right) + M_S^2 d \sqrt{\delta^2_{\tilde{\mathsf{V}}} M^d + 2 M_V M_\Gamma \delta_\psi M^d}\right), \tag{12b}$$

where $M_\Gamma := \sup_{\mathcal{C} \subseteq \mathcal{X}} \|\mu_{\mathbf{s}^{\overline{\mathcal{C}}} | \mathbf{s}^{\mathcal{C}}}\|^2_{\Gamma_{\mathbf{s}^{\mathcal{C}}}}$. Moreover, if k is the RBF kernel with length–scale l and Assumption 4.5 holds, then $\delta_\psi = 2 - 2\exp(-\varepsilon^2/(2l^2))$ and

$$\mathbb{E}\left[|J(\pi) - J(\tilde{\pi})|\right] \leq C_0 \left(2\left(\varepsilon M_S + \varepsilon^2 M_\phi\right) + M_S^2 d \sqrt{\delta^2_{\tilde{\mathsf{V}}} M^d + M_V \delta_\psi M^d}\right), \tag{13a}$$

$$\mathbb{E}\left[|J(\pi) - J(\tilde{\pi})|\right] \leq C_0 \left(2\left(\varepsilon M_S + \varepsilon^2 M_\phi\right) + M_S^2 d \sqrt{\delta^2_{\tilde{\mathsf{V}}} M^d + 2 M_V M_\Gamma \delta_\psi M^d}\right). \tag{13b}$$

Notably, $\delta_{\tilde{\psi}}$ and $\delta_{\tilde{\mathsf{V}}}$ correspond with the SHAP loss and value function loss, respectively. As long as the perturbation tends to zero, the difference $\mathbb{E}\left[J(\pi) - J(\tilde{\pi})\right]$ also tends to zero.

Step 2: Convergence without perturbation Next, we investigate the convergence of RSA2C without perturbations and derive non-asymptotic bounds that quantify its progression toward the final policy. This result is formally stated in Theorem 4.9; see the proof in Appendix C.4.

Theorem 4.9 (Non-asymptotic convergence of RSA2C)**.** *Let the stepsizes be polynomial $\alpha_t^{\mathsf{h}} = \alpha_0(t+1)^{-\sigma}$, $\beta_t = \beta_0(t+1)^{-\nu}$, $\sigma \in (0,1)$, $\nu \in (0,1)$, and choose the canonical coupling $\nu = \frac{2}{3}\sigma$. Suppose in addition that the implementation/noise schedules satisfy $n^{-1}_{\mathrm{eff},t} = \Theta((t+1)^{-2\sigma})$. Define a random index $\tilde{t} \in \{0, 1, \ldots, T-1\}$ with $\Pr(\tilde{t} = t) = \alpha_t^{\mathsf{h}} / \sum_{k=0}^{T-1} \alpha_k^{\mathsf{h}}$. Under Assumptions 4.3–4.7, there exists a constant $C > 0$ such that*

$$(1-\gamma)\mathbb{E}\big[J(\pi^\star) - J(\pi_{h_t})\big] \leq \begin{cases} \zeta_{\mathrm{approx}} + \mathcal{O}\left(T^{-(1-\sigma)}\right), & \sigma > \frac{3}{4}, \\ \zeta_{\mathrm{approx}} + \mathcal{O}\left(\log^2 T / T^{1/4}\right), & \sigma = \frac{3}{4}\ (\Rightarrow \nu = \frac{1}{2}), \\ \zeta_{\mathrm{approx}} + \mathcal{O}\left(\log T / T^{1-\frac{2}{3}\sigma}\right), & 0 < \sigma < \frac{3}{4}, \end{cases}$$

where ζ_{approx} is the (compatible) approximation error of A^π. In particular, with $\sigma = \frac{3}{4}$ and $\nu = \frac{1}{2}$, $(1-\gamma)\mathbb{E}\big[J(\pi^\star) - J(\pi_{h_t})\big] \leq \zeta_{\mathrm{approx}} + \mathcal{O}\left(\log^2 T / T^{1/4}\right)$, which implies $\mathcal{O}\left((1-\gamma)^{-5}\epsilon^{-4}\log^2 \frac{1}{\epsilon}\right)$ sample complexity for $\mathbb{E}[J(\pi^\star) - J(\pi_{h_t})] \leq \epsilon + \mathcal{O}(\zeta_{\mathrm{approx}})$.

Step 3: Convergence with perturbation Employing the Cauchy inequality, we derive a bound on the perturbation-induced error, ensuring that the learning process remains stable. The formal convergence result is presented in Theorem 4.10.

Theorem 4.10 (Non-asymptotic convergence under perturbations)**.** *Let $\tilde{s} \in \mathcal{B}(s) = \{\tilde{s} : \|\tilde{s} - s\|_2 \leq \varepsilon\}$ be the state perturbation and stepsizes $\alpha_t^h = \alpha_0(t+1)^{-3/4}$, $\beta_t = \beta_0(t+1)^{-1/2}$. Let $\tilde{t}$ be drawn with $\Pr(\tilde{t} = t) = \alpha_t^h / \sum_{k=0}^{T-1} \alpha_k^h$. Under Assumptions 4.3–4.7, for some constant $C_1 > 0$,*

$$\mathbb{E}\left[J(\pi^\star) - J\left(\tilde{\pi}_{h_{\tilde{t}}}\right)\right] \leq \zeta_{\text{approx}}/(1-\gamma) + C_1 \mathcal{B}_k(\varepsilon) + \mathcal{O}\left(\log^2 T/(T^{1/4}(1-\gamma))\right),$$

where $\mathcal{B}_k(\varepsilon)$ is the perturbation term from Theorem 4.8.

5 Simulation Results

In this section, we present simulations on three continuous-control environments, whose details are provided in Appendix D.1. The full set of hyperparameters is shown in Appendix D.2. Results for `Pendulum-v1` are shown below; the other environments `BipedalWalker-v3` and `Ant-v5` is reported in Appendix D.3-D.4. All curves and tables report results averaged over 10 random seeds.

We evaluate the efficiency, stability, and interpretability of RSA2C. For *efficiency*, we conduct ablations and report the overhead as FLOPs and per-update runtime. For *stability*, we evaluate robustness to state perturbations by injecting zero-mean noise with varying covariance into the states. For *interpretability*, we track the RKHS–SHAP score computed from the *Value Critic* by beeswarm and heatmap plots to examine how attribution routing affects learning stability and efficiency.

To verify the theoretical non-asymptotic rate, we additionally design a discounted linear–quadratic regulator (LQR) benchmark based on a linearized inverted pendulum, for which the optimal value function admits a closed-form solution via the discrete-time algebraic Riccati equation. In this LQR setting we can explicitly compute the cost gap $J(\pi^\star) - J(\pi)$ along training, and thus empirically examine the convergence behaviour of our non-asymptotic rate, as illustrated in Appendix D.5.

Evaluation on Efficiency. We conduct the ablation study including two variants (RSA2C-KME and RSA2C-CME), RSA2C without RKHS–SHAP (Advanced AC), and traditional AC under RKHS (Lever & Stafford, 2015) (RKHS-AC). As shown in Figure 2, RSA2C demonstrates super performance and more stable convergence on `Pendulum-v1` under SHAP assistance. Compared with the Advanced AC and RKHS-AC, the final average returns of RSA2C-CME and RSA2C-KME improve by approximately 47.6% and 49.2%, respectively. Notably, RSA2C-KME achieves faster convergence but exhibits larger fluctuations, whereas RSA2C-CME converges more smoothly. Figure 3 further compares RSA2C with standard deep RL algorithms, including soft Actor-Critic (SAC) (Haarnoja et al., 2018) and proximal policy optimization (PPO) (Schulman et al., 2017). Despite not relying on deep neural networks, both RSA2C-KME and RSA2C-CME demonstrate competitive returns on `Pendulum-v1`. SAC reaches a relatively high initial return but quickly saturates, and PPO exhibits slower improvement and larger fluctuations.

Table 1 reports the FLOPs and runtime measured on an Intel(R) Xeon(R) Gold 6248 processor. RSA2C-CME incur almost no additional computational overhead compared with RSA2C-KME. Although both variants substantially increase FLOPs, its wall-clock runtime rises by only up to 12.58%. It is worth emphasizing that our method does not rely on deep neural networks. The overall FLOPs and runtime are significantly lower than those of SAC and PPO (even when executed on GPU), underscoring the lightweight and interpretable nature of our approach.

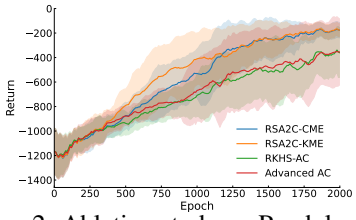


Figure 2: Ablation study on Pendulum-v1.

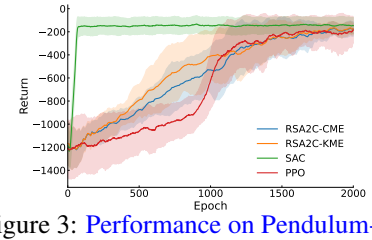


Figure 3: Performance on Pendulum-v1.

Evaluation on Stability. We evaluate the performance of RSA2C-KME and RSA2C-CME under noisy state perturbations with zero mean and varying variance in Table 2. The results show that RSA2C-CME maintains a standard deviation within the range of $[35, 51]$ across all noise levels, whereas RSA2C-KME exhibits much larger variability with standard deviations in $[225, 271]$. This

Table 1: FLOPs and Runtime (ms) on `Pendulum-v1`

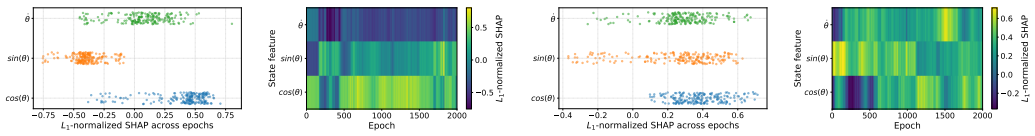
	RKHS-AC	Advanced AC	RSA2C-KME	RSA2C-CME	SAC (GPU)	PPO (GPU)
MFLOPs	5.248	12.439	14.612	14.960	2565.734	111.411
Runtime	355.223	608.955	667.661	685.572	7346.2	2713.3

indicates that RSA2C-CME achieves consistently stable performance under noise, while RSA2C-KME suffers from considerable fluctuations. Nevertheless, relative to Advanced AC (approximately -326.72), even the worst-case performance of RSA2C-KME (-316.48) surpasses the achievable performance of this baseline. We attribute this difference to the handling of state correlations. RSA2C-CME explicitly models feature dependencies, enabling it to adaptively redistribute importance when noise is introduced, thereby mitigating error amplification caused by single-dimension dominance. In contrast, RSA2C-KME ignores correlations, making it vulnerable to severe performance degradation when noise affects the dominant feature dimension. Specifically, both SAC and PPO degrade substantially and remain less stable than the proposed RSA2C variants.

Table 2: Stability under different scales of noise variance on `Pendulum-v1`

	0	0.001	0.005	0.01
RSA2C-KME	-164.82 ± 34.75	-316.48 ± 225.94	-296.74 ± 232.46	-268.61 ± 270.60
RSA2C-CME	-170.24 ± 41.61	-180.96 ± 34.91	-172.19 ± 44.66	-176.10 ± 50.70
SAC	-169.07 ± 89.39	-170.05 ± 88.13	-171.58 ± 91.05	-172.79 ± 92.68
PPO	-217.68 ± 116.82	-219.36 ± 119.34	-201.60 ± 119.16	-219.09 ± 116.21

Evaluation on Interpretability. We analyze the interpretability of RSA2C-KME/CME using beeswarm and heatmap visualizations in Figure 4. Since RSA2C-KME does not account for state correlations, its decisions are often dominated by single feature dimension, whereas RSA2C-CME explicitly models correlations, leading to more balanced allocation across features. In the early training stage, the importance ranking under RSA2C-KME is $\cos(\theta) > \dot{\theta} > \sin(\theta)$. Here, $\cos(\theta)$ and $\dot{\theta}$ jointly determine whether the pendulum can be swung upward, while $\sin(\theta)$ primarily determines the swing direction. Once effective swinging begins, maintaining a consistent direction becomes critical, and the weight of $\sin(\theta)$ rises. Upon convergence, the ranking shifts to $\cos(\theta) > \sin(\theta) > \dot{\theta}$, reflecting the greater importance of angle control over torque. By contrast, RSA2C-CME emphasizes $\sin(\theta)$ in the early stage due to its use of joint distributions, which helps the policy quickly identify the correct upward-swing direction. As training progresses, the weights become more evenly distributed across the three dimensions, with $\sin(\theta)$ and $\cos(\theta)$ jointly dominant and $\dot{\theta}$ playing a supportive role. This balanced allocation of importance enables RSA2C-CME to be smoother and more stable.



(a) Beeswarm under KME (b) Heatmap under KME (c) Beeswarm under CME (d) Heatmap under CME

Figure 4: Visualization on interpretability of RSA2C on `Pendulum-v1`.

6 Conclusion

We introduced *RSA2C*, an attribution-aware two–timescale AC algorithm that turns state feature importance into a Mahalanobis distance on an OVK-based policy. By instantiating Actor, Value Critic, and Advantage Critic in RKHSs and deriving adaptive signals from RKHS–SHAP (via KME for on-manifold expectations and CME for off-manifold expectations), RSA2C couples sample-efficient learning with interpretability. Our analysis provides a global, non-asymptotic convergence guarantee under state perturbations by decomposing the learning gap into an attribution-induced perturbation term and a refined two–timescale convergence term. Empirically, across three continuous-control environments, RSA2C remains efficient and stable under injected state perturbations. Attribution visualizations further reveal environment-relevant state features, evidencing interpretability. Extending RSA2C to high-dimensional or pixel-based observations via scalable kernel approximations (e.g., RFFs), and combining kernel-based attributions with deep representations to improve stability and expressiveness, are two promising research directions.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide a detailed description of the environment details in Appendix D.1. We also list all relevant parameters in Appendix D.2. If the paper is accepted, we will provide an open-source link in the camera-ready version.

REFERENCES

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift, 2020. URL https://arxiv.org/abs/1908.00261.

Gerrard Aissing and Hendrik J Monkhorst. Linear dependence in basis-set calculations for extended systems. *International journal of quantum chemistry*, 43(6):733–745, 1992.

Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. *Advances in neural information processing systems*, 31, 2018.

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Semih Cayci, Niao He, and R Srikant. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *Transactions on Machine Learning Research*, 2024.

Siu Lun Chau, Robert Hu, Javier Gonzalez, and Dino Sejdinovic. Rkhs-shap: Shapley values for kernel methods. *Advances in neural information processing systems*, 35:13050–13063, 2022.

Leonardo L Custode and Giovanni Iacca. Evolutionary learning of interpretable decision trees. *IEEE Access*, 11:6169–6184, 2023.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Kernel-based reinforcement learning: A finite-time analysis. In *International Conference on Machine Learning*, pp. 2783–2792. PMLR, 2021.

Qingxu Fu, Zhiqiang Pu, Yi Pan, Tenghai Qiu, and Jianqiang Yi. Fuzzy feedback multiagent reinforcement learning for adversarial dynamic multiteam competitions. *IEEE Transactions on Fuzzy Systems*, 32(5):2811–2824, 2024.

Jasmina Gajcin and Ivana Dusparic. Redefining counterfactual explanations for reinforcement learning: Overview, challenges and opportunities. *ACM Computing Surveys*, 56(9):1–33, 2024.

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.

Daniel Hein, Steffen Udluft, and Thomas A Runkler. Interpretable policies for reinforcement learning by genetic programming. *Engineering Applications of Artificial Intelligence*, 76:158–169, 2018.

Yuzheng Hu, Ziwei Ji, and Matus Telgarsky. Actor-critic is implicitly biased towards high entropy optimal policies. In *10th International Conference on Learning Representations, ICLR 2022*, 2022.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Guy Lever and Ronnie Stafford. Modelling policies in mdps in reproducing kernel hilbert space. In *Artificial intelligence and statistics*, pp. 590–598. PMLR, 2015.

Jiahui Li, Kun Kuang, Baoxiang Wang, Furui Liu, Long Chen, Fei Wu, and Jun Xiao. Shapley counterfactual credits for multi-agent reinforcement learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 934–942, 2021.

Yifan Li, Shuhan Qi, Xuan Wang, Jiajia Zhang, and Lei Cui. A novel tree-based method for interpretable reinforcement learning. *ACM Transactions on Knowledge Discovery from Data*, 18 (9):1–22, 2024.

Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random fourier features. In *International conference on machine learning*, pp. 3905–3914. PMLR, 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. Explainable reinforcement learning: A survey and comparative review. *ACM Computing Surveys*, 56(7):1–36, 2024.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Konstantinos D Polyzos, Qin Lu, Alireza Sadeghi, and Georgios B Giannakis. On-policy reinforcement learning via ensemble gaussian processes with application to resource allocation. In *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pp. 1018–1022. IEEE, 2021.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Angel Romero, Yunlong Song, and Davide Scaramuzza. Actor-critic model predictive control. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14777–14784. IEEE, 2024.

Matthias Rosynski, Frank Kirchner, and Matias Valdenegro-Toro. Are gradient-based saliency maps useful in deep reinforcement learning? In *"I Can't Believe It's Not Better!"NeurIPS 2020 workshop*, 2020.

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

Lloyd S Shapley. A value for n-person games. *Contribution to the Theory of Games*, 2, 1953.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. Programmatically interpretable reinforcement learning. In *International conference on machine learning*, pp. 5045–5054. PMLR, 2018.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*, 2020.

Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. *Advances in neural information processing systems*, 25, 2012.

Yongliang Yang, Hufei Zhu, Qichao Zhang, Bo Zhao, Zhenning Li, and Donald C Wunsch. Sparse online kernelized actor-critic learning in reproducing kernel hilbert space. *Artificial Intelligence Review*, 55(1):23–58, 2022.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037, 2020.

CONTENTS

1 Introduction **1**
1.1 Contribution . . . 2
1.2 Related Works . . . 3

2 Preliminaries **3**
2.1 Markov Decision Process . . . 3
2.2 RKHS-SHAP . . . 3

3 Algorithm Design of RSA2C **4**
3.1 Advanced Actor–Critic Framework . . . 6
3.2 SHAP Computing. . . . 7

4 Theoretical Guarantees **7**

5 Simulation Results **9**

6 Conclusion **10**

Appendix **14**

A Online Sparsification in RSA2C **16**

B Proof of RL update **17**
B.1 Advanced Actor update . . . 17
B.2 Compatible function approximation . . . 18

C Proof of theoretical guarantees **18**
C.1 Constants and Notation . . . 18
C.2 Supporting Lemmas . . . 19
C.3 Proof of Theorem 4.8 . . . 21
C.4 Proof of Theorem 4.9 . . . 25
C.5 Proof of supporting facts . . . 31
C.5.1 Proof of Proposition 4.3 . . . 31
C.5.2 Proof of Proposition B.2 . . . 31
C.5.3 Proof of Lemma C.2 . . . 32
C.5.4 Proof of Lemma C.3 . . . 34
C.5.5 Proof of Lemma C.4 . . . 36
C.5.6 Proof of Lemma C.5 . . . 37
C.5.7 Proof of Lemma C.6 . . . 38
C.5.8 Proof of Lemma C.7 . . . 39

C.5.9 Proof of Lemma C.8 39
C.5.10 Proof of Lemma C.9 40
C.5.11 Proof of Lemma C.11 41

D Additional Experiments **46**
D.1 Details for Environments 47
D.2 Hyper-parameters 47
D.3 Results of `BipedalWalker-v3` 47
D.4 Results of `Ant-v5` 48
D.5 Simulation on Theoretical non-asymptotic rate 51
D.5.1 Details for LQR 51
D.5.2 Results of theoretical gap 53

LLM USAGE

In preparing this paper, large language models (LLMs) were employed solely as assistive tools for minor language refinement and for literature retrieval and discovery (e.g., identifying related work). All technical contributions, results, and conclusions are entirely the work of the authors.

A ONLINE SPARSIFICATION IN RSA2C

Unlike parametric models with fixed-size representations, successively updating RSA2C by incrementally adding new components leads to increasingly complex function representations. While this enables the modeling of expressive policies and value or advantage functions, it also introduces substantial computational overhead, particularly during evaluation. To ensure tractability, it is therefore essential to control the complexity of the kernel expansions for both the Actor function $h(\cdot)$ and the Critic features $\psi(\cdot)$, as well as their gradients.

To address this, we introduce an online sparsification scheme that incrementally prunes the kernel representation in a data-efficient manner. This scheme is inspired by the sparse kernel learning framework in (Yang et al., 2022), and is applicable to both vector-valued and scalar-valued kernels.

We first describe the sparsification strategy for the advanced Actor. Given a kernel-based policy representation of the form

$$h(\mathbf{s}) = \frac{1}{N}\sum_j K(\mathbf{s}, \mathbf{s}_j)\mathbf{c}_j \in \mathcal{H}_K,$$

our goal is to obtain a sparse approximation:

$$\hat{h}(\mathbf{s}) = \frac{1}{q}\sum_{(\mathbf{s}_j,\mathbf{c}_j)\in\mathcal{D}_\mathsf{A}} K(\mathbf{s}, \mathbf{s}_j)\mathbf{c}_j \in \mathcal{H}_K,$$

where only $q \ll N$ coefficients are non-zero and $\mathcal{D}_\mathsf{A}$ denotes the dictionary of selected kernel centers and their corresponding coefficients.

To enable this, we adopt an online kernel Approximate Linear Dependence (ALD) method (Aissing & Monkhorst, 1992), which incrementally constructs a compact and informative dictionary during policy optimization in RKHS.

If the current dictionary $\mathcal{D}_\mathsf{A}$ has not yet reached the predefined size limit, a new sample $(\mathbf{s}_\iota, \mathbf{c}_\iota)$ is directly added. The coefficient $\mathbf{c}_\iota$ is computed as:

$$\mathbf{c}_\iota = A_{\overline{w}_\mathsf{A}}^{\pi_{h,\Sigma}}(\mathbf{s}_\iota, \mathbf{a}_\iota)\Sigma^{-1}(\mathbf{a}_\iota - h(\mathbf{s}_\iota)),$$

where $A_{\overline{w}_\mathsf{A}}^{\pi_{h,\Sigma}}(\cdot,\cdot)$ is the advantage-weighted Actor gradient term, and $h(\cdot)$ denotes the current RKHS-based policy.

Once the dictionary reaches its size constraint, for each policy update step (see Eq. 4), we must decide whether the new sample provides sufficiently novel information to warrant inclusion. This decision is made by solving the following kernel projection problem:

$$\min_{\{\mathbf{c}_j\}\in\mathcal{D}_\mathsf{A}} \left\| \sum_j K(\mathbf{s}_j, \cdot)\mathbf{c}_j - K(\mathbf{s}_\iota, \cdot)\mathbf{c}_\iota \right\|^2,$$

which yields the projection residual $\xi_{\mathrm{A},\iota}$. If this residual exceeds a predefined threshold η, the dictionary entry with the largest historical approximation error is replaced by $(\mathbf{s}_\iota, \mathbf{c}_\iota)$. The residual is then recomputed using the remaining $q-1$ dictionary elements, and the stored values are updated accordingly.

If the residual falls below the threshold, we retain the existing dictionary structure but update the coefficients $\{\mathbf{c}_j\}$ using the newly optimized projection, thereby refining the approximation under the current basis.

Once the dictionary $\mathcal{D}_\mathsf{A}$ is finalized, the associated Critic weights w_V can be directly computed using the resulting sparse kernel representation.

Finally, the sparsification process for Value Critic mirrors that of the Actor, with a key distinction: Value Critic employs a scalar-valued kernel to approximate the scalar value function $V(\mathbf{s})$, whereas the Actor uses a vector-valued kernel to represent the policy mapping $h(\mathbf{s})$ in the action space. Due to this structural similarity, we omit redundant derivations for Value Critic and refer the reader to the Actor sparsification procedure for details.

B PROOF OF RL UPDATE

B.1 ADVANCED ACTOR UPDATE

We now consider how to compute the steepest ascent direction of the return $J(\theta)$ in Eq. (4) w.r.t. h when the policy is modelled as in Eq. (4) and h is modelled non-parametrically in a (vector-valued) RKHS $\mathcal{H}_K \subseteq \mathcal{A}^{\mathcal{S}}$. Importantly, the gradient will be an entire function in $\mathcal{H}_K$. Recalling Eq. (4), to compute the steepest ascent direction we first need to compute the gradient of $\log \pi_{h,\Sigma}(\mathbf{a}_t \mid \mathbf{s}_t)$ which is then integrated together with the Q-function. This will be a functional gradient and the notion of the Fréchet derivative is sufficient for us.

Different from the goal of standard RL as maximizing the cumulative discounted reward $J(h,\Sigma) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t\right]$, we here consider a more general maximum objective, which is

$$J(h,\Sigma) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t\right]. \tag{14}$$

According to policy gradient theorem, there is

$$\nabla_h J(h,\Sigma) = \mathbb{E}_{\mathbf{s}\sim D^{\pi_{h,\Sigma}}, \mathbf{a}\sim\pi_{h,\Sigma}(\mathbf{a}|\mathbf{s})}\left[Q_\psi^{\pi_{h,\Sigma}}(\mathbf{s},\mathbf{a}) \nabla_h \log \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s})\right]. \tag{15}$$

The policy gradient w.r.t. h To derive the policy gradient w.r.t. h for Actor, we begin with the definition of Fréchet derivative.

Definition B.1. *The Fréchet derivative is the derivative for functions on a Banach space. Let $\mathcal{V}$ and $\mathcal{W}$ be Banach spaces, and $\mathcal{U} \subset \mathcal{V}$ be an open subset of $\mathcal{V}$. A function $f : \mathcal{U} \to \mathcal{W}$ is called Fréchet differentiable at $x \in \mathcal{U}$ if there exists a bounded linear operator $Df|_x : \mathcal{V} \to \mathcal{W}$ such that,*

$$\lim_{r\to 0} \frac{\|f(x+r) - f(x) - Df|_x(r)\|_{\mathcal{W}}}{\|r\|_{\mathcal{V}}} = 0.$$

According to the kernel-based policy in Eq. (4), we derive

$$\log \pi_{h,\Sigma}(\mathbf{a}_t \mid \mathbf{s}_t) = -\log((2\pi)^{\frac{m}{2}} (\det(\Sigma))^{\frac{1}{2}}) - \frac{1}{2}\left(\mathbf{a}_t - h(\mathbf{s}_t)\right)^\top \Sigma^{-1}\left(\mathbf{a}_t - h(\mathbf{s}_t)\right). \tag{16}$$

Proposition B.2. *The derivative of the map $f : \mathcal{H}_K \to \mathbb{R}$, $f : h \mapsto \log \pi_{h,\Sigma}(\mathbf{a}_t \mid \mathbf{s}_t)$, at h, is the bounded linear map $Df|_h : \mathcal{H}_K \mapsto \mathbb{R}$ defined by*

$$Df|_h : g \mapsto (\mathbf{a} - h(\mathbf{s}))\Sigma^{-1} g(\mathbf{s}) = \langle K(\mathbf{s},\cdot)\Sigma^{-1}(\mathbf{a} - h(\mathbf{s})), g\rangle_K. \tag{17}$$

Thus the direction of steepest ascent is the function

$$\nabla_h \log \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s}) = K(\mathbf{s},\cdot)\Sigma^{-1}(\mathbf{a} - h(\mathbf{s})) \in \mathcal{H}_K.$$

Finally, we obtain the gradient of Actor w.r.t. h as follows:

$$\begin{aligned}
\nabla_h J(h,\Sigma) =& \mathbb{E}_{\mathbf{s}\sim D^{\pi_{h,\Sigma}}, \mathbf{a}\sim\pi_{h,\Sigma}(\mathbf{a}|\mathbf{s})}\left[Q_\psi^{\pi_{h,\Sigma}}(\mathbf{s},\mathbf{a}) \nabla_h \log \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s})\right] \\
=& \mathbb{E}_{\mathbf{s}\sim D^{\pi_{h,\Sigma}}, \mathbf{a}\sim\pi_{h,\Sigma}(\mathbf{a}|\mathbf{s})}\left[Q_\psi^{\pi_{h,\Sigma}}(\mathbf{s},\mathbf{a}) K(\mathbf{s}_t,\cdot)\Sigma^{-1}(\mathbf{a} - h(\mathbf{s}))\right] \\
\approx& \frac{1}{n}\sum_{\iota}\left[Q_\psi^{\pi_{h,\Sigma}}(\mathbf{s}_\iota,\mathbf{a}_\iota) K(\mathbf{s}_\iota,\cdot)\Sigma^{-1}(\mathbf{a}_\iota - h(\mathbf{s}_\iota))\right]. \\
\approx& \frac{1}{n}\sum_{\iota}\left[\widehat{A}_\psi^{\pi_{h,\Sigma}}(\mathbf{s}_\iota,\mathbf{a}_\iota) K(\mathbf{s}_\iota,\cdot)\Sigma^{-1}(\mathbf{a}_\iota - h(\mathbf{s}_\iota))\right],
\end{aligned} \tag{18}$$

where the last line exists since value function $V_{w_\mathsf{V}}^{\pi_{h,\Sigma}}(\mathbf{s}_t)$ is not directly affected by action $\mathbf{a}$ and

$$\mathbb{E}_{\mathbf{a}\sim\pi_{h,\Sigma}(\mathbf{a}|\mathbf{s})}\left[\nabla \log \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s})\right] = \sum_{\mathbf{a}} \left[\nabla \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s})\right] = \nabla \left[\sum_{\mathbf{a}} \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s})\right] = \nabla[1] = 0.$$

Note that the steepest ascent direction is a function in $\mathcal{H}$. And we can estimate $\nabla_h J(h, \Sigma)$ by sampling n state-action pairs $\{\mathbf{s}_\iota, \mathbf{a}_\iota, r_\iota, \mathbf{s}_{\iota+1}\}$ and approximating with the average.

B.2 Compatible function approximation

In this section, we demonstrate that the Advantage Critic and Value Critic satisfied the compatible function architecture.

To begin with, we have

$$\widehat{Q}^{\pi_{h,\Sigma}}(\mathbf{s}, \mathbf{a}) = \widehat{A}^{\pi_{h,\Sigma}}(\mathbf{s}, \mathbf{a}) + \widehat{V}^{\pi_{h,\Sigma}}(\mathbf{s}) = A_{w_\mathsf{A}}^{\pi_{h,\Sigma}}(\mathbf{s}, \mathbf{a}) + V_{w_\mathsf{V}}^{\pi_{h,\Sigma}}(\mathbf{s})$$

along with function approximation.

