Finite-Sample Regret Analysis of Nash Q-Learning with Random-Feature Approximation

Yongshan Chen[1] **Yuchen Hou**[1] **Zhuowen Zou**[2] **Calvin Yeung**[2] **Mohsen Imani**[2] **Tian Lan**[3] **Mahdi Imani**[1]

[1]Department of Electrical and Computer Engineering, Northeastern University, Boston, Massachusetts, USA
[2]Department of Computer Science, University of California, Irvine, Irvine, California, USA
[3]Department of Electrical and Computer Engineering, George Washington University, Washington, DC, USA

Abstract

This paper establishes finite-sample equilibrium learning in episodic two-player zero-sum Markov games and evaluates performance by exploitability regret, the cumulative best-response gap over episodes. Existing finite-time analyses of Nash/Minimax-Q learning are largely confined to tabular models or fixed linear realizability and do not make explicit how scalable nonlinear value approximation impacts equilibrium performance. We develop an *optimistic* Nash Q-learning framework that uses random Fourier features to approximate a shift-invariant kernel (RFNashQ), enabling nonlinear generalization while preserving closed-form ridge-regression updates and per-state zero-sum stage-game computation. Our main result proves a high-probability exploitability-regret guarantee that separates statistical uncertainty—controlled by an effective-dimension (log-determinant) complexity of the induced feature class—from an explicit additive representation error due to finite random-feature approximation, which decreases with the feature dimension. This yields a principled approximation–estimation tradeoff and recovers known linear-style finite-sample behavior when the equilibrium Q-functions are realizable in the induced feature class. Experiments on standard episodic zero-sum benchmarks with discrete and continuous state spaces, averaged over multiple seeds with 95% confidence intervals, are consistent with these predictions: exploitability regret is competitive with tabular and linear baselines, and a representation-dimension ablation follows the predicted direction over part of its range. We report these as qualitative empirical support—noting nonmonotonic, high-variance behavior—rather than a claim of optimality.

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) studies sequential decision making among agents that interact strategically over time. A canonical and theoretically grounded model for competitive MARL is the episodic Markov (stochastic) game, where the goal is to learn an *equilibrium policy profile* rather than a single-agent optimum. We focus on *two-player zero-sum* (2P0S) Markov games, where Nash equilibria coincide with minimax saddle points and admit dynamic programming.

Despite this favorable structure, learning minimax equilibrium policies with *finite-sample* guarantees remains challenging beyond small tabular environments. Classical value-based methods such as Minimax-Q and Nash Q-learning reduce equilibrium learning to repeated *stage-game* solves: at each state and time step, agents solve the zero-sum matrix game induced by action-value estimates and bootstrap via a minimax Bellman recursion Littman [1994], Hu and Wellman [2003]. While sound in finite state–action spaces, tabular representations are infeasible in large or continuous state spaces, and approximation errors can propagate through both the Bellman recursion and the minimax stage-game operator across the horizon.

Recent theoretical progress has established finite-sample regret guarantees for Nash/Minimax-Q learning under structured function approximation, including linear realizability assumptions Liu et al. [2021], Cisneros-Velarde and Koyejo [2023]. These results replace asymptotic convergence with regret-based criteria, enabling finite-time statements for equilibrium learning. However, linear function classes can be too restrictive to capture the nonlinear structure of equilibrium action-value functions encountered in practice. Deep nonlinear approximators are expressive, but stability, sample efficiency, and sharp finite-sample characterization in adversarial Markov games remain limited.

We seek a Nash-Q–style algorithm that scales via nonlinear generalization yet retains the computational and statistical

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:1154–1180.

structure needed for sharp finite-sample analysis. Our key idea is *random-feature* value approximation: we use random Fourier features (RFF) Rahimi and Recht [2007] to construct a fixed-dimensional embedding that approximates a shift-invariant kernel. This yields nonlinear generalization while preserving ridge-style least-squares updates and the concentration tools used in modern reinforcement learning theory. The resulting high-dimensional embeddings are also compatible with hyperdimensional/vector-symbolic representations Kanerva [2009], Gayler [2004], Plate [1995], but our guarantees are driven by the random-feature approximation viewpoint.

Building on this representation, we develop an *optimistic* Nash Q-learning framework that combines ridge-style value estimation in random-feature space with per-state minimax stage-game solves. Our central theoretical contribution is a high-probability *exploitability-regret* guarantee with an explicit two-part structure: (i) a *statistical uncertainty* term controlled by an effective-dimension/log-determinant complexity measure of the induced random-feature class, and (ii) an explicit additive *approximation* term arising from finite-dimensional random-feature kernel approximation. Crucially, this approximation term decreases as the feature dimension increases and vanishes in the realizable regime. Technically, our analysis propagates finite-feature approximation through the minimax Bellman operator, yielding an explicit and controllable contribution to exploitability regret. This decomposition makes the approximation–estimation tradeoff explicit and extends regret-based equilibrium learning beyond fixed linear realizability while maintaining fixed-dimensional updates.

Our main contributions are:

- **Nash Q-learning with random features.** We introduce a Nash Q-learning algorithm for episodic 2P0S Markov games that uses random Fourier features to obtain nonlinear generalization while retaining ridge-style least-squares updates and per-state minimax stage-game solves.
- **Finite-sample exploitability-regret guarantees.** Under boundedness and martingale-noise assumptions, we establish high-probability exploitability-regret bounds that decompose into an effective-dimension/log-determinant statistical term and an explicit additive approximation term due to finite random features; the approximation term decreases with feature dimension and vanishes under realizability, yielding an explicit approximation–estimation tradeoff.
- **Empirical validation.** We evaluate on episodic zero-sum benchmarks spanning tabular, discrete, and continuous-state settings, reporting mean $\pm$ 95% CI over three seeds. The results are consistent with our theory—on Littman Soccer the normalized cumulative regret of RFNashQ decreases over training, and a feature-dimension ablation follows the predicted direction between $D{=}10^4$ and 5×10^4—while remaining competitive with tabular and linear baselines. We present these as qualitative checks: the ablation is non-monotonic with overlapping confidence intervals, and we do not claim minimax optimality or uniform superiority.

2 RELATED WORK

Nash/Minimax-Q learning in two-player zero-sum Markov games. Value-based equilibrium learning in two-player zero-sum Markov games builds on the dynamic-programming structure of stochastic games: Minimax-Q and Nash Q-learning reduce equilibrium learning to repeated zero-sum stage-game solves with Bellman-style bootstrapping Shapley [1953], Littman [1994], Hu and Wellman [2003]. Classical guarantees are tabular and asymptotic, motivating finite-sample analyses under approximation.

Finite-sample regret under function approximation. Recent work establishes sharp high-probability regret guarantees for Nash/Minimax-Q learning under structured function approximation, most notably linear realizability Liu et al. [2021], Cisneros-Velarde and Koyejo [2023]. These results form the closest foundation for our analysis, but they assume fixed linear feature classes and do not explicitly quantify the impact of scalable nonlinear approximation on exploitability regret.

Kernel and random-feature approximation. Random Fourier features provide a fixed-dimensional approximation to shift-invariant kernels with controllable error Rahimi and Recht [2007], enabling nonlinear generalization while retaining linear-algebraic estimation. Our work brings this approximation viewpoint into the Nash-Q regret setting by explicitly quantifying how finite-dimensional random-feature approximation contributes additively to exploitability regret.

General-sum Markov games and alternative equilibria. In general-sum Markov games, Nash learning can be complicated by non-uniqueness and non-contractive operators Daskalakis et al. [2009], Zinkevich et al. [2005], motivating complementary regret-based convergence results for correlated and coarse correlated equilibria (CE/CCE) via online learning dynamics Aumann [1974], Hart and Mas-Colell [2000], Erez et al. [2023], Mao et al. [2024]. These approaches target weaker equilibrium notions in general-sum settings, whereas we focus on exploitability regret with respect to minimax/Nash performance in two-player zero-sum games.

Positioning of this work. Building on optimism-based Nash/Minimax-Q analyses Liu et al. [2021], Cisneros-Velarde and Koyejo [2023], we develop optimistic Nash

Q-learning with finite-dimensional random-feature value approximation and establish finite-sample exploitability-regret guarantees that separate statistical uncertainty from approximation error. The resulting bound exposes an explicit approximation–estimation tradeoff in the representation dimension while retaining fixed-dimensional ridge-style updates and per-state minimax computation.

3 PRELIMINARIES

We give here only the notation and definitions needed for the main results. A complete version of the preliminaries, including the detailed kernel assumptions, random-feature concentration statements, confidence-radius construction, and auxiliary technical facts, is deferred to Appendix A.

Notation and problem setting. We study episodic two-player zero-sum Markov games with horizon H and discount factor $\gamma \in (0, 1]$. At step $h \in [H]$, the state is $s_h \in \mathcal{S}$, the two players choose actions $a_h^1 \in \mathcal{A}^1$ and $a_h^2 \in \mathcal{A}^2$, and we write

$$a_h = (a_h^1, a_h^2) \in \mathcal{A} := \mathcal{A}^1 \times \mathcal{A}^2.$$

Player 1 receives reward $r_h(s_h, a_h) \in [0, 1]$, player 2 receives $-r_h(s_h, a_h)$, and the next state is sampled from

$$s_{h+1} \sim P_h(\cdot \mid s_h, a_h).$$

For a policy profile $\pi = (\pi^1, \pi^2)$, let V_h^π and Q_h^π denote the usual finite-horizon discounted value and action-value functions for player 1, with $V_{H+1}^\pi \equiv 0$.

For any payoff matrix $M \in \mathbb{R}^{|\mathcal{A}^1| \times |\mathcal{A}^2|}$, define

$$\mathrm{Val}(M) := \max_{\mu \in \Delta(\mathcal{A}^1)} \min_{\nu \in \Delta(\mathcal{A}^2)} \mu^\top M \nu.$$

Given an action-value function Q_h, define the induced stage value

$$V_{Q,h}(s) := \mathrm{Val}\big(Q_h(s, \cdot)\big),$$

where $Q_h(s, \cdot)$ is viewed as a matrix indexed by (a^1, a^2). The Nash–Bellman operator is

$$(\mathcal{T}Q)_h(s, a) := r_h(s, a) + \gamma \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\big[V_{Q,h+1}(s')\big], \qquad V_{Q,H+1} \equiv 0.$$

The equilibrium action-value sequence $Q^\star = \{Q_h^\star\}_{h=1}^H$ is induced by this backward recursion, and

$$V_h^\star(s) := V_{Q^\star,h}(s).$$

We repeatedly use the standard stability property

$$\big|\mathrm{Val}(M) - \mathrm{Val}(M')\big| \le \|M - M'\|_\infty.$$

Random-feature Nash-Q approximation. We use a fixed embedding $\psi : \mathcal{S} \to \mathbb{R}^d$ and random Fourier features

$$\phi_D(u) = \frac{\sqrt{2}}{\sqrt{D}}\big[\cos(\omega_1^\top u + b_1), \ldots, \cos(\omega_D^\top u + b_D)\big]^\top,$$

where $\omega_j \sim \mathcal{N}(0, \sigma^{-2} I_d)$ and $b_j \sim \mathrm{Unif}[0, 2\pi]$. Thus

$$\mathbb{E}[\phi_D(u)^\top \phi_D(u')] = \exp\left(-\frac{\|u - u'\|_2^2}{2\sigma^2}\right).$$

We work with normalized features

$$\widetilde{\phi}(s) := \phi_D(\psi(s))/\sqrt{2}, \qquad \|\widetilde{\phi}(s)\|_2 \le 1.$$

Equivalently, this can be viewed as an HDC-style fixed high-dimensional randomized representation.

Following the notation used in the appendix, we parameterize the action-value estimate with an action-indexed weight vector:

$$\widehat{Q}_{k,h}(s, a) := \widetilde{\phi}(s)^\top \widehat{w}_{k,(h,a)}.$$

For each step h and joint action a, define the regularized design matrix

$$\Lambda_{k,(h,a)} := \lambda I_D + \sum_{\tau < k:\, a_{\tau,h} = a} \widetilde{\phi}(s_{\tau,h}) \widetilde{\phi}(s_{\tau,h})^\top.$$

The fitted parameter $\widehat{w}_{k,(h,a)}$ is obtained by ridge regression on the corresponding one-step optimistic Bellman targets

$$y_{\tau,h} := r_h(s_{\tau,h}, a_{\tau,h}) + \gamma V_{k,h+1}(s_{\tau,h+1}), \qquad V_{k,H+1} \equiv 0.$$

Given a confidence parameter $\beta_{k,(h,a)}$, define

$$\mathrm{rad}_{k,h}(s, a) := \beta_{k,(h,a)} \sqrt{\widetilde{\phi}(s)^\top \Lambda_{k,(h,a)}^{-1} \widetilde{\phi}(s)}.$$

The optimistic Nash-Q estimate is

$$Q_{k,h}(s, a) := \min\Big\{\widehat{Q}_{k,h}(s, a) + \mathrm{rad}_{k,h}(s, a), H - h + 1\Big\}.$$

The policy executed in episode k is obtained by solving the matrix game $Q_{k,h}(s, \cdot)$ at each visited state, and

$$V_{k,h}(s) := \mathrm{Val}\big(Q_{k,h}(s, \cdot)\big).$$

Kernel approximation scale and regret. For the approximation analysis, let K be the Gaussian kernel associated with the above random-feature construction, and let L be the corresponding kernel integral operator under a reference distribution ρ_X. We use the effective dimension

$$N(\lambda) := \mathrm{Tr}\big((L + \lambda I)^{-1} L\big).$$

For each step h and joint action a, define the population Bellman regression target

$$f_{h,a}^\star(s) := \mathbb{E}\big[r_h(s, a) + \gamma V_{h+1}^\star(s') \mid s, a\big], \qquad s' \sim P_h(\cdot \mid s, a).$$

We assume the standard source condition

$$f^{\star}_{h,a} = L^r g_{h,a}, \qquad r \geq \frac{1}{2}, \qquad g_{h,a} \in L^2(\rho_X).$$

We summarize the resulting one-step random-feature and kernel approximation error by

$$\eta_{k,h} := c_1 \sqrt{\frac{\log(1/\delta)}{D}} + c_2 \left(\sqrt{\frac{N(\lambda)}{n_h(k)}} + \lambda^r \right),$$

where $n_h(k)$ denotes the number of samples available at step h before episode k. The formal high-probability approximation statement is given in the appendix.

Let $\pi_k = (\pi_k^1, \pi_k^2)$ be the policy profile executed in episode k from the fixed initial state s_{init}. We measure performance by player 1's cumulative exploitability regret:

$$\text{Regret}(K) := \sum_{k=1}^{K} \left[V_1^{(\text{BR}(\pi_k^2), \pi_k^2)}(s_{\text{init}}) - V_1^{\pi_k}(s_{\text{init}}) \right],$$

where $\text{BR}(\pi_k^2)$ is a best response to player 2's policy. In two-player zero-sum games, the equivalent duality-gap formulation can also be used.

