Learning Two-Player Mixture Markov Games: Kernel Function Approximation and Correlated Equilibrium

Chris Junchi Li*
University of California, Berkeley
junchili@berkeley.edu

Dongruo Zhou*
University of California, Los Angeles
drzhou@cs.ucla.edu

Quanquan Gu
University of California, Los Angeles
qgu@cs.ucla.edu

Michael I. Jordan
University of California, Berkeley
jordan@cs.berkeley.edu

Abstract

We consider learning Nash equilibria in two-player zero-sum Markov Games with nonlinear function approximation, where the action-value function is approximated by a function in a Reproducing Kernel Hilbert Space (RKHS). The key challenge is how to do exploration in the high-dimensional function space. We propose a novel online learning algorithm to find a Nash equilibrium by minimizing the duality gap. At the core of our algorithms are upper and lower confidence bounds that are derived based on the principle of optimism in the face of uncertainty. We prove that our algorithm is able to attain an $O(\sqrt{T})$ regret with polynomial computational complexity, under very mild assumptions on the reward function and the underlying dynamic of the Markov Games. We also propose several extensions of our algorithm, including an algorithm with a Bernstein-type bonus that can achieve a tighter regret bound, and another algorithm for model misspecification that can be applied to neural network function approximation.

1 Introduction

Multi-agent reinforcement learning (MARL) has been the focus of research across a range of research communities [Shapley, 1953, Littman, 1994]. The case of two-player Markov Games (MG) has been of particular interest. In this case, two players select their actions based on the current state simultaneously and independently. One player (the max-player) aims to maximize the return based on the reward provided by the environment, while the other (the min-player) aims to minimize it. A series of recent results have established polynomial sample complexity/regret guarantees that depend on the cardinality of state/action spaces for two-player zero-sum MGs [Wei et al., 2017, Bai and Jin, 2020, Bai et al., 2020, Liu et al., 2021, Jia et al., 2019, Sidford et al., 2020, Cui and Yang, 2021, Lagoudakis and Parr, 2002, Perolat et al., 2015, Pérolat et al., 2016a,b, 2017, Jin et al., 2021b].

Meanwhile, most of the recent successful applications of MARL deal with *large state/action spaces* that may be continuous or a fine-grained discretization of a continuous space. Examples include GO [Silver et al., 2016], autonomous driving [Shalev-Shwartz et al., 2016], TexasHold'em poker [Brown and Sandholm, 2019], and AlphaStar for the game Starcraft [Vinyals et al., 2019]. In order to tackle problems with large state/action spaces, researchers have designed MARL algorithms based on *function approximation* which approximate the original high-dimensional value

*Equal contribution.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

function/policy by a function approximator. For instance, Xie et al. [2020] and Chen et al. [2022] studied RL for two-player zero-sum MGs with *linear function approximation*, where it is assumed that there are a set of *linear features* that span the transition kernel and reward function spaces. In contrast to RL with linear function approximation, RL with *nonlinear function approximation* (e.g., kernel and neural network approximation) aims to take advantage of the superior representational power of nonlinear function compared to linear parameterizations. For example, Jin et al. [2022] studied neural-network-based RL in the setting of *MGs with low multi-agent Bellman eluder dimension*, obtaining algorithms that have polynomial dependence on the complexity of the underlying function class. Although this yields a strong theoretical guarantee, the algorithm that they propose is not computationally efficient due to the constructed highly nonconvex confidence sets. The following question is still open: *Can we design a computationally and statistically efficient RL algorithm for learning two-player Markov Games with nonlinear function approximation?*

In this paper, we give an affirmative answer to this question for a class of episodic Markov Games, dubbed *mixture Markov Games*, when using a nonlinear approximation function in a Reproducing Kernel Hilbert Space (RKHS). We propose a novel kernel-based MARL algorithmic framework for general episodic two-player zero-sum MGs which provides provable regret guarantees. We summarize the contributions of our work as follows:

- We propose a `KernelCCE-VTR` algorithm for two-player zero-sum MGs. In particular, at each episode, `KernelCCE-VTR` uses kernel function approximation to approximate the optimal value function and constructs corresponding confidence sets, following the *"Optimism-in-Face-of-Uncertainty"* principle [Abbasi-Yadkori et al., 2011] to select an action based on the current state. In contrast to algorithms in Jin et al. [2022], which construct implicit confidence sets that are in general computationally intractable, our algorithm `KernelCCE-VTR` crafts a computationally efficient exploration bonus based on the gram matrix of the kernel function.
- Under the assumption that the transition dynamics belongs to some RKHS, we show that our algorithm `KernelCCE-VTR` is able to find a Nash equilibrium of the game with a $\widetilde{O}(d_{\mathcal{F}}H^2\sqrt{T})$ regret bound on the duality gap, where H is the horizon, T is the number of the episodes, and $d_{\mathcal{F}}$ represents the complexity of the function class $\mathcal{F}$. We also propose an extension of `KernelCCE-VTR` that utilizes *weighted kernel ridge regression* and a *Bernstein-type bonus* to achieve $\widetilde{O}(d_{\mathcal{F}}H^{3/2}\sqrt{T})$ regret. When $\mathcal{F}$ reduces to the d-dimensional linear function class, our regret reduces to $\widetilde{O}(dH^{3/2}\sqrt{T})$, which almost matches the lower bound in Chen et al. [2022].
- We also study the general case where the transition dynamics belongs to some RKHS up to a misspecification error. We show that our `KernelCCE-VTR` can achieve a similar regret as in the well-specified case. In particular, we study the neural network function approximation case which can be regarded as a special instance of the misspecified RKHS case and derive the corresponding regret bound.

Notation. We use lowercase letters to denote scalars, and lower and uppercase bold letters to denote vectors and matrices. We use $\|\cdot\|$ to indicate Euclidean norm, and for a semi-positive definite matrix $\mathbf{\Sigma}$ and any vector $\mathbf{x}$, $\|\mathbf{x}\|_{\mathbf{\Sigma}} := \|\mathbf{\Sigma}^{1/2}\mathbf{x}\| = \sqrt{\mathbf{x}^\top\mathbf{\Sigma}\mathbf{x}}$. For real t and interval $[a, b]$, we use $\Pi_{[a,b]}[t]$ to indicate the projection of t onto $[a, b]$, i.e. $\Pi_{[a,b]}[t] = \max(a, \min(b, t))$. For positive integer N we sometimes define $[N] = \{1, \dots, N\}$ for compactness. We also adopt the standard big-O and big-Ω notations: say $a_n = O(b_n)$ if and only if there exists $C > 0, N > 0$, for any $n > N$, $a_n \le Cb_n$; $a_n = \Omega(b_n)$ if $a_n \ge Cb_n$. The notations $\widetilde{O}$ and $\widetilde{\Omega}$ are adopted when the C above hides a polylogarithmic factor.

2 Related Work

Online RL with function approximation. MARL with function approximation can be seen as an extension of RL with function approximation on MDPs. There are several lines of work studying RL with function approximation. The first line of work studies the so-called linear MDP which assumes the reward function and transition dynamics are linear functions of a feature mapping defined on the state and action spaces [Yang and Wang, 2020, Jin et al., 2020, Zanette et al., 2020]. These works propose model-free algorithms with sublinear regret on the number of episodes K. The second line of work studies the linear mixture MDP which assumes the transition kernel is a linear combination of several base models [Modi et al., 2020, Jia et al., 2020, Zhou et al., 2021b,a]. These studies proposed model-based RL algorithms that estimate the transition kernel with finite sample complexity or sublinear regret guarantees. The third line of work studies general function approximation which assumes that either the value function or the transition kernel can be approximated by a general class of functions [Osband and Van Roy, 2014, Jiang et al., 2017, Sun et al., 2019, Wang et al., 2020, Yang et al., 2020, Du et al., 2021, Jin et al., 2021a]. Algorithms proposed in this line enjoy finite regret or sample complexity bounds that depend on some general complexity measures such as Eluder dimension [Russo and Van Roy, 2013, Osband and Van Roy, 2014], Bellman rank [Jiang et al., 2017], witness rank [Sun et al., 2019], information gain [Yang et al., 2020], bilinear class [Du et al., 2021] and Bellman eluder dimension [Jin et al., 2021a].