According to Theorem 2 of policy gradient with function approximation in Sutton et al. (1999), we just need to prove

$$\mathbb{E}_{\mathbf{s}\sim D^{\pi_{h,\Sigma}}, \mathbf{a}\sim\pi_{h,\Sigma}(\mathbf{a}|\mathbf{s})}\left[\left(Q^{\pi_{h,\Sigma}}(\mathbf{s}, \mathbf{a}) - \widehat{Q}^{\pi_{h,\Sigma}}(\mathbf{s}, \mathbf{a})\right) \nabla_{w_\mathsf{A}, w_\mathsf{V}} \widehat{Q}^{\pi_{h,\Sigma}}(\mathbf{s}, \mathbf{a})\right] = 0, \tag{19}$$

which shows that the error in $\widehat{Q}^{\pi_{h,\Sigma}}(\mathbf{s}, \mathbf{a})$ is orthogonal to the gradient of the policy parametrization.

Since value function $V_{w_\mathsf{V}}^{\pi_{h,\Sigma}}(\mathbf{s})$ is not directly affected by action $\mathbf{a}$, and

$$\mathbb{E}_{\mathbf{a}\sim\pi_{h,\Sigma}(\mathbf{a}|\mathbf{s})}\left[\nabla \log \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s})\right] = \sum_{\mathbf{a}} \left[\nabla \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s})\right] = \nabla \left[\sum_{\mathbf{a}} \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s})\right] = \nabla[1] = 0.$$

Therefore, we turn to derive

$$\mathbb{E}_{\mathbf{s}\sim D^{\pi_{h,\Sigma}}, \mathbf{a}\sim\pi_{h,\Sigma}(\mathbf{a}|\mathbf{s})}\left[\left(Q^{\pi_{h,\Sigma}}(\mathbf{s}, \mathbf{a}) - \widehat{Q}^{\pi_{h,\Sigma}}(\mathbf{s}, \mathbf{a})\right) \nabla_{w_\mathsf{A}} A_{w_\mathsf{A}}^{\pi_{h,\Sigma}}(\mathbf{s}, \mathbf{a})\right] = 0. \tag{20}$$

Combining

$$\nabla_{w_\mathsf{A}} A_{w_\mathsf{A}}^{\pi_{h,\Sigma}}(\mathbf{s}, \mathbf{a}) = \nu(\mathbf{s}, \mathbf{a}) = K(\mathbf{s}, \cdot)\Sigma^{-1}\left(\mathbf{a} - h(\mathbf{s})\right)$$

with Eq. (15), we complete the proof.

C Proof of theoretical guarantees

C.1 Constants and Notation

Unless otherwise stated, all constants below are positive and independent of the time index.

- $q \in \mathbb{N}$: (sparse) dictionary size.
- $C_k > 0$: minimum separation among dictionary centers.
- $M_k := \sup_{\mathbf{s}} k(\mathbf{s}, \mathbf{s})$: boundedness constant of a scalar kernel k.
- $M_K := \sup_{\mathbf{s}} \|K(\mathbf{s}, \mathbf{s})\|_{\mathrm{op}}$: boundedness of an operator-valued kernel K.
- C_{ev}: evaluation-operator bound, i.e., $\|f(\mathbf{s})\|_2 \le C_{\mathrm{ev}} \|f\|_{\mathcal{H}}$ for all f in the RKHS.
- M_Γ: uniform bound for second-order operators such as $\Gamma_s = \mathbb{E}[\psi(\mathbf{s}) \otimes \psi(\mathbf{s})]$ and $\Gamma_{2s} = \mathbb{E}[\psi(\mathbf{s}) \otimes \psi(\mathbf{s}')]$, so that $\|\Gamma_s\|_{\mathrm{op}}, \|\Gamma_{2s}\|_{\mathrm{op}} \le M_\Gamma$ (for scalar kernels one may take $M_\Gamma = M_k$).
- L_k: Lipschitz constant of the kernel with respect to inputs and we reuse the same symbol when applied to an entire Gram matrix.
- L_ψ: Lipschitz constant of the (scalar) feature map, $\|\psi(\mathbf{s}) - \psi(\tilde{\mathbf{s}})\|_{\mathcal{H}_k} \le L_\psi \|\mathbf{s} - \tilde{\mathbf{s}}\|_2$.

- L_{grad}: Lipschitz constant of the score-gradient $g(h) = \nabla_h \log \pi_h(\mathbf{a} \mid \mathbf{s})$ with respect to h; in our bounds one can take
$$L_{\text{grad}} = \sqrt{M_K}\|\Sigma^{-1}\|_{\text{op}}^{1/2} C_{\text{ev}}.$$
- $M_\Sigma \coloneqq \|\Sigma^{-1}\|_{\text{op}}$: boundedness of the Actor covariance.
- M_a: bound on action residuals, i.e., $\|\mathbf{a} - h(\mathbf{s})\|_2 \le M_a$.
- M_V: bound on the Value Critic norm in RKHS, $\|w_{\mathsf{V}}^\star\|_{\mathcal{H}_k} \le M_V$.
- $M_c \coloneqq \max_j \|\mathbf{c}_j\|_2$, $M_\eta \coloneqq \|\eta\|_2$: bounds on the Actor and Critic coefficient vectors.
- $\rho_{\mathsf{V}} \coloneqq \exp\left(-\frac{C_k^2}{2l^2}\right)$: coherence bound for Gaussian kernels (kernel lenth-scale l); hence $\lambda_{\min}(\mathbf{K}_{\mathsf{V},\mathsf{V}}) \ge 1 - (q-1)\rho_{\mathsf{V}}$.
- $M_S \coloneqq \sup_{\mathbf{s}} \|\mathbf{s}\|_2$: state-norm bound (used with Mahalanobis RBF perturbations).
- M: uniform bound for component kernels in product kernels ($k^{(j)} \le M$); implies $\|\mu_{\mathcal{C}}\|_{\mathcal{H}} \le M^{|\overline{\mathcal{C}}|/2}$.
- Sampling/concentration constants: n_{eff} (effective sample size), τ_{mix} (mixing time), confidence level $\delta \in (0,1)$, and Bernstein-type constants C_B, C_b.
- Two-timescale stepsizes: $\alpha_t^{\mathsf{v}} = \dfrac{C_{\mathsf{v}}}{(t+1)^\nu}$, $\alpha_t^{\mathsf{h}} = \dfrac{C_{\mathsf{h}}}{(t+1)^\sigma}$ with $0 < \nu < \sigma \le 1$.
- L_{V}: Lipschitz drift of the Value Critic w.r.t. actor parameters: $\|w_{\mathsf{V}}^\star(h') - w_{\mathsf{V}}^\star(h)\| \le L_{\mathsf{V}}\|h' - h\|$.

When only scalar kernels are involved, one may safely replace M_K, M_Γ by M_k. For operator-valued kernels / vector-valued RKHS, keep M_K (kernel boundedness) and M_Γ (second-order operator bound) separate.

C.2 Supporting Lemmas

Lemma C.1 (Theorem 5 in Zhang et al. (2020))**.** *Given a policy π for a non-adversarial MDP, its value function is $V_\pi(\mathbf{s})$. Under the adversary b in SA-MDP, for all $\mathbf{s} \in \mathcal{S}$ we have*
$$\max_{\mathbf{s}\in\mathcal{S}} \left\{V^\pi(\mathbf{s}) - V^{\widetilde{\pi}}(\mathbf{s})\right\} \le \alpha \max_{\mathbf{s}\in\mathcal{S}} \max_{\widetilde{\mathbf{s}}\in B(\mathbf{s})} \mathrm{D}_{\mathrm{TV}}(\pi(\cdot|\mathbf{s}), \widetilde{\pi}(\cdot|\widetilde{\mathbf{s}})), \tag{21}$$
where $\mathrm{D}_{\mathrm{TV}}(\pi(\cdot|\mathbf{s}), \widetilde{\pi}(\cdot|\widetilde{\mathbf{s}}))$ is the total variation distance between $\pi(\cdot|\mathbf{s})$ and $\widetilde{\pi}(\cdot|\widetilde{\mathbf{s}})$, and $\alpha \coloneqq 2[1 + \frac{\gamma}{(1-\gamma)^2}] \max_{(\mathbf{s},\mathbf{a},\mathbf{s}')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}} |r(\mathbf{s},\mathbf{a},\mathbf{s}')|$ is a constant that does not depend on π and $\widetilde{\pi}$.

Lemma C.2 (Bounded perturbation of RKHS–SHAP values)**.** *Suppose Assumptions 4.4, 4.1, and 4.2 hold. Let k be a product kernel with d bounded component kernels satisfying $|k^{(i)}(\mathbf{s},\mathbf{s})| \le M$ for all $i \in \mathcal{X}$, and assume $\|w_{\mathsf{V}}\|_{\mathcal{H}_k}^2 \le M_V$. Define $\delta_\psi \coloneqq \sup_{\mathcal{C}\subseteq\mathcal{X}} \|\psi(\mathbf{s}^{\mathcal{C}}) - \psi(\widetilde{\mathbf{s}}^{\mathcal{C}})\|_{\mathcal{H}_k}^2$, where $\widetilde{\mathbf{s}} \in B(\mathbf{s})$, and let $\delta_{\widetilde{\mathsf{V}}}$ be the projection/perturbation residual given in Lemma C.3. Then, for any feature i,*
$$\left\|\phi_i^{(\text{off})} - \widetilde{\phi}_i^{(\text{off})}\right\|_2^2 \le 4\delta_{\widetilde{\mathsf{V}}}^2 M^d + 4M_V\delta_\psi M^d, \tag{22a}$$
$$\left\|\phi_i^{(\text{on})} - \widetilde{\phi}_i^{(\text{on})}\right\|_2^2 \le 4\delta_{\widetilde{\mathsf{V}}}^2 M^d + 8M_V M_\Gamma\delta_\psi M^d, \tag{22b}$$
where $M_\Gamma \coloneqq \sup_{\mathcal{C}\subseteq\mathcal{X}} \|\mu_{\mathbf{s}^{\overline{\mathcal{C}}}|\mathbf{s}^{\mathcal{C}}}\|_{\Gamma_{\mathbf{s}^{\mathcal{C}}}}^2$. If, in addition, k is an RBF kernel with lengthscale l and Assumption 4.5 holds (for ρ_{V}), then $\delta_\psi = 2 - 2\exp\big(-\|\mathbf{s} - \widetilde{\mathbf{s}}\|_2^2/(2l^2)\big)$ and
$$\left\|\phi_i^{(\text{off})} - \widetilde{\phi}_i^{(\text{off})}\right\|_2^2 \le \frac{4M^d M_\eta^2 q^4 \varepsilon^2}{\big(1-(q-1)\rho_{\mathsf{V}}\big)^2}\left(\frac{L_k}{1-(q-1)\rho_{\mathsf{V}}} + M_k L_\psi + 2M_k^{1/2}L_k\right)^2 + 4M_V\left(1 - e^{-\frac{\varepsilon^2}{2l^2}}\right), \tag{23a}$$
$$\left\|\phi_i^{(\text{on})} - \widetilde{\phi}_i^{(\text{on})}\right\|_2^2 \le \frac{4M^d M_\eta^2 q^4 \varepsilon^2}{\big(1-(q-1)\rho_{\mathsf{V}}\big)^2}\left(\frac{L_k}{1-(q-1)\rho_{\mathsf{V}}} + M_k L_\psi + 2M_k^{1/2}L_k\right)^2 + 8M_V M_\Gamma\left(1 - e^{-\frac{\varepsilon^2}{2l^2}}\right), \tag{23b}$$
where $\varepsilon = \max\{\max_j \|\mathbf{s}_j - \widetilde{\mathbf{s}}_j\|,\ \|\mathbf{s}_\iota - \widetilde{\mathbf{s}}_\iota\|\}$, $M_\eta \coloneqq \sup_t \|\eta_t\|_2$, and L_k and L_ψ are the Lipschitz constants from Assumption 4.4.

Lemma C.3 (Perturbation error of the value representation)**.** *Under Assumptions 4.4 and 4.5, consider dictionaries $\mathcal{D}_{\mathsf{V}} = \{\mathbf{s}_j\}_{j=1}^q$ and $\widetilde{\mathcal{D}}_{\mathsf{V}} = \{\widetilde{\mathbf{s}}_j\}_{j=1}^q$, and a new center $\mathbf{s}_\iota$ with perturbations bounded by $\max\{\max_j \|\mathbf{s}_j - \widetilde{\mathbf{s}}_j\|,\ \|\mathbf{s}_\iota - \widetilde{\mathbf{s}}_\iota\|\} \le \varepsilon$. Let $\rho_{\mathsf{V}} = \exp(-C_k^2/(2l^2))$ and assume both Gram matrices are invertible. Then the residual*

$$\delta_{\tilde{\mathsf{V}}} := \left\| \sum_{j,j'=1}^{q} \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\eta_\iota \psi(\mathbf{s}_j) - \sum_{j,j'=1}^{q} \left[\widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\widetilde{\mathbf{s}}_{j'}, \widetilde{\mathbf{s}}_\iota)\eta_\iota \psi(\widetilde{\mathbf{s}}_j) \right\|_{\mathcal{H}_k}$$

obeys the linear bound

$$\delta_{\tilde{\mathsf{V}}} \le \frac{M_\eta q^2}{1-(q-1)\rho_{\mathsf{V}}} \left[\frac{L_k}{1-(q-1)\rho_{\mathsf{V}}} + M_k L_\psi + 2M_k^{1/2} L_k \right] \varepsilon.$$

Lemma C.4 (Projection error for the Value Critic update)**.** *Assume 4.4 and 4.5. For the projected update $\overline{w}_{t+1} = w_t - \alpha_t^{\mathsf{v}} \nabla J(w_t)$ and $w_{t+1} = \Pi_{\mathcal{H}_{\mathcal{D}_{\mathsf{V}},t}}(\overline{w}_{t+1})$, the one-step projection error satisfies*

$$\delta_{\mathsf{PV}} := \|\overline{w}_{t+1} - w_{t+1}\|_{\mathcal{H}_k} \le M_k^{1/2} q \varepsilon_{\mathsf{PV}},$$

where

$$\varepsilon_{\mathsf{PV}} = \max\left\{ \sup_{\psi_j \in \mathcal{D}_{\mathsf{V}}} \left| \sum_{j'} \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\eta_\iota - \eta_j \right|, \right.$$
$$\left. \sup_{\psi_j \in \{\psi_\iota\} \cup (\mathcal{D}_{\mathsf{V}} \setminus \{\psi_{j^\star}\})} \left| \sum_{j'} \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_{j^\star})\eta_{j^\star} - \eta_j \right| \right\}.$$

Lemma C.5 (Gradient growth for the Value Critic)**.** *Under Assumptions 4.7 and 4.4, for any $t \ge 0$,*

$$\left\|g_t(w_{\mathsf{V},t})\right\|_{\mathcal{H}_k}^2 \le C_1 \|w_{\mathsf{V},t} - w_{\mathsf{V},t}^\star\|_{\mathcal{H}_k}^2, \qquad C_1 := \big((1+\gamma)M_k + \lambda\big)^2.$$

Lemma C.6 (Actor drift induced by kernel change)**.** *Let $K(\mathbf{s}, \mathbf{s}') = \kappa_{\mathbf{W}}(\mathbf{s}, \mathbf{s}')\Sigma_{\mathbf{K}}$ with $\|\Sigma_{\mathbf{K}}\|_{\mathrm{op}} \le M_\Sigma$, and*

$$h(\cdot) = \sum_{j=1}^{q} K(\cdot, \mathbf{s}_j)\mathbf{c}_j, \qquad h'(\cdot) = \sum_{j=1}^{q} K'(\cdot, \mathbf{s}_j)\mathbf{c}_j,$$

where K' uses $\mathbf{W}'$ instead of $\mathbf{W}$ and $M_A := \max_j \|\mathbf{c}_j\|_2$. If $\|\mathbf{s}\|_2 \le M_S$ for all states and $\sup_{\mathbf{s},\mathcal{C}} \|\mu_{\mathcal{C}}(\mathbf{s})\|_{\mathcal{H}_k} \le C_\mu$, then

$$\delta_{\mathsf{K},h} \le 16 M_{\Sigma_K}^2 M_S^4 q^2 M_a^2 C_\mu^2 \|w_{\mathsf{V}}' - w_{\mathsf{V}}\|_{\mathcal{H}_k}^2.$$

Lemma C.7 (Smoothness lower bound with projection error)**.** *Let $G(h) := \mathbb{E}_{\nu_{\pi^\star}}[\log \pi_h(\mathbf{a} \mid \mathbf{s})]$ be $L_{\log \pi}$-smooth and concave on $(\mathcal{H}_K, \|\cdot\|_{\mathcal{H}_K})$. Assume $M_K := \sup_{\mathbf{s}} \|K(\mathbf{s}, \mathbf{s})\|_{\mathrm{op}} < \infty$ and the orthogonal projection onto the Actor dictionary subspace is well defined (Assumption 4.5). If h_{t+1} is obtained by projecting $\overline{h}_{t+1}$ and $\delta_{K,t} := \|h_{t+1} - \overline{h}_{t+1}\|_{\mathcal{H}_K}^2$, then*

$$\mathbb{E}_{\nu_{\pi^\star}}\left[\log \pi_{h_{t+1}} - \log \pi_{h_t}\right] \ge \mathbb{E}_{\nu_{\pi^\star}}\left[\nabla \log \pi_{h_t}^\top (h_{t+1} - h_t)\right] - \frac{L_{\mathrm{grad}}}{2}\|h_{t+1} - h_t\|_{\mathcal{H}_K}^2 - 2M_a^2 M_\Sigma^2 \delta_{\mathsf{K},h_t},$$

where $\delta_{K,t}$ can be further controlled via Lemma C.6.

Lemma C.8 (Projection error for the Actor mean)**.** *Assume $M_K := \sup_{\mathbf{s}} \|K(\mathbf{s}, \mathbf{s})\|_{\mathrm{op}} < \infty$ and the block Gram matrix on the Actor dictionary is invertible (Assumption 4.5). For the projected update $\overline{h}_{t+1} = h_t - \alpha_t^{\mathsf{h}} \nabla J(h_t)$ and $h_{t+1} = \Pi_{\mathcal{H}_{K,\mathcal{D}_{\mathsf{A}},t}}(\overline{h}_{t+1})$,*

$$\delta_{\mathsf{PA}} := \|\overline{h}_{t+1} - h_{t+1}\|_{\mathcal{H}_K} \le \sqrt{M_K} q \varepsilon_{\mathsf{PA}},$$

where

$$\varepsilon_{\mathsf{PA}} = \max\left\{ \sup_{K(\cdot,\mathbf{s}_j) \in \mathcal{D}_{\mathsf{A}}} \left\|[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\mathbf{K}_{\mathsf{A},\iota}]\mathbf{c}_\iota - \mathbf{c}_j\right\|_2, \sup_{K(\cdot,\mathbf{s}_j) \in \{\mathbf{s}_\iota\} \cup (\mathcal{D}_{\mathsf{A}} \setminus \{\mathbf{s}_{j^\star}\})} \left\|[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\mathbf{K}_{\mathsf{A},j^\star}]\mathbf{c}_{j^\star} - \mathbf{c}_j\right\|_2 \right\}.$$

Lemma C.9 (Lipschitz drift of the target Critic)**.** *Under Assumptions 4.7 and 4.4, for any $h, h' \in \mathcal{H}_K$,*

$$\|w^\star_{\mathsf{V}}(h') - w^\star_{\mathsf{V}}(h)\|_{\mathcal{H}_k} \le L_{\mathsf{V}} \|h' - h\|_{\mathcal{H}_K}, \qquad L_{\mathsf{V}} := C_\nu \sqrt{M_k} \left(\frac{1}{\lambda} + \frac{2(1+\gamma)M_k}{\lambda^2} \right).$$

Lemma C.10 (Performance Difference Lemma, discounted MDP)**.** *For any two policies π, π',*

$$J(\pi') - J(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{(\mathbf{s},\mathbf{a}) \sim \nu_{\pi'}} \left[A^\pi(\mathbf{s}, \mathbf{a}) \right].$$

Lemma C.11 (Refined tracking error bound for the RKHS Critic under two time-scales)**.** *Let $\mathcal{H}_\psi$ be the Critic RKHS with bounded feature map $\|\psi(s)\|^2_{\mathcal{H}_\psi} \le M_k$. Suppose the on-policy sampling process is geometric mixing so that the effective sample size is $n_{\text{eff}} \asymp n/\tau_{\text{mix}}$. Consider the two time-scale stepsizes $\alpha^{\mathsf{v}}_t > 0$ and $\alpha^{\mathsf{h}}_t > 0$. Then there exists $T_0 < \infty$, under polynomial stepsizes $\alpha^{\mathsf{v}}_t = \frac{C_\nu}{(t+1)^\nu}$ and $\alpha^{\mathsf{h}}_t = \frac{C_{\mathsf{h}}}{(t+1)^\sigma}$ with $0 < \nu < \sigma \le 1$, such that the decay rates satisfy*

$$\mathbb{E}\left[\left\| w_{\mathsf{V},t} - w^\star_{\mathsf{V},t} \right\|^2_{\mathcal{H}_\psi} \right] = \begin{cases} O\left((t+1)^{-\nu}\right) + M_k q^2 \varepsilon^2_{\mathsf{PV}} + O\left(\frac{1}{n_{\text{eff}}}\right), & \sigma > \nu, \\ O\left((t+1)^{-\nu} \log^2 t\right) + M_k q^2 \varepsilon^2_{\mathsf{PV}} + O\left(\frac{1}{n_{\text{eff}}}\right), & \sigma = \nu \text{ (borderline)}. \end{cases}$$

C.3 Proof of Theorem 4.8

According to Lemma C.1, we just need to bound $\mathrm{D}_{\mathrm{TV}}(\pi(\cdot|\mathbf{s}), \widetilde{\pi}(\cdot|\widetilde{\mathbf{s}}))$. Specifically, the policy of RSA2C is characterize as Gaussian policy. We upper bound the TV distance as

$$\begin{aligned} & \mathrm{D}_{\mathrm{TV}}(\pi(\cdot|\mathbf{s}), \widetilde{\pi}(\cdot|\widetilde{\mathbf{s}})) \\ \le & \sqrt{\frac{1}{2} \mathrm{D}_{\mathrm{KL}}(\pi(\cdot|\mathbf{s}) \| \widetilde{\pi}(\cdot|\widetilde{\mathbf{s}}))} \\ = & \sqrt{\frac{1}{4} \left(\log \left| \Sigma_{\widetilde{\mathbf{s}}} \Sigma_{\mathbf{s}}^{-1} \right| + \mathrm{tr}\left(\Sigma_{\widetilde{\mathbf{s}}}^{-1} \Sigma_{\mathbf{s}} \right) + \left(\widetilde{h}(\widetilde{\mathbf{s}}) - h(\mathbf{s}) \right)^\top \Sigma_{\widetilde{\mathbf{s}}}^{-1} \left(\widetilde{h}(\widetilde{\mathbf{s}}) - h(\mathbf{s}) \right) - m \right)} \\ = & \sqrt{\frac{1}{4} \left(\left(\widetilde{h}(\widetilde{\mathbf{s}}) - h(\mathbf{s}) \right)^\top \Sigma^{-1} \left(\widetilde{h}(\widetilde{\mathbf{s}}) - h(\mathbf{s}) \right) \right)}. \end{aligned} \tag{24}$$

Here, the last equality holds due to $\Sigma_{\mathbf{s}} = \Sigma_{\widetilde{\mathbf{s}}} = \Sigma$ followed from line 9 in Algorithm 1.

Armed with eigen-decomposition method, we further bound the result in Eq. (24) as

$$\mathrm{D}_{\mathrm{TV}}(\pi(\cdot|\mathbf{s}), \widetilde{\pi}(\cdot|\widetilde{\mathbf{s}})) \le \frac{\sqrt{\lambda_{\max}(\Sigma^{-1})}}{2} \left\| \widetilde{h}(\widetilde{\mathbf{s}}) - h(\mathbf{s}) \right\|_2 = \frac{M_\Sigma}{2} \left\| \widetilde{h}(\widetilde{\mathbf{s}}) - h(\mathbf{s}) \right\|_2, \tag{25}$$

where $M_\Sigma := \sqrt{\lambda_{\max}(\Sigma^{-1})}$.

Armed with the definition of mapping function and Eq. (3), we have

$$
\begin{aligned}
&\left\|\widetilde{h}(\widetilde{\mathbf{s}})-h(\mathbf{s})\right\|_2\\
=&\left\|\frac{1}{q}\sum_{(\mathbf{s}_j,\mathbf{c}_j)\in\mathcal{D}_{\mathsf{A}}}K(\mathbf{s},\mathbf{s}_j)\mathbf{c}_j-\frac{1}{q}\sum_{(\mathbf{s}_j,\mathbf{c}_j)\in\mathcal{D}_{\mathsf{A}}}\widetilde{K}(\widetilde{\mathbf{s}},\widetilde{\mathbf{s}}_j)\widetilde{\mathbf{c}}_j\right\|_2\\
=&\left\|\frac{1}{q}\sum_{(\mathbf{s}_j,\mathbf{c}_j)\in\mathcal{D}_{\mathsf{A}}}\left[K(\mathbf{s},\mathbf{s}_j)\mathbf{c}_j-\widetilde{K}(\widetilde{\mathbf{s}},\widetilde{\mathbf{s}}_j)\widetilde{\mathbf{c}}_j\right]\right\|_2\\
\overset{\text{(i)}}{\leq}&\frac{1}{q}\sum_{(\mathbf{s}_j,\mathbf{c}_j)\in\mathcal{D}_{\mathsf{A}}}\left\|K(\mathbf{s},\mathbf{s}_j)\mathbf{c}_j-\widetilde{K}(\widetilde{\mathbf{s}},\widetilde{\mathbf{s}}_j)\widetilde{\mathbf{c}}_j\right\|_2\\
=&\frac{1}{q}\sum_{(\mathbf{s}_j,\mathbf{c}_j)\in\mathcal{D}_{\mathsf{A}}}\left\|\kappa_\phi(\mathbf{s},\mathbf{s}_j)\Sigma_{\mathsf{K}}\mathbf{c}_j-\kappa_{\widetilde{\mathbf{W}}}(\widetilde{\mathbf{s}},\widetilde{\mathbf{s}}_j)\Sigma_{\mathsf{K}}\widetilde{\mathbf{c}}_j\right\|_2\\
\overset{\text{(ii)}}{\leq}&\frac{1}{q}\sum_{(\mathbf{s}_j,\mathbf{c}_j)\in\mathcal{D}_{\mathsf{A}}}\left[\left|\kappa_\phi(\mathbf{s},\mathbf{s}_j)-\kappa_{\widetilde{\mathbf{W}}}(\widetilde{\mathbf{s}},\widetilde{\mathbf{s}}_j)\right|\ \|\Sigma_{\mathsf{K}}\|_{\text{op}}\|\mathbf{c}_j\|_2+\left|\kappa_\phi(\mathbf{s},\mathbf{s}_j)\right|\ \|\Sigma_{\mathsf{K}}\|_{\text{op}}\|\mathbf{c}_j-\widetilde{\mathbf{c}}_j\|_2\right]\\
\overset{\text{(iii)}}{\leq}&\frac{1}{q}\sum_{(\mathbf{s}_j,\mathbf{c}_j)\in\mathcal{D}_{\mathsf{A}}}\left[\lambda_{\max}\left(\Sigma_{\mathsf{K}}\right)M_C\delta_{\tilde{\kappa},j}+M_\kappa\lambda_{\max}\left(\Sigma_{\mathsf{K}}\right)\delta_{\tilde{c},j}\right]\\
\leq&\lambda_{\max}\left(\Sigma_{\mathsf{K}}\right)M_C\delta_{\tilde{\kappa}}+M_\kappa\lambda_{\max}\left(\Sigma_{\mathsf{K}}\right)\delta_{\tilde{c}},
\end{aligned}
\tag{26}
$$

where (i) and (ii) hold due to Cauchy-Schwarz inequality, and (iii) comes from $\|\Sigma_{\mathsf{K}}\|_{\text{op}}=\lambda_{\max}\left(\Sigma_{\mathsf{K}}\right)$, $\|\kappa_\phi(\mathbf{s},\mathbf{s}_j)\|_2\leq M_\kappa$, and the last inequality exists with $\delta_{\tilde{\kappa}}\coloneqq\sup_j\delta_{\tilde{\kappa},j}$ and $\delta_{\tilde{c}}\coloneqq\sup_j\delta_{\tilde{c},j}$.

In this paper, we utilize RKHS-SHAP by RBF kernel as Eq. (3). Therefore, armed with $w=\mathsf{diag}\{\{\phi_i\}_{i=1}^d\}$, we further bound $\delta_{\tilde{\kappa}}$ and $\delta_{\tilde{c}}$ as follows.

Bound $\delta_{\tilde{\kappa}}$. For the difference of kernel function, we have

$$
\begin{aligned}
\delta_{\tilde{\kappa},j}=&\left|\exp\left(-\frac{1}{2}(\mathbf{s}-\mathbf{s}_j)^\top w(\mathbf{s}-\mathbf{s}_j)\right)-\exp\left(-\frac{1}{2}(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)^\top\widetilde{w}(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)\right)\right|\\
\leq&\frac{1}{2}\left|(\mathbf{s}-\mathbf{s}_j)^\top w(\mathbf{s}-\mathbf{s}_j)-(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)^\top\widetilde{w}(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)\right|\\
&\times\exp\left(-\frac{1}{2}\min\left((\mathbf{s}-\mathbf{s}_j)^\top w(\mathbf{s}-\mathbf{s}_j),(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)^\top\widetilde{w}(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)\right)\right)\\
\leq&\frac{1}{2}\exp\left(-\frac{1}{2}M_S^2\right)\left|(\mathbf{s}-\mathbf{s}_j)^\top w(\mathbf{s}-\mathbf{s}_j)-(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)^\top\widetilde{w}(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)\right|\\
\leq&\frac{1}{2}\exp\left(-\frac{1}{2}M_S^2\right)\underbrace{\left|(\mathbf{s}-\mathbf{s}_j)^\top w(\mathbf{s}-\mathbf{s}_j)-(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)^\top w(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)\right|}_{A_1}\\
&+\frac{1}{2}\exp\left(-\frac{1}{2}M_S^2\right)\underbrace{\left|(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)^\top w(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)-(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)^\top\widetilde{w}(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)\right|}_{A_2},
\end{aligned}
\tag{27}
$$

where the first inequality holds due to the mean value theorem, the second one comes from $\|\mathbf{s}-\mathbf{s}_j\|_2\leq M_S$, and the last inequality exists by Cauchy-Schwarz inequality.