4 ALGORITHM

Our algorithm summarizes the ridge-based optimistic Nash Q-learning method analyzed in Section 5. The algorithm uses a fixed random-feature map $\widetilde{\phi}$ over state for each action pairs, enabling closed-form ridge updates and per-state zero-sum stage-game computation. In Section 6, we also implement a scalable variant that replaces the ridge solve with stochastic-gradient updates and may use ϵ-greedy action sampling; unless stated otherwise, our formal guarantees are for the ridge-based update and (when specified) an ϵ-greedy/approximate stage-game sampling regime. We specialize to two-player zero-sum Markov games: it suffices to parameterize player 1's action-value function, with player 2 receiving the negated payoff. The minimax stage value is computed by applying $\text{val}(\cdot)$ to the payoff matrix $Q_{k,h}(s, \cdot)$. Each episode begins with a backward *planning* pass that constructs optimistic values used in TD targets. Pseudocode of the algorithm is provided in Appendix B.

5 MAIN RESULTS

In this section we state finite-sample guarantees for optimistic Nash Q-learning with random-feature value approximation in episodic two-player zero-sum Markov games. Our analysis combines random-feature kernel approximation, self-normalized confidence bounds for ridge TD regression, and an exploitability-regret bound with an explicit approximation–estimation decomposition. Proofs are in the appendix.

5.1 RANDOM FOURIER FEATURES AND THE GAUSSIAN KERNEL

We begin by recalling that the random-feature map we use (random Fourier features) is an unbiased finite-dimensional approximation of the Gaussian kernel. This viewpoint enables kernel-like nonlinear generalization while preserving linear-algebraic estimation and standard concentration tools.

Lemma 5.1. (*Random Feature Kernel Consistency*). *Let*

$$\phi_D(x) = \frac{1}{\sqrt{D}} \big(\sqrt{2}\cos(\omega_1^\top x + b_1), \ldots, \sqrt{2}\cos(\omega_D^\top x + b_D) \big),$$

where $\omega_i \sim \mathcal{N}(0, \sigma^{-2} I)$ *and* $b_i \sim \text{Unif}[0, 2\pi]$. *Then for any* x, x',

$$\mathbb{E}\big[\phi_D(x)^\top \phi_D(x')\big] = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2} \right).$$

Moreover, the feature map is uniformly bounded as $\|\phi_D(x)\|_2 \leq \sqrt{2}$ *for all* x.

Taking expectation over the random phase variables b_i eliminates cross terms and reduces each coordinate-wise product to a cosine of the difference $x - x'$. The remaining expectation over ω_i corresponds to the characteristic function of a Gaussian distribution, which recovers the Gaussian kernel. Uniform boundedness follows from $\cos(\cdot) \leq 1$ and the normalization by D. A complete proof is provided in Appendix C.

Lemma 5.1 shows that inner products in the random-feature space are an unbiased estimator of the Gaussian kernel. Accordingly, linear prediction with ϕ_D can be viewed as a finite-dimensional approximation to kernel methods, enabling kernel-style regularity assumptions while retaining ridge-style linear estimation.

5.2 SELF-NORMALIZED TD CONCENTRATION

We next control the noise accumulated by ridge temporal-difference (TD) regression in the random-feature or HDC space. Our main tool is a self-normalized concentration inequality that controls accumulated noise in adaptive data sequences.

Theorem 5.1 (Self-Normalized TD Concentration)**.** *Let*

$$\Lambda_t = \lambda I + \sum_{i=1}^{t} \phi_i \phi_i^\top, \qquad S_t = \sum_{i=1}^{t} \phi_i \varepsilon_i,$$

where ϕ_i *is* $\mathcal{F}_{i-1}$*-measurable and* $\|\phi_i\|_2 \leq 1$. *Then for any* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$,

$$\|S_t\|_{\Lambda_t^{-1}} \leq \sigma_\varepsilon \sqrt{2 \log \frac{\det(\Lambda_t)^{1/2}}{\det(\lambda I)^{1/2} \delta}}.$$

The proof constructs an exponential supermartingale for each fixed direction $u \in \mathbb{R}^D$ using the conditional sub-Gaussianity of TD noise. A Gaussian mixture over directions converts the scalar concentration into a vector-valued bound, yielding a self-normalized inequality involving Λ_t^{-1}. The determinant term naturally arises from evaluating the resulting Gaussian integral and corresponds to the classical elliptical potential. See Appendix D.1.2 for details.

5.3 POINTWISE CONFIDENCE BOUNDS FOR Q-VALUE ESTIMATION

Using the self-normalized inequality, we derive confidence intervals for linear Q-function estimates obtained via ridge-regularized TD learning.

Theorem 5.2 (Pointwise Confidence Bound)**.** *Let w_k denote the ridge TD estimator at episode k, and let $w^\dagger$ be the population least-squares projection of the true Q-function onto the linear span of ϕ. Then for any state-action pair x,*

$$\left|\phi(x)^\top (w_k - w^\dagger)\right| \le \beta_k \sqrt{\phi(x)^\top \Lambda_k^{-1} \phi(x)},$$

where

$$\beta_k = \sigma_\varepsilon \sqrt{2 \log \frac{\det(\Lambda_k)^{1/2}}{\det(\lambda I)^{1/2} \delta}} + \sqrt{\lambda} \|w^\dagger\|_2.$$

The ridge optimality conditions express the parameter estimation error as a sum of a self-normalized noise term and a regularization-induced bias. Left-multiplying by $\Lambda_k^{-1}\phi(x)$ and applying Cauchy–Schwarz yields the stated bound. The stochastic term is controlled using Theorem 5.1, while the bias term scales with $\sqrt{\lambda}$. See Appendix D.1.3.

Implication. Theorem 5.2 provides explicit confidence radii for Q-values, which form the basis of our optimism-based exploration strategy.

5.4 RANDOM FEATURE APPROXIMATION ERROR

In addition to statistical uncertainty, using a finite number of random features induces approximation error relative to the target kernel class. We quantify this error through standard kernel ridge regression arguments.

Adaptive data versus reference-measure approximation. The samples used by the algorithm are generated adaptively by the visitation distributions induced by the learned policies, and are not assumed to be i.i.d. from the reference measure ρ_X. The adaptive statistical estimation error is handled by the self-normalized martingale concentration in Theorem 5.1. The random-feature approximation error is instead imposed as a pathwise event over the inputs and Bellman targets evaluated by the algorithm. Lemma C.8 should therefore be viewed as a reference-measure KRR result that motivates the scale of this approximation term, rather than as a direct proof for the adaptive-visitation case.

Assumption 5.1 (Pathwise random-feature approximation)**.** *With probability at least $1-\delta$ over the random-feature draw and the sample path, for all $h \in [H]$ and $k \in [K]$,*

$$\epsilon_{\mathrm{rff,h,k}} := \sup_{\tau<k} \left| g_{k,h}(x_{\tau,h}) - \phi(x_{\tau,h})^\top w_{k,h}^\dagger \right| \le c_1 \sqrt{\frac{\log(1/\delta)}{D}} + c_2 \left(\sqrt{\frac{\mathcal{N}(\lambda)}{n_h(k)}} + \lambda^r \right),$$

where $x_{\tau,h} = (s_{\tau,h}, a_{\tau,h})$, $g_{k,h}(x) = \mathbb{E}[r_h(x) + \gamma U_{k,h+1}(s') \mid x]$, and $U_{k,h+1}$ is the value target used in the corresponding backward pass. In particular, $U_{k,h+1} = V_{h+1}^\star$ for the fixed Bellman target and $U_{k,h+1} = \widehat{V}_{k,h+1}^{\mathrm{opt}}$ for the optimistic variant.

The first term captures random feature approximation error, while the remaining terms correspond to the variance–bias decomposition of kernel ridge regression. This bound shows that increasing the HDC dimension D improves approximation accuracy without changing the linear structure of the algorithm. Formal derivations are given in Appendix D.2.

5.5 ONE-STEP ERROR DECOMPOSITION

A key ingredient in our analysis is a decomposition of the Bellman error that separates approximation, estimation, and value-propagation effects. This decomposition allows us to propagate confidence bounds through the Bellman operator and is essential for establishing optimism and regret guarantees.

Let $Q_h^\dagger(s,a) = (s)^\top w_{h,a}^\dagger$ denote the population linear projection of the true Q-function at step h, and let $\widehat{Q}_h$ be the learned estimator. We consider the one-step Bellman error $|(\mathcal{T}\widehat{Q})_h(s,a) - \widehat{Q}_h(s,a)|$ and decompose it as follows.

Lemma 5.2 (One-Step Bellman Error Decomposition)**.** *For any step h and state-action pair (s,a),*

$$\left|(\mathcal{T}\widehat{Q})_h(s,a) - \widehat{Q}_h(s,a)\right| \le \underbrace{\left|(\mathcal{T}Q_h^\dagger)(s,a) - Q_h^\dagger(s,a)\right|}_{\text{(i) approximation error}} + \underbrace{\left|Q_h^\dagger(s,a) - (s)^\top \widehat{w}_{k,(h,a)}\right|}_{\text{(ii) estimation error}} + \underbrace{\gamma\, \mathbb{E}_{s'} \left|\widehat{V}_{\widehat{Q},h+1}(s') - V_{Q^\dagger,h+1}(s')\right|}_{\text{(iii) value propagation}}.$$

Interpretation. Lemma 5.2 decomposes the Bellman error into three conceptually distinct components: (i) an *approximation error* due to representing the Bellman target

within a linear function class, (ii) an *estimation error* arising from finite-sample TD regression, and (iii) a *value-propagation error* that captures how errors at step $h+1$ influence the current step through the Bellman operator.

The approximation term can be controlled using the random feature analysis in Section D.2, while the estimation term is bounded by the confidence radius derived in Section D.1.3. The remaining challenge is to control the propagation of value errors across stages, which is addressed by the Lipschitz continuity of the stage-game value.

Lemma 5.3 (Lipschitz Continuity of Stage Value)**.** *For any two payoff matrices M and M',*

$$\big|\mathrm{Val}(M) - \mathrm{Val}(M')\big| \le \|M - M'\|_\infty.$$

Consequence. Lemma 5.3 implies that the stage value operator is 1-Lipschitz with respect to entrywise perturbations of the payoff matrix. As a result, value-function errors can be upper bounded by the maximum pointwise Q-function estimation error:

$$\big|\widehat{V}_h(s) - V_{Q^\dagger,h}(s)\big| \le \max_a \big|(s)^\top(\widehat{w}_{k,(h,a)} - w^\dagger_{h,a})\big|.$$

Together, Lemmas 5.2 and 5.3 provide a clean mechanism for propagating statistical confidence bounds through the Bellman recursion. This mechanism underlies the optimism guarantees in Section E.1 and the regret analysis in Section E.

5.6 BELLMAN ERROR DECOMPOSITION AND OPTIMISM

We now combine estimation and approximation errors to construct optimistic Q-functions and propagate confidence bounds through the Bellman operator.

Theorem 5.3 (Optimism)**.** *Define the optimistic Q-function*

$$\widehat{Q}^{\mathrm{opt}}_h(s,a) = \phi(s)^\top w_{k,(h,a)} + 2\,\mathrm{rad}_{k,h}(s,a) + \varepsilon_{\mathrm{rff},h,k}.$$

Then for all k, h and all state-action pairs evaluated along the sample path, $Q_h(s,a) \le \widehat{Q}^{\mathrm{opt}}_{k,h}(s,a)$. Consequently, for every visited state s at which the full stage-game payoff matrix is evaluated, $V_h(s) \le \widehat{V}^{\mathrm{opt}}_{k,h}(s)$. If a uniform approximation event over the whole domain is assumed, the same conclusion holds for all s, a.

The proof proceeds by backward induction on the horizon h. At the terminal step, the bound follows directly from Theorem 5.2 and Assumption 5.1. For intermediate steps, the Bellman target is upper bounded by an optimistic one-step target, and the Lipschitz continuity of the stage-game value operator ensures that optimism propagates from Q-functions to state values. Details are given in Appendix E.1.

5.7 REGRET BOUNDS UNDER TWO ACTION-SELECTION REGIMES

We finally state regret guarantees for our algorithm in two-player zero-sum Markov games. A key distinction is whether the agent selects actions via (i) a practical TD+ε-greedy scheme that only approximates the stage-game Nash solution, or (ii) an *optimistic variant* that explicitly solves the stage-game Nash equilibrium on the optimistic Q-function and uses the corresponding optimistic value as the TD target. These two regimes lead to different regret decompositions and hence different final bounds.

Regret definition. Let π_k denote the policy executed in episode k and let $\mathrm{br}(\pi_k)$ be a best response (for the opponent). We measure regret as the cumulative best-response gap (standard in zero-sum Markov games):

$$\mathrm{Regret}(K) = \sum_{k=1}^{K} \Big(V^{\mathrm{br}(\pi_k),\pi_k}(s_{k,1}) - V^{\pi_k,\pi_k}(s_{k,1})\Big),$$

where $s_{k,1}$ is the initial state of episode k. (Any equivalent episode-wise regret notion used in the appendix yields the same order bounds.)

High-probability event and effective dimension. Let $\mathrm{rad}_{k,(h,a)}(s)$ be the confidence radius induced by Theorem 5.2. For each step h and joint action a, let $\widehat{C}_{h,a}$ be the empirical feature covariance, with eigenvalues $\{\widehat{\mu}_{a,j}\}_{j=1}^D$. Define the empirical effective dimension

$$\widetilde{d}_{\mathrm{eff}}(\lambda) := \sum_{a\in\mathcal{A}} \mathrm{tr}\Big((\widehat{C}_{h,a} + \lambda I)^{-1}\widehat{C}_{h,a}\Big) = \sum_{a\in\mathcal{A}}\sum_{j=1}^{D} \frac{\widehat{\mu}_{a,j}}{\widehat{\mu}_{a,j}+\lambda}.$$

The bound $\widetilde{d}_{\mathrm{eff}}(\lambda) \le |\mathcal{A}|D$ is only a worst-case finite-rank upper bound. Since directions with $\widehat{\mu}_{a,j} \ll \lambda$ contribute only $\widehat{\mu}_{a,j}/\lambda$, the effective dimension is governed by the spectral decay of the induced covariance operator rather than by D alone. Under the capacity condition $N(\lambda) \le C_N\lambda^{-\alpha}$, and when the random-feature approximation is sufficiently accurate, this quantity scales as

$$\widetilde{d}_{\mathrm{eff}}(\lambda) \lesssim |\mathcal{A}|\,\min\{D,\, N(\lambda)\} \le |\mathcal{A}|\,\min\{D,\, C_N\lambda^{-\alpha}\},$$

up to logarithmic and concentration factors. This is the effective dimension that enters the elliptical-potential control of cumulative uncertainty.