Learning two-player MGs with function approximation. There is a large body of literature on MARL for two-player MGs with function approximation. These works can be generally categorized into MARL with *linear function approximation* and MARL with *general function approximation*. For example, for linear function approximation, Xie et al. [2020] studied zero-sum simultaneous-move MGs where both the reward and transition kernel can be parameterized as linear functions of some feature mappings. They proposed an OMVI-NI algorithm with an $\widetilde{O}(\sqrt{d^3H^3T})$ regret, where d is the number of the feature dimension, H is the episode length and T is the total number of rounds. Chen et al. [2022] studied the linear mixture MGs and proposed a nearly minimax optimal Nash-UCRL-VTR algorithm with an $\widetilde{O}(dH\sqrt{T})$ regret and an $\Omega(dH\sqrt{T})$ matching lower bound. In contrast to this work, our `KernelCCE-VTR` does not assume the underlying transition dynamic or reward function has a linear structure. For MARL with general function approximation, Jin et al. [2022] studied the two-player zero-sum MGs with low multi-agent Bellman Eluder dimension and proposed a "Golf with Exploiter" algorithm using a general function class. They showed their algorithm enjoys an $\widetilde{O}(H\sqrt{dK\log N})$ regret, where d is the multi-agent Bellman eluder dimension, K is the number of episodes. Huang et al. [2022] studied two-player MGs with a finite minimax Eluder dimension and proposed a method called ONEMG with an $\widetilde{O}(H\sqrt{dK\log N})$ regret, where d is the minimax Eluder dimension. To obtain the desired function approximator, both Golf with Exploiter and ONEMG need to solve a constrained optimization problem, which is computationally intractable even in the linear function approximation setting. In contrast to Jin et al. [2022] and Huang et al. [2022], our proposed algorithms are computationally efficient and nearly optimal by using the Bernstein-type bonus. Qiu et al. [2021] also studied kernel function approximation for two-player MGs. However, there are two key differences between our work and theirs. First, Qiu et al. [2021] studied MGs where the expectation of the value function is in some RKHS; we, on the other hand, assume that the transition dynamics of the MG lie in an RKHS. Second, while the regret result in Qiu et al. [2021] depends on the covering number of the function space, our regret is *independent* of the covering number.

3 Preliminaries

In this section, we present the necessary definitions that will be adopted throughout the paper. Section 3.1 describes simultaneous-move games in the setting of zero-sum two-player Markov Games (MG) and recaps the concepts of equilibrium and duality gap that are employed in the game theory literature. Section 3.2 provides necessary definitions and notation for approximating action value function with functions belonging to a reproducing kernel Hilbert space (RKHS) via modeling the transition probability.

3.1 Two-player Markov Games

A simpler instance of Markov Games, referred to as turn-based games, can be seen as a special case of simultaneous-move games.[2] In a zero-sum two-player simultaneous-move Markov Game, the dynamical structure can be captured by an MG, denoted $(\mathcal{S}, \mathcal{A}_1, \mathcal{A}_2, r, \mathbb{P}, H)$, where $\mathcal{S}$ is the space of available states of the environment, $\mathcal{A}_1$ is the action space of the first player and $\mathcal{A}_2$ is the action space of the second player. H is the time horizon representing the maximum step of each round of play. The reward function $r : \{r_h(x, a, b) : h \in [H]\}$ is a sequence of mappings from $\mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2$ to $[-1, 1]$. The transition matrix $\mathbb{P} : \{\mathbb{P}_h(\cdot|x, a, b) : h \in [H]\}$ gives for each state actions triplet (x, a, b) and at each time h the stochastic response of the environment to the next $x' \in \mathcal{S}$. Here by "simultaneous move" we refer to the setting where at each round of game the two players P_1 and P_2 take actions $a \in \mathcal{A}_1, b \in \mathcal{A}_2$ simultaneously at a given state $x \in \mathcal{S}$, in contrast with the turn-based

[2]We present a discussion of the implications of our results for turn-based games in the supplementary materials.

game where r_h and $\mathbb{P}_h$ are defined for a state-action pair (x, a) where the action can be taken by either players. In the context of this paper, for simplicity of notation we let $\mathcal{A}_1 = \mathcal{A}_2 = \mathcal{A}$, while the results can be easily generalized to the case when $\mathcal{A}_1 \neq \mathcal{A}_2$. Similar definitions of a zero-sum two-player simultaneous-move episodic Markov Games can be found in Wei et al. [2017], Perolat et al. [2018], Xie et al. [2020].

In the above setting, two players P_1 and P_2 take actions according to their own strategies. We use $\pi := \{\pi_h\}_{h\in[H]}$ to denote the stochastic policy of P_1 and use $\nu := \{\nu_h\}_{h\in[H]}$ to denote the stochastic policy of P_2. We note that at time h, $\pi_h : \mathcal{S} \mapsto \Delta_{\mathcal{A}}$ maps the current state x_h to a probability distribution of the actions, and similarly for ν_h. Given two agents' policies, π, ν, across h steps, the state value function is defined as the expected total reward through H steps where at step $h \in [H]$ player P_1 follows policy $\pi_h(\cdot|x_h)$ and player P_2 follows policy $\nu_h(\cdot|x_h)$,

$$V_h^{\pi,\nu}(x) := \mathbb{E}_{\pi,\nu}\left[\sum_{t=h}^{H} r_t(s_t, a_t, b_t) \,\middle|\, x_h = x\right],\ V_{H+1}^{\pi,\nu}(x) := 0,$$

and where $V^{\pi,\nu}(x) := V_1^{\pi,\nu}(x)$. Note that the expectation is taken over all stochasticity in π_h, ν_h and $\mathbb{P}_h$. The action value function is defined as

$$Q_h^{\pi,\nu}(x, a, b) := \mathbb{E}_{\pi,\nu}\left[\sum_{t=h}^{H} r_t(x_t, a_t, b_t) \,\middle|\, x_h = x, a_h = a, b_h = b\right],\ Q_{H+1}^{\pi,\nu}(x, a, b) := 0,$$

and $Q^{\pi,\nu}(x, a, b) := Q_1^{\pi,\nu}$. From the definition of two value functions, we observe that for any $x \in \mathcal{S}$, the state value function given policy pair (π, ν) is the expectation of the corresponding action value function

$$V_h^{\pi,\nu}(x) := \mathbb{E}_{(a,b)\sim(\pi,\nu)} Q_h^{\pi,\nu}(x, a, b),$$

where the expectation is taken over the action distribution induced by the policy pair. Throughout this paper, we use superscripts do denote the number of episodes and subscripts to denote the number of horizon steps.