Firstly, we consider the upper bound on the state perturbation term, denoted as A_1. Specifically, we have

$$
\begin{aligned}
A_1=&\left|(\mathbf{s}-\mathbf{s}_j)^\top w(\mathbf{s}-\mathbf{s}_j)-(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)^\top w(\widetilde{\mathbf{s}}-\widetilde{\mathbf{s}}_j)\right|\\
=&\left|(\mathbf{s}-\mathbf{s}_j)^\top w(\mathbf{s}-\mathbf{s}_j)-((\mathbf{s}-\mathbf{s}_j)-(\mathbf{s}-\widetilde{\mathbf{s}})+(\mathbf{s}_j-\widetilde{\mathbf{s}}_j))^\top w((\mathbf{s}-\mathbf{s}_j)-(\mathbf{s}-\widetilde{\mathbf{s}})+(\mathbf{s}_j-\widetilde{\mathbf{s}}_j))\right|\\
=&\left|2((\mathbf{s}-\widetilde{\mathbf{s}})-(\mathbf{s}_j-\widetilde{\mathbf{s}}_j))^\top w(\mathbf{s}-\mathbf{s}_j)-((\mathbf{s}-\widetilde{\mathbf{s}})-(\mathbf{s}_j-\widetilde{\mathbf{s}}_j))^\top w((\mathbf{s}-\widetilde{\mathbf{s}})-(\mathbf{s}_j-\widetilde{\mathbf{s}}_j))\right|\\
\leq&4\|\mathbf{s}-\widetilde{\mathbf{s}}\|_2M_S+4\|\mathbf{s}-\widetilde{\mathbf{s}}\|_2^2M_\phi
\end{aligned}
\tag{28}
$$

where the inequality follows from the Cauchy–Schwarz inequality.

Next, for A_2, which represents the RKHS-SHAP perturbation bound, we derive

$$\begin{aligned} A_2 &= \left|(\widetilde{\mathbf{s}} - \widetilde{\mathbf{s}}_j)^\top w(\widetilde{\mathbf{s}} - \widetilde{\mathbf{s}}_j) - (\widetilde{\mathbf{s}} - \widetilde{\mathbf{s}}_j)^\top \widetilde{w}(\widetilde{\mathbf{s}} - \widetilde{\mathbf{s}}_j)\right| \\ &= \left|(\widetilde{\mathbf{s}} - \widetilde{\mathbf{s}}_j)^\top (w - \widetilde{w}) (\widetilde{\mathbf{s}} - \widetilde{\mathbf{s}}_j)\right| \\ &\leq \sum_{i=1}^{d} M_S^2 \left|\phi_i - \widetilde{\phi}_i\right|. \end{aligned} \tag{29}$$

Therefore, we have

$$\delta_{\widetilde{\kappa}} \leq 2e^{-\frac{1}{2}M_S^2}\left(\varepsilon M_S + \varepsilon^2 M_\phi\right) + \frac{1}{2}e^{-\frac{1}{2}M_S^2}\sum_{i=1}^{d} M_S^2 \left|\phi_i - \widetilde{\phi}_i\right|. \tag{30}$$

Bound $\delta_{\widetilde{c}}$. For online sparsification, we consider the following optimization problem:

$$\min_{\{\mathbf{c}_j\}\in\mathcal{D}_{\mathsf{A}}} \left\| \sum_j K(\mathbf{s}_j, \cdot)\mathbf{c}_j - K(\mathbf{s}_\iota, \cdot)\mathbf{c}_\iota \right\|_{\mathcal{H}_K}^2. \tag{31}$$

By expanding the norm and applying the reproducing property of the RKHS, we obtain:

$$\begin{aligned} \mathcal{L}(\{\mathbf{c}_j\}) &= \left\langle \sum_j K(\mathbf{s}_j, \cdot)\mathbf{c}_j - K(\mathbf{s}_\iota, \cdot)\mathbf{c}_\iota, \sum_{j'} K(\mathbf{s}_{j'}, \cdot)\mathbf{c}_{j'} - K(\mathbf{s}_\iota, \cdot)\mathbf{c}_\iota \right\rangle_{\mathcal{H}_K} \\ &= \sum_{j,j'} \langle \mathbf{c}_j, K(\mathbf{s}_j, \mathbf{s}_{j'})\mathbf{c}_{j'}\rangle - 2\sum_j \langle \mathbf{c}_j, K(\mathbf{s}_j, \mathbf{s}_\iota)\mathbf{c}_\iota\rangle + \langle \mathbf{c}_\iota, K(\mathbf{s}_\iota, \mathbf{s}_\iota)\mathbf{c}_\iota\rangle. \end{aligned} \tag{32}$$

Let $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_q]^\top \in \mathbb{R}^{q\times d}$ denote the coefficient matrix, $\mathbf{K}_{\mathsf{A},\mathsf{A}} \in \mathbb{R}^{q\times q}$ the Gram matrix over the dictionary, and $\mathbf{K}_\iota \in \mathbb{R}^{q\times d}$ the cross-kernel matrix with j-th row given by $K(\mathbf{s}_j, \mathbf{s}_\iota)\mathbf{c}_\iota$. The objective function becomes:

$$\mathcal{L}(\mathbf{C}) = \mathrm{Tr}(\mathbf{C}^\top \mathbf{K}_{\mathsf{A},\mathsf{A}}\mathbf{C}) - 2\,\mathrm{Tr}(\mathbf{C}^\top \mathbf{K}_\iota) + \mathbf{c}_\iota^\top K(\mathbf{s}_\iota, \mathbf{s}_\iota)\mathbf{c}_\iota. \tag{33}$$

Setting the gradient $\nabla_{\mathbf{C}}\mathcal{L} = 0$ yields the closed-form optimal solution:

$$\mathbf{C}^* = \mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\mathbf{K}_\iota. \tag{34}$$

The optimal coefficient for each basis function is thus given by:

$$\mathbf{c}_j^* = \sum_{j'=1}^{q} \left[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} K(\mathbf{s}_{j'}, \mathbf{s}_\iota)\mathbf{c}_\iota. \tag{35}$$

This solution corresponds to the orthogonal projection of the new kernel function $K(\mathbf{s}_\iota, \cdot)\mathbf{c}_\iota$ onto the subspace spanned by the dictionary atoms $\{K(\mathbf{s}_j, \cdot)\}$.

Therefore, we have

$$\begin{aligned} \delta_{\widetilde{c},j} &= \left\| \sum_{j'=1}^{q} \left[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} K(\mathbf{s}_{j'}, \mathbf{s}_\iota)\mathbf{c}_\iota - \sum_{j'=1}^{q} \left[\widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} K(\widetilde{\mathbf{s}}_{j'}, \mathbf{s}_\iota)\mathbf{c}_\iota \right\|_2 \\ &\leq \sum_{j'=1}^{q} \left\| \left[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} K(\mathbf{s}_{j'}, \mathbf{s}_\iota)\mathbf{c}_\iota - \left[\widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} K(\widetilde{\mathbf{s}}_{j'}, \mathbf{s}_\iota)\mathbf{c}_\iota \right\|_2 \\ &= \sum_{j'=1}^{q} \left\| \left[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \kappa_\phi(\mathbf{s}_{j'}, \mathbf{s}_\iota)\Sigma_{\mathsf{K}}\mathbf{c}_\iota - \left[\widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \kappa_\phi(\widetilde{\mathbf{s}}_{j'}, \mathbf{s}_\iota)\Sigma_{\mathsf{K}}\mathbf{c}_\iota \right\|_2 \\ &\leq \sum_{j'=1}^{q} \left\| \left[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \kappa_\phi(\mathbf{s}_{j'}, \mathbf{s}_\iota) - \left[\widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \kappa_\phi(\widetilde{\mathbf{s}}_{j'}, \mathbf{s}_\iota) \right\|_{\mathcal{H}_K} \|\Sigma_{\mathsf{K}}\|_2 \|\mathbf{c}_\iota\|_2 \\ &\leq \lambda_{\max}(\Sigma_{\mathsf{K}})\, M_c \sum_{j'=1}^{q} \left\| \left[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \kappa_\phi(\mathbf{s}_{j'}, \mathbf{s}_\iota) - \left[\widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \kappa_\phi(\widetilde{\mathbf{s}}_{j'}, \mathbf{s}_\iota) \right\|_{\mathcal{H}_K}. \end{aligned} \tag{36}$$

We consider bounding the RKHS norm

$$\left\| \left[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \kappa_\phi(\mathbf{s}_{j'}, \mathbf{s}_\iota) - \left[\widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \kappa_\phi(\widetilde{\mathbf{s}}_{j'}, \mathbf{s}_\iota) \right\|_{\mathcal{H}_K}, \tag{37}$$

where $\mathbf{K}_{\mathsf{A},\mathsf{A}}$ and $\widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}$ denote Gram matrices over kernel dictionaries $\mathcal{D}_\mathsf{A} = \{\mathbf{s}_j\}$ and its perturbed version $\widetilde{\mathcal{D}}_\mathsf{A} = \{\widetilde{\mathbf{s}}_j\}$, respectively. Using the triangle inequality and RKHS norm properties, we decompose:

$$\begin{aligned}
&\left\| \left[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \kappa_\phi(\mathbf{s}_{j'}, \mathbf{s}_\iota) - \left[\widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \kappa_\phi(\widetilde{\mathbf{s}}_{j'}, \mathbf{s}_\iota) \right\|_{\mathcal{H}_K} \\
&\leq \left| \left[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1} - \widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \right| \cdot \|\kappa_\phi(\mathbf{s}_{j'}, \cdot)\|_{\mathcal{H}_K} + \left| \left[\widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \right| \cdot \|\kappa_\phi(\mathbf{s}_{j'}, \cdot) - \kappa_\phi(\widetilde{\mathbf{s}}_{j'}, \cdot)\|_{\mathcal{H}_K} \\
&\leq M_\kappa \left| \left[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1} - \widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \right| + \left| \left[\widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}^{-1}\right]_{jj'} \right| \cdot \|\kappa_\phi(\mathbf{s}_{j'}, \cdot) - \kappa_\phi(\widetilde{\mathbf{s}}_{j'}, \cdot)\|_{\mathcal{H}_K}.
\end{aligned} \tag{38}$$

Then, similar to the bound of δ_κ, we have

$$\|\kappa_\phi(\mathbf{s}_{j'}, \cdot) - \kappa_\phi(\widetilde{\mathbf{s}}_{j'}, \cdot)\|_{\mathcal{H}_K} \leq 2e^{-\frac{1}{2}M_S^2}\left(\varepsilon M_S + \varepsilon^2 M_\phi\right) + \frac{1}{2}e^{-\frac{1}{2}M_S^2}\sum_{i=1}^{d} M_S^2 \left\|\phi_i - \widetilde{\phi}_i\right\|_2. \tag{39}$$

Moreover, we apply the Banach perturbation lemma to the inverse kernel matrices. Then:

$$\left\|\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1} - \widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}^{-1}\right\| \leq \|\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\| \cdot \|\widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}} - \mathbf{K}_{\mathsf{A},\mathsf{A}}\| \cdot \|\widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}}^{-1}\|. \tag{40}$$

Suppose the dictionary centers satisfy a minimum separation $C_K > 0$ and κ_ϕ is Gaussian. Then off-diagonal terms are bounded by $\rho_\mathsf{A} \coloneqq \exp\left(-\frac{C_K^2}{2l^2}\right)$ and the smallest eigenvalue of $\mathbf{K}_{\mathsf{A},\mathsf{A}}$ satisfies

$$\lambda_{\min}(\mathbf{K}_{\mathsf{A},\mathsf{A}}) \geq 1 - (q-1)\rho_\mathsf{A}, \quad \Rightarrow \quad \|\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1}\| \leq \frac{1}{1-(q-1)\rho_\mathsf{A}}. \tag{41}$$

Combining with the kernel matrix perturbation

$$\|\widetilde{\mathbf{K}}_{\mathsf{A},\mathsf{A}} - \mathbf{K}_{\mathsf{A},\mathsf{A}}\| \leq 2e^{-\frac{1}{2}M_S^2}\left(\varepsilon M_S + \varepsilon^2 M_\phi\right) + \frac{1}{2}e^{-\frac{1}{2}M_S^2}\sum_{i=1}^{d} M_S^2 \left\|\phi_i - \widetilde{\phi}_i\right\|_2, \tag{42}$$

we obtain the final upper bound:

$$\delta_{\widetilde{c}} \leq \frac{e^{-\frac{1}{2}M_S^2} M_{\Sigma_\mathsf{K}} M_c q}{1-(q-1)\rho_\mathsf{A}}\left[\frac{M_\kappa}{1-(q-1)\rho_\mathsf{A}} + 1\right] \cdot \left[2\left(\varepsilon M_S + \varepsilon^2 M_\phi\right) + \frac{1}{2}\sum_{i=1}^{d} M_S^2 \left\|\phi_i - \widetilde{\phi}_i\right\|_2\right], \tag{43}$$

where $M_{\Sigma_\mathsf{K}} \coloneqq \lambda_{\max}(\Sigma_\mathsf{K})$. This result characterizes the sensitivity of the RKHS element to perturbations in both the kernel centers and the Gram matrix structure.

Combining two items. Thus, for the case of RBF kernel K, there is

$$\begin{aligned}
\left\|\widetilde{h}(\widetilde{\mathbf{s}}) - h(\mathbf{s})\right\|_2 \leq & M_{\Sigma_\mathsf{K}}\left(M_C + \frac{M_{\Sigma_\mathsf{K}} M_\kappa M_c q}{1-(q-1)\rho_\mathsf{A}}\left[\frac{M_\kappa}{1-(q-1)\rho_\mathsf{A}} + 1\right]\right) \\
& \times \left(2e^{-\frac{1}{2}M_S^2}\left(\varepsilon M_S + \varepsilon^2 M_\phi\right) + \frac{1}{2}e^{-\frac{1}{2}M_S^2}\sum_{i=1}^{d} M_S^2 \left\|\phi_i - \widetilde{\phi}_i\right\|_2\right).
\end{aligned} \tag{44}$$

According to Lemma C.2, we obtain the error of mapping function for general kernel k as

$$\left\|\widetilde{h}^{(\text{off})}(\widetilde{\mathbf{s}}) - h^{(\text{off})}(\mathbf{s})\right\|_2 \leq C_0\left(2\left(\varepsilon M_S + \varepsilon^2\right) + M_S^2 d\sqrt{\delta_\mathsf{V}^2 M^d + M_V \delta_\psi M^d}\right) \tag{45a}$$

$$\left\|\widetilde{h}^{(\text{on})}(\widetilde{\mathbf{s}}) - h^{(\text{on})}(\mathbf{s})\right\|_2 \leq C_0\left(2\left(\varepsilon M_S + \varepsilon^2\right) + M_S^2 d\sqrt{\delta_\mathsf{V}^2 M^d + 2M_V M_\Gamma \delta_\psi M^d}\right), \tag{45b}$$

where

$$C_0 = M_{\Sigma_\mathsf{K}} e^{-\frac{1}{2}M_S^2} \left(M_C + \frac{M_{\Sigma_\mathsf{K}} M_\kappa M_c q}{1-(q-1)\rho_\mathsf{A}} \left[\frac{M_\kappa}{1-(q-1)\rho_\mathsf{A}} + 1 \right] \right).$$

For the case of RBF kernel k, there is

$$\left\| \widetilde{h}^{(\mathrm{off})}(\widetilde{\mathbf{s}}) - h^{(\mathrm{off})}(\mathbf{s}) \right\|_2 \leq C_0 \left(2\left(\varepsilon M_S + \varepsilon^2 M_\phi\right) + M_S^2 d \sqrt{C_2 \varepsilon^2 + M_V \left(1 - \exp\left(-\frac{\varepsilon^2}{2l^2}\right)\right)} \right) \tag{46a}$$

$$\left\| \widetilde{h}^{(\mathrm{on})}(\widetilde{\mathbf{s}}) - h^{(\mathrm{on})}(\mathbf{s}) \right\|_2 \leq C_0 \left(2\left(\varepsilon M_S + \varepsilon^2 M_\phi\right) + M_S^2 d \sqrt{C_2 \varepsilon^2 + 2 M_V M_\Gamma \left(1 - \exp\left(-\frac{\varepsilon^2}{2l^2}\right)\right)} \right), \tag{46b}$$

where

$$C_2 = \frac{M^d M_w^2 L_k^2 q^2}{\left(1-(q-1)\rho_\mathsf{V}\right)^2} \left(\frac{1}{1-(q-1)\rho_\mathsf{V}} + 1 \right)^2.$$

Now, we are ready to establish Theorem 4.8. Recall Eq. (25), and we are ready to complete the proof. Let $J(\pi) = \mathbb{E}_{s_0 \sim \rho}\left[V^\pi(s_0)\right]$. Then, by Pinsker's inequality,

$$\mathbb{E}\left[J(\pi) - J(\tilde{\pi})\right] \leq \max_{\mathbf{s} \in \mathcal{S}} \left\{ V^\pi(\mathbf{s}) - V^{\widetilde{\pi}}(\mathbf{s}) \right\} \leq \alpha \max_s \max_{\tilde{s} \in \mathcal{B}(s)} D_{\mathrm{TV}}\left(\pi(\cdot \mid s), \tilde{\pi}(\cdot \mid \tilde{s})\right). \tag{47}$$

We complete the proof of Theorem 4.8.

C.4 Proof of Theorem 4.9

Proof sketch of Theorem 4.9. (i) *One–step improvement.* Using the compatible feature representation $A^{\pi_{h_t}} = \langle w_{\mathsf{A},t}, g(h_t) \rangle_{\mathcal{H}_K}$ and the performance–difference lemma, we study the functional $D(h) = \mathbb{E}_{\nu_{\pi^\star}}[\log \pi_h]$. By $L_{\log \pi}$-smoothness, the update h_{t+1} (unprojected step plus projection) yields a descent-type inequality for $D(h_t) - D(h_{t+1})$ that isolates: a smoothness penalty from $\|h_{t+1} - h_t\|$, a projection/sparsification error, a kernel–drift term from changing features, and stochastic terms coming from the score estimate $\hat{g}(h_t)$. (ii) *Bounding residuals.* Each remainder is controlled by standard ingredients: (a) critic tracking and the Lipschitz dependence of $w_\mathsf{V}^\star(h)$ bound the kernel–drift; (b) sparsification guarantees bound $\|h_{t+1} - \overline{h}_{t+1}\|$; (c) second–moment bounds for $\hat{g}(h_t)$ control variance; and (d) Young/Cauchy–Schwarz inequalities handle the bias and error–variance coupling. Altogether the per–step remainder scales like $(\alpha_t^\mathsf{h})^2 + \rho_t + q^2 \varepsilon_{\mathsf{PA}}^2 + n_{\mathrm{eff}}^{-1}$, with ρ_t capturing the critic–actor time–scale gap. (iii) *Telescoping and rates.* Summing the one–step inequality over t, choosing polynomial step sizes $\alpha_t^\mathsf{h} = \alpha_0 (t+1)^{-\sigma}$ and $\beta_t = \beta_0 (t+1)^{-\nu}$ with $0 < \nu < \sigma$, and letting the implementation noise terms decay as $\Theta((t+1)^{-2\sigma})$, the residual sums are of order $\sum_t (\alpha_t^\mathsf{h})^2$. Selecting the critical coupling $\nu = \frac{2}{3}\sigma$ yields the stated averaged suboptimality rates $\mathcal{O}\left((1-\gamma)^{-1} T^{-(1-\sigma)}\right)$, with the usual logarithmic refinements at the regime boundaries.

Step 1: Decomposing the one-step improvement According to the definition of Actor in Eq (4) and proof of Appendix B.1, we use the score feature in the RKHS $\mathcal{H}_K$ as

$$g(h)(\mathbf{s}, \mathbf{a}) \coloneqq \nabla_h \log \pi_h(\mathbf{a} \mid \mathbf{s}) = K(\mathbf{s}, \cdot) \Sigma^{-1} \left(\mathbf{a} - h(\mathbf{s})\right) \in \mathcal{H}_K.$$

With compatible approximation proved in Appendix B.2, we posit

$$A^{\pi_{h_t}}(\mathbf{s}, \mathbf{a}) = \left\langle w_{\mathsf{A},t}, g(h_t)(\mathbf{s}, \mathbf{a}) \right\rangle_{\mathcal{H}_K}, \qquad w_{\mathsf{A},t} \in \mathcal{H}_K. \tag{48}$$

By the performance-difference lemma introduced in Lemma C.10, for any two policies,

$$J(\pi^\star) - J(\pi_{h_t}) = \frac{1}{1-\gamma} \mathbb{E}_{(\mathbf{s},\mathbf{a}) \sim \nu_{\pi^\star}} \left[A^{\pi_{h_t}}(\mathbf{s}, \mathbf{a}) \right].$$

Substituting the compatible form of Eq. (48) yields

$$J(\pi^\star) - J(\pi_{h_t}) = \frac{1}{1-\gamma} \mathbb{E}_{\nu_{\pi^\star}} \left[\left\langle w_{\mathsf{A},t}, g(h_t) \right\rangle_{\mathcal{H}_K} \right] + \frac{1}{1-\gamma} \zeta_{\mathrm{approx}}. \tag{49}$$

Define the divergence functional as

$$D(h) \coloneqq \mathbb{E}_{\nu_{\pi^\star}}\left[\log \pi_h(\mathbf{a} \mid \mathbf{s})\right].$$

Recall $\log \pi_h(\mathbf{a} \mid \mathbf{s})$ is L_ψ-smooth in h in Lemma C.7 and $\nabla_h \log \pi_{h_t}(\mathbf{a} \mid \mathbf{s}) = g(h_t)$ in Proposition B.2. Then for any h_{t+1}, there is

$$\begin{aligned}
&D(h_t) - D(h_{t+1}) \\
=&\mathbb{E}_{\nu_{\pi^\star}}\Big[\log \pi_{h_t}(\mathbf{a} \mid \mathbf{s}) - \log \pi_{h_{t+1}}(\mathbf{a} \mid \mathbf{s})\Big] \\
\geq& \left\langle \mathbb{E}_{\nu_{\pi^\star}}\left[\nabla_h \log \pi_{h_t}(\mathbf{a} \mid \mathbf{s})\right], h_{t+1} - h_t \right\rangle_{\mathcal{H}_K} - \frac{L_{\text{grad}}}{2}\|h_{t+1} - h_t\|^2_{\mathcal{H}_K} - 2M_a^2 M_\Sigma^2 \delta_{\mathsf{K},h_t} \\
=& \left\langle \mathbb{E}_{\nu_{\pi^\star}}\left[g(h_t)\right], h_{t+1} - h_t \right\rangle_{\mathcal{H}_K} - \frac{L_{\text{grad}}}{2}\|h_{t+1} - h_t\|^2_{\mathcal{H}_K} - 2M_a^2 M_\Sigma^2 \delta_{\mathsf{K},h_t}.
\end{aligned} \tag{50}$$

The additional term δ_{K,h_t} above accounts for the kernel drift/mismatch incurred by Critic-induced feature changes (cf. Lemma C.6); M_A and M_Σ are uniform bounds on the compatible weight and $\|\Sigma^{-1}\|_2$ that make the reduction explicit.

Let $\overline{h}_{t+1}$ denote the unprojected mirror step and h_{t+1} the *implemented* update after the sparsification process. We have the update step takes the form

$$\overline{h}_{t+1} - h_t = \alpha_t^{\mathsf{h}} \hat{g}(h_t), \qquad \text{with a score estimate } \hat{g}(h_t) \in \mathcal{H}_K, \tag{51}$$

and decompose

$$h_{t+1} - h_t = \left(h_{t+1} - \overline{h}_{t+1}\right) + \left(\overline{h}_{t+1} - h_t\right) = \left(h_{t+1} - \overline{h}_{t+1}\right) + \alpha_t^{\mathsf{h}} \hat{g}(h_t).$$

Plugging this decomposition into Eq. (50) gives

$$\begin{aligned}
D(h_t) - D(h_{t+1}) = &\left\langle \mathbb{E}_{\nu_{\pi^\star}}\left[g(h_t)\right], h_{t+1} - \overline{h}_{t+1} \right\rangle_{\mathcal{H}_K} + \alpha_t^{\mathsf{h}} \left\langle \mathbb{E}_{\nu_{\pi^\star}}\left[g(h_t)\right], \hat{g}(h_t) \right\rangle_{\mathcal{H}_K} \\
&- \frac{L_\psi}{2}\|h_{t+1} - h_t\|^2_{\mathcal{H}_K} - 2M_A^2 M_\Sigma^2 \delta^2_{\mathsf{K},h}.
\end{aligned} \tag{52}$$

Expand the correlation with the stochastic score:

$$\begin{aligned}
\left\langle \mathbb{E}_{\nu_{\pi^\star}}\left[g(h_t)\right], \hat{g}(h_t) \right\rangle_{\mathcal{H}_K} &= \left\langle \mathbb{E}_{\nu_{\pi^\star}}\left[g(h_t)\right], g(h_t) \right\rangle_{\mathcal{H}_K} + \left\langle \mathbb{E}_{\nu_{\pi^\star}}\left[g(h_t)\right], \hat{g}(h_t) - g(h_t) \right\rangle_{\mathcal{H}_K} \\
&= \|\mathbb{E}_{\nu_{\pi^\star}}\left[g(h_t)\right]\|^2_{\mathcal{H}_K} + \mathbb{E}_{\nu_{\pi^\star}}\left[\langle g(h_t), \hat{g}(h_t) - g(h_t)\rangle_{\mathcal{H}_K}\right],
\end{aligned} \tag{53}$$

where we used linearity of expectation and inner product.

By Eq. (49), we have

$$\mathbb{E}_{\nu_{\pi^\star}}\left[\left\langle w_{\mathsf{A},t}, g(h_t) \right\rangle_{\mathcal{H}_K}\right] = (1-\gamma)\left(J(\pi^\star) - J(\pi_{h_t})\right).$$

Add and subtract $w_{\mathsf{A},t}$ inside the inner product of Eq. (53) to get

$$\langle g, \hat{g} - g\rangle = \langle w_{\mathsf{A},t}, \hat{g}\rangle - \langle w_{\mathsf{A},t}, g\rangle + \langle g - w_{\mathsf{A},t}, \hat{g} - g\rangle.$$

Take $\mathbb{E}_{\nu_{\pi^\star}}[\cdot]$ on both sides, and use the compatibility identity to obtain

$$\begin{aligned}
\langle \mathbb{E}_{\nu_{\pi^\star}}[g], \hat{g}\rangle =& \mathbb{E}_{\nu_{\pi^\star}}\left[\|g\|^2\right] + \mathbb{E}_{\nu_{\pi^\star}}\left[\langle w_{\mathsf{A},t}, \hat{g}\rangle\right] - (1-\gamma)\left(J(\pi^\star) - J(\pi_{h_t})\right) \\
&+ \mathbb{E}_{\nu_{\pi^\star}}\left[\langle g - w_{\mathsf{A},t}, \hat{g} - g\rangle\right].
\end{aligned} \tag{54}$$

For brevity we dropped the explicit (h_t) argument and the subscript $\mathcal{H}_K$ on inner products.

Substituting Eq. (54) into Eq. (52) and rearranging to isolate the performance gap on the left yields

$$\begin{aligned}
\alpha_t^{\mathsf{h}}(1-\gamma)\left(J(\pi^\star) - J(\pi_{h_t})\right) \leq& D(h_t) - D(h_{t+1}) + 2M_A^2 M_\Sigma^2 \delta^2_{\mathsf{K},h_t} + \frac{L_\psi}{2}\|h_{t+1} - h_t\|^2_{\mathcal{H}_K} \\
&+ \alpha_t^{\mathsf{h}} \mathbb{E}_{\nu_{\pi^\star}}\left[\langle w_{\mathsf{A},t}, \hat{g}(h_t)\rangle\right] - \left\langle \mathbb{E}_{\nu_{\pi^\star}}[g(h_t)], h_{t+1} - \overline{h}_{t+1} \right\rangle + \alpha_t^{\mathsf{h}} \zeta_{\text{approx}} \\
&- \alpha_t^{\mathsf{h}} \|\mathbb{E}_{\nu_{\pi^\star}}\left[g(h_t)\right]\|^2_{\mathcal{H}_K} - \alpha_t^{\mathsf{h}} \mathbb{E}_{\nu_{\pi^\star}}\left[\langle g(h_t) - w_{\mathsf{A},t}, g(h_t) - \hat{g}(h_t)\rangle\right].
\end{aligned} \tag{55}$$

Consequently, Eq. (55) is the desired one–step decomposition: it expresses the scaled performance gap $\alpha_t^{\mathsf{h}}(1-\gamma)\left(J(\pi^\star) - J(\pi_{h_t})\right)$ in terms of (i) the descent of $D(h_t)$, (ii) a kernel-drift term $\delta^2_{\mathsf{K},h+t}$, (iii) a smoothness penalty, (iv) a bias term involving $\mathbb{E}\langle w_{\mathsf{A},t}, \hat{g}(h_t)\rangle$, (v) the deviation between the ideal and implemented Actor steps $\langle \mathbb{E}[g(h_t)], h_{t+1} - \overline{h}_{t+1}\rangle$, and (vi) two variance–like corrections coming from $\|g(h_t)\|^2$ and the error–variance coupling $\langle g(h_t) - w_{\mathsf{A},t}, g(h_t) - \hat{g}(h_t)\rangle$. These residuals will be upper bounded in Step 2.