5.7.1 TD + ε-greedy (Approximate Stage-NE Sampling)

Theorem 5.4 (Regret Bound: TD + ε-greedy)**.** *For any $\delta \in (0,1)$, with probability at least $1-\delta$, the regret of the*

TD+ε-greedy variant satisfies

$$\mathrm{Regret}(K) \;\le C\, ar\beta\, H\sqrt{KH\,\widetilde{d}_{\mathrm{eff}}(\lambda)\,\log\!\Big(1+\tfrac{K}{\lambda}\Big)} \;+\; C'KH\,\epsilon_{\mathrm{approx}} \;+\; C''KH\,\varepsilon,$$

where $ar\beta$ denotes the typical scale of $\beta_{k,(h,a)}$ across (h,a), and the approximation term is

$$\epsilon_{\mathrm{approx}} = c_1 D^{-1/2} + c_2\left(\sqrt{\tfrac{N(\lambda)}{K}} + \lambda^r\right).$$

The starting point is a regret decomposition that upper bounds regret by (i) martingale difference terms from stochastic transitions/rewards, (ii) the sum of confidence radii along the realized trajectory, and (iii) the cumulative random-feature approximation error; additionally, TD+ε-greedy introduces a $KH\varepsilon$ term that accounts for deviations from exact stage-game Nash sampling. The martingale terms are bounded by Azuma–Hoeffding, while the cumulative confidence radii are controlled via an elliptical potential argument, yielding a dependence on $\widetilde{d}_{\mathrm{eff}}(\lambda)$. Finally, summing the per-step random-feature error bound gives the $KH\epsilon_{\mathrm{approx}}$ term. See Theorem E.1 for the detailed decomposition and bounds.

Choice of λ and D. The above bound makes the approximation–estimation tradeoff explicit. Under the capacity condition $N(\lambda) \le C_N \lambda^{-\alpha}$, the approximation term satisfies

$$\begin{aligned}\epsilon_{\mathrm{approx}} &= c_1 D^{-1/2} + c_2\left(\sqrt{\frac{N(\lambda)}{K}} + \lambda^r\right)\\ &\le c_1 D^{-1/2} + \widetilde{O}\Big(K^{-1/2}\lambda^{-\alpha/2} + \lambda^r\Big).\end{aligned}$$

Balancing the kernel variance term $K^{-1/2}\lambda^{-\alpha/2}$ and the kernel bias term λ^r gives $\lambda \asymp K^{-1/(2r+\alpha)}$. With this choice,

$$\sqrt{\frac{N(\lambda)}{K}} \asymp \lambda^r \asymp K^{-r/(2r+\alpha)}.$$

Choosing $D \asymp K^{2r/(2r+\alpha)}$ balances the finite random-feature error $D^{-1/2}$ with the kernel bias–variance rate. Therefore, $\epsilon_{\mathrm{approx}} = \widetilde{O}\big(K^{-r/(2r+\alpha)}\big)$. Moreover, since $N(\lambda) \lesssim K^{\alpha/(2r+\alpha)}$, the effective dimension term scales as $\widetilde{d}_{\mathrm{eff}}(\lambda) \lesssim |\mathcal{A}| K^{\alpha/(2r+\alpha)}$. Substituting this into the leading statistical term yields the representative scaling

$$\begin{aligned}\mathrm{Regret}(K) &= \widetilde{O}\Big(\bar{\beta} H\sqrt{H|\mathcal{A}|}K^{(r+\alpha)/(2r+\alpha)}\Big)\\ &\quad + \widetilde{O}\Big(HK^{(r+\alpha)/(2r+\alpha)}\Big) + O(KH\epsilon),\end{aligned}$$

up to constants and logarithmic factors. Hence the average regret decays as

$$\frac{\mathrm{Regret}(K)}{K} = \widetilde{O}\Big(K^{-r/(2r+\alpha)}\Big) + O(H\epsilon).$$

Table 1: Final-snapshot exploitability-regret proxy on Littman Soccer at episode 6000, and final-window TD loss on discrete predator–prey. Mean $\pm$ 95% CI over three seeds $\{0, 1009, 2018\}$.

Method	Littman Soccer regret @6000	Predator–prey TD loss
Tabular	7111.1 ± 3185.7	—
Linear Approx.	8872.2 ± 7886.9	7.22 ± 4.90
RFNashQ (Ours)	$\mathbf{3588.9 \pm 3874.8}$	$\mathbf{4.63 \pm 0.37}$

5.7.2 Optimistic Variant (Stage-NE on $\widehat{Q}^{\mathrm{opt}}$)

Theorem 5.5 (Regret Bound: Optimistic Variant)**.** *Assume actions are selected by solving the stage-game Nash equilibrium on the optimistic Q-function $\widehat{Q}^{\mathrm{opt}}$ and the TD targets use the corresponding optimistic value $\widehat{V}^{\mathrm{opt}}$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathrm{Regret}(K) \;\le\; \widetilde{O}\Big(\beta\, H\sqrt{KH} \;+\; \beta\, H\sqrt{K\,\widetilde{d}_{\mathrm{eff}}(\lambda)}\Big).$$

The key difference from Theorem 5.4 is *optimism*: the approximation and estimation errors are absorbed into the optimistic bonuses by construction, so the approximation error is absorbed into the optimistic bonuses, and therefore does not appear as a separate additive term beyond the cumulative confidence widths. The remaining analysis reduces to bounding martingale terms and the cumulative optimistic bonus, which is controlled by the same elliptical potential machinery, with the linear dimension replaced by the effective dimension $\widetilde{d}_{\mathrm{eff}}(\lambda) \le |A|D$. See Theorem E.2 for details.

6 EXPERIMENTS

In this section, we apply our Random-Feature Optimistic Nash Q-learning (RFNashQ) algorithm to several benchmark environments and compare its performance with multiple baselines, including tabular Nash Q-learning, linear function approximation, NQOVI, and deep Q-learning. We will release the code upon publication.

The discrete predator–prey environment is a discrete-state variant of MPE simple_tag_v3, where agents are randomly initialized and can choose among four actions: moving up, down, left, or right, each with a step size of one. The Pac-Man environment follows the classic Pac-Man setting, in which beans appear randomly on the map, and agents are rewarded for collecting them. Details of the Littman Soccer environment can be found in Littman [1994].

Throughout this section we report the mean $\pm$ 95% confidence interval over three independent seeds $0, 1009, 2018$. The tabular Q-learning run (200,000 iterations) is used only as an approximate reference for the surrogate-regret proxy, not as a baseline whose convergence speed we claim to beat. We treat these results as a qualitative check of the pre-

Table 2: **Average Training Time for Different Algorithms under Different Benchmarks.** We calculated the average time for the most common baselines under usual benchmark environments to train for 1000 episodes. Each setting is repeated 3 times and we report the mean wall-clock time. ✗indicates the method is not applicable for this environment.

Algorithm	discrete predator&prey (5x5)	Pacman (5x5)	Littman Soccer (4×5)	MPE simple_tag_v3
Tabular Q Learning	10m6s	49m33s	9m49s	✗
Linear Approximation	12m37s	52m8s	2m8s	27m11s
NQOVICisneros-Velarde and Koyejo [2023]	47h4m58s	56h6m31s	16h16m20s	57h33m16s
DQNVan Hasselt et al. [2016]	15h26m2s	46h56m37s	2h56m13s	48h17m27s
RFNashQ (Ours)	29m43s	2h35m6s	9m45s	1h47m38s

dicted approximation–estimation trade-off and do not claim minimax optimality.

Sub-Linear Regret Performance. In Figure 1, we report the normalized cumulative (average) regret $\widehat{R}(t)/t$ on Littman Soccer. On this environment RFNashQ's average regret decreases over training, which is consistent with the sub-linear trend our theory predicts; we note that this decreasing trend is observed on Littman Soccer and treat the other environments as sanity checks. Exact exploitability is unavailable in most benchmarks, so this quantity is a surrogate regret against a strong reference policy.

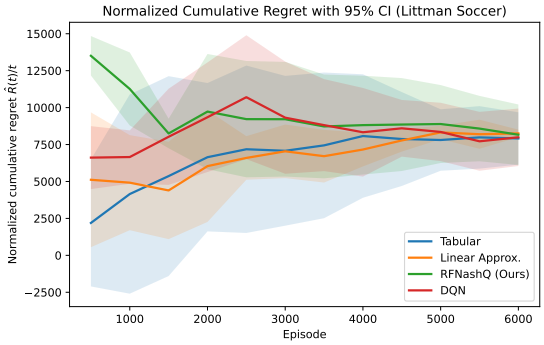


Figure 1: Normalized cumulative regret $\widehat{R}(t)/t$ on Littman Soccer, mean $\pm$ 95% CI over 3 seeds. RFNashQ's normalized cumulative regret decreases from $\approx 1.35 \times 10^4$ to $\approx 8.16 \times 10^3$ over training, consistent with the sub-linear average-regret trend predicted by our analysis.

In Figure 2, we present the approximate regret of different algorithms across multiple benchmarks, compared against the baseline tabular Q-learning algorithm, which was trained for 200,000 iterations to ensure convergence. These curves should be read as a qualitative sanity check rather than a claim of optimality. RFNashQ attains approximate regret broadly comparable to the tabular and linear baselines and no worse than DQN; on Littman Soccer its normalized cumulative regret decreases over training. We do not claim minimax optimality, and the wide confidence intervals require careful interpretation.

On Littman Soccer these results are consistent with sub-linear growth of the (approximate) cumulative regret of RFNashQ. We emphasize this is a qualitative corroboration of the theory on finite benchmarks, not a demonstration of uniform superiority: the gains are not driven by a single favorable seed, but the wide intervals and the non-monotonic dimension trend require careful reporting.

Table 3: RFF/HDC dimension ablation on Pac-Man: final cumulative approximate regret, mean $\pm$ 95% CI over three seeds. The trend is non-monotonic and the intervals overlap.

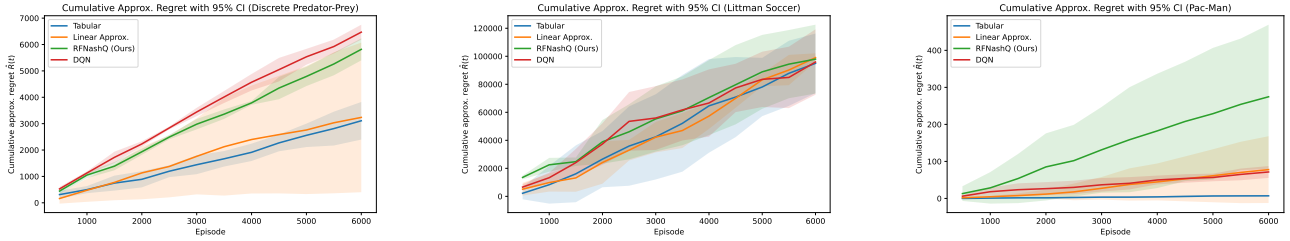
Dimension D	Final cumulative approx. regret
2000	83.2 ± 42.0
10000	274.6 ± 194.7
50000	97.6 ± 84.6

On Littman Soccer the final-snapshot regret proxy at episode 6000 is lowest for RFNashQ (3588.9) versus Tabular (7111.1) and Linear Approx. (8872.2); on discrete predator–prey the final-window TD loss is 4.63 ± 0.37 (RFNashQ) versus 7.22 ± 4.90 (Linear Approx.). Results are provided in Table 1

Ablation. Increasing D from 10000 to 50000 reduces the final cumulative approximate regret from 274.6 to 97.6, the direction predicted by the $D^{-1/2}$ approximation term. However, the trend is non-monotonic ($D = 2000$ attains 83.2, below $D = 10000$) and the 95% confidence intervals overlap substantially, so we do not read this as a statistically clean confirmation of the dimension scaling. We interpret it as the $D^{-1/2}$ term superimposed on large seed variance at these sample sizes, which the wide intervals make explicit.

Training Cost. In Table 2, we report the average computational time required by different algorithms across various benchmark environments. While maintaining competitive performance and broad applicability, our algorithm achieves training and computational costs comparable to classical baseline methods (e.g., tabular approaches) and remains significantly more efficient than alternative methods capable of handling both discrete and continuous environments, such as NQOVI and DQN.

All experiments are conducted on a machine equipped with an Intel Xeon w7-3455 CPU, 251 GB of system memory,



(a) Discrete predator&prey (b) Littman Soccer (c) Pacman

Figure 2: Cumulative approximate regret $\widehat{R}(t_m) = \sum_j (t_j - t_{j-1})\widehat{r}(t_j)$ on three benchmarks, mean $\pm$ 95% CI over 3 seeds, checkpoints $500 \rightarrow 6000$. The cumulative form is markedly smoother than the original per-checkpoint curves. NQOVI is omitted here as its cost (Table1) precludes a matched multi-seed run; it appears in Table1 for wall-clock comparison only.

and two NVIDIA RTX 6000 Ada GPUs running CUDA version 13.0. Each experiment is repeated at least three times, and the reported results are averaged over these runs. For DQN, we use a batch size of 128 and a replay buffer size of 200,000, and apply an early-stopping criterion during training.

7 CONCLUSION

We studied finite-sample equilibrium learning in episodic two-player zero-sum Markov games through the lens of Nash-Q style value learning with optimism. To scale beyond tabular and fixed linear representations, we introduced an optimistic Nash Q-learning framework that represents state–action values using a hyperdimensional random-feature embedding (instantiated via random Fourier features for a shift-invariant kernel), while retaining tractable ridge-style updates and per-state minimax stage-game computation. Our main theoretical contribution is a high-probability regret guarantee that makes the effect of representation explicit: the bound separates a statistical uncertainty term (governed by an effective-dimension complexity measure) from an additive kernel-approximation term that decreases with the random-feature dimension, yielding a principled approximation–estimation tradeoff. Together, these results provide an analyzable nonlinear extension of Nash-Q regret theory with explicit representation control in zero-sum Markov games.

Several directions arise naturally from our analysis. First, our guarantees apply to the ridge-based optimistic variant; extending comparable guarantees to practical stochastic-gradient and ϵ-greedy implementations would further narrow the remaining theory–practice gap. Second, our results assume exact per-state minimax computation; relaxing this requirement to approximate equilibrium oracles and quantifying the resulting degradation would broaden applicability. Third, it would be valuable to extend the same approximation–estimation decomposition beyond the two-player zero-sum regime while developing adaptive choices of kernel parameters and random-feature dimension based on data, thereby improving scalability without sacrificing explicit approximation control.

IMPACT STATEMENT

This work advances algorithmic and theoretical understanding of equilibrium learning in multi-agent reinforcement learning, with experiments conducted in simulated environments. Potential applications include improved coordination and decision-making in multi-agent systems, while misuse in competitive or adversarial settings remains possible. We encourage responsible consideration when deploying equilibrium learning methods in real-world systems.