Nash equilibrium and duality gap. In a zero-sum two-player Markov Game, P_1 wants to maximize the expected reward $V^{\pi,\nu}(x)$ via the choice of the policy π. On the contrary, P_2 wants to minimize $V^{\pi,\nu}(x)$ by properly choosing ν. For fixed ν, we define the best response policy with respect to V and ν as $\mathrm{br}(\nu)$ and define $V_h^{*,\nu} := V_h^{\mathrm{br}(\nu),\nu}$ and $Q_h^{*,\nu} := Q_h^{\mathrm{br}(\nu),\nu}$, We define $V_h^{\pi,*} := V_h^{\pi,\mathrm{br}(\pi)}$ and $Q_h^{\pi,*} := Q_h^{\pi,\mathrm{br}(\pi)}$ similarly. A Nash equilibrium is a pair of policies (π^*, ν^*) that are the best response policy for each other, which we write as $V^{\pi^*,*}(x) = V^{\pi^*,\nu^*}(x) = V^{*,\nu^*}(x)$. For notational simplicity we write $V^* := V^{\pi^*,\nu^*}, Q^* := Q^{\pi^*,\nu^*}$. By definition of the best response policy, we obtain weak duality:

$$V_h^{\pi,*}(x) \leq V_h^*(x) \leq V_h^{*,\nu}(x).$$

For any policy pair (π, ν), we define the duality gap as $V_1^{*,\nu^t}(x_1^t) - V_1^{\pi^t,*}(x_1^t)$. We call a pair an *$\epsilon$-approximate Nash equilibrium (NE)* if $V_1^{*,\nu^t}(x_1^t) - V_1^{\pi^t,*}(x_1^t) \leq \epsilon$. We also define the *regret* in the MG setting as follows:

$$\mathrm{Regret}(T) := \sum_{t=1}^{T} V_1^{*,\nu^t}(x_1^t) - V_1^{\pi^t,*}(x_1^t).$$

Coarse correlated equilibrium. We introduce the *Coarse Correlated Equilibrium (CCE)* notion which will be used in our proposed algorithms. Given payoff matrices $Q_1, Q_2 : \mathcal{S} \times \mathcal{A} \times \mathcal{A} \mapsto \mathbb{R}$ and the state x, we define the CCE of the game as a joint distribution σ on $\mathcal{A} \times \mathcal{A}$ satisfying:

$$\mathbb{E}_{(a,b)\sim\sigma}\left[Q_1(x, a, b)\right] \geq \mathbb{E}_{b\sim\mathcal{P}_2\sigma}\left[Q_1(x, a', b)\right], \quad \forall a' \in \mathcal{A}, \tag{3.1}$$

$$\mathbb{E}_{(a,b)\sim\sigma}\left[Q_2(x, a, b)\right] \leq \mathbb{E}_{a\sim\mathcal{P}_1\sigma}\left[Q_2(x, a, b')\right], \quad \forall b' \in \mathcal{A}, \tag{3.2}$$

where $\mathcal{P}_1\sigma$ denotes the marginal of σ on the first coordinate (min-player) and $\mathcal{P}_2\sigma$ denotes the marginal of σ on the second coordinate (max-player). We use $\texttt{FIND\_CCE}(Q_1, Q_2, x)$ to denote σ. When σ can be written as a product of two policies over action space $\mathcal{A}$, it is a Nash equilibrium [Xie et al., 2020]. To compute a CCE given Q_1, Q_2, x, please see Appendix I.

3.2 Nonlinear Function Approximation by Reproducing Kernel Hilbert Spaces

For simplicity of notation, we use $z = (x, a, b)$ to denote a state-action-action triplet (or state-action tuple) in $\mathcal{Z} := \mathcal{S} \times \mathcal{A} \times \mathcal{A}$. An RKHS $\mathcal{H}$ with kernel $\mathbf{K}(\cdot, \cdot) : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$ is a general form of linear function class. Every RKHS $\mathcal{H}$ consists of functions on $\mathcal{Z}$, with a feature mapping $\boldsymbol{\phi} : \mathcal{Z} \mapsto \mathcal{H}$, such that $\forall \boldsymbol{f} \in \mathcal{H}$ and $\forall z \in \mathcal{Z}$, $\boldsymbol{f}(z) = \langle \boldsymbol{f}, \boldsymbol{\phi}(z) \rangle_{\mathcal{H}}$. The kernel K is thus defined for every $x, y \in \mathcal{Z} \times \mathcal{Z}$ as $\mathbf{K}(x, y) = \langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(y) \rangle_{\mathcal{H}}$. We call $\boldsymbol{\phi}$ the *feature mapping* induced by the RKHS $\mathcal{H}$ with kernel K. In the following sections, we sometimes use $f^\top g$ as a simplification of $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{H}}$ when $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{H}$. We make no distinction in notation between the vector inner product and the product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$; the distinction can be read out from the nature of the two objects in the product. For every RKHS $\mathcal{H}$, there exists a natural eigenvalue decomposition in $\mathcal{L}^2(\mathcal{Z})$. RKHS approximation is a generalization of the linear function approximation of finite dimension d which can be infinite-dimensional. In the following, we define the so-called *kernel mixture MG*, which can be regarded as an extension from the linear mixture MDP [Jia et al., 2020, Ayoub et al., 2020, Zhou et al., 2021a] and linear mixture MG [Chen et al., 2022] to their kernel counterpart.

Kernel mixture MG. In a kernel mixture MG model, we model the transition probability $\mathbb{P}_h(s'|z) : \mathcal{Z} \mapsto \Delta(\mathcal{S})$ as an element in an RKHS $\mathcal{H}$ with feature mapping $\boldsymbol{\phi}(s'|z) : \mathcal{Z} \times \mathcal{S} \to \mathcal{H}$, such that for an unknown true parameter $\boldsymbol{\theta}_h^* \in \mathcal{H}$, $\mathbb{P}_h(s'|z) = \langle \boldsymbol{\phi}(s'|z), \boldsymbol{\theta}_h^* \rangle_{\mathcal{H}}$ for all $s' \in \mathcal{S}$ and $z \in \mathcal{Z}$. A similar MG structure called kernel MG has been studied by Qiu et al. [2021], which assumes that the transition probability satisfies $\mathbb{P}_h(s'|z) = \langle \boldsymbol{\phi}(z), \boldsymbol{\mu}_h(s') \rangle$ for some $\boldsymbol{\phi}(\cdot), \boldsymbol{\mu}_h(\cdot) \in \mathcal{H}$. The single-agent MDP counterparts of kernel MGs and kernel mixture MGs are linear MDPs and linear mixture MDPs. Zhou et al. [2021b] have shown that linear MDPs and linear mixture MDPs are different classes of MDPs and one cannot be covered by each other. Following a similar argument, we can also show that kernel mixture MGs and kernel MGs are different classes of MGs and cannot imply each other.

At time h, for any estimate of the value function $V_h(\cdot) : \mathcal{S} \mapsto \mathbb{R}$, we note that the expectation of value function at time $h+1$, $\mathbb{P}_h V_{h+1}$ is an element in the RKHS $\mathbb{P}_h V_{h+1}(z) = \langle \boldsymbol{\phi}_{V_{h+1}}(z), \boldsymbol{\theta}_h^* \rangle_{\mathcal{H}}$, where $\boldsymbol{\phi}_{V_{h+1}}(z) := \sum_{s' \in \mathcal{S}} V_{h+1}(s') \boldsymbol{\phi}(s'|z)$ integrates the product of the feature mapping with the estimated value of s' over $\mathcal{S}$. It is worth noting that the quantity $\boldsymbol{\phi}_V(\cdot)$ plays an important role in previous linear mixture model-based algorithms [Jia et al., 2020, Ayoub et al., 2020, Zhou et al., 2021a, Chen et al., 2022]. We assume that for any bounded value function $V(\cdot) : \mathcal{S} \mapsto [-1, 1]$ and any $z \in \mathcal{Z}$, $\|\boldsymbol{\phi}_V(z)\|_{\mathcal{H}} \le 1$. Given that the reward function $r_h(z)$ is known, we obtain through the Bellman equation that

$$Q_h^{*,\nu}(\cdot) = r_h(\cdot) + (\mathbb{P}_h V_{h+1}^{*,\nu})(\cdot) = r_h(\cdot) + \left\langle \boldsymbol{\phi}_{V_{h+1}^{*,\nu}}(\cdot), \boldsymbol{\theta}_h^* \right\rangle_{\mathcal{H}}, \tag{3.3}$$

$$Q_h^{\pi,*}(\cdot) = r_h(\cdot) + (\mathbb{P}_h V_{h+1}^{\pi,*})(\cdot) = r_h(\cdot) + \left\langle \boldsymbol{\phi}_{V_{h+1}^{\pi,*}}(\cdot), \boldsymbol{\theta}_h^* \right\rangle_{\mathcal{H}}. \tag{3.4}$$