Step 2: Bounding the residual terms. We proceed to upper bound each term in Eq. (55). Throughout, recall $g(h_t)(\mathbf{s},\mathbf{a}) = K(\mathbf{s},\cdot)\Sigma^{-1}(\mathbf{a} - h_t(\mathbf{s}))$ and define

$$\overline{K}_{\pi^\star} := \mathbb{E}_{\mathbf{s}\sim d^{\pi^\star}}[K(\mathbf{s},\mathbf{s})], \qquad M_K := \sup_{\mathbf{s}} K(\mathbf{s},\mathbf{s}), \qquad M_\Sigma := \|\Sigma^{-1}\|_2.$$

Bounding kernel drift δ_{K,h_t}. By Lemma C.6, there exists an absolute constant so that

$$\delta_{\mathsf{K},h_t} \le 16 M_{\Sigma_K}^2 M_S^4 q^2 M_a^2 C_\mu^2 \left\|w_{\mathsf{V},t+1} - w_{\mathsf{V},t}\right\|_{\mathcal{H}_k}^2. \tag{56}$$

Insert the intermediate optima $\{w^\star_{\mathsf{V},t}\}$ and apply the triangle inequality and parallelogram identity in $\mathcal{H}_k$:

$$\begin{aligned}\left\|w_{\mathsf{V},t+1} - w_{\mathsf{V},t}\right\|_{\mathcal{H}_k}^2 &= \left\|(w_{\mathsf{V},t+1} - w^\star_{\mathsf{V},t+1}) + (w^\star_{\mathsf{V},t+1} - w^\star_{\mathsf{V},t}) + (w^\star_{\mathsf{V},t} - w_{\mathsf{V},t})\right\|_{\mathcal{H}_k}^2 \\ &\le 3\left\|w_{\mathsf{V},t+1} - w^\star_{\mathsf{V},t+1}\right\|_{\mathcal{H}_k}^2 + 3\left\|w^\star_{\mathsf{V},t+1} - w^\star_{\mathsf{V},t}\right\|_{\mathcal{H}_k}^2 + 3\left\|w^\star_{\mathsf{V},t} - w_{\mathsf{V},t}\right\|_{\mathcal{H}_k}^2,\end{aligned} \tag{57}$$

where the first term corresponds the Critic error at $t+1$, the second term is shift of optimum, and the last term represents the Critic error at t. Taking expectations and invoking Lemma C.11 together with the sensitivity bound (161) for the shifted optimum yields, for constants $L_{\mathsf{V}}, G_h, C, C_{\mathrm{crit}}, M_k, C_b$,

$$\mathbb{E}\left[\left\|w_{\mathsf{V},t+1} - w_{\mathsf{V},t}\right\|_{\mathcal{H}_\psi}^2\right] \le \begin{cases} 3L_{\mathsf{V}}^2 G_h^2 (\alpha_t^{\mathsf{h}})^2 + \dfrac{6C}{(t+1)^\nu} + 6M_k q^2 \varepsilon_{\mathsf{PV}}^2 + 6\dfrac{C_b}{n_{\mathrm{eff}}}, & \sigma > \frac{3}{2}\nu, \\ 3L_{\mathsf{V}}^2 G_h^2 (\alpha_t^{\mathsf{h}})^2 + \dfrac{6C_{\mathrm{crit}}\log^2(t+1)}{(t+1)^\nu} + 6M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \dfrac{6C_b}{n_{\mathrm{eff}}}, & \sigma = \frac{3}{2}\nu, \\ 3L_{\mathsf{V}}^2 G_h^2 (\alpha_t^{\mathsf{h}})^2 + \dfrac{6C}{(t+1)^{2(\sigma-\nu)}} + 6M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \dfrac{6C_b}{n_{\mathrm{eff}}}, & \nu < \sigma < \frac{3}{2}\nu. \end{cases} \tag{58}$$

Combining Eq. (56)–Eq. (58) and letting $C_{\mathsf{K}} := 48 M_\Sigma^2 M_S^4 q^2 M_A^2 C_\mu^2$, we obtain

$$\mathbb{E}\left[\delta_{\mathsf{K},h_t}\right] \le C_{\mathsf{K}} \times \begin{cases} L_{\mathsf{V}}^2 G_h^2 (\alpha_t^{\mathsf{h}})^2 + \dfrac{2C}{(t+1)^\nu} + 2M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \dfrac{2C_b}{n_{\mathrm{eff}}}, & \sigma > \frac{3}{2}\nu, \\ L_{\mathsf{V}}^2 G_h^2 (\alpha_t^{\mathsf{h}})^2 + \dfrac{2C_{\mathrm{crit}}\log^2(t+1)}{(t+1)^\nu} + 2M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \dfrac{2C_b}{n_{\mathrm{eff}}}, & \sigma = \frac{3}{2}\nu, \\ L_{\mathsf{V}}^2 G_h^2 (\alpha_t^{\mathsf{h}})^2 + \dfrac{2C}{(t+1)^{2(\sigma-\nu)}} + 2M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \dfrac{2C_b}{n_{\mathrm{eff}}}, & \nu < \sigma < \frac{3}{2}\nu. \end{cases} \tag{59}$$

Bounding smoothness penalty $\frac{L_{\mathrm{grad}}}{2}\|h_{t+1} - h_t\|_{\mathcal{H}_K}^2$. By $h_{t+1} - h_t = (h_{t+1} - \overline{h}_{t+1}) + \alpha_t^{\mathsf{h}}\hat{g}(h_t)$ and $\|u+v\|^2 \le 2\|u\|^2 + 2\|v\|^2$,

$$\frac{L_{\mathrm{grad}}}{2}\|h_{t+1} - h_t\|_{\mathcal{H}_K}^2 \le L_{\mathrm{grad}}\|h_{t+1} - \overline{h}_{t+1}\|_{\mathcal{H}_K}^2 + L_{\mathrm{grad}}\left(\alpha_t^{\mathsf{h}}\right)^2 \|\hat{g}(h_t)\|_{\mathcal{H}_K}^2. \tag{60}$$

For the second factor, write $\hat{g}(h_t) = g(h_t) + (\hat{g}(h_t) - g(h_t))$ and use $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$:

$$\mathbb{E}_{\nu_{\pi^\star}}\left[\|\hat{g}(h_t)\|_{\mathcal{H}_K}^2\right] \le 2\mathbb{E}_{\nu_{\pi^\star}}\left[\|g(h_t)\|_{\mathcal{H}_K}^2\right] + 2\mathbb{E}_{\nu_{\pi^\star}}\left[\|g(h_t) - \hat{g}(h_t)\|_{\mathcal{H}_K}^2\right]. \tag{61}$$

By direct calculation, there is $\mathbb{E}_{\nu_{\pi^\star}}[\|g(h_t)\|_{\mathcal{H}_K}^2] = \overline{K}_{\pi^\star}\mathrm{tr}(\Sigma^{-1})$. We assume the score estimator has bounded second moment as

$$\mathbb{E}_{\nu_{\pi^\star}}\left[\|g(h_t) - \hat{g}(h_t)\|_{\mathcal{H}_K}^2\right] \le \frac{C_b}{n_{\mathrm{eff}}}. \tag{62}$$

Taking expectations in Eq. (60) and applying Eq. (61) and (62) gives

$$\mathbb{E}\left[\frac{L_{\mathrm{grad}}}{2}\|h_{t+1} - h_t\|_{\mathcal{H}_K}^2\right] \le L_{\mathrm{grad}}\mathbb{E}\left[\|h_{t+1} - \overline{h}_{t+1}\|_{\mathcal{H}_K}^2\right] + 2L_{\mathrm{grad}}\left(\alpha_t^{\mathsf{h}}\right)^2\left(\overline{K}_{\pi^\star}\mathrm{tr}(\Sigma^{-1}) + \frac{C_b}{n_{\mathrm{eff}}}\right). \tag{63}$$

If the sparsification with budget q and precision $\varepsilon_{\mathsf{PA}}$ ensures

$$\mathbb{E}\left[\|h_{t+1} - \overline{h}_{t+1}\|_{\mathcal{H}_K}^2\right] \le M_K q^2 \varepsilon_{\mathsf{PA}}^2 + \frac{C_b}{n_{\mathrm{eff}}}, \tag{64}$$

then

$$\mathbb{E}\left[\frac{L_{\mathrm{grad}}}{2}\|h_{t+1} - h_t\|_{\mathcal{H}_K}^2\right] \le L_{\mathrm{grad}}\left(M_K q^2 \varepsilon_{\mathsf{PA}}^2 + \frac{C_b}{n_{\mathrm{eff}}}\right) + 2L_{\mathrm{grad}}\left(\alpha_t^{\mathsf{h}}\right)^2\left(\overline{K}_{\pi^\star}\mathrm{tr}(\Sigma^{-1}) + \frac{C_b}{n_{\mathrm{eff}}}\right). \tag{65}$$

Bounding gradient energy $\|g(h_t)\|^2_{\mathcal{H}_K}$. By the reproducing property,

$$\|g(h_t)\|^2_{\mathcal{H}_K} = \left\langle K(\mathbf{s},\cdot)\Sigma^{-1}\left(\mathbf{a}-h_t(\mathbf{s})\right), K(\mathbf{s},\cdot)\Sigma^{-1}\left(\mathbf{a}-h_t(\mathbf{s})\right)\right\rangle_{\mathcal{H}_K}$$
$$= K(\mathbf{s},\mathbf{s})\left\|\Sigma^{-1}\left(\mathbf{a}-h_t(\mathbf{s})\right)\right\|_2^2. \tag{66}$$

Pointwise, for $\|\mathbf{a}\|_2 \le M_a$ and $\|h_t(\mathbf{s})\|_2 \le C_h$ a.s.,

$$\|g(h_t)\|^2_{\mathcal{H}_K} \le M_K M_\Sigma^2 (M_a + C_h)^2. \tag{67}$$

Conditioned on $\mathbf{s}$ with $\mathbf{a} \mid \mathbf{s} \sim \mathcal{N}(h_t(\mathbf{s}), \Sigma)$,

$$\mathbb{E}\left[\|g(h_t)\|^2_{\mathcal{H}_K} \mid \mathbf{s}\right] = K(\mathbf{s},\mathbf{s})\mathrm{tr}(\Sigma^{-1}), \qquad \Rightarrow \quad \mathbb{E}_{\nu_{\pi^\star}}\left[\|g(h_t)\|^2_{\mathcal{H}_K}\right] = \overline{K}_{\pi^\star}\mathrm{tr}(\Sigma^{-1}). \tag{68}$$

If $m_K \coloneqq \mathrm{essinf}_{\mathbf{s}} K(\mathbf{s},\mathbf{s}) > 0$, then the negative term in Eq. (55) admits the uniform lower bound

$$-\alpha_t^{\mathsf{h}}\mathbb{E}_{\nu_{\pi^\star}}\left[\|g(h_t)\|^2_{\mathcal{H}_K}\right] \le -\alpha_t^{\mathsf{h}} m_K \mathrm{tr}(\Sigma^{-1}) \le -\alpha_t^{\mathsf{h}} m_K \frac{d}{\lambda_{\max}(\Sigma)}. \tag{69}$$

Bounding bias term $\mathbb{E}_{\nu_{\pi^\star}}\langle w_{\mathsf{A},t}, \hat{g}(h_t)\rangle_{\mathcal{H}_K}$. Using $\hat{g}(h_t) = g(h_t) + (\hat{g}(h_t) - g(h_t))$ and the compatibility identity $\mathbb{E}_{\nu_{\pi^\star}}\langle w_{\mathsf{A},t}, g(h_t)\rangle = (1-\gamma)\left(J(\pi^\star) - J(\pi_{h_t})\right)$,

$$\mathbb{E}_{\nu_{\pi^\star}}\left[\langle w_{\mathsf{A},t}, \hat{g}(h_t)\rangle\right] = (1-\gamma)\left(J(\pi^\star) - J(\pi_{h_t})\right) + \mathbb{E}_{\nu_{\pi^\star}}\left[\langle w_{\mathsf{A},t}, \hat{g}(h_t) - g(h_t)\rangle\right]. \tag{70}$$

By Cauchy–Schwarz inequality and assuming $\|w_{\mathsf{A},t}\|_{\mathcal{H}_K} \le M_{Adv}$, armed with Eq. (62), there is

$$\mathbb{E}_{\nu_{\pi^\star}}\left[\langle w_{\mathsf{A},t}, \hat{g}(h_t)\rangle\right] \le (1-\gamma)\left(J(\pi^\star) - J(\pi_{h_t})\right) + M_{Adv}\sqrt{\frac{C_b}{n_{\mathrm{eff}}}}$$
$$\le (1-\gamma)\left(J(\pi^\star) - J(\pi_{h_t})\right) + \frac{\eta}{2}M_{Adv}^2 + \frac{1}{2\eta}\frac{C_b}{n_{\mathrm{eff}}}, \quad \forall \eta > 0. \tag{71}$$

Bounding deviation term. Let $D_t \coloneqq \langle \mathbb{E}_{\nu_{\pi^\star}}[g(h_t)], h_{t+1} - \overline{h}_{t+1}\rangle_{\mathcal{H}_K}$. By Cauchy–Schwarz and Jensen inequalitis, we derive

$$\mathbb{E}[D_t] \ge -\left\|\mathbb{E}_{\nu_{\pi^\star}}[g(h_t)]\right\|_{\mathcal{H}_K}\sqrt{\mathbb{E}\left[\|h_{t+1} - \overline{h}_{t+1}\|^2_{\mathcal{H}_K}\right]}$$
$$\ge -\sqrt{\mathbb{E}_{\nu_{\pi^\star}}\|g(h_t)\|^2_{\mathcal{H}_K}}\sqrt{\mathbb{E}\left[\|h_{t+1} - \overline{h}_{t+1}\|^2_{\mathcal{H}_K}\right]}$$
$$= -\sqrt{\overline{K}_{\pi^\star}\mathrm{tr}(\Sigma^{-1})}\sqrt{\mathbb{E}\left[\|h_{t+1} - \overline{h}_{t+1}\|^2_{\mathcal{H}_K}\right]}. \tag{72}$$

Using Eq. (64), this becomes

$$\mathbb{E}[D_t] \ge -\sqrt{\overline{K}_{\pi^\star}\mathrm{tr}(\Sigma^{-1})}\sqrt{M_K q^2 \varepsilon_{\mathsf{P}V}^2 + \frac{C_b}{n_{\mathrm{eff}}}}. \tag{73}$$

Bounding error–variance coupling. Define $T_t \coloneqq \mathbb{E}_{\nu_{\pi^\star}}\left[\langle g(h_t) - w_{\mathsf{A},t}, g(h_t) - \hat{g}(h_t)\rangle_{\mathcal{H}_K}\right]$. For any $\eta > 0$, Young's inequality gives

$$T_t \ge -\frac{\eta}{2}\mathbb{E}_{\nu_{\pi^\star}}\left[\|g(h_t) - w_{\mathsf{A},t}\|^2_{\mathcal{H}_K}\right] - \frac{1}{2\eta}\mathbb{E}_{\nu_{\pi^\star}}\left[\|g(h_t) - \hat{g}(h_t)\|^2_{\mathcal{H}_K}\right]. \tag{74}$$

Expanding the first expectation and using $\mathbb{E}_{\nu_{\pi^\star}}\langle w_{\mathsf{A},t}, g(h_t)\rangle = (1-\gamma)\left(J(\pi^\star) - J(\pi_{h_t})\right) \ge 0$ and Eq. (68), we derive

$$\mathbb{E}_{\nu_{\pi^\star}}\left[\|g(h_t) - w_{\mathsf{A},t}\|^2_{\mathcal{H}_K}\right] = \mathbb{E}_{\nu_{\pi^\star}}\left[\|g(h_t)\|^2_{\mathcal{H}_K}\right] + \mathbb{E}\left[\|w_{\mathsf{A},t}\|^2_{\mathcal{H}_K}\right] - 2\mathbb{E}_{\nu_{\pi^\star}}\left[\langle w_{\mathsf{A},t}, g(h_t)\rangle\right]$$
$$\le \overline{K}_{\pi^\star}\mathrm{tr}(\Sigma^{-1}) + M_{Adv}^2. \tag{75}$$

Together with Eqs. (62) and (74) yields the tunable bound

$$T_t \ge -\frac{\eta}{2}\left(\overline{K}_{\pi^\star}\mathrm{tr}(\Sigma^{-1}) + M_{Adv}^2\right) - \frac{1}{2\eta}\frac{C_b}{n_{\mathrm{eff}}}, \qquad \forall \eta > 0, \tag{76}$$

whose optimal choice $\eta^\star = \sqrt{\dfrac{C_b}{n_{\mathrm{eff}}\left(\overline{K}_{\pi^\star}\mathrm{tr}(\Sigma^{-1}) + M_{Adv}^2\right)}}$ gives the compact form

$$T_t \ge -\sqrt{\left(\overline{K}_{\pi^\star}\mathrm{tr}(\Sigma^{-1}) + M_{Adv}^2\right)\frac{C_b}{n_{\mathrm{eff}}}}. \tag{77}$$

Step 3: Convergence rates under polynomial stepsizes. Starting from the one–step decomposition in Eq. (55) and the bounds assembled in Step 2, we now telescope in time and convert the per-iteration inequality into rates. Throughout this step we use the Young form of the bias bound Eq. (71), and we retain the good negative term $-\alpha_t^{\mathsf{h}}\mathbb{E}\left[\|g(h_t)\|_{\mathcal{H}_K}^2\right]$ to absorb the deviation inner product via a Young's inequality.

Collecting Eqs. (59), (65), (71), (73), and (77) into Eq. (55), and dropping the good negative term Eq. (69) for an upper bound, we obtain

$$\begin{aligned}
\alpha_t^{\mathsf{h}}(1-\gamma)\left(J(\pi^\star)-J(\pi_{h_t})\right) \le\ & (D(h_t)-D(h_{t+1})) + 2M_a^2 M_{\Sigma_K}^2 \mathbb{E}[\delta_{\mathsf{K},h}] \\
& + L_{\text{grad}}\left(M_K q^2 \varepsilon_{\mathsf{PV}}^2 + \frac{C_b}{n_{\text{eff}}}\right) \\
& + 2L_{\text{grad}}\left(\alpha_t^{\mathsf{h}}\right)^2\left(\overline{K}_{\pi^\star}\operatorname{tr}(\Sigma^{-1}) + \frac{C_b}{n_{\text{eff}}}\right) \\
& + \alpha_t^{\mathsf{h}}\left[(1-\gamma)\left(J(\pi^\star)-J(\pi_{h_t})\right) + M_{Adv}\sqrt{\tfrac{C_b}{n_{\text{eff}}}}\right] \\
& + \sqrt{\overline{K}_{\pi^\star}\operatorname{tr}(\Sigma^{-1})}\sqrt{M_K q^2 \varepsilon_{\mathsf{PV}}^2 + \tfrac{C_b}{n_{\text{eff}}}} \\
& + \alpha_t^{\mathsf{h}}\sqrt{\left(\overline{K}_{\pi^\star}\operatorname{tr}(\Sigma^{-1}) + M_{Adv}^2\right)\tfrac{C_b}{n_{\text{eff}}}} + \alpha_t^{\mathsf{h}}\zeta_{\text{approx}}. \qquad (78)
\end{aligned}$$

Inequality Eq. (78) will be telescoped and simplified under step-size choices in Step 3.

Summing Eq. (55) from $t=0$ to $T-1$ and applying the bounds Eqs. (59), (65), (71), (72), and (77), we obtain, for any $\eta_t > 0$ and any $\varepsilon_t \in (0,1)$,

$$\begin{aligned}
& (1-\gamma)\sum_{t=0}^{T-1}\alpha_t^{\mathsf{h}}\mathbb{E}\big[J(\pi^\star)-J(\pi_{h_t})\big] \\
\le\ & D(h_0) - D(h_T) + 2M_a^2 M_{\Sigma_K}^2 \sum_{t=0}^{T-1}\mathbb{E}\left[\delta_{\mathsf{K},h_t}\right] + L_{\text{grad}}\sum_{t=0}^{T-1}\mathbb{E}\left[\|h_{t+1}-\overline{h}_{t+1}\|_{\mathcal{H}_K}^2\right] \\
& + 2L_{\text{grad}}\sum_{t=0}^{T-1}\left(\alpha_t^{\mathsf{h}}\right)^2\left(\overline{K}_{\pi^\star}\operatorname{tr}(\Sigma^{-1}) + \frac{C_b}{n_{\text{eff},t}}\right) + \sum_{t=0}^{T-1}\alpha_t^{\mathsf{h}}\left(\frac{\eta_t}{2}M_{Adv}^2 + \frac{C_b}{2\eta_t n_{\text{eff},t}}\right) \\
& + \sum_{t=0}^{T-1}\left[\frac{\varepsilon_t}{2}\alpha_t^{\mathsf{h}}\mathbb{E}\left[\|g(h_t)\|_{\mathcal{H}_K}^2\right] + \frac{1}{2\varepsilon_t\alpha_t^{\mathsf{h}}}\mathbb{E}\left[\|h_{t+1}-\overline{h}_{t+1}\|_{\mathcal{H}_K}^2\right]\right] \\
& + \sum_{t=0}^{T-1}\alpha_t^{\mathsf{h}}\sqrt{\left(\overline{K}_{\pi^\star}\operatorname{tr}(\Sigma^{-1}) + M_{Adv}^2\right)\frac{C_b}{n_{\text{eff},t}}} + \sum_{t=0}^{T-1}\alpha_t^{\mathsf{h}}\zeta_{\text{approx}} \\
\le\ & B_D + 2M_A^2 M_{\Sigma_K}^2 \sum_{t=0}^{T-1}\mathbb{E}\left[\delta_{\mathsf{K},h_t}\right] + L_{\text{grad}}\sum_{t=0}^{T-1}\mathbb{E}\left[\|h_{t+1}-\overline{h}_{t+1}\|_{\mathcal{H}_K}^2\right] \\
& + 2L_{\text{grad}}\sum_{t=0}^{T-1}\left(\alpha_t^{\mathsf{h}}\right)^2\left(\overline{K}_{\pi^\star}\operatorname{tr}(\Sigma^{-1}) + \frac{C_b}{n_{\text{eff},t}}\right) + \sum_{t=0}^{T-1}\alpha_t^{\mathsf{h}}\left(\frac{\eta_t}{2}M_{Adv}^2 + \frac{C_b}{2\eta_t n_{\text{eff},t}}\right) \\
& + \sum_{t=0}^{T-1}\left[\frac{\varepsilon_t}{2}\alpha_t^{\mathsf{h}}\mathbb{E}\left[\|g(h_t)\|_{\mathcal{H}_K}^2\right] + \frac{1}{2\varepsilon_t\alpha_t^{\mathsf{h}}}\mathbb{E}\left[\|h_{t+1}-\overline{h}_{t+1}\|_{\mathcal{H}_K}^2\right]\right] \\
& + \sum_{t=0}^{T-1}\alpha_t^{\mathsf{h}}\sqrt{\left(\overline{K}_{\pi^\star}\operatorname{tr}(\Sigma^{-1}) + M_{Adv}^2\right)\frac{C_b}{n_{\text{eff},t}}} + \sum_{t=0}^{T-1}\alpha_t^{\mathsf{h}}\zeta_{\text{approx}}. \qquad (79)
\end{aligned}$$

The negative term $-\alpha_t^{\mathsf{h}}\mathbb{E}\left[\|g(h_t)\|_{\mathcal{H}_K}^2\right]$ in Eq. (55) can be dropped. With this trick, the deviation contribution scales as $\frac{1}{\alpha_t^{\mathsf{h}}}\mathbb{E}\|h_{t+1}-\overline{h}_{t+1}\|^2$.

To make all residuals summable, we adopt the standard, square-summable schedules:

$$\alpha_t^{\mathsf{h}} = \alpha_0 (t+1)^{-\sigma}, \tag{80a}$$

$$\beta_t = \beta_0 (t+1)^{-\nu}, \tag{80b}$$

$$\frac{C_b}{n_{\mathrm{eff},t}} = \Theta\left((t+1)^{-2\sigma}\right), \tag{80c}$$

$$\mathbb{E}\left[\|h_{t+1} - \overline{h}_{t+1}\|_{\mathcal{H}_K}^2\right] = \Theta\left((t+1)^{-2\sigma}\right) \tag{80d}$$

with $\sigma \in (0,1)$ and $\nu \in (0,1)$. The last two relations are achieved, e.g., by using mini-batch sizes and sparsification budgets that grow like $(t+1)^{2\sigma}$. Under Eq. (80), the two implementation-induced sums in Eq. (79) are $\mathcal{O}\left(\sum_t (\alpha_t^{\mathsf{h}})^2\right)$.

Step 2 gives a bound on $\mathbb{E}[\delta_{\mathsf{K},h_t}]$, and we have

$$\mathbb{E}\left[\delta_{\mathsf{K},h_t}\right] = \mathcal{O}\left(\left(L_{\mathsf{V}} G_h \alpha_t^{\mathsf{h}}\right)^2 + \rho_t + q_t^2 \varepsilon_{\mathsf{PA},t}^2 + \frac{1}{n_{\mathrm{eff},t}}\right), \tag{81}$$

where

$$\rho_t \coloneqq \begin{cases} (t+1)^{-\nu}, & \sigma > \frac{3}{2}\nu, \\ (t+1)^{-\nu}\log^2(t+1), & \sigma = \frac{3}{2}\nu, \\ (t+1)^{-2(\sigma-\nu)}, & \nu < \sigma < \frac{3}{2}\nu. \end{cases}$$

Consequently,

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\delta_{\mathsf{K},h_t}\right] = \mathcal{O}\left(\sum_{t=0}^{T-1} (\alpha_t^{\mathsf{h}})^2 + \sum_{t=0}^{T-1} \rho_t + \sum_{t=0}^{T-1} q_t^2 \varepsilon_{\mathsf{P},t}^2 + \sum_{t=0}^{T-1} \frac{1}{n_{\mathrm{eff},t}}\right). \tag{82}$$

Under Eq. (80) and with $q_t^2 \varepsilon_{\mathsf{PV},t}^2 = \Theta((t+1)^{-\sigma})$, both the third and fourth sums are $\mathcal{O}\left(\sum_t (\alpha_t^{\mathsf{h}})^2\right)$.

For $p \in (0,1)$, $\sum_{t=0}^{T-1} (t+1)^{-p} = \Theta\left(T^{1-p}\right)$; for $p = 1$, it is $\Theta(\log T)$; for $p > 1$, it is $\Theta(1)$. Therefore, we have

$$S_\alpha(T) \coloneqq \sum_{t=0}^{T-1} \alpha_t^{\mathsf{h}} = \Theta\left(T^{1-\sigma}\right), \qquad S_{\alpha^2}(T) \coloneqq \sum_{t=0}^{T-1} (\alpha_t^{\mathsf{h}})^2 = \begin{cases} \Theta(1), & \sigma > \frac{1}{2}, \\ \Theta(\log T), & \sigma = \frac{1}{2}, \\ \Theta\left(T^{1-2\sigma}\right), & 0 < \sigma < \frac{1}{2}. \end{cases}$$

Moreover, there is

$$\sum_{t=0}^{T-1} \rho_t = \begin{cases} \Theta\left(T^{1-\nu}\right), & \sigma > \frac{3}{2}\nu,\ \nu < 1, \\ \Theta(\log T), & \sigma > \frac{3}{2}\nu,\ \nu = 1, \\ \Theta(1), & \sigma > \frac{3}{2}\nu,\ \nu > 1, \\ \Theta\left(\log^2 T \cdot T^{1-\nu}\right), & \sigma = \frac{3}{2}\nu,\ \nu < 1, \\ \Theta\left(T^{1-2(\sigma-\nu)}\right), & \nu < \sigma < \frac{3}{2}\nu,\ 2(\sigma-\nu) < 1, \\ \Theta(\log T), & \nu < \sigma < \frac{3}{2}\nu,\ 2(\sigma-\nu) = 1, \\ \Theta(1), & \nu < \sigma < \frac{3}{2}\nu,\ 2(\sigma-\nu) > 1. \end{cases}$$

Define a random index $\tilde{t} \in \{0, \ldots, T-1\}$ drawn with $\Pr(\tilde{t} = t) = \alpha_t^{\mathsf{h}} / S_\alpha(T)$. Then

$$\mathbb{E}\big[J(\pi^\star) - J(\pi_{h_{\tilde{t}}})\big] = \frac{\sum_{t=0}^{T-1} \alpha_t^{\mathsf{h}} \mathbb{E}\big[J(\pi^\star) - J(\pi_{h_t})\big]}{S_\alpha(T)}.$$

Dividing Eq. (79) by $(1-\gamma)S_\alpha(T)$ and using the estimates from results above yields

$$\begin{aligned} &\mathbb{E}\big[J(\pi^\star) - J(\pi_{h_{\tilde{t}}})\big] \\ \leq &\frac{B_D}{(1-\gamma)S_\alpha(T)} + \frac{C_1}{1-\gamma} \cdot \frac{S_{\alpha^2}(T)}{S_\alpha(T)} + \frac{C_2}{1-\gamma} \cdot \frac{\sum_{t=0}^{T-1} \rho_t}{S_\alpha(T)} + \frac{C_3}{1-\gamma} \cdot \frac{S_{\alpha^2}(T)}{S_\alpha(T)} + \frac{\zeta_{\mathrm{approx}}}{1-\gamma} \\ &+ \frac{C_4}{1-\gamma} \cdot \frac{\sum_{t=0}^{T-1} (\alpha_t^{\mathsf{h}}) \left(\frac{\eta_t}{2} M_{Adv}^2 + \frac{C_b}{2\eta_t n_{\mathrm{eff},t}} + \sqrt{(\overline{K}_{\pi^\star} \mathrm{tr}(\Sigma^{-1}) + M_{Adv}^2) \frac{C_b}{n_{\mathrm{eff},t}}}\right)}{S_\alpha(T)}. \end{aligned} \tag{83}$$

Here $C_1, \ldots, C_4$ absorb $M_K, M_\Sigma, M_{Adv}, L_{\text{grad}}, L_{\mathsf{V}}, G_h$ and the constants from Step 2. Choosing, e.g., $\eta_t \equiv \eta = \Theta(1)$ and the schedule introduced Eq. (80) so that $n_{\text{eff},t}^{-1} = \Theta((t+1)^{-2\sigma})$, the last fraction is $\mathcal{O}\left(S_{\alpha^2}(T)/S_\alpha(T)\right)$.

Following the two–timescale design, pick

$$\nu = \tfrac{2}{3}\sigma \qquad (\text{i.e., } \sigma = \tfrac{3}{2}\nu),$$

so that the Critic is sufficiently fast to track the Actor. In this case (the Critical regime) we have $\rho_t = (t+1)^{-\nu}\log^2(t+1)$ and $\sum_{t=0}^{T-1}\rho_t = \tilde{\Theta}\left(T^{1-\nu}\right)$ (where $\tilde{\Theta}$ hides polylog factors).

Plugging these sums into Eq. (83) gives the final rates. Collecting the dominant terms and keeping the three classical regimes of σ yields

$$\mathbb{E}\big[J(\pi^\star) - J(\pi_{h_{\tilde{t}}})\big] \le \frac{\zeta_{\text{approx}}}{1-\gamma} + \begin{cases} \mathcal{O}\left(\dfrac{1}{(1-\gamma)}\cdot\dfrac{1}{T^{1-\sigma}}\right), & \sigma > \frac{3}{4}, \\ \mathcal{O}\left(\dfrac{1}{(1-\gamma)}\cdot\dfrac{\log^2 T}{T^{1/2}}\right), & \sigma = \frac{3}{4}\ (\Rightarrow \nu = \frac{1}{2}), \\ \mathcal{O}\left(\dfrac{1}{(1-\gamma)}\cdot\dfrac{\log T}{T^{1-\frac{2}{3}\sigma}}\right), & 0 < \sigma < \frac{3}{4}. \end{cases} \tag{84}$$

If the sparsification is exact ($h_{t+1} \equiv \overline{h}_{t+1}$) and fresh data are used so that $n_{\text{eff},t}^{-1} = \mathcal{O}\left((t+1)^{-2\sigma}\right)$, then all implementation-driven terms fall into $S_{\alpha^2}(T)/S_\alpha(T)$ and are strictly dominated by the leading rates in Eq. (84). With fixed budgets (constant $q_t, \varepsilon_{\mathsf{PV},t}$ or $n_{\text{eff},t}$), the corresponding residuals create a floor of order $\mathcal{O}\left((1-\gamma)^{-1}T^{-\sigma}\right)$ or constant; the schedule in Eq. (80) (or taking the ideal update) removes this floor and recovers Eq. (84).

C.5 Proof of supporting facts

C.5.1 Proof of Proposition 4.3

Recall the policy is defined as a Gaussian distribution. Then the functional gradient of the log-policy w.r.t.h is given by

$$\nabla_h \log \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s}) = K(\mathbf{s}, \cdot)\Sigma^{-1}(\mathbf{a} - h(\mathbf{s})) \in \mathcal{H}_K, \tag{85}$$

under Proposition B.2. Therefore, the corresponding Fisher information operator is

$$\begin{aligned} F(h) &= \mathbb{E}_{\mathbf{s}\sim\nu_{\pi_h}, a\sim\pi_{h,\Sigma}(\cdot|\mathbf{s})}\left[\nabla_h \log \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s}) \otimes \nabla_h \log \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s})\right] \\ &= \mathbb{E}_{\mathbf{s}\sim\nu_{\pi_h}}\left[K(\mathbf{s},\cdot) \otimes K(\mathbf{s},\cdot)\cdot\Sigma^{-1}\right], \end{aligned} \tag{86}$$

where the last equation comes from $\mathbb{E}\left[(\mathbf{a} - h(\mathbf{s}))(\mathbf{a} - h(\mathbf{s}))^\top\right] = \Sigma$.