ACKNOWLEDGMENTS

The authors acknowledge the support of the National Science Foundation under Awards IIS-2311969 and ECCS-2431561, the Army Research Office under Award W911NF-24-2-0166, and the Office of Naval Research under Award N00014-23-1-2850.

References

Robert J Aumann. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1): 67–96, 1974.

Pedro Cisneros-Velarde and Sanmi Koyejo. Finite-sample guarantees for nash q-learning with linear function approximation. In *Uncertainty in Artificial Intelligence*, pages 424–432. PMLR, 2023.

Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009.

Liad Erez, Tal Lancewicki, Uri Sherman, Tomer Koren, and Yishay Mansour. Regret minimization and convergence to equilibria in general-sum markov games. In

International Conference on Machine Learning, pages 9343–9373. PMLR, 2023.

Ross W Gayler. Vector symbolic architectures answer jackendoff's challenges for cognitive neuroscience. *arXiv preprint cs/0412059*, 2004.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

Junling Hu and Michael P Wellman. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

Pentti Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive computation*, 1(2):139–159, 2009.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

Weichao Mao, Haoran Qiu, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, and Tamer Başar. $\widetilde{O}(t^{-1})$ convergence to (coarse) correlated equilibria in full-information general-sum markov games. In *6th Annual Learning for Dynamics & Control Conference*, pages 361–374. PMLR, 2024.

Tony A Plate. Holographic reduced representations. *IEEE Transactions on Neural networks*, 6(3):623–641, 1995.

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

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features, 2021. URL `https://arxiv.org/abs/1602.04474`.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

Joel A. Tropp. An introduction to matrix concentration inequalities, 2015. URL `https://arxiv.org/abs/1501.01571`.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Martin Zinkevich, Amy Greenwald, and Michael Littman. Cyclic equilibria in markov games. *Advances in neural information processing systems*, 18, 2005.

Supplementary Material

Yongshan Chen[1] **Yuchen Hou**[1] **Zhuowen Zou**[2] **Calvin Yeung**[2] **Mohsen Imani**[2] **Tian Lan**[3] **Mahdi Imani**[1]

[1]Department of Electrical and Computer Engineering, Northeastern University, Boston, Massachusetts, USA
[2]Department of Computer Science, University of California, Irvine, Irvine, California, USA
[3]Department of Electrical and Computer Engineering, George Washington University, Washington, DC, USA

A PRELIMINARIES

Setting. We consider episodic two-player zero-sum Markov games with horizon H and discount factor $\gamma \in (0, 1]$. At step $h \in [H]$, the state is $s_h \in \mathcal{S}$, the players choose actions $a_h^1 \in \mathcal{A}^1$ and $a_h^2 \in \mathcal{A}^2$, and we write $a_h = (a_h^1, a_h^2) \in \mathcal{A} := \mathcal{A}^1 \times \mathcal{A}^2$. Player 1 receives reward $r_h(s_h, a_h) \in [0, 1]$ and player 2 receives $-r_h(s_h, a_h)$; the next state is drawn as $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$. A (Markov) policy for player $i \in \{1, 2\}$ is $\pi^i = \{\pi_h^i\}_{h=1}^H$ with $\pi_h^i(\cdot \mid s) \in \Delta(\mathcal{A}^i)$, and $\pi = (\pi^1, \pi^2)$ denotes a policy profile. For player 1, define

$$V_h^\pi(s) := \mathbb{E}_\pi\Big[\sum_{t=h}^{H} \gamma^{t-h} r_t(s_t, a_t) \,\Big|\, s_h = s\Big],$$

$$Q_h^\pi(s, a) := \mathbb{E}_\pi\Big[\sum_{t=h}^{H} \gamma^{t-h} r_t(s_t, a_t) \,\Big|\, s_h = s,\, a_h = a\Big],$$

and set $V_{H+1}^\pi \equiv 0$.

Stage-game value and Nash–Bellman operator. For any payoff matrix $M \in \mathbb{R}^{|\mathcal{A}^1| \times |\mathcal{A}^2|}$, define the minimax value

$$\mathrm{Val}(M) := \max_{\mu \in \Delta(\mathcal{A}^1)} \min_{\nu \in \Delta(\mathcal{A}^2)} \mu^\top M \nu.$$

For any $Q_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, define the induced stage value $V_{Q,h}(s) := \mathrm{Val}\big(Q_h(s, \cdot)\big)$, where $Q_h(s, \cdot)$ is viewed as a matrix indexed by (a^1, a^2). In finite-action 2P0S games, $\mathrm{Val}(\cdot)$ can be computed by a linear program. Setting $V_{Q,H+1} \equiv 0$, define the Nash–Bellman operator $\mathcal{T}$ by

$$(\mathcal{T}Q)_h(s, a) := r_h(s, a) + \gamma\, \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\big[V_{Q,h+1}(s')\big].$$

In 2P0S games, $Q^\star = \{Q_h^\star\}_{h=1}^H$ denotes the equilibrium action-value sequence induced by this recursion, with $V_h^\star(s) := V_{Q^\star,h}(s)$. A key stability fact we will use repeatedly is

$$|\mathrm{Val}(M) - \mathrm{Val}(M')| \le \|M - M'\|_\infty.$$

Random-feature (HDC/RFF) embedding. Our regression inputs are state–joint-action pairs $x = (s, a^1, a^2) \in \mathcal{X} := \mathcal{S} \times \mathcal{A}^1 \times \mathcal{A}^2$. When $\mathcal{X}$ is not Euclidean, we map x into $\mathbb{R}^d$ via a fixed embedding $\psi : \mathcal{X} \to \mathbb{R}^d$ (e.g., concatenation of a state vector with one-hot action indicators). For $\sigma > 0$, random Fourier features draw $\omega_j \sim \mathcal{N}(0, \sigma^{-2} I_d)$ and $b_j \sim \mathrm{Unif}[0, 2\pi]$ and define

$$\phi_D(u) = \frac{\sqrt{2}}{\sqrt{D}} \big[\cos(\omega_1^\top u + b_1), \ldots, \cos(\omega_D^\top u + b_D)\big]^\top,$$

so that $\mathbb{E}[\phi_D(u)^\top \phi_D(u')] = \exp(-\|u-u'\|_2^2/(2\sigma^2))$ (e.g., Rahimi and Recht [2007]) and $\|\phi_D(u)\|_2 \le \sqrt{2}$. We use normalized features

$$\widetilde{\phi}(x) := \phi_D(\psi(x))/\sqrt{2}, \qquad \|\widetilde{\phi}(x)\|_2 \le 1,$$

and interpret $\widetilde{\phi}$ as an HDC-style embedding: a fixed, high-dimensional randomized representation supporting simple inner products and stable linear updates. The optimistic value used for stage-game evaluation and TD targets is $V_{k,h}(s) := V_{Q_{k,h},h}(s) = \mathrm{Val}(Q_{k,h}(s,\cdot))$.

Effective dimension and Source Condition Assumption For each player i and step h, we assume that the Bayes regression function $Q^*_{i,h}(\cdot,\cdot)$ belongs to the RKHS associated with the Gaussian kernel and satisfies the source condition with exponent $r \ge 1/2$, i.e., $Q^*_{i,h} = L^r g_{i,h}$ for some $g_{i,h} \in L^2(\rho_X)$. We adopt the standard effective dimension and source condition assumptions from Rudi and Rosasco [2021]. Let L be the kernel integral operator associated with the Gaussian kernel K and a reference marginal state distribution ρ_X, i.e.,

$$(Lf)(x) := \int K(x,x')f(x')\,d\rho_X(x').$$

Define the effective dimension $N(\lambda) := \mathrm{Tr}\big((L+\lambda I)^{-1}L\big)$. We assume a standard source condition from kernel-based learning theory: for some $r \ge \frac{1}{2}$ and for each step h and joint action a, the regression target

$$f^*_{h,a}(s) := \mathbb{E}\big[r_h(s,a) + \gamma V^*_{h+1}(s') \mid s,a\big]$$

belongs to the RKHS of K and admits the representation $f^*_{h,a} = L^r g_{h,a}$, $g_{h,a} \in L^2(\rho_X)$.

Fitted and optimistic Nash-Q estimates. We parameterize player 1 action-values linearly:

$$\widehat{Q}_{k,h}(s,a) := \widetilde{\phi}(s)^\top \widehat{w}_{k,(h,a)},$$

where $\widehat{w}_{k,h} \in \mathbb{R}^D$ is estimated from data collected up to (but excluding) episode k. Given any action-value estimate Q_h, we write $V_{Q,h}(s) = \mathrm{Val}(Q_h(s,\cdot))$; in particular $\widehat{V}_{k,h}(s) := V_{\widehat{Q}_{k,h},h}(s)$.

At time (k,h) we form the optimism-consistent Nash-Q target $y_{k,h} := r_h(s_{k,h},a_{k,h}) + \gamma\, V_{Q_{k,h+1},h+1}(s_{k,h+1})$, setting $V_{Q_{k,H+1},H+1} \equiv 0$, where $Q_{k,h+1}$ is the current optimistic estimate defined below.

Let $\mathcal{F}_{k,h}$ denote the filtration up to the beginning of (k,h), and define $\epsilon_{k,h} := y_{k,h} - \mathbb{E}[y_{k,h} \mid \mathcal{F}_{k,h}]$. We assume $\epsilon_{k,h}$ is conditionally σ_ϵ^2-sub-Gaussian. Since $r_h \in [0,1]$ and we clip optimistic values (below), we have $y_{k,h} \in [0, 1+\gamma H]$ and Hoeffding's lemma implies $\sigma_\epsilon \le (1+\gamma H)/2$.

For each step h and action pair a, define the design matrix and ridge estimator

$$\Lambda_{k,h} := \lambda I_D + \sum_{\tau<k} \widetilde{\phi}(s_{\tau,h})\widetilde{\phi}(s_{\tau,h})^\top,$$

$$\widehat{w}_{k,h} \in \arg\min_{w\in\mathbb{R}^D} \sum_{\tau<k} \big(y_{\tau,h} - \widetilde{\phi}(s_{\tau,h})^\top w\big)^2 + \lambda\|w\|_2^2.$$

Since $r_h \in [0,1]$, all value functions satisfy $V^\pi_h(s) \le H-h+1$. Given $\beta_{k,h} > 0$, define the self-normalized radius and optimistic estimate

$$\mathrm{rad}_{k,h}(s,a) := \beta_{k,h}\sqrt{\widetilde{\phi}(s)^\top \Lambda^{-1}_{k,(h,a)}\widetilde{\phi}(s)},$$

$$Q_{k,h}(s,a) := \min\{\widehat{Q}_{k,h}(s,a) + \mathrm{rad}_{k,h}(s,a),\, H-h+1\}. \tag{A.1}$$

A standard high-probability choice is

$$\beta_{k,(h,a)} := \sigma_\epsilon\sqrt{2\log\left(\frac{\det(\Lambda_{k,(h,a)})^{1/2}}{\det(\lambda I_D)^{1/2}\delta}\right)} + \sqrt{\lambda}\,\|w^\star_h\|_2,$$

$$w^\star_h \in \arg\min_w \mathbb{E}\big[(f^\star_h(x) - \widetilde{\phi}(x)^\top w)^2\big],$$

where $f^\star_h$ is the population one-step Bellman regression target (defined next). The optimistic value used for stage-game evaluation and TD targets is $V_{k,h}(s) := V_{Q_{k,h},h}(s) = \mathrm{Val}(Q_{k,h}(s,\cdot))$.

Kernel regularity and approximation scale. The regret analysis under exact realizability depends only on the finite-dimensional linear class induced by $\widetilde{\phi}$. To quantify representation error induced by random features and kernel bias–variance tradeoffs, we adopt standard kernel-learning assumptions (e.g., Rudi and Rosasco [2021]). Let ρ_X be a fixed reference distribution over $\mathcal{X}$ (used only to state approximation rates), and let L be the kernel integral operator induced by the Gaussian kernel on $\mathcal{X}$. Define the effective dimension $N(\lambda) := \mathrm{Tr}((L+\lambda I)^{-1}L)$. To make the scaling of the effective dimension explicit, we also use the standard capacity condition

$$N(\lambda) = \mathrm{Tr}\big((L+\lambda I)^{-1}L\big) \le C_N \lambda^{-\alpha}, \qquad \alpha \in (0,1],$$

for some problem-dependent constant $C_N > 0$. This condition is standard in kernel learning and captures the spectral decay of the kernel integral operator. The crude finite-dimensional bound scales with the ambient random-feature dimension D, whereas the sharper kernel-dependent complexity is governed by $N(\lambda)$. For kernels with faster spectral decay, such as Gaussian kernels on compact domains, $N(\lambda)$ can grow only polylogarithmically in $1/\lambda$, up to problem-dependent constants.

Define the Bellman regression target

$$f_h^\star(x) := \mathbb{E}\big[r_h(x) + \gamma V_{h+1}^\star(s') \mid x\big], \qquad s' \sim P_h(\cdot \mid x),$$

and assume the source condition $f_h^\star = L^r g_h$ for some $r \ge \frac{1}{2}$ and $g_h \in L^2(\rho_X)$. We summarize the resulting one-step approximation/bias scale by

$$\eta_{k,h} := c_1\sqrt{\frac{\log(1/\delta)}{D}} + c_2\Big(\sqrt{\frac{N(\lambda)}{n_h(k)}} + \lambda^r\Big),$$

where $n_h(k)$ denotes the number of samples available at step h up to (but excluding) episode k. Under standard RFF concentration, together with the RKHS conditions, upper bounds, the one-step approximation component in fitting $f_h^\star$ by the random-feature class along the relevant domain (e.g., visited inputs) occurs with high probability; the formal statement and constants are given in the appendix.

Regret (exploitability). Let $\pi_k = (\pi_k^1, \pi_k^2)$ denote the policy profile executed in episode k from initial state s_0. Define player 1's best response $\boldsymbol{r}(\pi_k^2) \in \arg\max_\nu V_1^{(\nu,\pi_k^2)}(s_0)$ and the cumulative regret

$$\mathrm{Regret}(K) := \sum_{k=1}^{K}\Big(V_1^{(\boldsymbol{r}(\pi_k^2),\pi_k^2)}(s_0) - V_1^{\pi_k}(s_0)\Big).$$

In 2P0S games this is a standard exploitability notion; one may alternatively use the duality gap $V_1^{\dagger,\pi_k^2}(s_0) - V_1^{\pi_k^1,\dagger}(s_0)$, and our analysis applies to either definition.

We emphasize the distinction between the number of episodes and the number of random features. Throughout the paper, K denotes the total number of learning episodes, while D denotes the number of random Fourier features used in the value-function representation. Thus, the dependence on K in our regret bounds is the standard cumulative-regret dependence over episodes, rather than a dependence on the number of Fourier components. The representation complexity enters through D and the effective dimension terms defined below.

Notation simplification in the following proof For notational simplicity, in the sequel we focus on player 1 and omit the superscript when it is clear from context.