Weighted kernel function. In this work, we consider a general RKHS $\mathcal{H}$ and do not assume that we can access the feature mapping $\boldsymbol{\phi}$ directly. Instead, we assume that we can access the *weighted kernel function* $\boldsymbol{k}_{V_1,V_2}(\cdot, \cdot)$, which is defined as follows:

Definition 1. For any function pairs $V_1, V_2 : \mathcal{S} \to [0, 1]$ which map states to real numbers, the weighted kernel function $\boldsymbol{k}_{V_1,V_2}(\cdot, \cdot)$ is defined as follows: $\forall z_1, z_2 \in \mathcal{Z}$,

$$\boldsymbol{k}_{V_1,V_2}(z_1, z_2) := \sum_{s_1, s_2 \in \mathcal{S}} V_1(s_1) V_2(s_2) \langle \boldsymbol{\phi}(s_1|z_1), \boldsymbol{\phi}(s_2|z_2) \rangle_{\mathcal{H}}.$$

It is easy to see from Definition 1 that

$$\boldsymbol{k}_{V_1,V_2}(z_1, z_2) = \left\langle \sum_{s_1 \in \mathcal{S}} V_1(s_1) \boldsymbol{\phi}(s_1|z_1), \sum_{s_2 \in \mathcal{S}} V_2(s_2) \boldsymbol{\phi}(s_2|z_2) \right\rangle_{\mathcal{H}} = \langle \boldsymbol{\phi}_{V_1}(z_1), \boldsymbol{\phi}_{V_2}(z_2) \rangle_{\mathcal{H}},$$

which suggests that the weighted kernel function $\boldsymbol{k}_{V_1,V_2}(\cdot, \cdot)$ indeed captures the interaction (in inner product relation) between $\boldsymbol{\phi}_{V_1}(z_1)$ and $\boldsymbol{\phi}_{V_2}(z_2)$. We assume that we can access an integration oracle that computes $\boldsymbol{k}_{V_1,V_2}(z_1, z_2)$ for any function V_1, V_2 and state-action tuples z_1, z_2 efficiently.

4 Algorithm

In this section, we introduce our value-targeted iteration algorithm for the zero-sum two-player Markov Game setting with RKHS function approximation. We follow the *value-targeted regression* framework and the confidence set design as in UCRL [Jia et al., 2020, Ayoub et al., 2020], and combine the CCE technique [Xie et al., 2020] to deal with the zero-sum sub-game brought by upper confidence bound (UCB) and lower confidence bound (LCB) value functions. These techniques enable us to adapt the results from the linear setting to the nonlinear RKHS regime [Chowdhury and Gopalan, 2017, Yang et al., 2020, Zhou et al., 2020] to get a structure-dependent regret bound that is both computationally simple and statistically efficient.

Algorithm 1 `KernelCCE-VTR`

Input: bonus parameter $\beta > 0$.
for episode $t = 1, 2, \ldots, T$ **do**
    for step $h = H, H-1, \ldots, 1$ **do**
        Calculate $\overline{Q}_h^t(\cdot,\cdot,\cdot), \underline{Q}_h^t(\cdot,\cdot,\cdot)$ as in (4.3)
        Let $\sigma_h^t(\cdot) = \texttt{FIND\_CCE}(\overline{Q}_h^t, \underline{Q}_h^t, \cdot)$
        Let $\overline{V}_h^t(\cdot) = \mathbb{E}_{(a,b)\sim\sigma_h^t(\cdot)} \overline{Q}_h^t(\cdot, a, b)$ and $\underline{V}_h^t(\cdot) = \mathbb{E}_{(a,b)\sim\sigma_h^t(\cdot)} \underline{Q}_h^t(\cdot, a, b)$
    end for
    Receive initial state x_1^t
    for step $h = 1, 2, \ldots, H$ **do**
        Sample $(a_h^t, b_h^t) \sim \sigma_h^t(x_h^t)$
        P_1 takes action a_h^t, P_2 takes action b_h^t
        Observe next state x_{h+1}^t
    end for
end for

To find an equilibrium (π^*, ν^*) of the value function $V_1^{\pi,\nu}(x_1)$, we design an algorithm using value-targeted regression (VTR) and upper/lower confidence bound-based exploration. As the min-player aims to minimize the value function while the max-player targets to maximize the value function, we use the upper confidence bound to encourage the exploration of the max-player and use the lower confidence bound to encourage the exploration of the min-player. Thus we need to define two value functions for the min/max-players respectively, i.e., $\overline{Q}_h^t, \underline{Q}_h^t, \overline{V}_h^t, \underline{V}_h^t$, where we adopt the overline notation for the over-estimation by the max-player and the underline notation for the under-estimation by the min-player. In the following, we only describe how to estimate the value functions for the max-player, while the value functions for the min-player can be estimated analogously. At each round of the game, we solve the following ridge regression problem for minimizing the Bellman error:

$$\overline{\boldsymbol{\theta}}_h^t = \min_{\boldsymbol{\theta}\in\mathcal{H}} \sum_{\tau=1}^{t-1} \left[\overline{V}_{h+1}^\tau(x_{h+1}^\tau) - \left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}(z_h^\tau), \boldsymbol{\theta} \right\rangle_{\mathcal{H}}\right]^2 + \lambda \|\boldsymbol{\theta}\|_{\mathcal{H}}^2 . \tag{4.1}$$

Note that in (4.1), $\overline{V}_{h+1}^\tau$ only depends on the previous trajectories $\left(x_i^j, a_i^j, b_i^j : j \in [\tau-1], i \in [H]\right)$. We denote the corresponding σ-algebra as $\mathcal{F}_{\tau-1}$, and thus we have $\overline{V}_{h+1}^\tau \in \mathcal{F}_{\tau-1}$. As each $\overline{V}_{h+1}^\tau(x_{h+1}^\tau)$ can be seen as a stochastic sample of $(\mathbb{P}_h \overline{V}_{h+1}^\tau)(z_h^\tau)$, the regularized regression problem of the max-player in (4.1) can be seen as solving a linear bandit problem with context $\boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}(z_h^\tau)$, reward function $(\mathbb{P}_h \overline{V}_{h+1}^\tau)(z_h^\tau)$ and noise term $\overline{V}_{h+1}^\tau(x_{h+1}^\tau) - (\mathbb{P}_h \overline{V}_{h+1}^\tau)(z_h^\tau)$. From the solution to the ridge regression problem (4.1), we can define the upper/lower confidence bound of the action-value functions $Q_h^{*,\nu}, Q_h^{\pi,*}$ respectively. For the simplicity of notation, we define the vectors $\overline{\boldsymbol{\Psi}}_h^t := \left(\boldsymbol{\phi}_{\overline{V}_{h+1}^1}(z_h^1), \ldots \boldsymbol{\phi}_{\overline{V}_{h+1}^{t-1}}(z_h^{t-1})\right)^\top \in \mathcal{H}^{t-1}$.