We assume a coercivity condition that there exists a constant $\lambda_F > 0$ such that

$$\langle f, F(h) f\rangle_{\mathcal{H}_K} \ge \lambda_F \cdot \|f\|_{\mathcal{H}_K}^2, \quad \forall f \in \mathcal{H}_K. \tag{87}$$

This ensures that the Fisher operator $F(h)$ is positive definite on $\mathcal{H}_K$, and enables natural gradient updates in RKHS.

C.5.2 Proof of Proposition B.2

According to the definition of f, we have

$$\begin{aligned} f(h) &= \log \pi_{h,\Sigma}(\mathbf{a} \mid \mathbf{s}) \\ &= -\log((2\pi)^{\frac{m}{2}}(\det(\Sigma))^{\frac{1}{2}}) - \frac{1}{2}\left(\mathbf{a} - h(\mathbf{s})\right)^\top \Sigma^{-1}\left(\mathbf{a} - h(\mathbf{s})\right); \\ f(h+g) &= \log \pi_{h+g,\Sigma}(\mathbf{a} \mid \mathbf{s}) \\ &= -\log((2\pi)^{\frac{m}{2}}(\det(\Sigma))^{\frac{1}{2}}) - \frac{1}{2}\left(\mathbf{a} - h(\mathbf{s}) - g(\mathbf{s})\right)^\top \Sigma^{-1}\left(\mathbf{a} - h(\mathbf{s}) - g(\mathbf{s})\right). \end{aligned}$$

Therefore, we extend the Banach spaces $\mathcal{V}$ and $\mathcal{W}$ to Hilbert spaces to obtain

$$\begin{aligned}\lim_{g\to 0}\frac{\|f(h+g)-f(h)-Df|_h(g)\|_2}{\|g\|_{\mathcal{H}_K}} &= \lim_{g\to 0}\frac{\left\|g(\mathbf{s})^\top\Sigma^{-1}g(\mathbf{s})\right\|_2}{\|g\|_{\mathcal{H}_K}}\\ &= \lim_{g\to 0}\frac{\langle g, K(\mathbf{s},\cdot)\Sigma^{-1}g(\mathbf{s})\rangle_{\mathcal{H}_K}}{\|g\|_{\mathcal{H}_K}}\\ &\le \lim_{g\to 0}\frac{\|g\|_{\mathcal{H}_K}\left(\Sigma^{-1}g(\mathbf{s})\right)^\top K(\mathbf{s},\cdot)\Sigma^{-1}g(\mathbf{s})}{\|g\|_{\mathcal{H}_K}}\\ &= \lim_{g\to 0}\left(\Sigma^{-1}g(\mathbf{s})\right)^\top K(\mathbf{s},\cdot)\Sigma^{-1}g(\mathbf{s}) \to 0,\end{aligned}$$

where the inequality comes from the Cauchy-Schwarz inequality.

C.5.3 Proof of Lemma C.2

Since the Shapley functional is a linear combination of bounded linear functionals (value functionals), it admits a Riesz representer in the RKHS. Therefore, given a value functional v indexed by input $\mathbf{s}$ and coalition $\mathcal{C}$, the Shapley functional $\phi_{\mathbf{s},i}:\mathcal{H}_k\to\mathbb{R}$ such that $\phi_{\mathbf{s},i}(v)$ is the ith Shapley values on input $\mathbf{s}$, can be written as the following linear combination of value functionals

$$\phi_{\mathbf{s},i} = \frac{1}{d}\sum_{\mathcal{C}\subseteq\mathcal{X}\setminus\{i\}}\binom{d-1}{|C|}^{-1}\left(v_{\mathcal{C}\cup\{i\}}(\mathbf{s})-v_{\mathcal{C}}(\mathbf{s})\right). \tag{88}$$

To prove that Shapley functionals between two observations $\mathbf{s}$ and $\widetilde{\mathbf{s}}$ are δ close when the two points are close, we proceed as the following three parts.

Part 1: Bound the feature maps We show the results of special case as the usual product RBF kernel, leading to bound the distance of the feature maps as a function of δ.

When we pick $\mathbf{s},\widetilde{\mathbf{s}}\in\mathbb{R}^d$ and with a product RBF kernel i.e., $k(\mathbf{s},\widetilde{\mathbf{s}})=\prod_{j=1}^d k^{(j)}(\mathbf{s}_j,\widetilde{\mathbf{s}}_j)$, where $k^{(j)}$ are RBF kernels. For simplicity, we assume they all share the same lengthscale l. Since

$$\|\psi(\mathbf{s})-\psi(\widetilde{\mathbf{s}})\|_{\mathcal{H}_k}^2 = k(\mathbf{s},\mathbf{s})+k(\widetilde{\mathbf{s}},\widetilde{\mathbf{s}})-2k(\mathbf{s},\widetilde{\mathbf{s}}),$$

the first two terms ($k(\mathbf{s},\mathbf{s})$ and $k(\widetilde{\mathbf{s}},\widetilde{\mathbf{s}})$) are 1 and we can bound the last term according to the basic form of RBF kernel as follows.

$$k(\mathbf{s},\widetilde{\mathbf{s}}) = \exp\left(-\frac{\sum_{j=1}^d|\mathbf{s}_j-\widetilde{\mathbf{s}}_j|^2}{2l^2}\right) = \exp\left(-\frac{\|\mathbf{s}-\widetilde{\mathbf{s}}\|_2^2}{2l^2}\right)\ge\exp\left(-\frac{\varepsilon^2}{2l^2}\right), \tag{89}$$

where $\varepsilon=\|\mathbf{s}-\tilde{\mathbf{s}}\|_2$ and $k(\mathbf{s},\mathbf{s})=1$ for RBF kernel. Then we can bound the difference in feature maps as follows,

$$\|\psi(\mathbf{s})-\psi(\widetilde{\mathbf{s}})\|_{\mathcal{H}_k}^2 \le 2-2\exp\left(-\frac{\varepsilon^2}{2l^2}\right). \tag{90}$$

Therefore, the distance in feature maps $\|\psi(\mathbf{s})-\psi(\widetilde{\mathbf{s}})\|_{\mathcal{H}_k}$ can be expressed by the distance between $\mathbf{s}$ and $\widetilde{\mathbf{s}}$ in the RBF kernel. Different bounds can be derived for different kernels and we only show the special RBF case for illustration purposes.

Part 2: Bound value functionals We then upper bound the value functionals and show that this bound can be relaxed so that it is independent with the choice of coalition. We first proceed with the interventional case and move on to observational afterwards. Using the result of Propositions A.2 and A.4 in Chau et al. (2022), we have

- Off-manifold value functionals: For any coalition $\mathcal{C}$, define $T_{\mathcal{C}}^{(\text{off})} = \left\|\mu_{\mathcal{C}}^{(\text{off})}(\mathbf{s})-\mu_{\mathcal{C}}^{(\text{off})}(\widetilde{\mathbf{s}})\right\|_{\mathcal{H}_k}^2$. Under the definition of $\mu_{\mathcal{C}}^{(\text{off})}(\mathbf{s})$ in Eq. (10), there is

$T_{\mathcal{C}}^{(\text{off})} \leq \left\|\psi(\mathbf{s}^{\mathcal{C}}) - \psi(\widetilde{\mathbf{s}}^{\mathcal{C}})\right\|_{\mathcal{H}_{k_{\mathcal{C}}}}^2 \left\|\mu_{\mathbf{S}^{\overline{\mathcal{C}}}}\right\|_{\mathcal{H}_{k_{\overline{\mathcal{C}}}}}^2$ with Cauchy-Schwarz inequality. Let $\delta_\psi := \sup_{\mathcal{C}\subseteq\mathcal{X}} \left\|\psi(\mathbf{s}^{\mathcal{C}}) - \psi(\widetilde{\mathbf{s}}^{\mathcal{C}})\right\|_{\mathcal{H}_{k_{\mathcal{C}}}}^2$ and assume kernels are all bounded per dimension by M, i.e $k^{(j)}(\mathbf{s}, \widetilde{\mathbf{s}}) \leq M$ for all $j \in \{1, 2, \cdots, d\}$. Armed with $\sup_{\mathcal{C}\subseteq\mathcal{X}} M^{|\mathcal{C}|} = M^d$, then the bound can be further loosen up as

$$\begin{aligned} T_{\mathcal{C}}^{(\text{off})} &= \left\|\mu_{\mathcal{C}}^{(\text{off})}(\mathbf{s}) - \mu_{\mathcal{C}}^{(\text{off})}(\widetilde{\mathbf{s}})\right\|_{\mathcal{H}_k}^2 \\ &= \left\|\psi(\mathbf{s}^{\mathcal{C}}) \otimes \mu_{\mathbf{S}^{\overline{\mathcal{C}}}} - \psi(\widetilde{\mathbf{s}}^{\mathcal{C}}) \otimes \mu_{\mathbf{S}^{\overline{\mathcal{C}}}}\right\|_{\mathcal{H}_k}^2 \\ &= \left\|\psi(\mathbf{s}^{\mathcal{C}}) - \psi(\widetilde{\mathbf{s}}^{\mathcal{C}})\right\|_{\mathcal{H}_{k_{\mathcal{C}}}}^2 \left\|\mu_{\mathbf{S}^{\overline{\mathcal{C}}}}\right\|_{\mathcal{H}_{k_{\overline{\mathcal{C}}}}}^2 \\ &\leq \delta_\psi \sup_{\mathcal{C}\subseteq\mathcal{X}} M^{|\mathcal{C}|} = \delta_\psi M^d, \end{aligned} \tag{91}$$

where $\left\|\mu_{\mathbf{S}^{\overline{\mathcal{C}}}}\right\|_{\mathcal{H}_{k_{\overline{\mathcal{C}}}}}^2 = \left\|\mathbb{E}[k(\mathbf{S}^{\overline{\mathcal{C}}}, \widetilde{\mathbf{S}}^{\overline{\mathcal{C}}})]\right\|^2 \leq M^{|\overline{\mathcal{C}}|}$.

- On-manifold value functionals: For any coalition $\mathcal{C}$, define $T_{\mathcal{C}}^{(\text{on})} = \left\|\mu_{\mathcal{C}}^{(\text{on})}(\mathbf{s}) - \mu_{\mathcal{C}}^{(\text{on})}(\widetilde{\mathbf{s}})\right\|_{\mathcal{H}_k}^2$. Under the definition of $\mu_{\mathcal{C}}^{(\text{on})}(\mathbf{s})$ in Eq. (10), there is $T_{\mathcal{C}}^{(\text{on})} \leq \left\|\psi(\mathbf{s}^{\mathcal{C}}) - \psi(\widetilde{\mathbf{s}}^{\mathcal{C}})\right\|_{\mathcal{H}_{k_{\mathcal{C}}}}^2 \left\|\mu_{\mathbf{S}^{\overline{\mathcal{C}}}|\mathbf{S}^{\mathcal{C}}}\right\|_{\mathcal{H}_{\Gamma_{\mathbf{S}^{\mathcal{C}}}}}^2 \left(\left\|\psi(\mathbf{s}^{\mathcal{C}})\right\|_{\mathcal{H}_{k_{\mathcal{C}}}}^2 + \left\|\psi(\widetilde{\mathbf{s}}^{\mathcal{C}})\right\|_{\mathcal{H}_{k_{\mathcal{C}}}}^2\right)$, where $\mathcal{H}_{\Gamma_{\mathbf{S}^{\mathcal{C}}}}$ is the $\mathcal{H}_{k_{\overline{\mathcal{C}}}}$-valued RKHS. Let $\delta_\psi := \sup_{\mathcal{C}\subseteq\mathcal{X}} \left\|\psi(\mathbf{s}^{\mathcal{C}}) - \psi(\widetilde{\mathbf{s}}^{\mathcal{C}})\right\|_{\mathcal{H}_{k_{\mathcal{C}}}}^2$ and $M_\Gamma = \sup_{\mathcal{C}\subseteq\mathcal{X}} \left\|\mu_{\mathbf{S}^{\overline{\mathcal{C}}}|\mathbf{S}^{\mathcal{C}}}\right\|_{\mathcal{H}_{\Gamma_{\mathbf{S}^{\mathcal{C}}}}}^2$. Then $T_{\mathcal{C}}^{(\text{on})} \leq 2M_\Gamma M^d \delta_\psi$ for all coalition $\mathcal{C}$.

Part 3: Bound the Shapley functionals Finally, we bound the loss of the value functionals under the definition of Eq. (9) as

$$\begin{aligned} |v_{\mathcal{C}}(\mathbf{s}) - v_{\mathcal{C}}(\widetilde{\mathbf{s}})|^2 =& |\langle w_{\mathsf{V}}, \mu_{\mathcal{C}}(\mathbf{s})\rangle_{\mathcal{H}_k} - \langle \widetilde{w}_{\mathsf{V}}, \mu_{\mathcal{C}}(\widetilde{\mathbf{s}})\rangle_{\mathcal{H}_k}|^2 \\ \leq& 2\left|\langle w_{\mathsf{V}} - \widetilde{w}_{\mathsf{V}}, \mu_{\mathcal{C}}(\mathbf{s})\rangle_{\mathcal{H}_k}\right|^2 + 2\left|\langle \widetilde{w}_{\mathsf{V}}, \mu_{\mathcal{C}}(\mathbf{s}) - \mu_{\mathcal{C}}(\widetilde{\mathbf{s}})\rangle_{\mathcal{H}_k}\right|^2 \\ \leq& 2\left\|w_{\mathsf{V}} - \widetilde{w}_{\mathsf{V}}\right\|_{\mathcal{H}_k}^2 \left\|\mu_{\mathcal{C}}(\mathbf{s})\right\|_{\mathcal{H}_k}^2 + 2\left\|\widetilde{w}_{\mathsf{V}}\right\|_{\mathcal{H}_k}^2 \left\|\mu_{\mathcal{C}}(\mathbf{s}) - \mu_{\mathcal{C}}(\widetilde{\mathbf{s}})\right\|_{\mathcal{H}_k}^2 \\ \leq& 2\delta_{\mathsf{V}}^2 M^d + 2M_V \left\|\mu_{\mathcal{C}}(\mathbf{s}) - \mu_{\mathcal{C}}(\widetilde{\mathbf{s}})\right\|_{\mathcal{H}_k}^2, \end{aligned} \tag{92}$$

where $\delta_{\tilde{\mathsf{V}}}$ is the loss of value function caused by perturbation, $\left\|\mu_{\mathcal{C}}(\mathbf{s})\right\|_{\mathcal{H}_k}^2 = \left\|\mathbb{E}\left[k(\mathbf{S}^{\mathcal{C}}, \mathbf{S}'^{\mathcal{C}})\right]\right\|_{\mathcal{H}_k}^2$, and M_V denotes the upper bound of $\left\|w_{\mathsf{V}}\right\|_{\mathcal{H}_k}^2$.

Since Shapley values are the expectation of differences of value functions, by devising a coalition independent bound for the difference in value functions, the expectation disappears in our bound. Therefore, along with Lemma C.3 of

$$\delta_{\tilde{\mathsf{V}}} \leq \frac{M_\eta q^2}{1-(q-1)\rho_{\mathsf{V}}} \left[\frac{L_k}{1-(q-1)\rho_{\mathsf{V}}} + M_k L_\psi + 2\sqrt{M_k} L_k\right] \varepsilon,$$

we have

$$\begin{aligned} \left\|\phi_i - \widetilde{\phi}_i\right\|_2^2 &= \left\|\frac{1}{d} \sum_{\mathcal{C}\subseteq\mathcal{X}\setminus\{i\}} \binom{d-1}{|\mathcal{C}|}^{-1} \left[v_{\mathcal{C}\cup i}(\mathbf{s}) - v_{\mathcal{C}}(\mathbf{s}) - (v_{\mathcal{C}\cup i}(\widetilde{\mathbf{s}}) - v_{\mathcal{C}}(\widetilde{\mathbf{s}}))\right]\right\|_2^2 \\ &\leq \frac{2}{d} \sum_{\mathcal{C}\subseteq\mathcal{X}\setminus\{i\}} \binom{d-1}{|\mathcal{C}|}^{-1} \left[\left\|v_{\mathcal{C}}(\mathbf{s}) - v_{\mathcal{C}}(\widetilde{\mathbf{s}})\right\|_2^2 + \left\|v_{\mathcal{C}\cup\{i\}}(\mathbf{s}) - v_{\mathcal{C}\cup\{i\}}(\widetilde{\mathbf{s}})\right\|_2^2\right] \\ &= 2\mathbb{E}_{\mathcal{C}}\left[\left\|v_{\mathcal{C}}(\mathbf{s}) - v_{\mathcal{C}}(\widetilde{\mathbf{s}})\right\|_2^2 + \left\|v_{\mathcal{C}\cup\{i\}}(\mathbf{s}) - v_{\mathcal{C}\cup\{i\}}(\widetilde{\mathbf{s}})\right\|_2^2\right]. \end{aligned} \tag{93}$$

Since we have proven bounds for $T_{\mathcal{C}}^{(\mathrm{on})} = \left\|\mu_{\mathcal{C}}^{(\mathrm{on})}(\mathbf{s}) - \mu_{\mathcal{C}}^{(\mathrm{on})}(\widetilde{\mathbf{s}})\right\|_{\mathcal{H}_k}^2$ and $T_{\mathcal{C}}^{(\mathrm{off})} = \left\|\mu_{\mathcal{C}}^{(\mathrm{off})}(\mathbf{s}) - \mu_{\mathcal{C}}^{(\mathrm{off})}(\widetilde{\mathbf{s}})\right\|_{\mathcal{H}_k}^2$ that is coalition independent, we can directly substitute the bound inside the expectation. Therefore, the loss of RKHS-SHAP values caused by $B(\mathbf{s})$ is

$$\left\|\phi_i^{(\mathrm{off})} - \widetilde{\phi}_i^{(\mathrm{off})}\right\|_2^2 \leq 4\delta_{\mathsf{V}}^2 M^d + 4M_V \delta_\psi M^d \tag{94a}$$

$$\left\|\phi_i^{(\mathrm{on})} - \widetilde{\phi}_i^{(\mathrm{on})}\right\|_2^2 \leq 4\delta_{\mathsf{V}}^2 M^d + 8M_V M_\Gamma \delta_\psi M^d. \tag{94b}$$

In the case when we pick k as a product RBF kernel, we have $\delta_\psi = 2 - 2\exp\left(-\frac{\varepsilon^2}{2l^2}\right)$ and $M = 1$. Then, we can simplify the result as

$$\begin{aligned}\left\|\phi_i^{(\mathrm{off})} - \widetilde{\phi}_i^{(\mathrm{off})}\right\|_2^2 \leq & \frac{4M^d M_\eta^2 q^4 \varepsilon^2}{(1-(q-1)\rho_{\mathsf{V}})^2}\left(\frac{L_k}{1-(q-1)\rho_{\mathsf{V}}} + M_k L_\psi + 2\sqrt{M_k} L_k\right)^2 \\ & + 4M_V\left(1 - \exp\left(-\frac{\varepsilon^2}{2l^2}\right)\right)\end{aligned} \tag{95a}$$

$$\begin{aligned}\left\|\phi_i^{(\mathrm{on})} - \widetilde{\phi}_i^{(\mathrm{on})}\right\|_2^2 \leq & \frac{4M^d M_\eta^2 q^4 \varepsilon^2}{(1-(q-1)\rho_{\mathsf{V}})^2}\left(\frac{L_k}{1-(q-1)\rho_{\mathsf{V}}} + M_k L_\psi + 2\sqrt{M_k} L_k\right)^2 \\ & + 8M_V M_\Gamma\left(1 - \exp\left(-\frac{\varepsilon^2}{2l^2}\right)\right).\end{aligned} \tag{95b}$$

C.5.4 Proof of Lemma C.3

According to the projection residual, the sparsification process can be divided into two cases.

Case 1: Maintain the dictionary $\mathcal{D}_{\mathsf{V}}$ For online sparsification, we consider the following optimization problem:

$$\min_{\{\eta_j\}\in\mathcal{D}_{\mathsf{V}}} \left\|\sum_j \eta_j \psi(\mathbf{s}_j) - \eta_\iota \psi(\mathbf{s}_\iota)\right\|_{\mathcal{H}_k}^2. \tag{96}$$

By expanding the norm and applying the reproducing property of the RKHS, we obtain:

$$\begin{aligned}\mathcal{L}(\{\eta_j\}) &= \left\langle \sum_j \eta_j \psi(\mathbf{s}_j) - \eta_\iota \psi(\mathbf{s}_\iota), \sum_{j'} \eta_{j'} \psi(\mathbf{s}_{j'}) - \eta_\iota \psi(\mathbf{s}_\iota)\right\rangle_{\mathcal{H}_k} \\ &= \sum_{j,j'} \eta_j k(\mathbf{s}_j, \mathbf{s}_{j'}) \eta_{j'} - 2\sum_j \eta_j k(\mathbf{s}_j, \mathbf{s}_\iota)\eta_\iota + \eta_\iota^2 k(\mathbf{s}_\iota, \mathbf{s}_\iota).\end{aligned} \tag{97}$$

Let $\eta_{\mathsf{V}} = [\eta_1, \ldots, \eta_q]^\top \in \mathbb{R}^q$ denote the coefficient vector, and define the Gram matrix $\mathbf{K}_{\mathsf{V},\mathsf{V}} \in \mathbb{R}^{q\times q}$ with entries $k(\mathbf{s}_j, \mathbf{s}_{j'})$, and the cross-kernel vector $\mathbf{k}_\iota \in \mathbb{R}^q$ with entries $k(\mathbf{s}_j, \mathbf{s}_\iota)\eta_\iota$. Then, the objective function becomes:

$$\mathcal{L}(\eta_{\mathsf{V}}) = \eta_{\mathsf{V}}^\top \mathbf{K}_{\mathsf{V},\mathsf{V}} \eta_{\mathsf{V}} - 2\eta_{\mathsf{V}}^\top \mathbf{k}_\iota + \eta_\iota^2 k(\mathbf{s}_\iota, \mathbf{s}_\iota). \tag{98}$$

Setting the gradient $\nabla_{\eta_{\mathsf{V}}}\mathcal{L} = 0$ yields the closed-form optimal solution:

$$\eta_{\mathsf{V}}^* = \mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\mathbf{k}_\iota. \tag{99}$$

Thus, the optimal coefficient for each basis function $k(\mathbf{s}_j, \cdot)$ is given by:

$$\eta_j^* = \sum_{j'} \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\eta_\iota. \tag{100}$$

This solution corresponds to the approximate orthogonal projection of the new kernel function $\eta_\iota \psi(\mathbf{s}_\iota)$ onto the subspace spanned by the kernel atoms $\{\psi(\mathbf{s}_j)\}_{j=1}^{q}$ in $\mathcal{H}_k$, and is used to decide whether the new basis $\psi(\mathbf{s}_\iota)$ is sufficiently novel to be added to the sparse dictionary $\mathcal{D}_{\mathsf{V}}$.

Therefore, we have

$$\begin{aligned}
\delta_{\tilde{\mathsf{V}}} =& \left\| \sum_j \sum_{j'} \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\eta_\iota \psi(\mathbf{s}_j) - \sum_j \sum_{j'} \left[\widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\widetilde{\mathbf{s}}_{j'}, \widetilde{\mathbf{s}}_\iota)\eta_\iota \psi(\widetilde{\mathbf{s}}_j) \right\|_{\mathcal{H}_k} \\
\leq& \sum_j \sum_{j'} \left\| \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\eta_\iota \psi(\mathbf{s}_j) - \left[\widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\widetilde{\mathbf{s}}_{j'}, \widetilde{\mathbf{s}}_\iota)\eta_\iota \psi(\widetilde{\mathbf{s}}_j) \right\|_{\mathcal{H}_k} \\
\leq& \sum_j \sum_{j'} \left\| \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\psi(\mathbf{s}_j) - \left[\widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\widetilde{\mathbf{s}}_{j'}, \widetilde{\mathbf{s}}_\iota)\psi(\widetilde{\mathbf{s}}_j) \right\|_{\mathcal{H}_k} \|\eta_\iota\|_2 \\
\leq& M_\eta \sum_j \sum_{j'} \left\| \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\psi(\mathbf{s}_j) - \left[\widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\widetilde{\mathbf{s}}_{j'}, \widetilde{\mathbf{s}}_\iota)\psi(\widetilde{\mathbf{s}}_j) \right\|_{\mathcal{H}_k}. \qquad (101)
\end{aligned}$$

We consider bounding the RKHS norm

$$\left\| \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\psi(\mathbf{s}_j) - \left[\widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\widetilde{\mathbf{s}}_{j'}, \widetilde{\mathbf{s}}_\iota)\psi(\widetilde{\mathbf{s}}_j) \right\|_{\mathcal{H}_k}, \qquad (102)$$

where $\mathbf{K}_{\mathsf{V},\mathsf{V}}$ and $\widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}$ denote Gram matrices over kernel dictionaries $\mathcal{D}_{\mathsf{V}} = \{\mathbf{s}_j\}$ and its perturbed version $\widetilde{\mathcal{D}}_{\mathsf{V}} = \{\widetilde{\mathbf{s}}_j\}$, respectively. Using the triangle inequality and RKHS norm properties, we decompose:

$$\begin{aligned}
& \left\| \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\psi(\mathbf{s}_j) - \left[\widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\widetilde{\mathbf{s}}_{j'}, \widetilde{\mathbf{s}}_\iota)\psi(\widetilde{\mathbf{s}}_j) \right\|_{\mathcal{H}_k} \\
\leq& \left| \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1} - \widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} \right| \cdot \|k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\|_2 \cdot \|\psi(\mathbf{s}_j)\|_{\mathcal{H}_k} \\
&+ \left| \left[\widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} \right| \cdot \|k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\psi(\mathbf{s}_j) - k(\widetilde{\mathbf{s}}_{j'}, \widetilde{\mathbf{s}}_\iota)\psi(\widetilde{\mathbf{s}}_j)\|_{\mathcal{H}_k} \\
\leq& \sqrt{M_k^3} \left| \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1} - \widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} \right| + \left| \left[\widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} \right| \cdot \|k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\psi(\mathbf{s}_j) - k(\widetilde{\mathbf{s}}_{j'}, \widetilde{\mathbf{s}}_\iota)\psi(\widetilde{\mathbf{s}}_j)\|_{\mathcal{H}_k}, \qquad (103)
\end{aligned}$$

where $\|k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\|_2 \leq M_k$ and $\|\psi(\mathbf{s}_j)\|_{\mathcal{H}_k} \leq \sqrt{M_k}$. Then, with the Lipschitz continuous of RBF k and ψ, we have

$$\begin{aligned}
& \|k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\psi(\mathbf{s}_j) - k(\widetilde{\mathbf{s}}_{j'}, \widetilde{\mathbf{s}}_\iota)\psi(\widetilde{\mathbf{s}}_j)\|_{\mathcal{H}_k} \\
\leq& |k(\mathbf{s}_{j'}, \mathbf{s}_\iota)| \cdot \|\psi(\mathbf{s}_j) - \psi(\widetilde{\mathbf{s}}_j)\|_{\mathcal{H}_k} + |k(\mathbf{s}_{j'}, \mathbf{s}_\iota) - k(\widetilde{\mathbf{s}}_{j'}, \widetilde{\mathbf{s}}_\iota)| \cdot \|\psi(\widetilde{\mathbf{s}}_j)\|_{\mathcal{H}_k} \\
\leq& M_k L_\psi \varepsilon + 2\sqrt{M_k} L_k \varepsilon. \qquad (104)
\end{aligned}$$

Moreover, we apply the Banach perturbation lemma to the inverse kernel matrices. Then:

$$\left\| \mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1} - \widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}^{-1} \right\|_{\mathrm{op}} \leq \|\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\|_{\mathrm{op}} \cdot \|\widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}} - \mathbf{K}_{\mathsf{V},\mathsf{V}}\|_{\mathrm{op}} \cdot \|\widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}}^{-1}\|_{\mathrm{op}}. \qquad (105)$$

Suppose the dictionary centers satisfy a minimum separation $C_k > 0$ and k is Gaussian. Then off-diagonal terms are bounded by $\rho_{\mathsf{V}} \coloneqq \exp\left(-\frac{C_k^2}{2l^2}\right)$ and the smallest eigenvalue of $\mathbf{K}_{\mathsf{V},\mathsf{V}}$ satisfies

$$\lambda_{\min}(\mathbf{K}_{\mathsf{V},\mathsf{V}}) \geq 1 - (q-1)\rho_{\mathsf{V}}, \quad \Rightarrow \quad \|\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\|_{\mathrm{op}} \leq \frac{1}{1-(q-1)\rho_{\mathsf{V}}}. \qquad (106)$$

Combining with the kernel matrix perturbation

$$\|\widetilde{\mathbf{K}}_{\mathsf{V},\mathsf{V}} - \mathbf{K}_{\mathsf{V},\mathsf{V}}\| \leq L_k \varepsilon, \qquad (107)$$

where we interpret L_k as the Lipschitz constant for the operator norm on the entire Gram matrix. Then, we obtain the final upper bound:

$$\delta_{\tilde{\mathsf{V}}} \leq \frac{M_\eta q^2}{1-(q-1)\rho_{\mathsf{V}}} \left[\frac{L_k}{1-(q-1)\rho_{\mathsf{V}}} + M_k L_\psi + 2\sqrt{M_k} L_k \right] \varepsilon. \qquad (108)$$

Case 2: Update the dictionary $\mathcal{D}_{\mathsf{V}}$ Without loss of generality, we assume $j^\star$ is replaced by ι. Therefore, the optimization problem becomes

$$\min_{\{\eta_j\}\in\mathcal{D}'_{\mathsf{V}}} \left\| \sum_j \eta_j \psi(\mathbf{s}_j) - \eta_{j^\star}\psi(\mathbf{s}_{j^\star}) \right\|_{\mathcal{H}_k}^2 . \tag{109}$$

By expanding the norm and applying the reproducing property of the RKHS, we obtain:

$$\begin{aligned}\mathcal{L}(\{\eta_j\}) &= \left\langle \sum_j \eta_j \psi(\mathbf{s}_j) - \eta_{j^\star}\psi(\mathbf{s}_{j^\star}), \sum_{j'} \eta_{j'}\psi(\mathbf{s}_{j'}) - \eta_{j^\star}\psi(\mathbf{s}_{j^\star}) \right\rangle_{\mathcal{H}_k} \\ &= \sum_{j,j'} \eta_j k(\mathbf{s}_j,\mathbf{s}_{j'})\eta_{j'} - 2\sum_j \eta_j k(\mathbf{s}_j,\mathbf{s}_{j^\star})\eta_{j^\star} + \eta_{j^\star}^2 k(\mathbf{s}_{j^\star},\mathbf{s}_{j^\star}). \end{aligned} \tag{110}$$

Similar to the proof in case 1, we obtain the final upper bound:

$$\delta_{\tilde{\mathsf{V}}} \leq \frac{M_\eta q^2}{1-(q-1)\rho_{\mathsf{V}}} \left[\frac{L_k}{1-(q-1)\rho_{\mathsf{V}}} + M_k L_\psi + 2\sqrt{M_k} L_k \right] \varepsilon. \tag{111}$$

This result characterizes the sensitivity of the RKHS element to perturbations in both the kernel centers and the Gram matrix structure.