B PSEUDOCODE OF ALGORITHM

Algorithm 1 Random-feature optimistic Nash Q-learning (RFNashQ)

Input: horizon H, episodes K, discount γ, feature map $\widetilde{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^D$, regularization λ, bonus schedule $\{\beta_{k,h}\}$ (or Eq. (A.1)), zero-sum stage-game solver $\mathcal{S}$, initial-state distribution ρ.
Initialize: for all $h \in [H]$, set $\Lambda_{1,h} \leftarrow \lambda I_D$ and $b_{1,h} \leftarrow 0 \in \mathbb{R}^D$.
for $k = 1, 2, \ldots, K$ **do**
  (Planning) Set $V_{k,H+1}(\cdot) \equiv 0$. For $h = H, H-1, \ldots, 1$:
    $\widehat{w}_{k,(h,a)} \leftarrow \Lambda_{k,(h,a)}^{-1} b_{k,h}$ and $\widehat{Q}_{k,h}(s,a) \leftarrow \widetilde{\phi}(s)^\top \widehat{w}_{k,(h,a)}$.
    Construct $Q_{k,h}$ and $V_{k,h}(s) = \mathrm{val}(Q_{k,h}(s,\cdot))$ via Eq. (A.1).
  Sample initial state $s_{k,1} \sim \rho$.
  for $h = 1, 2, \ldots, H$ **do**
    $(\mu_{k,h}(s_{k,h}), \nu_{k,h}(s_{k,h})) \leftarrow \mathcal{S}(Q_{k,h}(s_{k,h}, \cdot))$.
    Sample $a_{k,h}^1 \sim \mu_{k,h}(s_{k,h})$, $a_{k,h}^2 \sim \nu_{k,h}(s_{k,h})$; set $a_{k,h} = (a_{k,h}^1, a_{k,h}^2)$.
    Observe reward $r_h(s_{k,h}, a_{k,h})$ (player 1) and next state $s_{k,h+1}$.
    $y_{k,h} \leftarrow r_h(s_{k,h}, a_{k,h}) + \gamma V_{k,h+1}(s_{k,h+1})$.
    $\Lambda_{k+1,h} \leftarrow \Lambda_{k,h} + \widetilde{\phi}(s_{k,h})\widetilde{\phi}(s_{k,h})^\top$.
    $b_{k+1,h} \leftarrow b_{k,h} + \widetilde{\phi}(s_{k,h})\, y_{k,h}$.
  end for
end for

C PROOF OF LEMMA 5.1

In this section, we introduce some lemmas used in the later proof.

Lemma 5.1. (*Random Feature Kernel Consistency*). *Let*

$$\phi_D(x) = \frac{1}{\sqrt{D}}\big(\sqrt{2}\cos(\omega_1^\top x + b_1), \ldots, \sqrt{2}\cos(\omega_D^\top x + b_D)\big),$$

where $\omega_i \sim \mathcal{N}(0, \sigma^{-2} I)$ *and* $b_i \sim \mathrm{Unif}[0, 2\pi]$. *Then for any* x, x',

$$\mathbb{E}\big[\phi_D(x)^\top \phi_D(x')\big] = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right).$$

Moreover, the feature map is uniformly bounded as $\|\phi_D(x)\|_2 \le \sqrt{2}$ *for all* x.

Proof. The expectation over b gives

$$\mathbb{E}_b\Big[\tfrac{\sqrt{2}}{\sqrt{D}}\cos(u + b)\, \tfrac{\sqrt{2}}{\sqrt{D}}\cos(v + b)\Big] = \frac{1}{D}\cos(u - v).$$

For $\omega \sim \mathcal{N}(0, \sigma^{-2} I)$ and $\Delta = x - x'$,

$$\mathbb{E}_\omega[\cos(\omega^\top \Delta)] = \exp\big(-\tfrac{1}{2}\mathrm{Var}(\omega^\top \Delta)\big) = \exp\left(-\frac{\|\Delta\|_2^2}{2\sigma^2}\right),$$

since $\mathrm{Var}(\omega^\top \Delta) = \|\Delta\|_2^2/\sigma^2$. Combining the two expectations yields the result. □

Lemma C.1. *For any* x, $\|\phi_D(x)\|_2 \le \sqrt{2}$.

Proof. $\cos(\cdot) \le 1$, $\phi_M(x) = D^{-\frac{1}{2}}\left(\sqrt{2}\cos\left(\omega_1^\intercal x + b_1\right), \cdots, \sqrt{2}\cos\left(\omega_D^\intercal x + b_D\right)\right)$, hence $\|\phi(x)\|_2 \le \sqrt{2}$. □

D PROOF OF APPROXIMATION ERRORS

In this section, we prove a bound on the estimation error of our HDC-based linear approximation.

D.1 SELF-NORMALIZED REGRESSION BOUND AND RIDGE ERROR

In this section, we work at a fixed step h and (joint) action $\boldsymbol{a}$. For clarity, we suppress the index $(h, \boldsymbol{a})$ from most symbols when no confusion arises.

We first restate and define some notations for this subsection:

Filtration: $\{\mathcal{F}_t\}_{t\geq 0}$ is the history σ-algebra up to the beginning of round t.

Features and design:

$$\phi_t \coloneqq \phi\left(\boldsymbol{s}_t\right) \in \mathbb{R}^D, G_t \coloneqq \sum_{i=1}^{t} \phi_i \phi_i^{\mathsf{T}}, \Lambda_t \coloneqq \lambda I + G_t \ (\lambda > 0)\,.$$

TD Target and centered noise:

$$y_t \coloneqq r_t + \gamma \widehat{V}_{t+1}\left(\boldsymbol{s}\right), \epsilon_t \coloneqq y_t - \mathbb{E}\left[y_t | \mathcal{F}_t\right].$$

Self-normalized sum:

$$S_t \coloneqq \sum_{i=1}^{t} \phi_i \epsilon_i \in \mathbb{R}^D.$$

D.1.1 Exponential super-martingale (scalar-to-vector lifting)

Lemma D.1. *For any fixed vector $u \in \mathbb{R}^D$ and any $t \geq 1$,*

$$\mathbb{E}\left[\exp\left(\langle u, \phi_t\rangle \epsilon_t\right) | \mathcal{F}_t\right] \leq \exp\left(\frac{1}{2}\sigma_\epsilon^2 \langle u, \phi_t\rangle^2\right) \ a.s.$$

Proof. Condition on $\mathcal{F}_t$. The random variable ϵ is conditionally mean-zero and sub-Gaussian with proxy σ_ϵ^2 by (SG), so

$$\mathbb{E}\left[\exp\left(\lambda \epsilon_t\right) | \mathcal{F}_t\right] \leq \exp\left(\lambda^2 \sigma_\epsilon^2 / 2\right).$$

Then we apply this with $\lambda = \langle u, \phi_t\rangle$ to get the result. □

Lemma D.2. *For any fixed $u \in \mathbb{R}^D$, the process*

$$M_t\left(u\right) \coloneqq \exp\left(\underbrace{\langle u, S_t\rangle}_{\sum_{i\leq t}\langle u, \phi_i\rangle \epsilon_i} - \frac{1}{2}\sigma_\epsilon^2 \sum_{i=1}^{t} \langle u, \phi_i\rangle^2\right), t \geq 0,$$

is a nonnegative supermartingale w.r.t.$\{\mathcal{F}_t\}$, with $\mathbb{E}\left[M_t\left(u\right)\right] \leq 1$ for all t.

Proof. Using Lemma D.1 and the tower property, we have

$$\mathbb{E}\left[M_t\left(u\right) | \mathcal{F}_{t-1}\right] = M_{t-1}\left(u\right) \mathbb{E}\left[\exp\left(\langle u, \phi_t\rangle \epsilon_t - \frac{1}{2}\sigma_\epsilon^2 \langle u, \phi_t\rangle^2\right) | \mathcal{F}_t\right] \leq M_{t-1}\left(u\right).$$

Iterate to obtain

$$\mathbb{E}\left[M_t\left(u\right)\right] \leq \mathbb{E}\left[M_0\left(u\right)\right] = 1.$$

□

Also, the exponent can be rewritten as $\langle u, S_t\rangle - \frac{1}{2}\sigma_\epsilon^2\, u^{\mathsf{T}} G_t u$.

D.1.2 Vector self-normalized mgf via the mixture method

We now convert the family $\{M_t(u)\}_{u\in\mathbb{R}^D}$ into a single bound on the vector S_t via a Gaussian mixture over u. The trick is to integrate $M_t(u)$ against a data-dependent Gaussian density and evaluate the integral by completing the square.

Lemma D.3. *For any positive-definite matrix $A \succ 0$ and any vector $S \in \mathbb{R}^D$,*

$$\int_{\mathbb{R}^D} \exp\left(\langle u, S\rangle - \frac{1}{2}u^\mathsf{T} A u\right) du = (2\pi)^{\frac{D}{2}} (\det A)^{-\frac{1}{2}} \exp\left(\frac{1}{2}S^\mathsf{T} A^{-1} S\right).$$

Proof. Complete the square:

$$\langle u, S\rangle - \frac{1}{2}u^\mathsf{T} A u = -\frac{1}{2}\big(u - A^{-1}S\big)^\mathsf{T} A\big(u - A^{-1}S\big) + \frac{1}{2}S^\mathsf{T} A^{-1} S.$$

Integrating the centered Gaussian kernel gives

$$\int_{\mathbb{R}^D} \exp\left(\langle u, S\rangle - \frac{1}{2}u^\mathsf{T} A u\right) du = (2\pi)^{D/2} \det(A)^{-1/2} \exp\left(\frac{1}{2}S^\mathsf{T} A^{-1} S\right).$$

□

Now we apply Lemma D.3 with $A = \sigma_\epsilon^2 \Lambda_t$ and $S = S_t$.

Lemma D.4. *For all $t \geq 0$,*

$$\mathbb{E}\left[\exp\left(\frac{1}{2\sigma_\epsilon^2}\|S_t\|^2_{\Lambda_t^{-1}}\right)\right] \leq \frac{\det(\Lambda_t)^{1/2}}{\det(\lambda I)^{1/2}}.$$

Proof. Define

$$I_t = \int_{\mathbb{R}^D} \exp\left(\langle u, S_t\rangle - \frac{1}{2}\sigma_\epsilon^2 u^\mathsf{T}\Lambda_t u\right) du.$$

Then by Lemma D.3,

$$I_t = (2\pi)^{\frac{D}{2}} \sigma_\epsilon^{-D} (\det \Lambda_t)^{-\frac{1}{2}} \exp\left(\frac{1}{2\sigma_\epsilon^2} S_t^\mathsf{T}\Lambda_t^{-1} S_t\right). \tag{B.1.2.1}$$

On the other hand, we write

$$\langle u, S_t\rangle - \frac{1}{2}\sigma_\epsilon^2 u^\mathsf{T}\Lambda_t u = \underbrace{\left(\langle u, S_t\rangle - \frac{1}{2}\sigma_\epsilon^2 u^\mathsf{T} G_t u\right)}_{\log M_t(u)} - \frac{1}{2}\sigma_\epsilon^2 \lambda \|u\|_2^2.$$

Hence, for every fixed u,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\langle u, S_t\rangle - \frac{1}{2}\sigma_\epsilon^2 u^\mathsf{T}\Lambda_t u\right)\right] &\leq \mathbb{E}[M_t(u)] \exp\left(-\frac{1}{2}\sigma_\epsilon^2\lambda\|u\|^2\right)\\ &\leq \exp\left(-\frac{1}{2}\sigma_\epsilon^2\lambda\|u\|^2\right).\end{aligned}$$

By Lemma D.2, Integrate both sides over $u \in \mathbb{R}^D$ and evaluate the right-hand Gaussian integral:

$$\mathbb{E}[I_t] \leq \int \exp\left(-\frac{1}{2}\sigma_\epsilon^2\lambda\|u\|^2\right) du = (2\pi)^{\frac{D}{2}} \sigma_\epsilon^{-D} (\det(\lambda I))^{-\frac{1}{2}}. \tag{B.1.2.2}$$

Combine (B.1.2.1) and (B.1.2.2), cancel the common constant $(2\pi)^{\frac{D}{2}}\sigma_\epsilon^{-D}$, and apply Jensen linearity; we have

$$\mathbb{E}\left[(\det(\Lambda_t))^{-\frac{1}{2}}\exp\left(\frac{1}{2\sigma_\epsilon^2}\|S_t\|^2_{\Lambda_t^{-1}}\right)\right] \le (\det(\lambda I))^{-\frac{1}{2}}.$$

Since $(\det(\Lambda_t))^{-\frac{1}{2}}$ is $\mathcal{F}_t$-measurable and positive, we can multiply both sides by $(\det(\Lambda_t))^{\frac{1}{2}}$ and then use the tower property to obtain

$$\mathbb{E}\left[\exp\left(\frac{1}{2\sigma_\epsilon^2}\|S_t\|^2_{\Lambda_t^{-1}}\right)\right] \le \mathbb{E}\left[\frac{\det(\Lambda_t)^{1/2}}{\det(\lambda I)^{1/2}}\right].$$

□

Corollary D.1. *For any $\delta \in (0,1)$, with probability of at least $1-\delta$,*

$$\|S_t\|_{\Lambda_t^{-1}} \le \sigma_\epsilon\sqrt{2\log\frac{\det(\Lambda_t)^{1/2}}{\det(\lambda I)^{1/2}\delta}}.$$

Proof. We apply Markov's inequality to Lemma D.4, and get

$$\Pr\left(\frac{1}{2\sigma_\epsilon^2}\|S_t\|^2_{\Lambda_t^{-1}} > \log\frac{\det(\Lambda_t)^{1/2}}{\det(\lambda I)^{1/2}\delta}\right) \le \frac{\mathbb{E}\left[\exp\left(\frac{1}{2\sigma_\epsilon^2}\|S_t\|^2_{\Lambda_t^{-1}}\right)\right]}{\mathbb{E}\left[\frac{\det(\Lambda_t)^{1/2}}{\det(\lambda I)^{1/2}\delta}\right]} \le \delta.$$

□

D.1.3 Ridge normal equations and pointwise prediction error

Now let $w^\dagger$ denote the population least-squares projection of the regression target

$$f^\star(s) \coloneqq \mathbb{E}\left[r + \gamma V(\boldsymbol{s}')\,|\boldsymbol{s},\boldsymbol{a}\right]$$

onto the linear span of $\phi(\cdot)$. That is,

$$w^\dagger \in \operatorname*{argmin}_{w\in\mathbb{R}^D}\mathbb{E}\left(\phi(\boldsymbol{s})^\mathsf{T} w - f^\star(\boldsymbol{s})\right)^2.$$

Let w^k be the ridge solution using samples $i<k$:

$$w^k \in \operatorname*{argmin}_w \sum_{i<k}\left(y_i - \phi_i^\mathsf{T} w\right)^2 + \lambda\|w\|_2^2.$$

Now we have the following lemma:

Lemma D.5. *The first-order optimality condition yields*

$$\Lambda_k\left(w^k - w^\dagger\right) = S_{k-1} - \lambda w^\dagger.$$

Proof. The gradient of the objective at w^k is zero:

$$\sum_{i<k}\phi_i\left(\phi^\mathsf{T} w^k - y_i\right) + \lambda w^k = 0.$$

Add and subtract $\phi_i\left(\phi_i^\mathsf{T} w^\dagger - \mathbb{E}\left[y_i|\mathcal{F}_i\right]\right)$ on the LHS, group terms, since $w^\dagger$ is the ridge optimal and use $S_{k-1} = \sum_{i<k}\phi_i\epsilon_i$ to get

$$\left(\sum_{i<k}\phi_i\phi_i^\mathsf{T} + \lambda I\right)\left(w^k - w^\dagger\right) = \sum_{i<k}\phi_i\left(y_i - \mathbb{E}\left[y_i|\mathcal{F}_i\right]\right) - \lambda w^\dagger.$$

Recognize Λ_k and S_{k-1}.