For a positive parameter $\beta_t > 0$ that will be chosen in later analysis, the confidence region centered at $\overline{\boldsymbol{\theta}}_h^t$ in the RKHS $\mathcal{H}$ is defined as

$$\overline{\mathcal{C}}_h^t = \left\{\boldsymbol{\theta} : \sqrt{\lambda \left\|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_h^t\right\|_{\mathcal{H}}^2 + \left\|\left\langle \overline{\boldsymbol{\Psi}}_h^t, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_h^t \right\rangle_{\mathcal{H}}\right\|^2} \le \beta_t \right\}, \tag{4.2}$$

We omit the definition of $\underline{\mathcal{C}}_h^t$ which is an analogue of Eq. (4.2) by changing all overline symbols to underline ones. Based on the confidence regions, we construct an optimistic/pessimistic estimate of $Q_h^{*,\nu}$ as

$$\overline{Q}_h^t := \Pi_{[-H,H]}\left[r_h + \max_{\boldsymbol{\theta}\in\overline{\mathcal{C}}_h^t} \left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^t}, \boldsymbol{\theta}\right\rangle_{\mathcal{H}}\right], \quad \underline{Q}_h^t := \Pi_{[-H,H]}\left[r_h + \min_{\boldsymbol{\theta}\in\underline{\mathcal{C}}_h^t} \left\langle \boldsymbol{\phi}_{\underline{V}_{h+1}^t}, \boldsymbol{\theta}\right\rangle_{\mathcal{H}}\right], \tag{4.3}$$

where $\Pi_{[-H,H]}$ is the projection operator onto $[-H, H]$, which is by definition the range of value functions. For the convenience of conducting an induction argument we define $\overline{V}^t_{H+1} = \underline{V}^t_{H+1} = 0$, and also $V^{\pi,\nu}_{H+1}(x) = 0$ and $V^{*,\nu^t}_{H+1} = V^{\pi^t,*}_{H+1} = 0$, since there is no more future steps starting from $h = H + 1$. Given the estimation of $\overline{Q}^t_h, \underline{Q}^t_h$, the next step is to estimate the corresponding state value functions $\overline{V}^t_h, \underline{V}^t_h$. We utilize the `FIND_CCE` algorithm introduced recently in Xie et al. [2020] to find a coarse-correlated equilibrium of the payoff pair $(\overline{Q}^t_h(z), \underline{Q}^t_h(z))$.

Computational efficiency. By substituting the closed-form solutions to the maximization/minimization problems in (4.3), we can derive the analytic-form for $\overline{Q}^t_h$ and $\underline{Q}^t_h$. Take $\overline{Q}^t_h$ as an example, we have

$$\overline{Q}^t_h(z) = \Pi_{[-H,H]}\Big[r_h(z) + \overline{\boldsymbol{k}}^t_h(z)^\top(\overline{\mathbf{K}}^t_h + \lambda\mathbf{I})^{-1}\overline{\boldsymbol{y}}^t_h + \beta_t \cdot \overline{w}^t_h(z)\Big], \tag{4.4}$$

where the gram matrix $\overline{\mathbf{K}}^t_h$ and vector-valued function $\overline{\boldsymbol{k}}^t_h$ are defined as

$$\overline{\mathbf{K}}^t_h = \left(\overline{\boldsymbol{\Psi}}^t_h\right)\left(\overline{\boldsymbol{\Psi}}^t_h\right)^\top \in \mathbb{R}^{(t-1)\times(t-1)}, \quad \overline{\boldsymbol{k}}^t_h = \left(\overline{\boldsymbol{\Psi}}^t_h\right)\boldsymbol{\phi}_{\overline{V}^t_{h+1}}(z) = \left(\boldsymbol{k}_{\overline{V}^i_{h+1},\overline{V}^t_{h+1}}(z^i_h, z)\right)_i \in \mathbb{R}^{t-1}.$$

Also, we have $\overline{\boldsymbol{y}}^t_h := \left[\overline{V}^1_{h+1}(x^1_h), \ldots, \overline{V}^{t-1}_{h+1}(x^{t-1}_h)\right]^\top$ and $\overline{w}^t_h(z) = \lambda^{-1/2}\Big[\boldsymbol{k}_{\overline{V}^t_{h+1},\overline{V}^t_{h+1}}(z,z) - \overline{\boldsymbol{k}}^t_h(z)^\top(\overline{\mathbf{K}}^t_h + \lambda\mathbf{I})^{-1}\overline{\boldsymbol{k}}^t_h(z)\Big]^{1/2}$. Therefore, by the assumption that the weighted kernel function $\boldsymbol{k}_{V_1,V_2}$ can be evaluated efficiently, and $\overline{Q}^t_h$ and $\underline{Q}^t_h$ can also be computed efficiently. Furthermore, given $\overline{Q}^t_h$ and $\underline{Q}^t_h$, `FIND_CCE` can also be implemented efficiently [Xie et al., 2020]. Thus, Algorithm 1 is computationally efficient.

5 Main Results

In this section, we present the regret bound of our algorithm for the kernel mixture Markov Game. Recall that for the linear function class, the regret upper bound is characterized by the dimension of the linear function, the horizon of the game, and the number of episodes [Chen et al., 2022]. Our analysis in the RKHS function approximation setting aligns with the linear function approximation setting when $\mathbf{K}(z, z') = \boldsymbol{\phi}(z)^\top\boldsymbol{\phi}(z')$.

When considering the nonlinear function class as an approximator of the value function, we need to develop a new concept analogous to the dimension d that characterizes the intrinsic complexity of the function class $\mathcal{F}$. We do so by making use of the maximal information gain, $\Gamma_{\mathbf{K}}(T, \lambda)$ [Srinivas et al., 2010], where T is the episode number and H is the time horizon. In particular, we define the *effective dimension* of the RKHS $\mathcal{H}$ with respect to the mixture MG as follows:

Definition 2. We define the effective dimension $\Gamma_{\mathbf{K}}(T, \lambda)$ as follows:

$$\Gamma_{\mathbf{K}}(T, \lambda) := \sup_{(V_i)_i,(z_i)_i} \frac{1}{2}\log\det\left(\mathbf{I} + \mathbf{K}(\{V_i\}_i, \{z_i\}_i)/\lambda\right),$$

for any $1 \le i \le T$, $V_i : \mathcal{S} \to [-H, H]$, $z_i \in \mathcal{Z}$, where V_i's are functions mapping from $\mathcal{S}$ to $[-H, H]$ and z_i's are state-action tuples. Here, $\mathbf{K}(\{V_i\}_i, \{z_i\}_i) \in \mathbb{R}^{T\times T}$ and its (p, q)-th entry for any $1 \le p, q \le T$ is $[\mathbf{K}(\{V_i\}_i, \{z_i\}_i)]_{(p,q)} = \boldsymbol{k}_{V_p,V_q}(z_p, z_q)$.