C.5.5 Proof of Lemma C.4

After the gradient obtained, RSA2C will execute the online sparsification, which can be divided into two cases. In this part, we ignore the time step t without causing misunderstandings.

Case 1: Maintain the dictionary $\mathcal{D}_{\mathsf{V}}$ For online sparsification, we consider the following optimization problem:

$$\min_{\{\eta_j\}\in\mathcal{D}_{\mathsf{V}}} \left\| \sum_j \eta_j \psi(\mathbf{s}_j) - \eta_\iota \psi(\mathbf{s}_\iota) \right\|_{\mathcal{H}_k}^2 . \tag{112}$$

By expanding the norm and applying the reproducing property of the RKHS, we obtain:

$$\begin{aligned}\mathcal{L}(\{\eta_j\}) &= \left\langle \sum_j \eta_j \psi(\mathbf{s}_j) - \eta_\iota\psi(\mathbf{s}_\iota), \sum_{j'} \eta_{j'}\psi(\mathbf{s}_{j'}) - \eta_\iota\psi(\mathbf{s}_\iota) \right\rangle_{\mathcal{H}_k} \\ &= \sum_{j,j'} \eta_j k(\mathbf{s}_j,\mathbf{s}_{j'})\eta_{j'} - 2\sum_j \eta_j k(\mathbf{s}_j,\mathbf{s}_\iota)\eta_\iota + \eta_\iota^2 k(\mathbf{s}_\iota,\mathbf{s}_\iota). \end{aligned} \tag{113}$$

Let $\eta_{\mathsf{V}} = [\eta_1,\ldots,\eta_q]^\top \in \mathbb{R}^q$ denote the coefficient vector, and define the Gram matrix $\mathbf{K}_{\mathsf{V},\mathsf{V}} \in \mathbb{R}^{q\times q}$ with entries $k(\mathbf{s}_j,\mathbf{s}_{j'})$, and the cross-kernel vector $\mathbf{k}_\iota \in \mathbb{R}^q$ with entries $k(\mathbf{s}_j,\mathbf{s}_\iota)\eta_\iota$. Then, the objective function becomes:

$$\mathcal{L}(\eta_{\mathsf{V}}) = \eta_{\mathsf{V}}^\top \mathbf{K}_{\mathsf{V},\mathsf{V}}\eta_{\mathsf{V}} - 2\eta_{\mathsf{V}}^\top \mathbf{k}_\iota + \eta_\iota^2 k(\mathbf{s}_\iota,\mathbf{s}_\iota). \tag{114}$$

Setting the gradient $\nabla_{\eta_{\mathsf{V}}}\mathcal{L} = 0$ yields the closed-form optimal solution:

$$\eta_{\mathsf{V}}^* = \mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\mathbf{k}_\iota. \tag{115}$$

Thus, the optimal coefficient for each basis function $k(\mathbf{s}_j,\cdot)$ is given by:

$$\eta_j^* = \sum_{j'} \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'},\mathbf{s}_\iota)\eta_\iota. \tag{116}$$

This solution corresponds to the approximate orthogonal projection of the new kernel function $\eta_\iota\psi(\mathbf{s}_\iota)$ onto the subspace spanned by the kernel atoms $\{\psi(\mathbf{s}_j)\}_{j=1}^q$ in $\mathcal{H}_k$, and is used to decide whether the new basis $\psi(\mathbf{s}_\iota)$ is sufficiently novel to be added to the sparse dictionary $\mathcal{D}_{\mathsf{V}}$.

Therefore, we have

$$\begin{aligned}\delta_{\mathsf{PV}} &= \left\| \sum_j \eta_j^\star \psi(\mathbf{s}_j) - \sum_j \eta_j \psi(\mathbf{s}_j) \right\|_{\mathcal{H}_k} \\ &= \left\| \sum_j \sum_{j'} \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\eta_\iota \psi(\mathbf{s}_j) - \sum_j \eta_j \psi(\mathbf{s}_j) \right\|_{\mathcal{H}_k} \\ &\leq \sum_j \left| \sum_{j'} \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\eta_\iota - \eta_j \right| \cdot \|\psi(\mathbf{s}_j)\|_{\mathcal{H}_k} \\ &\leq \sqrt{M_k} q \varepsilon_{\mathsf{PV}}, \end{aligned} \tag{117}$$

where $\varepsilon_{\mathsf{PV}} \coloneqq \sup_j \left| \sum_{j'} \left[\mathbf{K}_{\mathsf{V},\mathsf{V}}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\eta_\iota - \eta_j \right|$ and $\|\psi(\mathbf{s}_j)\|_{\mathcal{H}_k} \leq \sqrt{M_k}$.

Case 2: Update the dictionary $\mathcal{D}_{\mathsf{V}}$ Without loss of generality, we assume $j^\star$ is replaced by ι. Therefore, the optimization problem becomes

$$\min_{\{\eta_j\} \in \mathcal{D}'_{\mathsf{V}}} \left\| \sum_j \eta_j \psi(\mathbf{s}_j) - \eta_{j^\star} \psi(\mathbf{s}_{j^\star}) \right\|_{\mathcal{H}_k}^2 . \tag{118}$$

By expanding the norm and applying the reproducing property of the RKHS, we obtain:

$$\begin{aligned}\mathcal{L}(\{\eta_j\}) &= \left\langle \sum_j \eta_j \psi(\mathbf{s}_j) - \eta_{j^\star}\psi(\mathbf{s}_{j^\star}), \sum_{j'} \eta_{j'} \psi(\mathbf{s}_{j'}) - \eta_{j^\star}\psi(\mathbf{s}_{j^\star}) \right\rangle_{\mathcal{H}_k} \\ &= \sum_{j,j'} \eta_j k(\mathbf{s}_j, \mathbf{s}_{j'})\eta_{j'} - 2\sum_j \eta_j k(\mathbf{s}_j, \mathbf{s}_\iota)\eta_{j^\star} + \eta_{j^\star}^2 k(\mathbf{s}_{j^\star}, \mathbf{s}_{j^\star}). \end{aligned} \tag{119}$$

Similar to the proof in case 1, we obtain the final upper bound:

$$\delta_{\mathsf{PV}} \leq \sqrt{M_k} q \varepsilon_{\mathsf{PV}}, \tag{120}$$

where $\varepsilon_{\mathsf{PV}} \coloneqq \sup_j \left| \sum_{j'} \left[\mathbf{K}_{\mathsf{V}',\mathsf{V}'}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_{j^\star})\eta_{j^\star} - \eta_j \right|$. Therefore, we have

$$\begin{aligned}\varepsilon_{\mathsf{PV}} \coloneqq \max \Bigg\{ & \sup_{\psi_j \in \mathcal{D}_{\mathsf{V}}} \left| \sum_{j'} \left[\mathbf{K}_{\mathsf{V}',\mathsf{V}'}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_\iota)\eta_\iota - \eta_j \right|, \\ & \sup_{\psi_j \in \{\psi_\iota\} \cup (\mathcal{D}_{\mathsf{V}} \setminus \{\psi_{j^\star}\})} \left| \sum_{j'} \left[\mathbf{K}_{\mathsf{V}',\mathsf{V}'}^{-1}\right]_{jj'} k(\mathbf{s}_{j'}, \mathbf{s}_{j^\star})\eta_{j^\star} - \eta_j \right| \Bigg\}. \end{aligned} \tag{121}$$

C.5.6 Proof of Lemma C.5

The population gradient is

$$g_t\left(w_{\mathsf{V},t}\right) = \mathbb{E}_{\mathbf{s}\sim D^{\pi_{h,\Sigma}}, \mathbf{a}\sim\pi_{h,\Sigma}(\cdot|\mathbf{s})} \left[\left(V_{w_{\mathsf{V},t}}(\mathbf{s}) - r(\mathbf{s},\mathbf{a}) - \gamma V_{w_{\mathsf{V},t}}(\mathbf{s}') \right) \psi(\mathbf{s}) \right] + \lambda w_{\mathsf{V},t}, \tag{122}$$

and the optimality condition gives $g_t\left(w_{\mathsf{V},t}^\star\right) = 0$. Subtracting yields

$$\begin{aligned} g_t\left(w_{\mathsf{V},t}\right) &= \mathbb{E}\left[\left(V_{w_{\mathsf{V},t}}(\mathbf{s}) - V_{w_{\mathsf{V},t}^\star}(\mathbf{s}) - \gamma \left(V_{w_{\mathsf{V},t}}(\mathbf{s}') - V_{w_{\mathsf{V},t}^\star}(\mathbf{s}') \right) \right) \psi(\mathbf{s}) \right] + \lambda \left(w_{\mathsf{V},t} - w_{\mathsf{V},t}^\star \right) \\ &= \left(\Gamma_s - \gamma \Gamma_{2s} + \lambda \mathbf{I} \right) \left(w_{\mathsf{V},t} - w_{\mathsf{V},t}^\star \right), \end{aligned} \tag{123}$$

where $\Gamma_s \coloneqq \mathbb{E}[\psi(\mathbf{s}) \otimes \psi(\mathbf{s})]$ and $\Gamma_{2s} \coloneqq \mathbb{E}[\psi(\mathbf{s}) \otimes \psi(\mathbf{s}')]$. Hence, by submultiplicativity of operator norm,

$$\|g_t(w_{\mathsf{V},t})\|_{\mathcal{H}_k}^2 \leq \|\Gamma_s - \gamma\Gamma_{2s} + \lambda\mathbf{I}\|_{\mathrm{op}}^2 \left\|w_{\mathsf{V},t} - w_{\mathsf{V},t}^\star\right\|_{\mathcal{H}_k}^2. \tag{124}$$

We now bound $\|\Gamma_s - \gamma\Gamma_{2s} + \lambda\mathbf{I}\|_{\mathrm{op}}$. Assume the kernel is bounded as $\|\psi(\mathbf{s})\|_{\mathcal{H}_k}^2 = k(\mathbf{s},\mathbf{s}) \leq M_k$; then $\|\Gamma_s\|_{\mathrm{op}} \leq \mathbb{E}\|\psi(\mathbf{s})\|^2 \leq M_k$ and $\|\Gamma_{2s}\|_{\mathrm{op}} \leq \mathbb{E}\left[\|\psi(\mathbf{s})\|\|\psi(\mathbf{s}')\|\right] \leq M_k$ (by Cauchy–Schwarz and boundedness). Therefore,

$$\|\Gamma_s - \gamma\Gamma_{2s} + \lambda\mathbf{I}\|_{\mathrm{op}} \leq \|\Gamma_s\|_{\mathrm{op}} + \gamma\|\Gamma_{2s}\|_{\mathrm{op}} + \lambda \leq (1+\gamma)M_k + \lambda, \tag{125}$$

and consequently

$$\|g_t(w_{\mathsf{V},t})\|_{\mathcal{H}_k}^2 \leq C_1 \left\|w_{\mathsf{V},t} - w_{\mathsf{V},t}^\star\right\|_{\mathcal{H}_k}^2, \qquad C_1 \coloneqq \left((1+\gamma)M_k + \lambda\right)^2. \tag{126}$$

C.5.7 Proof of Lemma C.6

Using $\|AB\|_{\mathrm{op}} \leq \|A\|_{\mathrm{op}}\|B\|_{\mathrm{op}}$ and $\|\Sigma_{\mathbf{K}}\|_{\mathrm{op}} \leq M_{\Sigma_K}$, we have

$$\begin{aligned}
\|K'(\mathbf{s},\cdot) - K(\mathbf{s},\cdot)\|_{\mathrm{op}} =& \|\kappa'_\phi(\mathbf{s},\cdot)\Sigma_{\mathsf{K}} - \kappa_\phi(\mathbf{s},\cdot)\Sigma_{\mathsf{K}}\|_{\mathrm{op}} \\
\leq& \|\kappa'_\phi(\mathbf{s},\cdot) - \kappa_\phi(\mathbf{s},\cdot)\|_{\mathrm{op}}\|\Sigma_{\mathsf{K}}\|_{\mathrm{op}} \\
\leq& M_{\Sigma_K}\|\kappa'_\phi(\mathbf{s},\cdot) - \kappa_\phi(\mathbf{s},\cdot)\|_{\mathrm{op}} \\
=& M_{\Sigma_K}\left\|\exp\left(-\frac{1}{2}(\mathbf{s}-\cdot)^\top w'(\mathbf{s}-\cdot)\right) - \exp\left(-\frac{1}{2}(\mathbf{s}-\cdot)^\top w(\mathbf{s}-\cdot)\right)\right\|_{\mathrm{op}}.
\end{aligned} \tag{127}$$

For the Mahalanobis RBF $\kappa_{\mathbf{W}}(\mathbf{s},\mathbf{s}') = \exp(-\frac{1}{2}\mathbf{x}^\top\mathbf{W}\mathbf{x})$ with $\mathbf{x} \coloneqq \mathbf{s} - \mathbf{s}'$, the mean-value theorem gives, for some $\widetilde{\mathbf{W}}$ on the segment between $\mathbf{W}$ and $\mathbf{W}'$,

$$|\kappa_{\mathbf{W}'} - \kappa_{\mathbf{W}}| \leq \tfrac{1}{2}e^{-\frac{1}{2}\mathbf{x}^\top\widetilde{\mathbf{W}}\mathbf{x}}|\mathbf{x}^\top(\mathbf{W}' - \mathbf{W})\mathbf{x}| \leq \tfrac{1}{2}\|\mathbf{W}' - \mathbf{W}\|_{\mathrm{op}}\|\mathbf{x}\|_2^2.$$

If $\|\mathbf{s}\|_2 \leq M_S$ then $\|\mathbf{x}\|_2 \leq 2M_S$, hence $\sup_{\mathbf{s}}\|K'(\mathbf{s},\cdot) - K(\mathbf{s},\cdot)\|_{\mathrm{op}} \leq 2M_{\Sigma_K}M_S^2\|\mathbf{W}' - \mathbf{W}\|_{\mathrm{op}}$, i.e.,

$$\delta_{\mathsf{K},h} \leq q^2 M_c^2 \left(\sup_{\mathbf{s}}\|K'(\mathbf{s},\cdot) - K(\mathbf{s},\cdot)\|_{\mathrm{op}}\right)^2, \qquad M_c \coloneqq \max_j \|\mathbf{c}_j\|_2, \tag{128}$$

$$\sup_{\mathbf{s}}\|K'(\mathbf{s},\cdot) - K(\mathbf{s},\cdot)\|_{\mathrm{op}} \leq 2M_{\Sigma_K}M_S^2\|\mathbf{W}' - \mathbf{W}\|_{\mathrm{op}}. \tag{129}$$

Combining Eq. (128) and Eq. (129) yields

$$\delta_{\mathsf{K},h} \leq 4M_{\Sigma_K}^2 M_S^4 q^2 M_c^2\|\mathbf{W}' - \mathbf{W}\|_{\mathrm{op}}^2. \tag{130}$$

By Eq. (9) and Cauchy–Schwarz in $\mathcal{H}_k$,

$$\begin{aligned}
|v'_{\mathcal{C}}(\mathbf{s}) - v_{\mathcal{C}}(\mathbf{s})| &= |\langle w'_{\mathsf{V}} - w_{\mathsf{V}}, \mu_{\mathcal{C}}(\mathbf{s})\rangle| \\
&\leq \|w'_{\mathsf{V}} - w_{\mathsf{V}}\|_{\mathcal{H}_k}\|\mu_{\mathcal{C}}(\mathbf{s})\|_{\mathcal{H}_k} \\
&\leq \|w'_{\mathsf{V}} - w_{\mathsf{V}}\|_{\mathcal{H}_k}C_\mu,
\end{aligned} \tag{131}$$

we have

$$\begin{aligned}
\sup_{\mathbf{s},\mathcal{C}}\left|v'_{\mathcal{C}}(\mathbf{s}) - v_{\mathcal{C}}(\mathbf{s})\right| &= \sup_{\mathbf{s},\mathcal{C}}\left|\langle w'_{\mathsf{V}} - w_{\mathsf{V}}, \mu_{\mathcal{C}}(\mathbf{s})\rangle_{\mathcal{H}_k}\right| \\
&\leq \|w'_{\mathsf{V}} - w_{\mathsf{V}}\|_{\mathcal{H}_k}\sup_{\mathbf{s},\mathcal{C}}\|\mu_{\mathcal{C}}(\mathbf{s})\|_{\mathcal{H}_k} \\
&\leq \|w'_{\mathsf{V}} - w_{\mathsf{V}}\|_{\mathcal{H}_k}C_\mu.
\end{aligned} \tag{132}$$

The Shapley aggregation Eq. (11) is an average (with weights summing to one) of differences of the form $f_{\mathcal{C}\cup\{i\}} - f_{\mathcal{C}}$, so Jensen's inequality gives

$$\|\boldsymbol{\phi}' - \boldsymbol{\phi}\|_\infty \leq \frac{2}{d}\sum_{\mathcal{C}}\binom{d-1}{|\mathcal{C}|}^{-1}\sup_{\mathbf{s},\mathcal{C}}\left|v'_{\mathcal{C}}(\mathbf{s}) - v_{\mathcal{C}}(\mathbf{s})\right| \leq 2\|w'_{\mathsf{V}} - w_{\mathsf{V}}\|_{\mathcal{H}_k}C_\mu. \tag{133}$$

Substituting Eq. (133) into Eq. (130) yields

$$\delta_{\mathsf{K},h} \leq 16 M_{\Sigma_K}^2 M_S^4 q^2 M_c^2 C_\mu^2 \|w'_{\mathsf{V}} - w_{\mathsf{V}}\|_{\mathcal{H}_k}^2. \tag{134}$$

Finally, for tensor features in Eq. (10), with each component map bounded by M, the product structure implies $C_\mu \leq M^{d/2}$, so

$$\delta_{\mathsf{K},h} \leq 16 M_{\Sigma_K}^2 M_S^4 q^2 M_c^2 M^d \|w'_{\mathsf{V}} - w_{\mathsf{V}}\|_{\mathcal{H}_k}^2. \tag{135}$$

C.5.8 Proof of Lemma C.7

Let $g(h) \coloneqq \nabla \log \pi_h(\mathbf{a} \mid \mathbf{s}), \forall h \in \mathcal{H}_K$. According to Proposition B.2, we get the gradient expression as

$$g(h) = \nabla_h \log \pi_h(\mathbf{a} \mid \mathbf{s}) = K(\mathbf{s}, \cdot)\Sigma^{-1}\left(\mathbf{a} - h(\mathbf{s})\right). \tag{136}$$

For any $h, h' \in \mathcal{H}_K$,

$$\begin{aligned}
\|g(h') - g(h)\|_{\mathcal{H}_K} &= \left\|K'(\mathbf{s}, \cdot)\Sigma^{-1}\left(\mathbf{a} - h'(\mathbf{s})\right) - K(\mathbf{s}, \cdot)\Sigma^{-1}\left(\mathbf{a} - h(\mathbf{s})\right)\right\|_{\mathcal{H}_K} \\
&\leq \left\|K'(\mathbf{s}, \cdot)\Sigma^{-1}\left(\mathbf{a} - h'(\mathbf{s})\right) - K'(\mathbf{s}, \cdot)\Sigma^{-1}\left(\mathbf{a} - h(\mathbf{s})\right)\right\|_{\mathcal{H}_K} \\
&\quad + \left\|K'(\mathbf{s}, \cdot)\Sigma^{-1}\left(\mathbf{a} - h(\mathbf{s})\right) - K(\mathbf{s}, \cdot)\Sigma^{-1}\left(\mathbf{a} - h(\mathbf{s})\right)\right\|_{\mathcal{H}_K} \\
&= \left\|K'(\mathbf{s}, \cdot)\Sigma^{-1}\left(h(\mathbf{s}) - h'(\mathbf{s})\right)\right\|_{\mathcal{H}_K} + \left\|\left(K'(\mathbf{s}, \cdot) - K(\mathbf{s}, \cdot)\right)\Sigma^{-1}\left(\mathbf{a} - h(\mathbf{s})\right)\right\|_{\mathcal{H}_K} \\
&\leq \|K'(\mathbf{s}, \cdot)\|_{\mathrm{op}} \|\Sigma^{-1}\|_{\mathrm{op}} \|h'(\mathbf{s}) - h(\mathbf{s})\|_2 \\
&\quad + \|K'(\mathbf{s}, \cdot) - K(\mathbf{s}, \cdot)\|_{\mathrm{op}} \|\Sigma^{-1}\|_{\mathrm{op}} \|\mathbf{a} - h(\mathbf{s})\|_2 \\
&\leq \sqrt{M_K} \|\Sigma^{-1}\|_{\mathrm{op}} \sup_{\mathbf{s}} \|h'(\mathbf{s}) - h(\mathbf{s})\|_2 + 2M_a \|K'(\mathbf{s}, \cdot) - K(\mathbf{s}, \cdot)\|_{\mathrm{op}} \|\Sigma^{-1}\|_{\mathrm{op}} \\
&\leq \underbrace{\left(\sqrt{M_K} \|\Sigma^{-1}\|_{\mathrm{op}} C_{\mathrm{ev}}\right)}_{=: L_{\mathrm{grad}}} \|h' - h\|_{\mathcal{H}_K} + 2M_a M_\Sigma \delta_{\mathsf{K},h},
\end{aligned}$$

where $M_K \coloneqq \sup_{\mathbf{s}} \|K(\mathbf{s}, \mathbf{s})\|_{\mathrm{op}}$ and C_{ev} is the evaluation operator bound $\|v(\mathbf{s})\|_2 \leq C_{\mathrm{ev}} \|v\|_{\mathcal{H}_K}$ for all $v \in \mathcal{H}_K$. Besides, $\delta_{\mathsf{K},h}$ is defined in Lemma C.6.

Thus, we obtain

$$\log \pi_{h_{t+1}}(\mathbf{a} \mid \mathbf{s}) \geq \log \pi_{h_t}(\mathbf{a} \mid \mathbf{s}) + \langle g(h_t), h_{t+1} - h_t \rangle_{\mathcal{H}_K} - \frac{L_{\mathrm{grad}}}{2} \|h_{t+1} - h_t\|_{\mathcal{H}_K}^2 - 2M_a^2 M_\Sigma^2 \delta_{\mathsf{K},h_t}. \tag{137}$$

C.5.9 Proof of Lemma C.8

In this part, we prove Lemma C.8 by considering two different cases.

Case 1: Maintain the dictionary $\mathcal{D}_{\mathsf{A}}$ For the vector-valued RKHS $\mathcal{H}_K$ associated with the operator-valued kernel $K : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^{m \times m}$, the Actor mean function takes the form

$$h(\mathbf{s}) = \sum_{j=1}^{q} K(\mathbf{s}, \mathbf{s}_j)\mathbf{c}_j,$$

where $\mathbf{c}_j \in \mathbb{R}^m$ are coefficient vectors. Given a new candidate atom $K(\cdot, \mathbf{s}_\iota)\mathbf{c}_\iota$, online sparsification projects it onto the subspace spanned by $\{K(\cdot, \mathbf{s}_j)\mathbf{c} : \mathbf{c} \in \mathbb{R}^m, j = 1, \ldots, q\}$:

$$\min_{\{\mathbf{c}_j\}} \left\| \sum_{j=1}^{q} K(\cdot, \mathbf{s}_j)\boldsymbol{c}_j - K(\cdot, \mathbf{s}_\iota)\mathbf{c}_\iota \right\|_{\mathcal{H}_K}^2. \tag{138}$$

Using the reproducing property of vector-valued RKHSs,

$$\mathcal{L}(\{\boldsymbol{c}_j\}) = \sum_{j,j'=1}^{q} \boldsymbol{c}_j^\top K(\mathbf{s}_j, \mathbf{s}_{j'})\boldsymbol{c}_{j'} - 2\sum_{j=1}^{q} \boldsymbol{c}_j^\top K(\mathbf{s}_j, \mathbf{s}_\iota)\mathbf{c}_\iota + \mathbf{c}_\iota^\top K(\mathbf{s}_\iota, \mathbf{s}_\iota)\mathbf{c}_\iota. \tag{139}$$

Let $\mathbf{c}_{\mathsf{A}} = \mathrm{col}(\mathbf{c}_1, \ldots, \mathbf{c}_q) \in \mathbb{R}^{qm}$ denote the stacked coefficient vector, and define the block Gram matrix $\mathbf{K}_{\mathsf{A},\mathsf{A}} \in \mathbb{R}^{qm \times qm}$ with (j, j')-block $K(\mathbf{s}_j, \mathbf{s}_{j'})$, and the block cross-kernel $\mathbf{K}_{\mathsf{A},\iota} \in \mathbb{R}^{qm \times m}$ with block $K(\mathbf{s}_j, \mathbf{s}_\iota)$. Then,

$$\mathcal{L}(\boldsymbol{c}_{\mathsf{A}}) = \boldsymbol{c}_{\mathsf{A}}^\top \mathbf{K}_{\mathsf{A},\mathsf{A}} \boldsymbol{c}_{\mathsf{A}} - 2\boldsymbol{c}_{\mathsf{A}}^\top \mathbf{K}_{\mathsf{A},\iota} \mathbf{c}_\iota + \mathbf{c}_\iota^\top K(\mathbf{s}_\iota, \mathbf{s}_\iota) \mathbf{c}_\iota. \tag{140}$$

Setting $\nabla_{\boldsymbol{c}_{\mathsf{A}}} \mathcal{L} = 0$ yields the closed-form optimal solution:

$$\boldsymbol{c}_{\mathsf{A}}^\star = \mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1} \mathbf{K}_{\mathsf{A},\iota} \mathbf{c}_\iota. \tag{141}$$

The approximation error is bounded as

$$\begin{aligned} \delta_{\mathsf{PA}} &\coloneqq \left\| \sum_{j=1}^{q} K(\cdot, \mathbf{s}_j) \boldsymbol{c}_j^\star - \sum_{j=1}^{q} K(\cdot, \mathbf{s}_j) \mathbf{c}_j \right\|_{\mathcal{H}_K} \\ &\leq \sum_{j=1}^{q} \| K(\cdot, \mathbf{s}_j)(\boldsymbol{c}_j^\star - \mathbf{c}_j) \|_{\mathcal{H}_K} \leq \sqrt{M_K} q \varepsilon_{\mathsf{PA}}, \end{aligned} \tag{142}$$

where $M_K \coloneqq \sup_{\mathbf{s}} \|K(\mathbf{s}, \mathbf{s})\|_{\mathrm{op}}$ and $\varepsilon_{\mathsf{PA}} \coloneqq \sup_j \|\boldsymbol{c}_j^\star - \mathbf{c}_j\|_2$.

Case 2: Update the dictionary $\mathcal{D}_{\mathsf{A}}$ Suppose an existing dictionary element $\mathbf{s}_{j^\star}$ is replaced by $\mathbf{s}_\iota$, yielding $\mathcal{D}'_{\mathsf{A}}$. The projection problem becomes

$$\min_{\{\boldsymbol{c}_j\} \in \mathcal{D}'_{\mathsf{A}}} \left\| \sum_{j \in \mathcal{D}'_{\mathsf{A}}} K(\cdot, \mathbf{s}_j) \boldsymbol{c}_j - K(\cdot, \mathbf{s}_{j^\star}) \mathbf{c}_{j^\star} \right\|_{\mathcal{H}_K}^2. \tag{143}$$

Similar to Case 1, with the block Gram matrix $\mathbf{K}_{\mathsf{A}',\mathsf{A}'}$ and cross-kernel $\mathbf{K}_{\mathsf{A}',j^\star}$, the optimal coefficients are

$$\boldsymbol{c}_{\mathsf{A}'}^\star = \mathbf{K}_{\mathsf{A}',\mathsf{A}'}^{-1} \mathbf{K}_{\mathsf{A}',j^\star} \mathbf{c}_{j^\star}. \tag{144}$$

The error bound is

$$\delta_{\mathsf{PA}} \leq \sqrt{M_K} q \varepsilon_{\mathsf{PA}}, \tag{145}$$

where

$$\begin{aligned} \varepsilon_{\mathsf{PA}} \coloneqq \max \Big\{ & \sup_{\psi_j \in \mathcal{D}_{\mathsf{A}}} \|[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1} \mathbf{K}_{\mathsf{A},\iota}] \mathbf{c}_\iota - \mathbf{c}_j\|_2, \\ & \sup_{\psi_j \in \{\mathbf{s}_\iota\} \cup (\mathcal{D}_{\mathsf{A}} \setminus \{\mathbf{s}_{j^\star}\})} \|[\mathbf{K}_{\mathsf{A},\mathsf{A}}^{-1} \mathbf{K}_{\mathsf{A},j^\star}] \mathbf{c}_{j^\star} - \mathbf{c}_j\|_2 \Big\}. \end{aligned} \tag{146}$$

C.5.10 Proof of Lemma C.9

Recall $w_{\mathsf{V}}^\star(h) = (B(h) + \lambda \mathbf{I})^{-1} b(h)$ where $B(h) \coloneqq \Gamma_s(h) - \gamma \Gamma_{2s}(h)$ and $b(h) \coloneqq \mathbb{E}[r(\mathbf{s}, \mathbf{a}) \psi(s)]$, with h denoting the Actor parameters. Using the inverse perturbation identity $(X + \Delta)^{-1} - X^{-1} = -X^{-1} \Delta (X + \Delta)^{-1}$ and operator submultiplicativity,

$$\begin{aligned} \|w_{\mathsf{V}}^\star(h') - w_{\mathsf{V}}^\star(h)\| &\leq \|(B(h') + \lambda \mathbf{I})^{-1}\| \|b(h') - b(h)\| + \|(B(h') + \lambda \mathbf{I})^{-1} - (B(h) + \lambda \mathbf{I})^{-1}\| \|b(h)\| \\ &\leq \frac{1}{\lambda} \|b(h') - b(h)\| + \frac{1}{\lambda^2} \|B(h') - B(h)\|_{\mathrm{op}} \|b(h)\|. \end{aligned} \tag{147}$$

Assume bounded feature map $\|\psi(s)\|_{\mathcal{H}_\psi}^2 \leq M$ and bounded reward $r \in [0, 1]$. Let C_ν be the Lipschitz constant of the on-policy visitation distribution w.r.t. h (i.e., $\|\nu_{\pi_h} - \nu_{\pi_{h'}}\|_{\mathrm{TV}} \leq C_\nu \|h - h'\|$), which is standard under the ergodicity/Lipschitz policy assumptions. Then

$$\begin{aligned} \|b(h') - b(h)\| &\leq \sqrt{M} C_\nu \|h' - h\| \\ \|B(h') - B(h)\|_{\mathrm{op}} &\leq 2(1 + \gamma) M C_\nu \|h' - h\|, \\ \|b(h)\| &\leq \mathbb{E}[|r||\psi|] \leq \sqrt{M}. \end{aligned}$$

Combining gives the Lipschitz drift of the target Critic:

$$\|w_{\mathsf{V}}^\star(h') - w_{\mathsf{V}}^\star(h)\| \leq L_{\mathsf{V}} \|h' - h\|, \qquad L_{\mathsf{V}} \coloneqq C_\nu \sqrt{M} \left(\frac{1}{\lambda} + \frac{2(1 + \gamma) M}{\lambda^2} \right). \tag{148}$$

C.5.11 PROOF OF LEMMA C.11

We provide the proof of Lemma C.11 in three major steps.