□

Lemma D.6. *For any $\boldsymbol{x} \in \mathcal{S}$ with feature $\phi(x)$,*

$$\left|\phi(\boldsymbol{x})^{\mathsf{T}}\left(w^k - w^\dagger\right)\right| \leq \left(\|S_{k-1}\|_{\Lambda_k^{-1}} + \sqrt{\lambda}\left\|w^\dagger\right\|_2\right)\sqrt{\phi(x)^{\mathsf{T}}\Lambda_k^{-1}\phi(x)}.$$

Proof. Left-multiply Lemma D.5 by $\phi(x)^{\mathsf{T}}\Lambda_k^{-1}$ and take absolute values:

$$\begin{aligned}\left|\phi^{\mathsf{T}}\left(w^k - w^\dagger\right)\right| &= \left|\phi^{\mathsf{T}}\Lambda_k^{-1}S_{k-1} - \lambda\phi^{\mathsf{T}}\Lambda_k^{-1}w^\dagger\right| \\ &\leq \|\phi\|_{\Lambda_k^{-1}}\|S_{k-1}\|_{\Lambda_k^{-1}} + \lambda\|\phi\|_{\Lambda_k^{-1}}\left\|w^\dagger\right\|_{\Lambda_k^{-1}}.\end{aligned}$$

We then use Cauchy-Schwarz in the Λ_k^{-1} norm and $\left\|w^\dagger\right\|_{\Lambda_k^{-1}} \leq \lambda^{-\frac{1}{2}}\left\|w^\dagger\right\|_2$, which yields the result. □

Corollary D.2. *Combining Corollary D.1 with Lemma D.6, define*

$$\beta_k \coloneqq \sigma_\epsilon\sqrt{2\log\frac{\det(\Lambda_t)^{1/2}}{\det(\lambda I)^{1/2}\delta}} + \sqrt{\lambda}\left\|w^\dagger\right\|_2, \mathrm{rad}_k(x) \coloneqq \beta_k\sqrt{\phi(x)^{\mathsf{T}}\Lambda_k^{-1}\phi(x)}.$$

Then, with probability at least $1-\delta$,

$$\left|\phi(x)^{\mathsf{T}}\left(w^k - w^\dagger\right)\right| \leq \mathrm{rad}_k(x), \forall x.$$

D.2 RANDOM FEATURE APPROXIMATION ERROR

For each step $h \in [H]$ and player $i \in [N_p]$, the Bayesian regression target is the optimal action–value function

$$Q_{i,h}^*(\boldsymbol{s},\boldsymbol{a}) = \mathbb{E}\big[r_{i,h}(\boldsymbol{s},\boldsymbol{a}) + \gamma V_{i,h+1}^*(\boldsymbol{s}') \,\big|\, \boldsymbol{s},\boldsymbol{a}\big], \quad (s,a) \in S \times A.$$

Given the normalized HDC feature map $\phi : S \times A \to \mathbb{R}^D$, we define the per-step random-feature approximation error along the sample path as

$$\varepsilon_{\mathrm{rff},h,k} \coloneqq \sup_{k'<k}\left|Q_{i,h}^*(s_{k',h},a_{k',h}) - \phi(s_{k',h},a_{k',h})^\top w_{i,h}^*\right|. \tag{B.1}$$

where

$$w_{i,h}^* \in \arg\min_{w\in\mathbb{R}^D}\mathbb{E}\big[(\phi(s,a)^\top w - Q_{i,h}^*(s,a))^2\big].$$

Assumption D.1 (HDC random-feature approximation error)**.** *Let $n_h(k)$ denote the number of visits to step h prior to episode k, and assume there exist constants $c_-, c_+ > 0$ such that*

$$c_- k \;\leq\; n_h(k) \;\leq\; c_+ k \qquad \textit{for all } h \in [H],\ k \in [K].$$

Under the effective-dimension and source conditions in Section A, there exist constants $c_1, c_2 > 0$ such that, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of the HDC/random features and the sample path, the following holds for all $h \in [H]$ and $k \in [K]$:

$$\varepsilon_{\mathrm{rff},h,k} \;\leq\; c_1\sqrt{\frac{\log(1/\delta)}{D}} \;+\; c_2\Big(\sqrt{\frac{N(\lambda)}{n_h(k)}} + \lambda^r\Big). \tag{B.2}$$

Here $N(\lambda)$ is the effective dimension defined in Section A and $r \geq \frac{1}{2}$ is the source exponent.

For a finite D-dimensional kernel, we have the following lemma:

Lemma D.7. *For any $(\boldsymbol{s},\boldsymbol{a})\,(\boldsymbol{s}',\boldsymbol{a}')$ and $\forall 0 < \delta < 1$, with a probability of at least $1-\delta$ over RFF sampling,*

$$\left|\widetilde{\phi}(\boldsymbol{s},\boldsymbol{a})^{\mathsf{T}}\widetilde{\phi}(\boldsymbol{s}',\boldsymbol{a}') - K\left((\boldsymbol{s},\boldsymbol{a}),(\boldsymbol{s}',\boldsymbol{a}')\right)\right| \leq c_1\sqrt{\frac{\log(1/\delta)}{D}}.$$

Proof. By definition in Definition A, we have that

$$\widetilde{\phi}(\boldsymbol{s},\boldsymbol{a})^{\mathsf{T}}\widetilde{\phi}(\boldsymbol{s}',\boldsymbol{a}')=\frac{1}{D}\sum_{j=1}^{D}Z_j,$$

where $Z_j = 2\left(\cos(\omega_j^\top x+b_j)\right)\cos(\omega_j^\top x+b_j)\in[-2,2]$. It's mean $\mu_{Z_j} = K\left((\boldsymbol{s},\boldsymbol{a}),(\boldsymbol{s}',\boldsymbol{a}')\right)$. So with Hoeffding's inequality, we have

$$\left|\widetilde{\phi}(\boldsymbol{s},\boldsymbol{a})^{\mathsf{T}}\widetilde{\phi}(\boldsymbol{s}',\boldsymbol{a}')-K\left((\boldsymbol{s},\boldsymbol{a}),(\boldsymbol{s}',\boldsymbol{a}')\right)\right|\leq c_1\sqrt{\frac{\log(1/\delta)}{D}}.$$

□

Then we have the lemma for KRR variance & bias.

Lemma D.8. *With assumption Assumption A, we have that for a sample size n, let the (population) ridge solution $f_\lambda \in H$ solve $(C+\lambda I)f_\lambda = Cf^*_{h,a}$, and the empirical KRR estimator*

$$\widehat{f}_\lambda \in \operatorname*{argmin}_{f\in H}\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i)-y_i\right)^2+\lambda\|f\|_H^2.$$

Then with probability at least $1-\delta$,

$$\left\|\widehat{f}_\lambda - f^*_{h,a}\right\|_{L^2(\rho X)}\leq\widetilde{O}\left(\sqrt{\frac{N(\lambda)}{n}}\right)+\widetilde{O}\left(\lambda^r\right).$$

In particular, the variance term is $\widetilde{O}\left(\sqrt{\frac{N(\lambda)}{n}}\right)$, and the bias term is $\widetilde{O}\left(\lambda^r\right)$.

Proof. We firstly define the sampling operator $S_n : H\to\mathbb{R}^n$ by

$$(S_n f)_i = f(x_i),$$

and its adjoint $S_n^*:\mathbb{R}^n\to H$ via

$$S_n^* a=\sum_{i=1}^{n}a_i K_{x_i}.$$

The empirical covariance $C_n \coloneqq \frac{1}{n}S_n^*S_n : H\to H$ and the population covariance $C=T^*T$ share the nonzero spectrum with $L=TT^*$. We use the $L^2(\rho X)$-norm $\|h\|_{L^2}=\|Th\|_{L^2}$ for $h\in H$.

Step 1: Normal equations and the fundamental decomposition.

KRR normal equation gives that

$$(C_n+\lambda I)\widehat{f}_\lambda=\frac{1}{n}S_n^*y,(C+\lambda I)f_\lambda=Cf_\star.$$

Subtracting,

$$\begin{aligned}(C_n+\lambda I)\left(\widehat{f}_\lambda-f_\lambda\right)=\frac{1}{n}S_n^*\left(y-f_\star(x)\right)+(C_n-C)(f_\lambda-f_\star)\\+(C_n-C)f_\star-(C-C)f_\star.\end{aligned}$$

where we grouped the terms so that the first is the noise term and the second multiplies the bias vector $(f_\lambda - f_\star)$, Solve for $\widehat{f}_\lambda - f_\lambda$to get

$$\widehat{f}_\lambda-f_\lambda=(C_n+\lambda I)^{-1}\Big[\underbrace{\tfrac{1}{n}S_n^{\xi}}_{\text{noise}}+\underbrace{(C_n-C)(f_\lambda-f_\star)}_{\text{covariance deviation on bias}}\Big].$$

Step 2: On $C + \lambda I$

We introduce the resolvent sandwich

$$\left\|(C+\lambda I)^{\frac{1}{2}}\left(\widehat{f}_\lambda - f_\lambda\right)\right\|_H \leq \left\|(C+\lambda I)^{\frac{1}{2}}(C+\lambda I)^{-1}(C+\lambda I)^{\frac{1}{2}}\right\| \cdot \left\|(C+\lambda I)^{-\frac{1}{2}}[\cdots]\right\|_H.$$

On the high probability event,

$$\left\|(C+\lambda I)^{\frac{1}{2}}(C_n - C)(C+\lambda I)^{-\frac{1}{2}}\right\| \leq \frac{1}{2},$$

The Neumann series yields

$$\left\|(C+\lambda I)^{\frac{1}{2}}(C_n+\lambda I)^{-1}(C+\lambda I)^{\frac{1}{2}}\right\| \leq 2.$$

Hence

$$\begin{aligned}\left\|T\left(\widehat{f}_\lambda - f_\lambda\right)\right\|_{L^2} &= \left\|\left(\widehat{f}_\lambda - f_\lambda\right)\right\|_{L^2} \\ &\leq 2\left\|(C+\lambda I)^{-\frac{1}{2}}\left[\frac{1}{n}S_n^*\xi + (C_n - C)(f_\lambda - f_\star)\right]\right\|_H.\end{aligned}$$

Step 3: Variance term

For the noise part, condition on $\{x_i\}_{i=1}^n$, we have

$$Z \coloneqq \left\|(C+\lambda I)^{-\frac{1}{2}}\frac{1}{n}S_n^*\xi\right\|_H^2, G_\lambda \coloneqq S_n(C+\lambda I)^{-1}S_n^*.$$

Since ξ is conditionally σ^2-sub-Gaussian and $G_\lambda \succeq 0$, standard Hanson–Wright/chi-square tails give (e.g., by the second moment and Bernstein)

$$Z \leq \frac{2\sigma^2}{n^2}Tr(G_\lambda)\log\left(\frac{2}{\delta}\right) \text{ w.p. } \geq 1 - \frac{\delta}{2}.$$

Moreover,

$$\begin{aligned}Tr(G_\lambda) =& Tr\left((C+\lambda I)^{-1}S_n^*S_n\right) \\ =& nTr\left((C+\lambda I)^{-1}C_n\right) \\ \leq& nTr\left((C+\lambda I)^{-1}C\right) \\ =& nN(\lambda).\end{aligned}$$

Thus,

$$\begin{aligned}\left\|(C+\lambda I)^{-\frac{1}{2}}\frac{1}{n}S_n^*\xi\right\|_H \leq& \sigma\sqrt{\frac{2N(\lambda)\log\left(\frac{2}{\delta}\right)}{n}} \\ =& \widetilde{O}\left(\sqrt{\frac{N(\lambda)}{n}}\right).\end{aligned}$$

Step 4: Bias vector and covariance deviation

First, compute the population bias:

$$\begin{aligned}f_\lambda - f_\star &= (C+\lambda I)^{-1}Cf_\star - f_\star \\ &= -\lambda(C+\lambda I)^{-1}f_\star.\end{aligned}$$

By spectral calculus with $f_\star = L^r g$ and the unitary equivalence of C and L,

$$\begin{aligned}\left\|(C+\lambda I)^{-\frac{1}{2}}(f_\lambda - f_\star)\right\|_H &= \lambda \left\|(C+\lambda I)^{\frac{1}{2}} f_\star\right\|_H \\ &\le \lambda^r \|g\|_{L^2(\rho X)}.\end{aligned}$$

(Indeed, on each eigenvalue μ of C, $\lambda\mu^r/(\mu+\lambda) \le \lambda^r$ for $r \ge \frac{1}{2}$.)

Next, bound the empirical covariance deviation:

$$\left\|(C+\lambda I)^{-\frac{1}{2}}(C_n - C)(f_\lambda - f_\star)\right\|_H \le \left\|(C+\lambda I)^{-\frac{1}{2}}(C_n - C)(C+\lambda I)^{-\frac{1}{2}}\right\| \left\|(C+\lambda I)^{\frac{1}{2}}(f_\lambda - f_\star)\right\|_H.$$

By step 2 and the bias bound above,

$$\left\|(C+\lambda I)^{-\frac{1}{2}}(C_n - C)(f_\lambda - f_\star)\right\|_H \le \frac{1}{2}\cdot\lambda^r \|g\|_{L^2} = \widetilde{O}(\lambda^r).$$

Step 5: Putting together and translating to L^2-error

Combine steps 2,3,4, we get

$$\left\|\widehat{f}_\lambda - f_\lambda\right\|_{L^2} \le 2\cdot\widetilde{O}\left(\sqrt{\frac{N(\lambda)}{n}}\right) + 2\cdot\widetilde{O}(\lambda^r) = \widetilde{O}\left(\sqrt{\frac{N(\lambda)}{n}}\right) + \widetilde{O}(\lambda^r).$$

Finally, we get

$$\left\|\widehat{f}_\lambda - f_\star\right\|_{L^2} \le \left\|\widehat{f}_\lambda - f_\lambda\right\|_{L^2} + \|f_\lambda - f_\star\|_{L^2} \le \widetilde{O}\left(\sqrt{\frac{N(\lambda)}{n}}\right) + \widetilde{O}(\lambda^r),$$

Since $\|f_\lambda - f_\star\|_{L^2} \le \left\|(C+\lambda I)^{\frac{1}{2}}(f_\lambda - f_\star)\right\|_H \le \lambda^r \|g\|$. □

Remark D.1 (Role of Lemma C.8)**.** *Lemma C.8 is stated for i.i.d. samples from the reference measure ρ_X and controls the KRR error in $L^2(\rho_X)$. Since the algorithm collects data under adaptive visitation distributions $d_{k,h}$, Lemma C.8 does not by itself imply Assumption 5.1. It only motivates the scale of the approximation term. To derive the pathwise event from a reference-measure bound, one would need an additional coverage/concentrability condition or a uniform approximation argument. Our regret proof is stated conditional on Assumption 5.1.*

Here is the detailed proof of the lemma of the standard concentration event used in Step 2.