By the boundedness of $\boldsymbol{\phi}_V$ as in Section 3.2, it is easy to verify that both the tabular MG and the linear mixture MG enjoy a finite effective dimension. Specifically, for finite RKHS $\mathcal{H}$ with rank d, $\Gamma_{\mathbf{K}}(T, \lambda) = O(d \cdot \log T)$ approximately depicts the rank of $\mathcal{H}$. Via a concentration argument, we first present our main lemma for bounding the estimation error when choosing $\beta_t = \beta$ for all $t \ge 1$:

Lemma 3. Assuming that for any $h \in [H]$, $\|\boldsymbol{\theta}^*_h\|_{\mathcal{H}} \le B$. Let $\lambda = 1 + 1/T$ and β satisfies $(\beta/H)^2 \ge 2\Gamma_{\mathbf{K}}(T, \lambda) + 2 + 4 \cdot \log(1/\delta) + 2\lambda\,(B/H)^2$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ the following holds for any $(t, h) \in [T] \times [H]$ and any $(x, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{A}$:

$$\left|\left\langle \boldsymbol{\phi}_{\overline{V}^t_{h+1}}(x, a, b), \overline{\boldsymbol{\theta}}^t_h - \boldsymbol{\theta}^*_h\right\rangle_{\mathcal{H}}\right| \le \beta \cdot \overline{w}^t_h(x, a, b), \ \left|\left\langle \boldsymbol{\phi}_{\underline{V}^t_{h+1}}(x, a, b), \underline{\boldsymbol{\theta}}^t_h - \boldsymbol{\theta}^*_h\right\rangle_{\mathcal{H}}\right| \le \beta \cdot \underline{w}^t_h(x, a, b).$$

We are now ready to present our main theorem.

Theorem 4 (RKHS function approximation)**.** Under the same conditions as Lemma 3, with probability at least $1-\delta$, `KernelCCE-VTR` has the following regret

$$\text{Regret}(T) = O\left(\beta H \sqrt{T \cdot \Gamma_{\mathbf{K}}(T, \lambda)} + 1\right).$$

Remark 5. Theorem 4 suggests that by treating the norm B as a constant, `KernelCCE-VTR` achieves an $\widetilde{O}(\Gamma_{\mathbf{K}}(T,\lambda)H^2\sqrt{T})$ regret bound. When the RKHS degenerates to the Euclidean space, the regret bound reduces to $\widetilde{O}(dH^2\sqrt{T})$, which matches the $\widetilde{O}(dH^{3/2}\sqrt{T})$ regret for linear mixture MGs presented by Chen et al. [2022] up to a $\sqrt{H}$ factor.

Similar to Xie et al. [2020], by using a standard online-to-batch conversion technique, we can convert the regret bound in Theorem 4 to a PAC bound. For simplicity, let the initial states of each episode be the same, i.e., $x_1^t = x_1$. After T episodes, we select $t_0 \in [T]$ such that

$$t_0 = \operatorname*{argmin}_{t \in [T]} \left\{ \overline{V}_1^t(x_1) - \underline{V}_1^t(x_1) \right\}, \tag{5.1}$$

which yields the following sample complexity guarantee for finding an ϵ-approximate NE policy pair (π^{t_0}, ν^{t_0}).

Corollary 6 (Sample complexity)**.** Under the same condition as Theorem 4, by setting $T = O\big(\beta^2 H^2 \Gamma_{\mathbf{K}}(T,\lambda)/\epsilon^2\big) = \widetilde{O}\big(H^4 \Gamma_{\mathbf{K}}^2(T,\lambda)/\epsilon^2\big)$ and selecting t_0 as in (5.1), the policy pair (π^{t_0}, ν^{t_0}) is an ϵ-approximate NE.

6 Bernstein-type Bonus, Misspecification, and Neural Function Approximation

In this section, we propose several extensions of `KernelCCE-VTR`. Section 6.1 introduces `KernelCCE-VTR` with Bernstein-type bonus. Section 6.2 discusses kernel function approximation with misspecification. We also specialize the kernel function approximation with misspecification to the neural function approximation setting, which is deferred to Appendix C.

6.1 `KernelCCE-VTR` with Bernstein-type Bonus

Recall that in `KernelCCE-VTR`, we need to choose β in order to calculate the optimistic and pessimistic state-action value functions $\overline{Q}_h^t(\cdot), \underline{Q}_h^t(\cdot)$ defined in (4.3). The theoretical value of β is defined in Lemma 3, which controls the uncertainty of the action-value estimate. Such a choice of β is due to a Hoeffding-type concentration used in the proof of Lemma 3. It has been shown in Zhou et al. [2021a] that by using a Bernstein-type bonus and a sharp analysis based on the total variance lemma, one can obtain an improved algorithm with a tighter regret bound. Following this idea, we propose a `KernelCCE-VTR+` algorithm, which replaces the Hoeffding-type bonus with a Bernstein-type bonus. To demonstrate the construction of the Bernstein-type bonus, we take the max player for example. In particular, we solve the following weighted kernel ridge regression problem:

$$\overline{\boldsymbol{\theta}}_{h,1}^t = \min_{\boldsymbol{\theta} \in \mathcal{H}} \sum_{\tau=1}^{t-1} \left[\overline{V}_{h+1}^\tau(x_{h+1}^\tau) - \left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}(z_h^\tau), \boldsymbol{\theta} \right\rangle_{\mathcal{H}} \right]^2 \Big/ \left(\overline{R}_h^\tau \right)^2 + \lambda_1 \|\boldsymbol{\theta}\|_{\mathcal{H}}^2, \tag{6.1}$$

where the input is the normalized feature mapping $\boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}(z_h^\tau)/\overline{R}_h^\tau$, the output is the normalized value function $\overline{V}_{h+1}^\tau(x_{h+1}^\tau)/\overline{R}_h^\tau$, and the normalization factor $\overline{R}_h^\tau$ is an upper bound on the conditional variance of $\overline{V}_{h+1}^\tau(x_{h+1}^\tau)$. It is straightforward to verify that (6.1) admits a closed-form solution. Given that solution, we can compute the upper confidence bound of the action-value functions $Q_h^{*,\nu}$. In detail, we define $\overline{\boldsymbol{\Psi}}_{h,1}^t := \left(\boldsymbol{\phi}_{\overline{V}_{h+1}^1}(z_h^1)/\overline{R}_h^1, \ldots, \boldsymbol{\phi}_{\overline{V}_{h+1}^{t-1}}(z_h^{t-1})/\overline{R}_h^{t-1} \right)^\top \in \mathcal{H}^{t-1}$. The gram matrix $\overline{\mathbf{K}}_{h,1}^t$, vector-valued function $\overline{\boldsymbol{k}}_{h,1}^t$ and the confidence region centered at $\overline{\boldsymbol{\theta}}_{h,1}^t$ in the RKHS $\mathcal{H}$ can be calculated the same as in Algorithm 1, except that $\overline{\boldsymbol{\Psi}}_h^t$ is replaced by $\overline{\boldsymbol{\Psi}}_{h,1}^t$. Then the optimistic estimate of the action-value function $Q_h^{*,\nu}$ has the following form:

$$\overline{Q}_h^t(z) = \Pi_{[-H,H]} \left[r_h(z) + \overline{\boldsymbol{k}}_{h,1}^t(z)^\top (\overline{\mathbf{K}}_{h,1}^t + \lambda \mathbf{I})^{-1} \overline{\boldsymbol{y}}_{h,1}^t + \beta_t \cdot \overline{w}_{h,1}^t(z) \right], \tag{6.2}$$

where $\overline{\boldsymbol{y}}_{h,1}^t := \left[\overline{V}_{h+1}^1(x_h^1)/\overline{R}_h^1, \dots \overline{V}_{h+1}^{t-1}(x_h^{t-1})/\overline{R}_h^{t-1}\right]^\top$ and

$$\overline{w}_{h,1}^t(z) = \lambda_1^{-1/2} \cdot \left[\boldsymbol{k}_{\overline{V}_{h+1}^t, \overline{V}_{h+1}^t}(z,z) - \overline{\boldsymbol{k}}_{h,1}^t(z)^\top \left(\overline{\mathbf{K}}_{h,1}^t + \lambda_1 \cdot \mathbf{I}\right)^{-1} \overline{\boldsymbol{k}}_{h,1}^t(z)\right]^{1/2}.$$

Due to space limit, we defer the conditional variance estimator $\overline{R}_h^t$ to Appendix B. Similarly, we can construct the pessimistic estimate of the action-value function $Q_h^{\pi,*}$ for the min player. We have the following informal result for `KernelCCE-VTR+`. The full algorithm and its formal result can be found in Appendix B.