Proof sketch of Lemma C.11. The proof of Lemma C.11 consists of three steps as we briefly describe as follows.

Firstly, we conduct step 1 to decompose tracking error. We decompose the tracking error $\left\|w_{\mathsf{V},t+1} - w^{\star}_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k}$ into a projection error term, an exponentially decaying term, a variance term, a bias error term, a fixed-point shift error term, and a slow drift error term. Next, we handle step 2 to bound these error terms. And finally, we have step 3 to recursively refine tracking error bound. We further show that the slow-drift error term diminishes as the tracking error diminishes. By recursively substituting the preliminary bound of $\left\|w_{\mathsf{V},t} - w^{\star}_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k}$ into the slow-drift term, we obtain the refined decay rate of the tracking error.

Step 1: Decomposing tracking error. We define the tracking error as $w_{\mathsf{V},t+1} - w^{\star}_{\mathsf{V}}$ under dictionary $\mathcal{D}_{\mathsf{V},t+1}$ and $\mathcal{D}^{\star}_{\mathsf{V}}$. Then, we bound the recursion of the tracking error as follows. For any $t \geq 0$, we derive

$$
\begin{aligned}
&\left\|w_{\mathsf{V},t+1} - w^{\star}_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k} \\
=& \left\|w_{\mathsf{V},t+1} - \overline{w}_{\mathsf{V},t+1} + \overline{w}_{\mathsf{V},t+1} - w^{\star}_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k} \\
\leq& \left\|w_{\mathsf{V},t+1} - \overline{w}_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k} + \left\|\overline{w}_{\mathsf{V},t+1} - w^{\star}_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k} \\
=& \left\|w_{\mathsf{V},t+1} - \overline{w}_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k} + \left\|w_{\mathsf{V},t} + \alpha^{\mathsf{v}}_t \hat{g}_t\left(w_{\mathsf{V},t}\right) - w^{\star}_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k} \\
=& \left\|w_{\mathsf{V},t+1} - \overline{w}_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k} + \left\|w_{\mathsf{V},t} - w^{\star}_{\mathsf{V},t} + \alpha^{\mathsf{v}}_t \hat{g}_t\left(w_{\mathsf{V},t}\right) + w^{\star}_{\mathsf{V},t} - w^{\star}_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k} \\
\leq& \left\|w_{\mathsf{V},t} - w^{\star}_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} + \left\|w_{\mathsf{V},t+1} - \overline{w}_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k} + \left(\alpha^{\mathsf{v}}_t\right)^2 \left\|\hat{g}_t\left(w_{\mathsf{V},t}\right)\right\|^2_{\mathcal{H}_k} + \left\|w^{\star}_{\mathsf{V},t} - w^{\star}_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k} \\
&+ 2\alpha^{\mathsf{v}}_t \langle \hat{g}_t\left(w_{\mathsf{V},t}\right), w_{\mathsf{V},t} - w^{\star}_{\mathsf{V},t}\rangle_{\mathcal{H}_k} + 2\alpha^{\mathsf{v}}_t \langle \hat{g}_t\left(w_{\mathsf{V},t}\right), w^{\star}_{\mathsf{V},t} - w^{\star}_{\mathsf{V},t+1}\rangle_{\mathcal{H}_k} \\
&+ \langle w_{\mathsf{V},t} - w^{\star}_{\mathsf{V},t}, w^{\star}_{\mathsf{V},t} - w^{\star}_{\mathsf{V},t+1}\rangle_{\mathcal{H}_k},
\end{aligned} \tag{149}
$$

where the first line holds with $\overline{w}_{\mathsf{V},t+1} = w_{\mathsf{V},t} + \alpha^{\mathsf{v}}_t \hat{g}_t\left(w_{\mathsf{V},t}\right)$ before sparsification and the second inequation comes from triangle inequality.

Step 2: Bounding each terms. In this part, we bound the five error terms.

Bounding the projection error term $\left\|w_{\mathsf{V},t+1} - \overline{w}_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k}$. According to Lemma C.4, we have

$$
\left\|\overline{w}_{t+1} - w_{t+1}\right\|^2_{\mathcal{H}_k} \leq M_k q^2 \varepsilon^2_{\mathsf{PV}}. \tag{150}
$$

Bounding the variance term $\left\|\hat{g}_t\left(w_{\mathsf{V},t}\right)\right\|^2_{\mathcal{H}_k}$. We bound $\left\|\hat{g}_t\left(w_{\mathsf{V},t}\right)\right\|^2_{\mathcal{H}_k}$ by first decomposing it into the sum of a sampling part and a population part as

$$
\begin{aligned}
\left\|\hat{g}_t\left(w_{\mathsf{V},t}\right)\right\|^2_{\mathcal{H}_k} &= \left\|\hat{g}_t\left(w_{\mathsf{V},t}\right) - g_t\left(w_{\mathsf{V},t}\right) + g_t\left(w_{\mathsf{V},t}\right)\right\|^2_{\mathcal{H}_k} \\
&\leq \left\|\hat{g}_t\left(w_{\mathsf{V},t}\right) - g_t\left(w_{\mathsf{V},t}\right)\right\|^2_{\mathcal{H}_k} + \left\|g_t\left(w_{\mathsf{V},t}\right)\right\|^2_{\mathcal{H}_k},
\end{aligned} \tag{151}
$$

where we use triangle inequality. According to Lemma C.5, the population gradient is bounded as

$$
\left\|g_t\left(w_{\mathsf{V},t}\right)\right\|^2_{\mathcal{H}_k} \leq C_1 \left\|w_{\mathsf{V},t} - w^{\star}_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k}, \qquad C_1 \coloneqq \left((1+\gamma)M_k + \lambda\right)^2. \tag{152}
$$

We next control the sampling term. Write the empirical gradient as

$$
\hat{g}_t\left(w_{\mathsf{V},t}\right) = \frac{1}{n}\sum_{\iota=1}^{n}\left[\left(\Gamma_s(\mathbf{s}_\iota) - \gamma\Gamma_{2s}(\mathbf{s}_\iota, \mathbf{s}'_\iota)\right) w_{\mathsf{V},t} - r(\mathbf{s}_\iota, \mathbf{a}_\iota)\psi(\mathbf{s}_\iota)\right], \tag{153}
$$

and define $B_t \coloneqq \Gamma_s - \gamma\Gamma_{2s}$, $\hat{B}_t \coloneqq \Gamma_s(\mathbf{s}_\iota) - \gamma\Gamma_{2s}(\mathbf{s}_\iota, \mathbf{s}'_\iota)$, $b_t \coloneqq \mathbb{E}[r(\mathbf{s}, \mathbf{a})\psi(\mathbf{s})]$, $\hat{b}_t \coloneqq r(\mathbf{s}_\iota, \mathbf{a}_\iota)\psi(\mathbf{s}_\iota)$. Then

$$\begin{aligned}\hat{g}_t\left(w_{\mathsf{V},t}\right) - g_t\left(w_{\mathsf{V},t}\right) &= (\hat{B}_t - B_t)w_{\mathsf{V},t} - (\hat{b}_t - b_t) \\ &= (\hat{B}_t - B_t)(w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}) + (\hat{B}_t - B_t)w^\star_{\mathsf{V},t} - (\hat{b}_t - b_t),\end{aligned} \tag{154}$$

and by triangle inequality and submultiplicativity of $\|\cdot\|_{\mathrm{op}}$,

$$\begin{aligned}\|\hat{g}_t\left(w_{\mathsf{V},t}\right) - g_t\left(w_{\mathsf{V},t}\right)\|^2_{\mathcal{H}_k} \leq 3\Big(&\|\hat{B}_t - B_t\|^2_{\mathrm{op}} \left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} \\ &+ \|\hat{B}_t - B_t\|^2_{\mathrm{op}}\|w^\star_{\mathsf{V},t}\|^2_{\mathcal{H}_k} + \|\hat{b}_t - b_t\|^2_{\mathcal{H}_k}\Big).\end{aligned} \tag{155}$$

Assume on-policy samples come from a geometric mixing Markov chain with effective sample size $n_{\mathrm{eff}} \asymp n/\tau_{\mathrm{mix}}$; using a Hilbert-space (operator-valued) Bernstein inequality, there exists a universal constant choice such that, with probability at least $1 - \delta$,

$$\|\hat{B}_t - B_t\|_{\mathrm{op}} \leq \Delta_B \coloneqq (1+\gamma)M_\Gamma \left(\sqrt{\frac{2\ln(6/\delta)}{n_{\mathrm{eff}}}} + \frac{2\ln(6/\delta)}{3n_{\mathrm{eff}}}\right), \tag{156}$$

$$\|\hat{b}_t - b_t\|_{\mathcal{H}_k} \leq \Delta_b \coloneqq \sqrt{M_\Gamma} \left(\sqrt{\frac{2\ln(6/\delta)}{n_{\mathrm{eff}}}} + \frac{2\ln(6/\delta)}{3n_{\mathrm{eff}}}\right), \tag{157}$$

where we define $M_\Gamma \coloneqq \sup \|\Gamma - \lambda\mathbf{I}\|_{\mathrm{op}}$.

Note that the $(1+\gamma)$ factor in Δ_B comes from $\hat{B}_t - B_t = (\hat{\Gamma}_s - \Gamma_s) - \gamma(\hat{\Gamma}_{2s} - \Gamma_{2s})$ and the triangle inequality. Besides, M_Γ is the almost-sure bound on the operator norm of each summand. And the $(2/3)$ factor appears in the standard Bernstein tail bound for bounded summands. Consequently, using $\|w^\star_{\mathsf{V},t}\|_{\mathcal{H}_k} \leq M_V$,

$$\begin{aligned}\|\hat{g}_t\left(w_{\mathsf{V},t}\right) - g_t\left(w_{\mathsf{V},t}\right)\|^2_{\mathcal{H}_k} \leq& 3\left(\Delta_B^2 \left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} + \Delta_B^2 M_V^2 + \Delta_b^2\right) \\ =& \frac{C_B}{n_{\mathrm{eff}}}\left(\left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} + M_V^2\right) + \frac{C_b}{n_{\mathrm{eff}}}.\end{aligned} \tag{158}$$

Armed with $r \in [0, 1]$, we can take

$$C_B = 6(1+\gamma)^2 M_\Gamma^2 \left(2\ln\frac{6}{\delta}\right) \quad \text{and} \quad C_b = 6M_\Gamma\left(2\ln\frac{6}{\delta}\right). \tag{159}$$

Finally, combining with the population bound gives

$$\begin{aligned}\|\hat{g}_t\left(w_{\mathsf{V},t}\right)\|^2_{\mathcal{H}_k} \leq& C_1\left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} + \frac{C_B}{n_{\mathrm{eff}}}\left(\left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} + M_V^2\right) + \frac{C_b}{n_{\mathrm{eff}}} \\ =& \left(C_1 + \frac{C_B}{n_{\mathrm{eff}}}\right)\left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} + \frac{C_B}{n_{\mathrm{eff}}}M_V^2 + \frac{C_b}{n_{\mathrm{eff}}}\end{aligned} \tag{160}$$

with $C_1 = ((1+\gamma)M_k + \lambda)^2$.

Bounding the slow drift error term $\langle w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}, w^\star_{\mathsf{V},t} - w^\star_{\mathsf{V},t+1}\rangle_{\mathcal{H}_k}$. In two time-scales, the Actor update satisfies $\|h_{t+1} - h_t\| \leq G_h\alpha_t$ for some bound G_h on the natural/vanilla policy gradient step (this follows from bounded score function and step-size choice). Hence

$$\|w^\star_{\mathsf{V},t} - w^\star_{\mathsf{V},t+1}\|_{\mathcal{H}_k} \leq L_{\mathsf{V}} G_h \alpha_t^{\mathsf{h}}. \tag{161}$$

By Cauchy–Schwarz and the Young inequality,

$$\begin{aligned}\langle w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}, w^\star_{\mathsf{V},t} - w^\star_{\mathsf{V},t+1}\rangle_{\mathcal{H}_k} &\leq \left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|_{\mathcal{H}_k}\left\|w^\star_{\mathsf{V},t} - w^\star_{\mathsf{V},t+1}\right\|_{\mathcal{H}_k} \\ &\leq \left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|_{\mathcal{H}_k} L_{\mathsf{V}} G_h \alpha_t^{\mathsf{h}} \\ &\leq \left(\left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} + 1\right) L_{\mathsf{V}} G_h \alpha_t^{\mathsf{h}}.\end{aligned} \tag{162}$$

Therefore, the slow drift error term, which is caused by the two-scale nature of the algorithm, produces a slow-drift penalty $O(\alpha_t^{\mathsf{h}})$ and diminishes as the tracking error diminishes.

Bounding the bias error term $\langle \hat{g}_t(w_{\mathsf{V},t}), w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\rangle_{\mathcal{H}_k}$. To begin with, we decompose the bias error term as

$$
\begin{aligned}
&\langle \hat{g}_t(w_{\mathsf{V},t}), w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\rangle_{\mathcal{H}_k} \\
\leq& \|\hat{g}_t(w_{\mathsf{V},t})\|_{\mathcal{H}_k} \left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|_{\mathcal{H}_k} \\
\leq& \frac{1}{2}\|\hat{g}_t(w_{\mathsf{V},t})\|^2_{\mathcal{H}_k} + \frac{1}{2}\left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} \\
\leq& \frac{1}{2}\left(\left(C_1 + \frac{C_B}{n_{\text{eff}}}\right)\left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} + \frac{C_B}{n_{\text{eff}}}M_V^2 + \frac{C_b}{n_{\text{eff}}}\right) + \frac{1}{2}\left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} \\
=& \frac{1}{2}\left(C_1 + \frac{C_B}{n_{\text{eff}}} + 1\right)\left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} + \frac{1}{2}\left(\frac{C_B}{n_{\text{eff}}}M_V^2 + \frac{C_b}{n_{\text{eff}}}\right),
\end{aligned} \tag{163}
$$

where the last inequality comes from the results in Eq. (160).

Bounding the fixed-point drift error term $\langle \hat{g}_t(w_{\mathsf{V},t}), w^\star_{\mathsf{V},t} - w^\star_{\mathsf{V},t+1}\rangle_{\mathcal{H}_k}$ and $\left\|w^\star_{\mathsf{V},t} - w^\star_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k}$. On the one hand, we bound $\langle \hat{g}_t(w_{\mathsf{V},t}), w^\star_{\mathsf{V},t} - w^\star_{\mathsf{V},t+1}\rangle_{\mathcal{H}_k}$ based on the results above as follows.

$$
\begin{aligned}
&\langle \hat{g}_t(w_{\mathsf{V},t}), w^\star_{\mathsf{V},t} - w^\star_{\mathsf{V},t+1}\rangle_{\mathcal{H}_k} \\
\leq& \|\hat{g}_t(w_{\mathsf{V},t})\|_{\mathcal{H}_k} \left\|w^\star_{\mathsf{V},t} - w^\star_{\mathsf{V},t+1}\right\|_{\mathcal{H}_k} \\
\leq& \frac{1}{2}\left(\|\hat{g}_t(w_{\mathsf{V},t})\|^2_{\mathcal{H}_k} + 1\right)\left\|w^\star_{\mathsf{V},t} - w^\star_{\mathsf{V},t+1}\right\|_{\mathcal{H}_k} \\
\leq& \frac{1}{2}\left(\left(C_1 + \frac{C_B}{n_{\text{eff}}}\right)\left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} + \frac{C_B}{n_{\text{eff}}}M_V^2 + \frac{C_b}{n_{\text{eff}}} + 1\right) L_{\mathsf{V}} G_h \alpha_t^{\mathsf{h}} \\
=& \frac{1}{2}L_{\mathsf{V}} G_h \alpha_t^{\mathsf{h}}\left(C_1 + \frac{C_B}{n_{\text{eff}}}\right)\left\|w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}\right\|^2_{\mathcal{H}_k} + \frac{1}{2}\left(\frac{C_B}{n_{\text{eff}}}M_V^2 + \frac{C_b}{n_{\text{eff}}} + 1\right) L_{\mathsf{V}} G_h \alpha_t^{\mathsf{h}},
\end{aligned} \tag{164}
$$

where the last line holds due to Eqs. (160) and (161).

On the other hand, we bound $\left\|w^\star_{\mathsf{V},t} - w^\star_{\mathsf{V},t+1}\right\|^2_{\mathcal{H}_k} \leq L_{\mathsf{V}}^2 G_h^2 (\alpha_t^{\mathsf{h}})^2$ directly from Eq. (161).

Summing up these results. Let $e_t \coloneqq w_{\mathsf{V},t} - w^\star_{\mathsf{V},t}$. Assume the regularised population operator $B_\lambda \coloneqq \Gamma_s - \gamma\Gamma_{2s} + \lambda\mathbf{I}$ is λ-strongly monotone, i.e., $\langle B_\lambda x, x\rangle \geq \lambda\|x\|^2$. Summing up the results in each term, we obtain, conditioning on $\mathcal{F}_{t-\tau_t}$,

$$
\begin{aligned}
&\mathbb{E}\left[\|e_{t+1}\|^2_{\mathcal{H}_\psi} \mid \mathcal{F}_{t-\tau_t}\right] \\
\leq& \|e_t\|^2_{\mathcal{H}_\psi} + M_k q^2 \varepsilon^2_{\mathsf{PV}} \\
&+ (\alpha_t^{\mathsf{v}})^2\left(\left(C_1 + \frac{C_B}{n_{\text{eff}}}\right)\|e_t\|^2_{\mathcal{H}_\psi} + \frac{C_B}{n_{\text{eff}}}M_V^2 + \frac{C_b}{n_{\text{eff}}}\right) + L_{\mathsf{V}}^2 G_h^2 (\alpha_t^{\mathsf{h}})^2 \\
&+ 2\alpha_t^{\mathsf{v}}\left(\frac{1}{2}\left(C_1 + \frac{C_B}{n_{\text{eff}}} + 1\right)\|e_t\|^2_{\mathcal{H}_\psi} + \frac{1}{2}\left(\frac{C_B}{n_{\text{eff}}}M_V^2 + \frac{C_b}{n_{\text{eff}}}\right)\right) \\
&+ 2\alpha_t^{\mathsf{v}}\left(\frac{1}{2}L_{\mathsf{V}} G_h \alpha_t^{\mathsf{h}}\left(C_1 + \frac{C_B}{n_{\text{eff}}}\right)\|e_t\|^2_{\mathcal{H}_\psi} + \frac{1}{2}\left(\frac{C_B}{n_{\text{eff}}}M_V^2 + \frac{C_b}{n_{\text{eff}}} + 1\right) L_{\mathsf{V}} G_h \alpha_t^{\mathsf{h}}\right) \\
&+ \left(\|e_t\|^2_{\mathcal{H}_\psi} + 1\right) L_{\mathsf{V}} G_h \alpha_t^{\mathsf{h}} \\
=& \left(1 - 2\lambda\alpha_t^{\mathsf{v}} + c_B(\alpha_t^{\mathsf{v}})^2 + c_\times \alpha_t^{\mathsf{v}}\alpha_t^{\mathsf{h}} + c_M \alpha_t^{\mathsf{h}}\right)\|e_t\|^2_{\mathcal{H}_\psi} + B_t,
\end{aligned} \tag{165}
$$

where the coefficients are collected as

$$c_B := C_1 + \frac{C_B}{n_{\text{eff}}}, \qquad c_\times := \left(C_1 + \frac{C_B}{n_{\text{eff}}}\right) L_{\mathsf{V}} G_h, \qquad c_M := L_{\mathsf{V}} G_h,$$

$$\begin{aligned} B_t :=& M_k q^2 \varepsilon_{\mathsf{PV}}^2 + (\alpha_t^{\mathsf{v}} + (\alpha_t^{\mathsf{v}})^2) \left(\frac{C_B}{n_{\text{eff}}} M_V^2 + \frac{C_b}{n_{\text{eff}}}\right) \\ &+ \alpha_t^{\mathsf{v}} \alpha_t^{\mathsf{h}} L_{\mathsf{V}} G_h \left(\frac{C_B}{n_{\text{eff}}} M_V^2 + \frac{C_b}{n_{\text{eff}}} + 1\right) + L_{\mathsf{V}} G_h \alpha_t^{\mathsf{h}} + L_{\mathsf{V}}^2 G_h^2 (\alpha_t^{\mathsf{h}})^2. \end{aligned}$$

Define the contraction coefficient

$$\eta_t := 2\lambda \alpha_t^{\mathsf{v}} - c_B (\alpha_t^{\mathsf{v}})^2 - c_\times \alpha_t^{\mathsf{v}} \alpha_t^{\mathsf{h}} - c_M \alpha_t^{\mathsf{h}}.$$

Hence, Eq. (165) becomes the standard supermartingale-type recursion

$$\begin{aligned} \mathbb{E}\left[\|e_{t+1}\|^2 \mid \mathcal{F}_{t-\tau_t}\right] \leq& (1-\eta_t) \|e_t\|_{\mathcal{H}_\psi}^2 + B_t \\ \leq& \Pi_{i=\hat{t}}^t (1-\eta_i) \|e_{\hat{t}}\|_{\mathcal{H}_\psi}^2 + \sum_{i=\hat{t}}^t \left[\Pi_{j=i+1}^t (1-\eta_j)\right] B_i, \end{aligned} \tag{166}$$

where we have used that η_i, B_i are $\mathcal{F}_{i-\tau_i}$-measurable (hence independent of the inner conditional expectation at step i). We now bound the two terms on the right-hand side.

Assume the two time-scale stepsizes $\alpha_t^{\mathsf{v}} = \frac{C_{\mathsf{v}}}{(t+1)^\nu}$, $\alpha_t^{\mathsf{h}} = \frac{C_{\mathsf{h}}}{(t+1)^\sigma}$ with $0 < \nu < \sigma \leq 1$. By definition,

$$\eta_t = 2\lambda \alpha_t^{\mathsf{v}} - c_B (\alpha_t^{\mathsf{v}})^2 - c_\times \alpha_t^{\mathsf{v}} \alpha_t^{\mathsf{h}} - c_M \alpha_t^{\mathsf{h}} = \frac{2\lambda C_{\mathsf{v}}}{(t+1)^\nu} - \frac{c_B C_{\mathsf{v}}^2}{(t+1)^{2\nu}} - \frac{c_\times C_{\mathsf{v}} C_{\mathsf{h}}}{(t+1)^{\nu+\sigma}} - \frac{c_M C_{\mathsf{h}}}{(t+1)^\sigma}.$$

Hence, there exists $T_0 < \infty$ and $\tilde{c} \in (0, \lambda C_{\mathsf{v}}]$ such that for all $t \geq T_0$,

$$\eta_t \geq \frac{\tilde{c}}{(t+1)^\nu}. \tag{167}$$

Using $\log(1-x) \leq -x$ for $x \in (0,1)$ and Eq. (167),

$$\prod_{i=\hat{t}}^{t-1} (1-\eta_i) \leq \exp\left(-\sum_{i=\hat{t}}^{t-1} \eta_i\right) \leq \exp\left(-\tilde{c} \sum_{i=\hat{t}}^{t-1} \frac{1}{(i+1)^\nu}\right). \tag{168}$$

If $0 < \nu < 1$, then $\sum_{i=\hat{t}}^{t-1} (i+1)^{-\nu} \geq \frac{(t+1)^{1-\nu} - (\hat{t}+1)^{1-\nu}}{1-\nu}$, so

$$\prod_{i=\hat{t}}^{t-1} (1-\eta_i) \leq \exp\left(-\frac{\tilde{c}}{1-\nu}\left[(t+1)^{1-\nu} - (\hat{t}+1)^{1-\nu}\right]\right). \tag{169}$$

When $\nu = 1$, we obtain $\prod_{i=\hat{t}}^{t-1} (1-\eta_i) \leq \left(\frac{\hat{t}+1}{t+1}\right)^{\tilde{c}}$. In both cases, the product decays at least polynomially.

Write B_i in the separated form

$$\begin{aligned} B_i =& \underbrace{M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \tfrac{C_b}{n_{\text{eff}}}}_{b_0} + \underbrace{\frac{C_B}{n_{\text{eff}}} M_V^2}_{b_1} \alpha_i^{\mathsf{v}} + \underbrace{\left(\frac{C_B}{n_{\text{eff}}} \frac{C_B}{n_{\text{eff}}} M_V^2 + \tfrac{C_b}{n_{\text{eff}}} + 1\right) L_{\mathsf{V}} G_h}_{b_2} \alpha_i^{\mathsf{v}} \alpha_i^{\mathsf{h}} + \underbrace{L_{\mathsf{V}} G_h}_{b_3} \alpha_i^{\mathsf{h}} \\ &+ \underbrace{\left(\frac{C_B}{n_{\text{eff}}} M_V^2 + \tfrac{C_b}{n_{\text{eff}}}\right)}_{b_4} (\alpha_i^{\mathsf{v}})^2 + \underbrace{L_{\mathsf{V}}^2 G_h^2}_{b_5} (\alpha_i^{\mathsf{h}})^2. \end{aligned} \tag{170}$$

Define the weights $W_{i,t} := \prod_{j=i+1}^{t-1} (1-\eta_j)$. Using the same argument as in Eq. (169),

$$W_{i,t} \leq \exp\left(-\frac{\tilde{c}}{1-\nu}\left[(t+1)^{1-\nu} - (i+1)^{1-\nu}\right]\right). \tag{171}$$

Next, we ignore constant factors and focus solely on the decay induced by the learning rate. For ease of exposition, we take the step size to follow the schedule $(i+1)^{-\rho}$.

We will repeatedly use the following calculus-type estimate (change variable $u = (i+1)^{1-\nu}$, compare sum with integral): for any $\rho > 0$ there exist finite constants $C_1(\rho), C_2(\rho)$ (independent of t) such that for $\rho > \nu$

$$\sum_{i=\hat{t}}^{t-1} W_{i,t} \frac{1}{(i+1)^{\rho}} \le C_1(\rho) \exp\left(-\frac{\tilde{c}}{2(1-\nu)}\left[(t+1)^{1-\nu} - (\hat{t}+1)^{1-\nu}\right]\right) + \frac{C_2(\rho)}{(t+1)^{\rho-\nu}}), \quad (172)$$

and for $\rho = \nu$ the second term becomes $\frac{K_2(\nu)\log(t+1)}{(t+1)^{\epsilon}}$ for any fixed $\epsilon \in (0, \nu)$ (or equivalently a log-loss). Applying Eq. (172) to each summand in B_i with

$$\alpha_i^{\mathsf{v}} = \frac{C_{\mathsf{v}}}{(i+1)^{\nu}},\ \alpha_i^{\mathsf{h}} = \frac{C_{\mathsf{h}}}{(i+1)^{\sigma}},\ \alpha_i^{\mathsf{v}}\alpha_i^{\mathsf{h}} = \frac{C_{\mathsf{v}}C_{\mathsf{h}}}{(i+1)^{\nu+\sigma}},\ (\alpha_i^{\mathsf{v}})^2 = \frac{C_{\mathsf{v}}^2}{(i+1)^{2\nu}},\ (\alpha_i^{\mathsf{h}})^2 = \frac{C_{\mathsf{h}}^2}{(i+1)^{2\sigma}},$$

and noting $0 < \nu < \sigma$, we obtain that there exists constants $\bar{C}_0, \ldots, \bar{C}_5 < \infty$ (depending only on $\tilde{c}, \nu, \sigma, C_{\mathsf{h}}, C_{\mathsf{v}}$ and b_ℓ's) such that

$$\begin{aligned}\sum_{i=\hat{t}}^{t-1} W_{i,t} B_i &\le \bar{C}_0 \exp\left(-\frac{\tilde{c}}{2(1-\nu)}\left[(t+1)^{1-\nu} - (\hat{t}+1)^{1-\nu}\right]\right) \\ &\quad + \frac{\bar{C}_1}{(t+1)^{\nu}} + \frac{\bar{C}_2}{(t+1)^{\nu+\sigma-\nu}} + \frac{\bar{C}_3}{(t+1)^{\sigma-\nu}} + \frac{\bar{C}_4}{(t+1)^{2\nu-\nu}} + \frac{\bar{C}_5}{(t+1)^{2\sigma-\nu}} + b_0^{\star} \\ &\le \bar{C}_0 \exp\left(-\frac{\tilde{c}}{2(1-\nu)}\left[(t+1)^{1-\nu} - (\hat{t}+1)^{1-\nu}\right]\right) + \frac{C_6}{(t+1)^{\nu}} + b_0^{\star}, \end{aligned} \quad (173)$$

where in the last inequality we use $\sigma > \nu$ so that the slowest-decaying polynomial term is $(t+1)^{-\nu}$, and we set $b_0^{\star} := M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \frac{C_b}{n_{\text{eff}}}$ (the non-vanishing bias floor).

Plugging Eq. (169) and Eq. (173) into Eq. (166) yields, for $t \ge T_0$,

$$\begin{aligned}\mathbb{E}\left[\|e_t\|_{\mathcal{H}_\psi}^2\right] &\le \exp\left(-\frac{\tilde{c}}{1-\nu}\left[(t+1)^{1-\nu} - (\hat{t}+1)^{1-\nu}\right]\right)\mathbb{E}\|e_{\hat{t}}\|_{\mathcal{H}_\psi}^2 \\ &\quad + \bar{C}_0 \exp\left(-\frac{\tilde{c}}{2(1-\nu)}\left[(t+1)^{1-\nu} - (\hat{t}+1)^{1-\nu}\right]\right) + \frac{C_6}{(t+1)^{\nu}} + b_0^{\star}.\end{aligned}$$

As a result, we conclude that there exist finite constants $\widetilde{C}$ and $\widetilde{C}_b$ such that

$$\mathbb{E}\left[\|e_t\|_{\mathcal{H}_\psi}^2\right] \le \frac{\widetilde{C}}{(t+1)^{\nu}} + M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \frac{\widetilde{C}_b}{n_{\text{eff}}}, \qquad t \ge T_0,$$

which is the decay bound under the two time-scale stepsizes with $\sigma > \nu$. The borderline case $\nu = 1$ follows similarly, replacing Eq. (169) and Eq. (172) by their logarithmic counterparts, incurring at most a polylogarithmic loss.