Lemma D.9. *Under the assumptions of Lemma D.8 and $\sup_x K(x,x) \le \kappa^2$, there is a universal $c > 0$ such that with probability at least $1-\delta$,*

$$\left\|(C+\lambda I)^{-\frac{1}{2}}(C - n - C)(C+\lambda I)^{-\frac{1}{2}}\right\| \le c\left(\sqrt{\frac{N(\lambda) + \log\frac{1}{\delta}}{n}} + \frac{\log\frac{1}{\delta}}{n\lambda}\right).$$

In particular, for n large enough so that the RHS $\le \frac{1}{2}$, the event in Proof D.2 step 2 holds.

Proof. Set the centered, self-adjoint summands in $\mathcal{L}(\mathcal{H})$:

$$Y_i \coloneqq (C+\lambda I)^{-\frac{1}{2}}(K_{x_i}\otimes K_{x_i} - C)(C+\lambda I)^{-\frac{1}{2}}, i = 1, \cdots, n.$$

Then $\mathbb{E}[Y_i] = 0$ and

$$Z_n \coloneqq (C+\lambda I)^{-\frac{1}{2}}(C_n - C)(C+\lambda I)^{-\frac{1}{2}} = \frac{1}{n}\sum_{i=1}^n Y_i.$$

We will apply Tropp's matrix Bernstein inequality with intrinsic dimension. For that, we bound the single–summand norm $\|Y_i\|$, and the variance operator $V \coloneqq \sum_{i=1}^{n} \mathbb{E}\left[Y_i^2\right]$ through its operator norm $\|V\|$ and its trace $Tr\left(V\right)$.

We let

$$g_x \coloneqq \left(C + \lambda I\right)^{-\frac{1}{2}} K_x \in \mathcal{H}.$$

Then

$$\left(C + \lambda I\right)^{-\frac{1}{2}} \left(K_x \otimes K_x\right) \left(C + \lambda I\right)^{-\frac{1}{2}} = g_x \otimes g_x,$$
$$\|g_x\|_H^2 = \langle K_x, \left(C + \lambda I\right)^{-1} K_x \rangle_H \leq \frac{K\left(x,x\right)}{\lambda} \leq \frac{\kappa^2}{\lambda}.$$

Also, we set

$$B \coloneqq \left(C + \lambda I\right)^{-\frac{1}{2}} C \left(C + \lambda I\right)^{-\frac{1}{2}},$$

then $0 \preceq B \preceq I$, so $\|B\| \leq 1$. Hence

$$\begin{aligned}
\|Y_i\| &= \|g_{x_i} \otimes g_{x_i} - B\| \\
&\leq \|g_{x_i} \otimes g_{x_i}\| + \|B\| \\
&\leq \frac{\kappa^2}{\lambda} + 1 \\
&\leq \frac{2\kappa^2}{\lambda},
\end{aligned}$$

absorbing constants for convenience. We write $A_x \coloneqq g_x \otimes g_x$ (rank-one, with $A_x^2 = \|g_x\|_H^2 A_x$). Since $Y_i = A_{x_i} - B$,

$$\begin{aligned}
Y_i^2 &= \left(A_{x_i} - B\right)^2 \\
&\preceq 2\left(A_{x_i}^2 + B^2\right) \\
&\preceq 2\left(\|g_{x_i}\|_H^2 A_{x_i} + B\right) \\
&\preceq 2\left(\frac{\kappa^2}{\lambda}\right) B.
\end{aligned}$$

Taking expectation and summing,

$$\|V\| \leq 2n\left(\frac{\kappa^2}{\lambda} + 1\right)\|B\| \leq 2n\left(\frac{\kappa^2}{\lambda} + 1\right),$$
$$Tr\left(V\right) \leq 2n\left(\frac{\kappa^2}{\lambda} + 1\right) Tr\left(B\right).$$

Note that $Tr\left(B\right) = Tr\left(\left(C + \lambda I\right)^{-1} C\right) = N\left(\lambda\right)$.

Let $S \coloneqq \sum_{i=1}^{n} Y_i$, Tropp [2015](6.1.4) shows that for zero-mean, independent, self-adjoint Y_i, with $L \coloneqq \max_i \|Y_i\|$ and $V \coloneqq \sum \mathbb{E}\left[Y_i^2\right]$, for $\forall t > 0$,

$$\Pr\left\{\|S\| \geq t\right\} \leq 4\frac{Tr\left(V\right)}{\|V\|} \exp\left(-\frac{t^2/2}{\|V\| + Lt/3}\right).$$

So absorbing all the constants, we have that $L \lesssim \kappa^2/\lambda$, $\|V\| \lesssim n\left(\kappa^2/\lambda + 1\right)$, and $\frac{Tr(V)}{\|V\|} \lesssim N\left(\lambda\right)$.

Set $t = n\epsilon$ to bound $\|Z_n\| = \|S\|/n$. The previous inequality yields

$$\Pr\left\{\|Z_n\| \geq \epsilon\right\} \leq 4c_0 N\left(\lambda\right) \exp\left(-\frac{n\epsilon^2/2}{c_1\left(\kappa^2/\lambda + 1\right)} + c_2\left(\kappa^2/\lambda\right)\epsilon\right).$$

Choosing ϵ so that the RHS is $\leq \delta$ and solving the quadratic in ϵ gives the standard Bernstein form

$$\|Z_n\| \leq c\left(\sqrt{\frac{N\left(\lambda\right) + \log\left(\frac{1}{\delta}\right)}{n}} + \frac{\log\left(\frac{1}{\delta}\right)}{n\lambda}\right).$$

So we have the proof. □

D.3 ONE-STEP DECOMPOSITION AND LIPSCHITZ OF STAGE VALUE

In this subsection, we present the lemmas on one-step decomposition and the properties of the Stage Value.

Lemma D.10. *Let $Q_h^\dagger(\boldsymbol{s},\boldsymbol{a}) = \widetilde{\phi}(\boldsymbol{s})^\mathsf{T} w_{h,\boldsymbol{a}}^\dagger$, for any $(\boldsymbol{s},\boldsymbol{a})$, we can write*

$$\left|\left(\mathcal{T}\widehat{Q}\right)_h(\boldsymbol{s},\boldsymbol{a}) - \widehat{Q}_h(\boldsymbol{s},\boldsymbol{a})\right| \le \underbrace{\left|\left(\mathcal{T}Q^\dagger\right)_h(\boldsymbol{s},\boldsymbol{a}) - Q_h^\dagger(\boldsymbol{s},\boldsymbol{a})\right|}_{A} + \underbrace{\left|Q_h^\dagger(\boldsymbol{s},\boldsymbol{a}) - \widetilde{\phi}(\boldsymbol{s})^\mathsf{T}\widehat{\omega}_{k,(h,\boldsymbol{a})}\right|}_{B} + \underbrace{\gamma\mathbb{E}_{\boldsymbol{s}'}\left|V_{\widehat{Q},h+1}(\boldsymbol{s}') - V_{Q^\dagger,h+1}(\boldsymbol{s}')\right|}_{C}.$$

Moreover, we have $A \le \eta_{k,h}$ by results in Proof D.2, and

$$\left|V_{\widehat{Q},h}(\boldsymbol{s}) - V_{Q^\dagger,h}(s)\right| \le \max_{\boldsymbol{a}}\left|\phi(\boldsymbol{s})^\mathsf{T}\left(\widehat{w}_{k,h} - w_h^\dagger\right)\right| \le \max_{\boldsymbol{a}}\mathrm{rad}_{k,(h,\boldsymbol{a})}(\boldsymbol{s}).$$

Proof. Add and subtract $Q^\dagger$, and use the triangle inequality to obtain the result. Let A be the approximation/statistical error of the one-step regression target; bound it by D.1.

For the value-difference term, we use the 1-Lipschitz property of the stage value (Lemma D.11) in the $\|\cdot\|_\infty$ norm and apply Corollary D.1. □

Lemma D.11. *For any two payoff matrices M, M',*

$$|\mathrm{Val}(M) - \mathrm{Val}(M')| \le \|M - M'\|_\infty.$$

Proof. Let (p^*, q^*) be an equilibrium for M. Then we have

$$\begin{aligned}\mathrm{Val}(M') &\ge p^{*\mathsf{T}} M' q^* \\ &= \mathrm{Val}(M) + p^{*\mathsf{T}}(M - M')q^* \\ &\ge \mathrm{Val}(M) - \|M - M'\|_\infty.\end{aligned}$$

Swap M, M' to obtain the other direction. □

E PROOF OF REGRET GUARANTEES

In this section, we prove the sub-linearity of the regret that our algorithm reaches in 2-player 0-sum games.

E.1 OPTIMISM (FOR THE OPTIMISTIC VARIANT)

Definition E.1. *For each step h, state $\boldsymbol{s}$ and joint action $\boldsymbol{a}$, we define the optimistic Q $\widehat{Q}_h^{opt}(\boldsymbol{s},\boldsymbol{a})$ and value $\widehat{V}_h^{opt}(\boldsymbol{s})$to be*

$$\begin{aligned}\widehat{Q}_h^{opt}(\boldsymbol{s},\boldsymbol{a}) &\coloneqq \phi(\boldsymbol{s})^\mathsf{T}\widehat{w}_{k,(h,\boldsymbol{a})} + 2\mathrm{rad}_{k,(h,\boldsymbol{a})}(\boldsymbol{s}) + \varepsilon_{\mathrm{rff},k,h}, \\ \widehat{V}_h^{opt}(\boldsymbol{s}) &\coloneqq \min\left\{H, \mathrm{Val}\left(\widehat{Q}_h^{opt}(\boldsymbol{s},\cdot)\right)\right\}.\end{aligned}$$

Our object is to prove the following lemma:

Lemma E.1. *For all $k, h, \boldsymbol{s}, \boldsymbol{a}$ and for each player i,*

$$Q_h^{i,br\left(\pi_{-i}^{opt}\right),\pi_{-i}^{opt}}(\boldsymbol{s},\boldsymbol{a}) \le \widehat{Q}_h^{i,opt}(\boldsymbol{s},\boldsymbol{a}).$$

In particular, for player 1's payoff, $Q_h(\boldsymbol{s},\boldsymbol{a}) \le \widehat{Q}_h^{opt}(\boldsymbol{s},\boldsymbol{a})$, and thus have $V_h(\boldsymbol{s}) \le \widehat{V}_h^{opt}(\boldsymbol{s})$.

The proof proceeds by backward induction on h.

Proof. At the terminal step, we have

$$Q_H(\boldsymbol{s},\boldsymbol{a}) = r_H(\boldsymbol{s},\boldsymbol{a}).$$

Let

$$f^{\dagger}_{H,\boldsymbol{a}}(\boldsymbol{s}) \coloneqq r_H(\boldsymbol{s},\boldsymbol{a}), Q^{\dagger}_H(\boldsymbol{s},\boldsymbol{a}) \coloneqq \phi(\boldsymbol{s})^{\mathsf{T}} w^{\dagger}_H$$

be the population regression target and its best-in-class linear predictor (in the RFF-induced linear class). We have

$$\left|Q^{\dagger}_H(\boldsymbol{s},\boldsymbol{a}) - f^{\dagger}_{H,\boldsymbol{a}}(\boldsymbol{s})\right| \le \varepsilon_{\mathrm{rff},k,H}, \left|\phi(\boldsymbol{s})^{\mathsf{T}}\left(\widehat{w}_{k(H,\boldsymbol{a})} - w^{\dagger}_{H,\boldsymbol{a}}\right)\right| \le \mathrm{rad}_{k,(H,\boldsymbol{a})}(\boldsymbol{s}).$$

So we have

$$\begin{aligned} Q_H(\boldsymbol{s},\boldsymbol{a}) = f^{\dagger}_{H,\boldsymbol{a}}(\boldsymbol{s}) &\le Q^{\dagger}_H(\boldsymbol{s},\boldsymbol{a}) + \varepsilon_{\mathrm{rff},k,H} \\ &\le \phi(\boldsymbol{s})^{\mathsf{T}}\widehat{w}_{k,(H,\boldsymbol{a})} + \mathrm{rad}_{k,(H,\boldsymbol{a})}(\boldsymbol{s}) + \varepsilon_{\mathrm{rff},k,H} \\ &\le \phi(\boldsymbol{s})^{\mathsf{T}}\widehat{w}_{k,(H,\boldsymbol{a})} + 2\mathrm{rad}_{k,(H,\boldsymbol{a})}(\boldsymbol{s}) + \varepsilon_{\mathrm{rff},k,H} \\ &= \widehat{Q}^{opt}_H(\boldsymbol{s},\boldsymbol{a}). \end{aligned}$$

Thus $Q_H \le \widehat{Q}^{opt}_H$. By monotonicity of $\mathrm{Val}(\cdot)$ under entry wise order and since

$$V_H(\boldsymbol{s}) = \mathrm{Val}(Q_H(\boldsymbol{s},\cdot)),$$

we get

$$V_H(\boldsymbol{s}) \le \widehat{V}^{opt}_H(\boldsymbol{s}).$$

We assume that for some $h+1$ with $(1 \le h \le H)$ that for all states s and action a,

$$Q_{h+1}(\boldsymbol{s},\boldsymbol{a}) \le \widehat{Q}^{opt}_{h+1}(\boldsymbol{s},\boldsymbol{a}).$$

Then, by the 1-Lipschitz property of the stage value in $\|\cdot\|_\infty$ (Lemma D.11),

$$V_{h+1}(\boldsymbol{s}') = \mathrm{Val}(Q_{h+1}(\boldsymbol{s}',\cdot)) \le \mathrm{Val}\left(\widehat{Q}^{opt}_{h+1}(\boldsymbol{s}',\cdot)\right) = \widehat{V}^{opt}_{h+1}(\boldsymbol{s}').$$