Theorem 7 (Informal)**.** Let $d_{\text{eff}} = \Gamma_{\mathbf{K}}(T, \lambda)$, with proper choice of $\overline{R}_h^t$, $\underline{R}_h^t$ and β_t, with probability at least $1-\delta$, `KernelCCE-VTR+` has the following regret

$$\text{Regret}(T) = \widetilde{O}\bigg(d_{\text{eff}}^2 H^3 + \sqrt{d_{\text{eff}} H^4 + d_{\text{eff}}^2 H^3}\sqrt{T} + \left(d_{\text{eff}}^7 H^7 + d_{\text{eff}}^4 H^9\right)^{1/4} T^{1/4}\bigg).$$

Remark 8. When T is sufficiently large and $\Gamma_{\mathbf{K}}(T,\lambda)$ is larger than H, the regret bound in Theorem 7 is dominated by $\widetilde{O}\left(\Gamma_{\mathbf{K}}(T,\lambda)H^{3/2}\sqrt{T}\right)$, which improves the $\widetilde{O}\left(\Gamma_{\mathbf{K}}(T,\lambda)H^2\sqrt{T}\right)$ regret derived in Theorem 4 by a factor of $\sqrt{H}$. Compared with the $\widetilde{\Omega}\left(dH^{3/2}\sqrt{T}\right)$ lower bound proposed in Chen et al. [2022], our `KernelCCE-VTR+` algorithm is almost optimal when it reduces to the linear mixture MG.

6.2 Kernel Function Approximation with Misspecification

In this subsection, we consider the case where the function class may not be confined to an RKHS, but instead the distance to it can be bounded. This can be formulated as kernel function approximation with misspecification. We assume that there exists a misspecification error between the RKHS $\mathcal{H}$ and the true transition probability $\mathbb{P}_h(s'|z)$.

Assumption 9. There exists an $\iota_{\text{mis}} > 0$, an RKHS $\mathcal{H}$ with feature mapping $\boldsymbol{\phi} : \mathcal{Z} \mapsto \mathcal{S} \times \mathcal{H}$, and an unknown true parameter $\boldsymbol{\theta}_h^* \in \mathcal{H}$ satisfying $\|\boldsymbol{\theta}_h^*\|_{\mathcal{H}} \le B$ such that for any $h \in [H]$, the distance of the transition probability $\mathbb{P}_h$ to $\mathcal{H}$ can be bounded by ι_{mis}, which is $\left\|\mathbb{P}_h(\cdot|z) - \langle \boldsymbol{\phi}(\cdot|z), \boldsymbol{\theta}_h^*\rangle_{\mathcal{H}}\right\|_{\text{TV}} \le \iota_{\text{mis}}$.

In order to deal with model misspecification, the key idea is to enlarge β_t in the definition of the optimistic action-value function in (4.3). More specifically, we will add an extra $\mathcal{O}(H\iota_{\text{mis}}\sqrt{t})$ term brought by misspecification error to β specified in Lemma 3. We can show that `KernelCCE-VTR` with such enlarged β will have a sublinear regret in the presence of misspecification.

Theorem 10 (RKHS function approximation with misspecification)**.** Assuming that for any $h \in [H]$, $\|\boldsymbol{\theta}_h^*\|_{\mathcal{H}} \le B$. Set $\lambda = 1 + 1/T$ in the `KernelCCE-VTR` Algorithm. For any $\delta > 0$ and any β_t satisfying $(\beta_t/H)^2 \ge 2\Gamma_{\mathbf{K}}(T,\lambda) + 3 + 6 \cdot \log(1/\delta) + 3\lambda(B/H)^2 + 3\iota_{\text{mis}}^2 t$, there exists a global constant $c > 0$ such that with probability at least $1-\delta$, we have

$$\text{Regret}(T) \le c\left(\beta_T H\sqrt{T \cdot \Gamma_{\mathbf{K}}(T,\lambda)} + 1 + H^2 T\iota_{\text{mis}}\right).$$

In words, Theorem 10 suggests that in the misspeficified case, `KernelCCE-VTR` can achieve the same regret as that in the well-specified case up to an $O\left(\sqrt{\Gamma_{\mathbf{K}}(T,\lambda)}H^2 T\iota_{\text{mis}}\right)$ error. Such a linear dependence on ι_{mis} matches the result of single agent RL for the finite dimensional case [Jin et al., 2020, Zanette et al., 2020].

7 Conclusions

In this work, we studied learning for two-player mixture MGs using a kernel function approximation. We introduced a new formulation of kernel mixture MGs and proposed an algorithm `KernelCCE-VTR` that exploits the kernel function of the MG. We show that our `KernelCCE-VTR` is able to achieve a sublinear $\widetilde{O}(d_{\mathbf{K}}H^2\sqrt{T})$ regret. We further improve our algorithm with a *Bernstein-type bonus* and *weighted kernel ridge regression*, which enjoys a better $\widetilde{O}(d_{\mathbf{K}}H^{3/2}\sqrt{T})$ regret and nearly matches the regret lower bound in Chen et al. [2022] when reducing to linear mixture MGs. Finally, we extend our analysis of the basic RKHS setting to a more general nonlinear function approximation setting with misspecification errors and demonstrate that neural networks can be treated as a special instance of this misspecification framework. We believe our framework and analysis greatly broadens the expressiveness of the function classes used for MGs. We leave the study of learning in general-sum MGs by kernel function approximation as future work.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers and area chair for their helpful comments. DZ and QG are supported in part by the National Science Foundation CAREER Award 1906169 and the Sloan Research Fellowship. MIJ is supported in part by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764 and by the Vannevar Bush Faculty Fellowship program under grant number N00014-21-1-2941 and NSF grant IIS-1901252. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin F Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pages 551–560. PMLR, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, 365(6456): 885–890, 2019.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pages 10836–10846, 2019.

Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player zero-sum linear mixture Markov games. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167, pages 227–261. PMLR, 2022.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

Michael B Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. *Journal of the ACM (JACM)*, 68(1):1–39, 2021.

Qiwen Cui and Lin F Yang. Minimax sample complexity for turn-based stochastic game. In *Uncertainty in Artificial Intelligence*, pages 1496–1504. PMLR, 2021.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in RL. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

Baihe Huang, Jason D Lee, Zhaoran Wang, and Zhuoran Yang. Towards general function approximation in zero-sum Markov games. In *International Conference on Learning Representations*, 2022.

Zeyu Jia, Lin F. Yang, and Mengdi Wang. Feature-based Q-learning for two-player stochastic games. *arXiv preprint arXiv:1906.00423*, 2019.

Zeyu Jia, Lin Yang, Csaba Szepesvari, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pages 666–686. PMLR, 2020.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. In *Advances in Neural Information Processing Systems*, volume 34, 2021a.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent RL. *arXiv preprint arXiv:2110.14555*, 2021b.

Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent RL in large state spaces. In *International Conference on Machine Learning*, pages 10251–10279. PMLR, 2022.

Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311, 1984.

Michail G. Lagoudakis and Ronald Parr. Value function approximation in zero-sum Markov games. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 283–292. Morgan Kaufmann Publishers Inc., 2002.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning*, 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.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning*, pages 1321–1329, 2015.