Step 3: Recursively refining tracking error bound Starting from the one–step supermartingale recursion

$$\mathbb{E}\left[\|e_{t+1}\|_{\mathcal{H}_\psi}^2 \mid \mathcal{F}_{t-\tau_t}\right] \le (1-\eta_t)\|e_t\|_{\mathcal{H}_\psi}^2 + B_t + \underbrace{2\langle e_t, w_{\mathsf{V},t}^{\star} - w_{\mathsf{V},t+1}^{\star}\rangle_{\mathcal{H}_k}}_{\text{slow-drift term}}. \quad (174)$$

Ignoring the slow-drift term in Eq. (174) and unrolling via the discrete Grönwall inequality, we obtain

$$\mathbb{E}\left[\|e_t\|_{\mathcal{H}_\psi}^2\right] \le \exp\left(-\tfrac{\tilde{c}}{1-\nu}\left[(t+1)^{1-\nu} - (\hat{t}+1)^{1-\nu}\right]\right)\mathbb{E}\|e_{\hat{t}}\|_{\mathcal{H}_\psi}^2 + \sum_{i=\hat{t}}^{t-1} W_{i,t}\mathbb{E}\left[B_i\right], \quad (175)$$

where

$$W_{i,t} = \prod_{j=i+1}^{t-1}(1-\eta_j) \le \exp\left(-\tfrac{\tilde{c}}{1-\nu}[(t+1)^{1-\nu} - (i+1)^{1-\nu}]\right).$$

B_i is built from linear, quadratic, and cross terms of α_i^{v} and α_i^{h}, plus a bias floor. We estimate the sums by comparing them with integrals via the substitution $u = (i+1)^{1-\nu}$. For large t, this gives finite constants D_0, D_b that do not depend on t as follows.

$$\mathbb{E}\left[\|e_t\|_{\mathcal{H}_\psi}^2\right] \le \frac{D_0}{(t+1)^\nu} + M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \frac{D_b}{n_{\text{eff}}}, \tag{176}$$

i.e., an initial polynomial decay at rate $(t+1)^{-\nu}$ up to the irreducible bias floor.

By Lipschitz continuity of the population fixed point in the Actor parameter on the slow time-scale, there exists $L_{\mathsf{V}} < \infty$ such that

$$\|w_{\mathsf{V},t+1}^\star - w_{\mathsf{V},t}^\star\|_{\mathcal{H}_\psi} \le L_{\mathsf{V}} \|h_{t+1} - h_t\| \le L_{\mathsf{V}} G_h \alpha_t^{\mathsf{h}} = \frac{L_{\mathsf{V}} G_h C_\alpha}{(t+1)^\sigma}. \tag{177}$$

Conditioning on $\mathcal{F}_{t-\tau_t}$ and using Cauchy–Schwarz,

$$\begin{aligned}\mathbb{E}\big[2\langle e_t, w_{\mathsf{V},t}^\star - w_{\mathsf{V},t+1}^\star\rangle_{\mathcal{H}_k} \mid \mathcal{F}_{t-\tau_t}\big] &\le 2\sqrt{\mathbb{E}\left[\|e_t\|_{\mathcal{H}_\psi}^2\right]} \cdot \frac{L_{\mathsf{V}} G_h C_\alpha}{(t+1)^\sigma} \\ &\le \frac{C_{\text{sd}}}{(t+1)^{\sigma+\nu/2}} + \frac{C_{\text{sd}}}{(t+1)^\sigma}\left(M_k^{1/2} q \varepsilon_{\mathsf{PV}} + n_{\text{eff}}^{-1/2}\right),\end{aligned} \tag{178}$$

where Eq. (176) was used in the last inequality.

Returning to Eq. (174), taking total expectation and unrolling as before yields Eq. (175) with an augmented noise term

$$\tilde{B}_t \coloneqq B_t + \frac{C_{\text{sd}}}{(t+1)^{\sigma+\nu/2}} + \frac{C_{\text{sd}}}{(t+1)^\sigma}\left(M_k^{1/2} q \varepsilon_{\mathsf{PV}} + n_{\text{eff}}^{-1/2}\right).$$

Applying the same weighted-sum estimate gives

$$\mathbb{E}\left[\|e_t\|_{\mathcal{H}_\psi}^2\right] \le \frac{D_1}{(t+1)^\nu} + \frac{D_2}{(t+1)^{\sigma-\nu}} + \frac{D_3}{(t+1)^{\sigma+\nu/2-\nu}} + M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \frac{D_b'}{n_{\text{eff}}}. \tag{179}$$

- **Fast slow–time scale:** $\sigma > \frac{3}{2}\nu$**.**

$$\mathbb{E}\left[\|e_t\|_{\mathcal{H}_\psi}^2\right] \le \frac{C}{(t+1)^\nu} + M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \frac{C_b}{n_{\text{eff}}}.$$

- **Critical case:** $\sigma = \frac{3}{2}\nu$**.**

$$\mathbb{E}\left[\|e_t\|_{\mathcal{H}_\psi}^2\right] \le \frac{C_{\text{crit}} \log^2(t+1)}{(t+1)^\nu} + M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \frac{C_b}{n_{\text{eff}}}.$$

- **Slow slow–time scale:** $\nu < \sigma < \frac{3}{2}\nu$**.**

$$\mathbb{E}\left[\|e_t\|_{\mathcal{H}_\psi}^2\right] \le \frac{C}{(t+1)^{2(\sigma-\nu)}} + M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \frac{C_b}{n_{\text{eff}}}.$$

Combining the above cases, we obtain the unified bound for large t as follows. Let $\nu \in (0,1)$ and $\sigma > \nu$ be the Critic and Actor step-size exponents, respectively. The tracking error satisfies

$$\mathbb{E}\left[\left\|w_{\mathsf{V},t} - w_{\mathsf{V},t}^\star\right\|_{\mathcal{H}_\psi}^2\right] = \mathbb{E}\left[\|e_t\|_{\mathcal{H}_\psi}^2\right] \le \begin{cases} \frac{C}{(t+1)^\nu} + M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \frac{C_b}{n_{\text{eff}}}, & \sigma > \frac{3}{2}\nu, \\ \frac{C_{\text{crit}} \log^2(t+1)}{(t+1)^\nu} + M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \frac{C_b}{n_{\text{eff}}}, & \sigma = \frac{3}{2}\nu, \\ \frac{C}{(t+1)^{2(\sigma-\nu)}} + M_k q^2 \varepsilon_{\mathsf{PV}}^2 + \frac{C_b}{n_{\text{eff}}}, & \nu < \sigma < \frac{3}{2}\nu, \end{cases} \tag{180}$$

for some finite constants C, C_{crit}, and C_b independent of t.

D Additional Experiments

We evaluate on three continuous-control environments from Gymnasium, including `Pendulum-v1` and `BipedalWalker-v3`.

D.1 Details for Environments

We use the default observation and action spaces and keep the environment settings consistent across methods. A concise summary of each environment is given in Tables 3–5.

Table 3: `Pendulum-v1` summary

Observation	$(\cos\theta,\ \sin\theta,\ \dot{\theta}) \in [-1,1] \times [-1,1] \times [-8,8]$
Action	Torque $\tau \in [-2,2]$ (1D)
Reward	$-(\theta^2 + 0.1\dot{\theta}^2 + 0.001\tau^2)$
Horizon	200 steps

Table 4: `BipedalWalker-v3` summary

Observation	24D proprioception + 10 lidar ranges
Action	Motor speeds in $[-1,1]^4$
Reward	Forward progress (small torque cost); -100 on fall
Horizon	Up to $\sim$1600 steps or on failure

Table 5: `Ant-v5` summary

Observation	105D state (proprioception, contacts, joint angles/velocities)
Action	8D joint torques in $[-1,1]^8$
Reward	Forward progress + alive bonus − control and contact costs
Horizon	1000 steps or termination on fall

D.2 Hyper-parameters

The hyper-parameters configured in experiments are summarized in Tables 6–8.

D.3 Results of `BipedalWalker-v3`

Evaluation on Efficiency. In the continuous control environment with multi-dimensional actions (`BipedalWalker-v3`), we compare two variants of RSA2C (RSA2C-CME and RSA2C-KME), our Advanced AC framework without SHAP, and the classical RKHS-AC. As shown in Figure 5, Advanced AC exhibits faster improvement in the early stage, but suffers from larger variance and slightly lower returns in the later phase compared to RSA2C. In contrast, both RSA2C variants achieve more stable trajectories and higher final returns, indicating that the main benefit of SHAP lies in stabilizing training and enhancing long-term convergence rather than accelerating the initial learning. Among them, CME yields smoother curves than KME.

Figure 6 compares RSA2C with the deep RL algorithms SAC and PPO. SAC reaches high returns quickly but fluctuates substantially throughout training, while PPO shows slower and less stable progress. The RSA2C variants eventually catch up and achieve comparable or superior asymptotic performance despite their lightweight non-neural architecture. This confirms that RSA2C can remain competitive with deep RL even in more challenging multi-dimensional action spaces.

Regarding computational complexity illustrated Table 9, RSA2C requires approximately 1.427 GFLOPs for CME setting and 0.381 GFLOPs for KME setting, which is higher than Advanced AC with 0.346 GFLOPs and RKHS-AC with 0.264 GFLOPs, but all methods remain within the same millisecond runtime scale. We can found similar trend on runtime. Specifically, CME increases theoretical FLOPs, while the empirical runtime grows by approximately 6.9%. This discrepancy arises because FLOPs measure the theoretical number of floating-point operations, ignoring parallelized matrix computations in practice. Importantly, both RSA2C variants remain orders of magnitude faster and lighter than SAC, which rely on deep neural networks and require far more computation even when executed on GPU.

Table 6: Hyper-parameters on `Pendulum-v1`

Description	Value
Epoch	2000
Discount factor γ	0.99
Initial Σ	$0.35\mathbf{I}$
Final Σ	$0.25\mathbf{I}$
Variance of RBF	0.8
Max size of dictionary $\mathcal{D}_{\mathsf{V}}$	384
Max size of dictionary $\mathcal{D}_{\mathsf{A}}$	384
Episodes of evaluation	5
Learning rate of Value Critic ν	0.01
Learning rate of Actor σ	1.0

Table 7: Hyper-parameters on `BipedalWalker-v3`

Description	Value
Epoch	1500
Discount factor γ	0.99
Initial Σ	$0.9\mathbf{I}$
Final Σ	$0.4\mathbf{I}$
Variance of RBF kernel	3.0
Max size of dictionary $\mathcal{D}_{\mathsf{V}}$	1024
Max size of dictionary $\mathcal{D}_{\mathsf{A}}$	1024
Eval episodes	5
Learning rate of Value Critic ν	0.005
Learning rate of Actor σ	0.08

Evaluation on Stability. We evaluate RSA2C-KME and RSA2C-CME on `BipedalWalker-v3` under zero-mean state noise with varying variance, illustrated in Table 10. Across nonzero noise levels, RSA2C-CME shows markedly lower variability, with standard deviations from 24.75 to 43.06, whereas RSA2C-KME ranges from 46.74 to 56.03. For the mean return, the two variants are similar at light noise 0 and 0.001, and CME outperforms KME at higher noise 0.005 and 0.01, yielding 264.59 versus 242.89 and 254.85 versus 242.05, respectively. Both variants exceed the Advanced AC baseline of 229.89 without perturbation. We attribute CME's superior stability to its explicit modeling of feature dependencies, which enables adaptive reweighting under noise and mitigates error amplification from single-dimension dominance, whereas KME ignores correlations and is more vulnerable when noise strikes dominant features.

Evaluation on Interpretability. In the `BipedalWalker-v3` environment, we compare the feature importance of KME and CME using beeswarm and heatmap visualizations, as shown in Fig. 7. The SHAP distribution of KME is generally small and concentrated, with most values near zero except for a few dimensions such as vel x, vel y, and joint speeds, exhibiting a sparse and nearly fixed pattern. Its heatmap remains largely yellow in the later training stages, indicating weak temporal dynamics. In contrast, CME demonstrates a wider SHAP range with both positive and negative contributions, consistently capturing the effects of hip and knee joints and their velocities, body posture (hull angle and angle speed), and leg contacts across multiple dimensions. Moreover, several key features are gradually reinforced in the second half of training, reflecting stronger temporal adaptability. This richer structure is also consistent with the noise-perturbation results, where SAC and PPO exhibit substantially larger fluctuations across all noise scales, whereas RSA2C maintains stable performance. Overall, compared with KME, CME provides richer, direction-sensitive, and training-evolving explanatory signals, with a greater emphasis on the coupling mechanisms among state features.

D.4 Results of `Ant-v5`

Evaluation on Efficiency. In the high-dimensional control task `Ant-v5`, the behavior of RSA2C differs markedly from its performance on lower-dimensional environments. As shown in Figure 8, all

Table 8: Hyper-parameters on `Ant-v5`

Description	Value
Epoch	2000
Discount factor γ	0.99
Initial Σ	$\mathbf{I}$
Final Σ	$0.3\mathbf{I}$
Variance of RBF kernel	10.0
Max size of dictionary $\mathcal{D}_{\mathsf{V}}$	4096
Max size of dictionary $\mathcal{D}_{\mathsf{A}}$	4096
Eval episodes	5
Learning rate of Value Critic ν	0.05
Learning rate of Actor σ	0.25

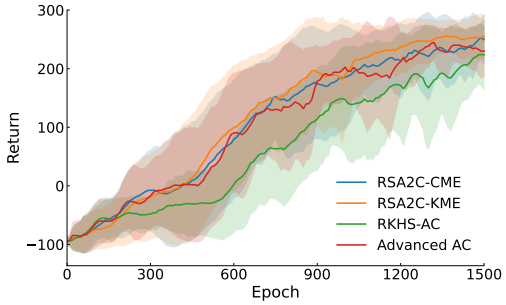


Figure 5: Ablation study on BipedalWalker-v3.

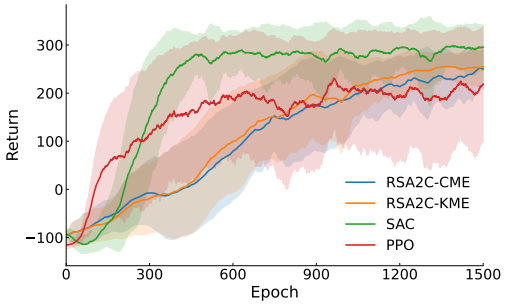


Figure 6: Performance on BipedalWalker-v3.

Table 9: FLOPs and Runtime (ms) of `BipedalWalker-v3`

	RKHS-AC	Advanced AC	RSA2C-KME	RSA2C-CME	SAC (GPU)	PPO (GPU)
GFLOPs	0.264	0.346	0.381	1.427	14.057	0.741
Runtime	778.109	1305.472	1354.726	1448.258	23982.756	8119.774

Table 10: Stability under different scales of noise variance of `BipedalWalker-v3`

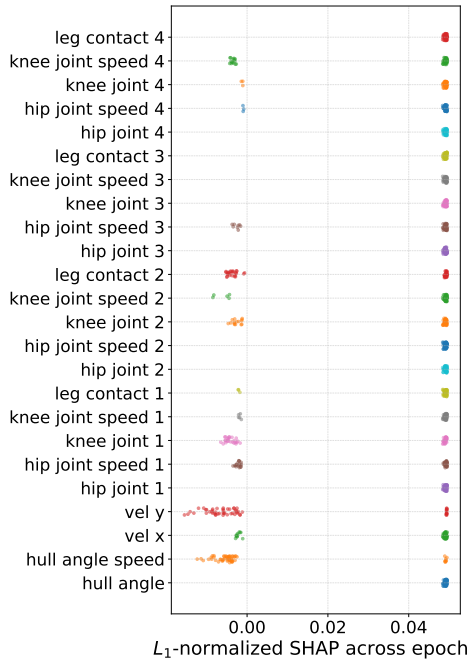
	0	0.001	0.005	0.01
RSA2C-KME	255.09 ± 37.19	249.72 ± 56.03	242.89 ± 54.94	242.05 ± 46.74
RSA2C-CME	249.72 ± 41.47	246.65 ± 43.06	264.59 ± 24.75	254.85 ± 27.31
SAC	277.59 ± 93.18	292.42 ± 57.25	299.98 ± 37.43	234.80 ± 121.49
PPO	241.41 ± 135.30	174.59 ± 129.96	259.03 ± 96.26	136.56 ± 169.12

Table 11: FLOPs and Runtime (ms) of `Ant-v5`

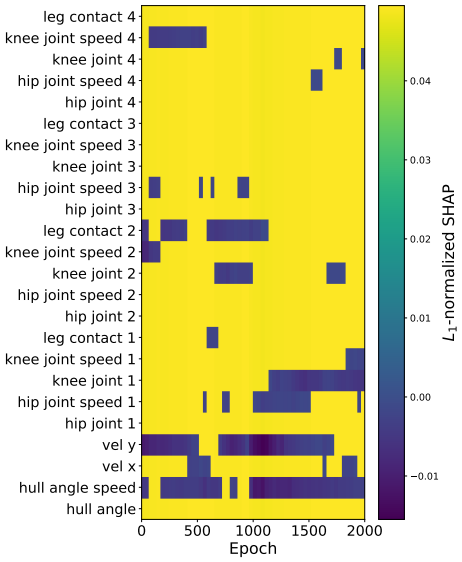
	RKHS-AC	Advanced AC	RSA2C-KME	RSA2C-CME	SAC (GPU)	PPO (GPU)
GFLOPs	0.5499	1.097	2.154	3.853	29.281	2.273
Runtime	826.865	1210.891	1358.706	1700.358	9623.472	1183.011

RKHS-based variants converge to a similar performance plateau around 1000. This indicates that RSA2C remains stable but becomes trapped in a local optimum when the state dimensionality is large and the underlying dynamics are highly nonlinear. Figure 9 shows that deep RL algorithms are able to discover significantly better solutions in high-dimensional state space. SAC continues to improve and surpasses 4000 return, clearly outperforming all RKHS-based methods. The limited gap among variants further suggests that, under such complexity, the linear RKHS function approximation combined with RBF kernels cannot fully capture the richer structure required for escaping suboptimal regions, thus diminishing the relative advantage of SHAP-guided feature weighting.

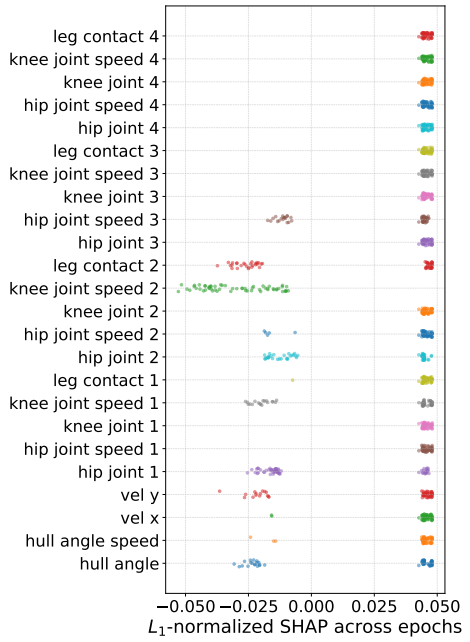
Table 11 summarizes the computational cost. RSA2C-KME (2.154 GFLOPs) and RSA2C-CME (3.853 GFLOPs) require more computation than RKHS-AC and Advanced AC, but they remain



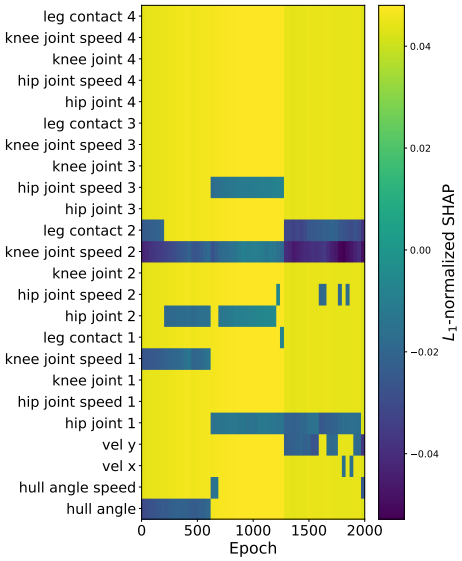
(a) Beeswarm under KME



(b) Heatmap under KME



(c) Beeswarm under CME



(d) Heatmap under CME

Figure 7: Visualization on interpretability of RSA2C on `BipedalWalker-v3`.

substantially lighter than deep RL baselines. SAC demands 29.28 GFLOPs and over 9600 ms per iteration on GPU, whereas PPO requires 2.27 GFLOPs but still incurs higher runtime than RSA2C-KME. Although CME is more costly than KME, both RSA2C variants maintain a favorable efficiency–stability trade-off. Importantly, the computational results reinforce the earlier observation that RSA2C is stable and lightweight, but its linear-RKHS formulation limits representational capacity in high-dimensional dynamics.

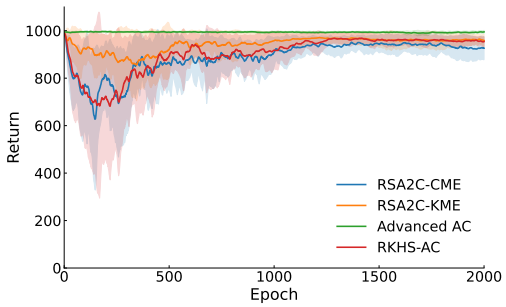


Figure 8: Ablation study on Ant-v5.

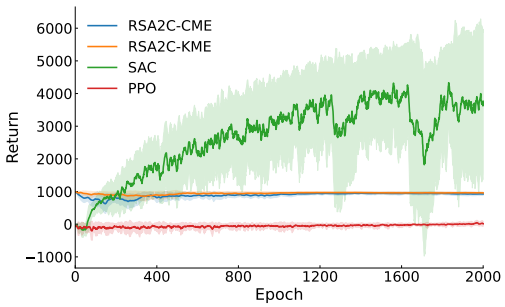


Figure 9: Performance on Ant-v5.

Table 12: Stability under different scales of noise variance of Ant-v5

	0	0.001	0.005	0.01
RSA2C-KME	960.36 ± 22.09	957.37 ± 13.85	954.13 ± 23.67	958.54 ± 21.90
RSA2C-CME	934.15 ± 44.39	953.84 ± 17.65	948.86 ± 26.02	959.65 ± 21.86
SAC	4477.33 ± 1950.69	3761.98 ± 2276.18	4067.61 ± 1410.60	2619.72 ± 1516.74
PPO	17.34 ± 45.12	155.64 ± 260.74	32.47 ± 65.07	-11.15 ± 32.11

Evaluation on Stability. Table 12 reports the robustness of different algorithms under zero-mean Gaussian perturbations with varying noise variance. Both RSA2C-KME and RSA2C-CME remain remarkably stable across all noise levels, with returns consistently around 940–960 and standard deviations below 45, indicating that the RKHS-based linear approximation maintains strong resistance to state corruption even in high-dimensional tasks. In contrast, SAC exhibits extremely large variance at every noise setting, with fluctuations exceeding 1400–2200, reflecting high sensitivity of deep RL. PPO is even more unstable. Although it performs moderately under small noise (variance 0.001), its return collapses toward zero or even negative values under larger perturbations.

Evaluation on Interpretability. Figure 10 visualizes the SHAP-based feature attributions of KME and CME using beeswarm and temporal heatmaps. For RSA2C-KME, the SHAP scores are highly sparse and concentrated around zero, with only a few contact-related dimensions exhibiting noticeable contributions. Moreover, the temporal heatmap under RSA2C-KME remains mostly uniform and lacks clear structure, indicating that the learned importance is nearly static throughout training. This suggests that RSA2C-KME fails to capture the intricate dependencies among high-dimensional proprioceptive and contact features, resulting in weak and non-evolving explanatory signals. In contrast, RSA2C-CME yields substantially richer and more diverse attributions. The beeswarm plot reveals distinct positive and negative contributions across many joint, contact, and orientation dimensions. The heatmap further shows clear temporal patterns that key contact forces and ankle/hip interactions activate selectively at different training stages, and several features—such as joint torques, angular velocities, and contact normals—exhibit gradually strengthened importance. These structured and evolving patterns indicate that RSA2C-CME successfully models feature correlations and adapts attribution as the policy improves.

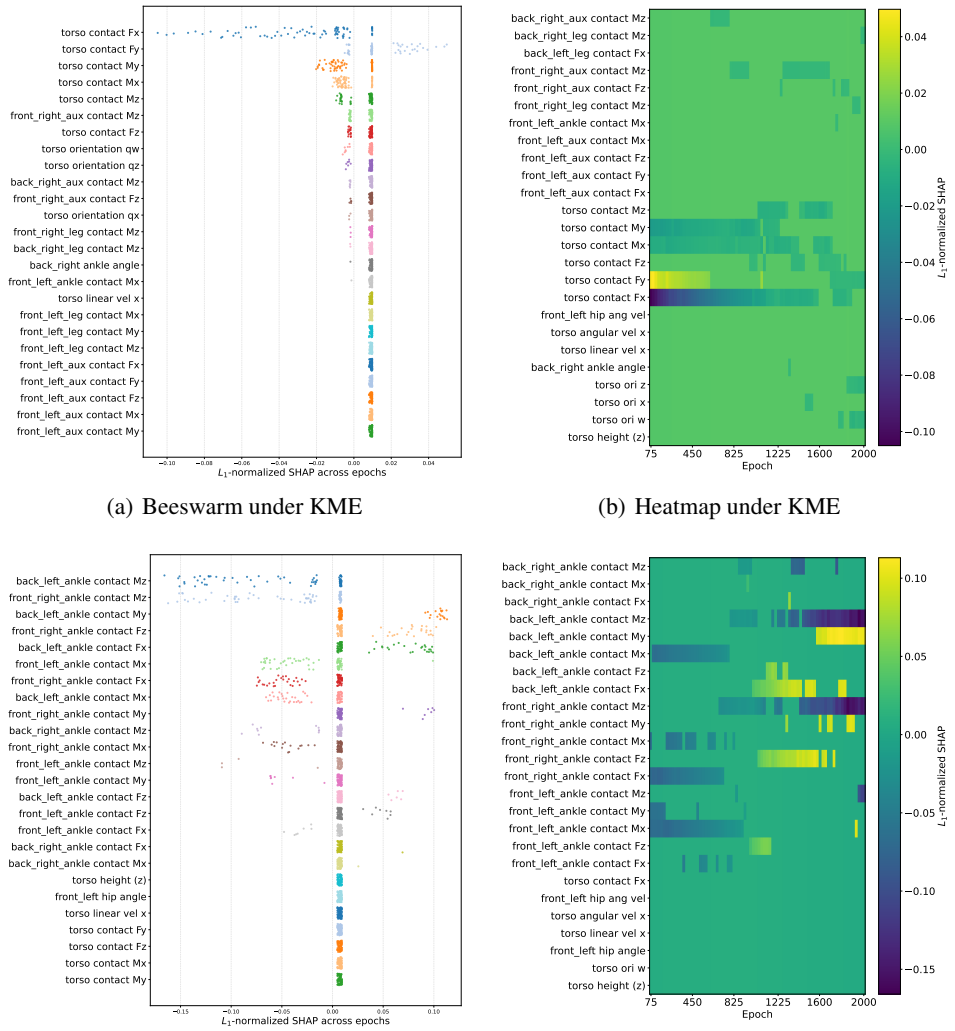
D.5 Simulation on Theoretical non-asymptotic rate

D.5.1 Details for LQR

To numerically examine the convergence guarantees in a controlled setting, we introduce a discounted linear–quadratic regulator (LQR) environment constructed from the linearization of the classic cart–pole (inverted pendulum) around the upright equilibrium. The continuous-time dynamics are linearized and then discretized with step size $\Delta t = 0.02$, yielding a four-dimensional state $\mathbf{s}_t = [x, \dot{x}, \theta, \dot{\theta}]^\top$ and a scalar control input $\mathbf{a}_t$. Specifically, x_t and $\dot{x}_t$ are the cart position and velocity, and θ_t and $\dot{\theta}_t$ are the pole angle (measured from the upright) and angular velocity, respectively.

The instantaneous quadratic cost is given by

$$r(\mathbf{s}_t, \mathbf{a}_t) = \mathbf{s}_t^\top P_1 \mathbf{s}_t + \mathbf{a}_t^\top P_2 \mathbf{a}_t,$$



(a) Beeswarm under KME

(b) Heatmap under KME

(c) Beeswarm under CME

(d) Heatmap under CME

Figure 10: Visualization on interpretability of RSA2C on `Ant-v5`.

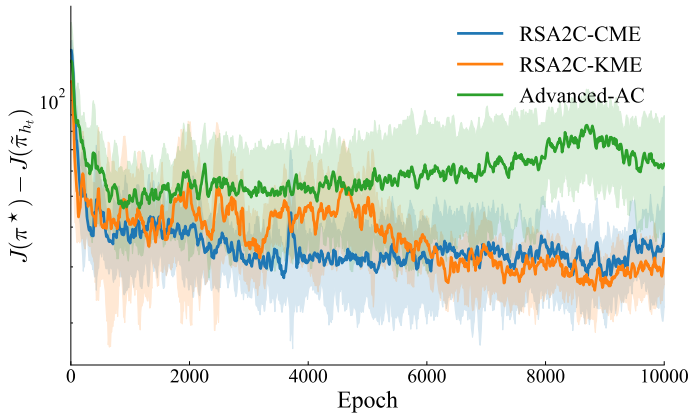


Figure 11: onvergence of the result gap $J(\pi) - J(\pi^\star)$ on the LQR environment

with $P_1 = \mathrm{diag}(1, 0.1, 10, 0.1)$ and $P_2 = 0.01$, and the return is defined as the discounted cumulative cost with factor $\gamma = 0.99$. Since this is a linear–quadratic system, the optimal value function takes the form $J^\star(\mathbf{s}) = \mathbf{s}^\top P_3 \mathbf{s}$, where P_3 is obtained by solving the discrete-time algebraic Riccati equation associated with the discounted dynamics. This allows us to compute, for specific discounted visitiation distribution, the gap $J(\pi) - J(\pi^\star)$ of the learned policy and to track the decay of this gap over training epochs, providing a quantitative convergence testbed for our non-asymptotic theory.

D.5.2 Results of theoretical gap

Figure 11 illustrates the convergence of the optimality gap $J(\pi^\star - J(\tilde{\pi}_{h_t})$ in the LQR setting, where the optimal value function is available in closed form via the Riccati equation. Both RSA2C-CME and RSA2C-KME exhibit a rapid and consistent reduction of the gap, significantly outperforming the Advanced-AC baseline in terms of convergence speed and stability. In particular, RSA2C-CME achieves the lowest steady-state gap with markedly reduced variance, indicating that its CME-based structural attribution leads to more reliable and directionally accurate policy updates. RSA2C-KME also improves substantially over the baseline, demonstrating the benefit of gradient-based feature weighting. Overall, the results confirm that RSA2C provides a more stable and sample-efficient path toward the optimal LQR controller.