Then for the current step h, we define the **optimistic one-step target** to be

$$f^{opt}_{h,\boldsymbol{a}}(\boldsymbol{s}) \coloneqq \mathbb{E}\left[r_h(\boldsymbol{s},\boldsymbol{a}) + \gamma\widehat{V}^{opt}_{h+1}(\boldsymbol{s}')\,|\boldsymbol{s},\boldsymbol{a}\right],$$

also be $(T^{opt})_h(\boldsymbol{s},\boldsymbol{a})$. Let $w^{\dagger,opt}_{h,\boldsymbol{a}}$ be the best-in-class linear predictor of $f^{opt}_{h,\boldsymbol{a}}$, i.e.

$$w^{\dagger,opt}_{h,\boldsymbol{a}} \in \operatorname*{argmin}_w \mathbb{E}\left(\phi(\boldsymbol{s})^{\mathsf{T}} w - f^{opt}_{h,\boldsymbol{a}}(\boldsymbol{s})\right)^2.$$

With the same RFF analysis we used before (applied conditionally on the history, so the target is fixed), we have

$$\begin{aligned} \left|f^{opt}_{h,\boldsymbol{a}}(\boldsymbol{s}) - \phi(\boldsymbol{s})^{\mathsf{T}} w^{\dagger,opt}_{h,\boldsymbol{a}}\right| &\le \varepsilon_{\mathrm{rff},k,h}, \\ \left|\phi(\boldsymbol{s})^{\mathsf{T}}\left(\widehat{w}_{k,(h,\boldsymbol{a})} - w^{\dagger,opt}_{h,\boldsymbol{a}}\right)\right| &\le \mathrm{rad}_{k,(h,\boldsymbol{a})}(\boldsymbol{s}), \end{aligned}$$

hence,

$$\left(T^{opt}\right)_h(\boldsymbol{s},\boldsymbol{a}) \le \phi(\boldsymbol{s})^{\mathsf{T}}\widehat{w}_{k,(h,\boldsymbol{a})} + \varepsilon_{\mathrm{rff},k,h} + \mathrm{rad}_{k,(h,\boldsymbol{a})}(\boldsymbol{s}). \tag{C.1.1}$$

The true one-step Bellman target is

$$(TQ)_h(\boldsymbol{s},\boldsymbol{a}) = \mathbb{E}\left[r_h(\boldsymbol{s},\boldsymbol{a}) + \gamma V_{h+1}(\boldsymbol{s}')\,|\boldsymbol{a},\boldsymbol{a}\right].$$

By the inductive hypothesis $V_{h+1}(\boldsymbol{s}') \leq \widehat{V}_{h+1}^{opt}(\boldsymbol{s}')$ for all $\boldsymbol{s}'$, so

$$(TQ)_h(\boldsymbol{s},\boldsymbol{a}) \leq \left(T^{opt}\right)_h(\boldsymbol{s},\boldsymbol{a}.) \tag{C.1.2}$$

Combining (C.1.1) and (C.1.2), we have

$$\begin{aligned} Q_h(\boldsymbol{s},\boldsymbol{a}) &= (TQ)_h(\boldsymbol{s},\boldsymbol{a}) \\ &\leq \left(T^{opt}\right)_h(\boldsymbol{s},\boldsymbol{a}) \\ &\leq \phi(\boldsymbol{s})^{\mathsf{T}} \widehat{w}_{k,(h,\boldsymbol{a})} + \varepsilon_{\mathrm{rff},k,h} + \mathrm{rad}_{k,(h,\boldsymbol{a})}(\boldsymbol{s}). \end{aligned}$$

Adding another $\mathrm{rad}_{k,(h,\boldsymbol{a})}(\boldsymbol{s})$ to the RHS and recalling the definition of $\widehat{Q}_h^{opt}$,

$$Q_h(\boldsymbol{s},\boldsymbol{a}) \leq \phi(\boldsymbol{s})^{\mathsf{T}} \widehat{w}_{k,(h,\boldsymbol{a})} + \varepsilon_{\mathrm{rff},k,h} + 2\mathrm{rad}_{k,(h,\boldsymbol{a})}(\boldsymbol{s}) = \widehat{Q}_h^{opt}(\boldsymbol{s},\boldsymbol{a}).$$

Therefore, we have $Q_h \leq \widehat{Q}_h^{opt}$ for all $(\boldsymbol{s},\boldsymbol{a})$.

By monotonicity and 1-Lipschitzness of $\mathrm{Val}(\cdot)$,

$$V_h(\boldsymbol{s}) = \mathrm{Val}(Q_h(\boldsymbol{s},\cdot)) \leq \mathrm{Val}\left(\widehat{Q}_h^{opt}(\boldsymbol{s},\cdot)\right) = \widehat{V}_h^{opt}(\boldsymbol{s}),$$

and the truncation of H defined of $\widehat{V}_h^{opt}$ helps to keep the inequality valid since $V_h(\boldsymbol{s}) \leq H$.

Finally, for each play er i, replacing the opponent by a best response can only increase the player i value when the payoff matrix increases entrywise; therefore

$$Q_h^{i,br\left(\pi_{-i}^{opt}\right),\pi_{-i}^{opt}}(\boldsymbol{s},\boldsymbol{a}) \leq \widehat{Q}_h^{i,opt}(\boldsymbol{s},\boldsymbol{a}).$$

which concludes the induction and the proof. □

E.2 REGRET DECOMPOSITION

We first define the following Martingale-difference terms:

Definition E.2. *We define*

$$\begin{aligned} \delta_{k,h} &\coloneqq \mathbb{E}_{\boldsymbol{a}\sim\pi_{k,h}}\left[\widehat{Q}_{k,h}(\boldsymbol{s}_{k,h},\boldsymbol{a})\right] - \widehat{Q}_{k,h}(\boldsymbol{s}_{k,h},\boldsymbol{a}_{k,h}), \\ \xi_{k,h+1} &\coloneqq \mathbb{E}\left[\widehat{V}_{k,h+1}(\boldsymbol{s}_{k,h+1})\,|\boldsymbol{s}_{k,h},\boldsymbol{a}_{k,h}\right] - \widehat{V}_{k,h+1}(\boldsymbol{s}_{k,h+1}). \end{aligned}$$

Then we have the following lemma:

Lemma E.2. *On the high-probability event,*

$$\mathrm{Regret}(K) \leq \sum_{k,h}\xi_{k,h} + \sum_{k,h}\delta_{k,h} + 2\sum_{k,h}\mathrm{rad}_{k,(h,\boldsymbol{a}_{k,h})}(\boldsymbol{s}_{k,h}) + \sum_{k,h}\varepsilon_{\mathrm{rff},k,h}.$$

Proof. By Lemma E.1 or its one-step surrogate bound, upper bound the true best-response gap by the difference in optimistic values.

We then expand

$$\widehat{Q} = \phi^{\mathsf{T}}\widehat{w} + 2\mathrm{rad} + \varepsilon,$$

then sum it over k,h. The linear part can be telescoped via

$$\widehat{V}_h = \mathbb{E}_{\boldsymbol{a}\sim\pi_h}\left[\widehat{Q}_h\right] - \delta_{k,h}, \mathbb{E}\left[\widehat{V}_{h+1}(\boldsymbol{s}_{h+1})\right] = \widehat{V}_{h+1}(\boldsymbol{s}_{h+1}) + \xi_{k,h+1}.$$

□

E.3 BOUNDS ON EACH SUM

In this subsection, we discuss the (high-probability) bound of each sum given out in Subsection E.2

Lemma E.3. *The sequence $\{\xi_{k,h}\}$ and $\{\delta_{k,h}\}$ are bounded martingale differences with respect to the rolling filtration. Therefore, we have that for $\forall\delta$, with probability at least $1-\delta$,*

$$\sum_{k,h}\xi_{k,h}\le c_\xi H\sqrt{KH\log\left(\frac{1}{\delta}\right)},\ \sum_{k,h}\delta_{k,h}\le c_\delta H\sqrt{KH\log\left(\frac{1}{\delta}\right)}.$$

Proof. By definition, we have $\mathbb{E}\left[\xi_{k,h}|\mathcal{F}_{k,h}\right]=0$ and $|\xi_{k,h}|\le H$, since $\widehat{V}\in[0,H]$. Similarly we have $\mathbb{E}\left[\delta_{k,h}|\mathcal{F}_{k,h}\right]=0$ and $|\delta_{k,h}|\le 2H$ since $\widehat{Q}\in[0,H]$. We then apply the Azuma–Hoeffding to the sum over KH terms. □

Lemma E.4. *Fix $(h,\boldsymbol{a})$, then*

$$\sum_k\phi\left(\boldsymbol{s}_{k,h}\right)^{\mathsf{T}}\left(\Lambda_{k,(h,\boldsymbol{a})}\right)^{-1}\phi\left(\boldsymbol{s}_{k,h}\right)\le 2\log\frac{\det\left(\Lambda_{K+1,h,\boldsymbol{a}}\right)}{\det\left(\lambda I\right)}.$$

Proof. By the matrix determinant lemma,

$$\det\left(\Lambda+\phi\phi^{\mathsf{T}}\right)=\det\left(\Lambda\right)\left(1+\phi^{\mathsf{T}}\Lambda^{-1}\phi\right).$$

terate and take logarithms to obtain

$$\log\frac{\left(\Lambda_{K+1}\right)}{\det\left(\lambda I\right)}=\sum_k\log\left(1+\phi^{\mathsf{T}}\Lambda^{-1}\phi\right).$$

Using $\log(1+x)\ge\frac{x}{2}$ for $x\in[0,1]$ and that $\phi_k^{\mathsf{T}}\left(\Lambda_k\right)^{-1}\phi_k\le 1$ (since $\Lambda_k\succeq\lambda I$), we conclude the claim. □

Corollary E.1. *Let $\widehat{C}_{h,\boldsymbol{a}}\coloneqq\frac{1}{n_{h,\boldsymbol{a}}}\sum\phi\phi^{\mathsf{T}}$, then*

$$\begin{aligned}\log\frac{\det\left(\Lambda_{K+1,h,\boldsymbol{a}}\right)}{\det\left(\lambda I\right)}&=\log\det\left(I+\lambda^{-1}\sum\phi\phi^{\mathsf{T}}\right)\\&\le Tr\left(\left(\widehat{C}_{h,\boldsymbol{a}}+\lambda I\right)^{-1}\widehat{C}_{h,\boldsymbol{a}}\right)\log\left(1+n_{h,\boldsymbol{a}}/\lambda\right).\end{aligned}$$

Then define

$$\begin{aligned}d_{\mathrm{eff},(h,\boldsymbol{a})}\left(\lambda\right)&\coloneqq Tr\left(\left(\widehat{C}_{h,\boldsymbol{a}}+\lambda I\right)^{-1}\widehat{C}_{h,\boldsymbol{a}}\right)\le D,\\\widetilde{d}_{\mathrm{eff}}\left(\lambda\right)&\coloneqq\sum_{\boldsymbol{a}}d_{\mathrm{eff},(h,\boldsymbol{a})}\left(\lambda\right)\le|\mathcal{A}|\,D.\end{aligned}$$

Then by Cauchy-Schwarz, we have

$$\sum_{k,h}\sqrt{\phi^{\mathsf{T}}\left(\Lambda_{k,(h,\boldsymbol{a}_{k,h})}\right)^{-1}\phi}\le c_{EP}H\sqrt{K\widetilde{d}_{\mathrm{eff}}\left(\lambda\right)\log\left(1+K/\lambda\right)}.$$

Lemma E.5. *· For the generic case, we take $\lambda\asymp(KH)^{-1/2}$, $D\asymp\sqrt{KH}\log(KH)$ and set $\varepsilon_{\mathrm{rff},k,h}\equiv c_\varepsilon\sqrt{\log\left(KH/\delta\right)/K}$. Then we have*

$$\sum_{k,h}\varepsilon_{\mathrm{rff},k,h}\le c'_\varepsilon H\sqrt{K\log\left(KH/\delta\right)}.$$

· For the benign case, if $N\left(\lambda\right)\lesssim\lambda^{-\gamma}$ and $r\ge\frac{1}{2}$, take $\lambda\asymp(KH)^{-1/(2r+\gamma)}$, $D\asymp N\left(\lambda\right)\log N\left(\lambda\right)$ and $\varepsilon_{\mathrm{rff},k,h}=c_\varepsilon\sqrt{N\left(\lambda/K\right)}+\lambda^r$. Then $\sum_{k,h}\varepsilon_{\mathrm{rff},k,h}$ is dominated by the elliptical term.

Proof. Plug Assumption D.1 and then sum over KH terms; choose λ, D to balance variance/bias term and keep ε on the same or lower order than the elliptical potential term. □

E.4 MAIN RESULTS

In this subsection, we discuss the regret bound for our algorithm, both on TD+ϵ-greedy and optimistic variant versions.

Theorem E.1 (TD + ϵ-greedy)**.** *For $\forall \delta > 0$, with probability at least $1-\delta$,*

$$\text{Regret}\,(K) \leq C\bar{\beta}H\sqrt{KH\widetilde{d}_{\text{eff}}\,(\lambda)\log\left(1+\frac{K}{\lambda}\right)} + C'KH\epsilon_{\text{approx}} + C''KH\epsilon.$$

where $\bar{\beta}$ represents the typical scale of $\beta_{k,(h,\boldsymbol{a})}$ across $(h,\boldsymbol{a})$ and $\epsilon_{\text{approx}} = c_1/\sqrt{D} + c_2\left(\sqrt{N\,(\lambda)\,/K} + \lambda^r\right)$.

Proof. Combine Lemma E.2 with Lemma E.3, Corollary E.1, and Lemma E.5. The ϵ-greedy term $KH\epsilon$ accounts for deviations from the exact stage NE sampling. □

Theorem E.2 (Optimistic variant)**.** *If actions are selected by solving the stage NE on $\widehat{Q}^{opt}$ and TD targets using $\widehat{V}^{opt}$, then with probability at least $1-\delta$,*

$$\text{Regret}\,(K) \leq \widetilde{O}\left(\beta H\sqrt{KH} + \beta H\sqrt{K\widetilde{d}_{\text{eff}}\,(\lambda)}\right).$$

matching the linear-approximation rate after replacing the linear dimension by $\widetilde{d}_{\text{eff}}\,(\lambda) \leq |\mathcal{A}|\,D$.

Proof. By Lemma E.1, the estimation and approximation errors are absorbed into the optimistic bonuses, so the independent $\sum\varepsilon$ terms and the $KH\epsilon$ terms vanish. The rest follows as in Theorem E.1. □

Explanation: Probability Union and Constants

We split the failure probability δ equally among: (i) self-normalized concentration, (ii) RFF approx- imation / KRR generalization, (iii) martingale deviations. A union bound yields a global event of probability at least $1-\delta$. All absolute constants and log factors are absorbed into $\widetilde{Q}\,(\cdot)$ as standard.