Julien Pérolat, Bilal Piot, Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. Softened approximate policy iteration for Markov games. In *International Conference on Machine Learning*, pages 1860–1868. PMLR, 2016a.

Julien Pérolat, Bilal Piot, Bruno Scherrer, and Olivier Pietquin. On the use of non-stationary strategies for solving two-player zero-sum Markov games. In *International Conference on Artificial Intelligence and Statistics*, pages 893–901, 2016b.

Julien Pérolat, Florian Strub, Bilal Piot, and Olivier Pietquin. Learning Nash equilibrium for general-sum Markov games from batch data. In *International Conference on Artificial Intelligence and Statistics*, pages 232–241. PMLR, 2017.

Julien Perolat, Bilal Piot, and Olivier Pietquin. Actor-critic fictitious play in simultaneous move multistage games. In *International Conference on Artificial Intelligence and Statistics*, pages 919–928, 2018.

Shuang Qiu, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. On reward-free RL with kernel and neural function approximations: Single-agent MDP and Markov game. In *International Conference on Machine Learning*, pages 8737–8747. PMLR, 2021.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, volume 26, pages 2256–2264, 2013.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Lloyd S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100, 1953.

Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 2992–3002. PMLR, 2020.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pages 2898–2933. PMLR, 2019.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, pages 4987–4997, 2017.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pages 3674–3682. PMLR, 2020.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces. *arXiv preprint arXiv:2011.04622*, 2020.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. In *International Conference on Machine Learning*, pages 11492–11502. PMLR, 2020.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Conference on Learning Theory*. PMLR, 2021a.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted MDPs with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021b.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Section 7
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

The appendix is organized as follows. In Appendix A we introduce basic properties of RKHS. In Appendix B we discuss the implementation details of KernelCCE-VTR+. In Appendix C we introduce more details for applying our algorithm to the neural function approximation setting. In Appendix D we prove results for KernelCCE-VTR. In Appendix E we prove results for KernelCCE-VTR+. In Appendix F we prove results for KernelCCE-VTR with misspecification. In Appendix G we prove results for KernelCCE-VTR with neural function approximation. In Appendix H we prove the remaining auxiliary lemmas. Finally, in Appendix I we discuss the implementation details of FIND_CCE as an instance of linear programming.

A Properties of the Reproducing Kernel Hilbert Spaces

Recall that in Section 4, we define the update rule of $\overline{Q}_h^t, \underline{Q}_h^t$ in Eq. (4.3), where each term is defined in the sense of computational accessibility. For convenience of theoretical analysis, in this section we provide the equivalent forms of the Q-update on the RKHS. We have the following simple facts:

Lemma 11. Define covariance matrices $\overline{\mathbf{\Lambda}}_h^t, \underline{\mathbf{\Lambda}}_h^t : \mathcal{H} \mapsto \mathcal{H}$ as

$$\overline{\mathbf{\Lambda}}_h^t := \lambda \mathbf{I}_{\mathcal{H}} + \left(\overline{\mathbf{\Psi}}_h^t\right)^\top \left(\overline{\mathbf{\Psi}}_h^t\right), \quad \underline{\mathbf{\Lambda}}_h^t := \lambda \mathbf{I}_{\mathcal{H}} + \left(\underline{\mathbf{\Psi}}_h^t\right)^\top \left(\underline{\mathbf{\Psi}}_h^t\right), \tag{A.1}$$

where $\mathbf{I}_{\mathcal{H}}$ is the identity mapping on $\mathcal{H}$. Then the following holds:

(a) $\overline{\boldsymbol{\theta}}_h^t := \left(\overline{\mathbf{\Psi}}_h^t\right)^\top \left[\overline{\mathbf{K}}_h^t + \lambda \mathbf{I}\right]^{-1} \overline{\mathbf{y}}_h^t = \left(\overline{\mathbf{\Lambda}}_h^t\right)^{-1} \left(\overline{\mathbf{\Psi}}_h^t\right)^\top \overline{\mathbf{y}}_h^t \in \mathcal{H}$ and the same holds for $\underline{\boldsymbol{\theta}}_h^t$;

(b) $\overline{w}_h^t = \left[\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top \overline{\mathbf{\Lambda}}_h^t \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)\right]^{1/2}$ and the same holds for $\underline{w}_h^t$;

(c) $\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z) = \left(\overline{\mathbf{\Psi}}_h^t\right)^\top (\overline{\mathbf{K}}_h^t + \lambda \mathbf{I})^{-1} \overline{\boldsymbol{k}}_h^t(z) + \lambda \cdot (\overline{\mathbf{\Lambda}}_h^t)^{-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)$.

Proof. We prove the statements as follows.

(a) By definition of $\overline{\mathbf{K}}_h^t$ in Section 4, we note that

$$\left(\overline{\mathbf{\Psi}}_h^t\right)^\top \left[\overline{\mathbf{K}}_h^t + \lambda \mathbf{I}\right] = \left(\overline{\mathbf{\Psi}}_h^t\right)^\top \left[\left(\overline{\mathbf{\Psi}}_h^t\right)\left(\overline{\mathbf{\Psi}}_h^t\right)^\top + \lambda \mathbf{I}\right] = \left[\left(\overline{\mathbf{\Psi}}_h^t\right)^\top \left(\overline{\mathbf{\Psi}}_h^t\right) + \lambda \mathbf{I}_{\mathcal{H}}\right] \left(\overline{\mathbf{\Psi}}_h^t\right)^\top.$$

Taking the inverse operation on both sides of the second equality, we conclude

$$\left[\left(\overline{\mathbf{\Psi}}_h^t\right)\left(\overline{\mathbf{\Psi}}_h^t\right)^\top + \lambda \mathbf{I}\right]^{-1} \left(\overline{\mathbf{\Psi}}_h^t\right)^{-\top} = \left(\overline{\mathbf{\Psi}}_h^t\right)^{-\top} \left[\left(\overline{\mathbf{\Psi}}_h^t\right)^\top \left(\overline{\mathbf{\Psi}}_h^t\right) + \lambda \mathbf{I}_{\mathcal{H}}\right]^{-1},$$

and hence we arrive at an equality on space $\mathcal{H} \times \mathbb{R}^t$ that:

$$\begin{aligned}\left(\overline{\mathbf{\Psi}}_h^t\right)^\top \left[\overline{\mathbf{K}}_h^t + \lambda \mathbf{I}\right]^{-1} &= \left(\overline{\mathbf{\Psi}}_h^t\right)^\top \left[\left(\overline{\mathbf{\Psi}}_h^t\right)\left(\overline{\mathbf{\Psi}}_h^t\right)^\top + \lambda \mathbf{I}\right]^{-1} \\ &= \left[\left(\overline{\mathbf{\Psi}}_h^t\right)^\top \left(\overline{\mathbf{\Psi}}_h^t\right) + \lambda \mathbf{I}_{\mathcal{H}}\right]^{-1} \left(\overline{\mathbf{\Psi}}_h^t\right)^\top = \left(\overline{\mathbf{\Lambda}}_h^t\right)^{-1} \left(\overline{\mathbf{\Psi}}_h^t\right)^\top.\end{aligned}$$

Multiplying both sides by $\overline{\mathbf{y}}_h^t$ we have that the close form solution of Eq. (4.1) satisfies

$$\overline{\boldsymbol{\theta}}_h^t := \left(\overline{\mathbf{\Psi}}_h^t\right)^\top \left[\overline{\mathbf{K}}_h^t + \lambda \mathbf{I}\right]^{-1} \overline{\mathbf{y}}_h^t = \left(\overline{\mathbf{\Lambda}}_h^t\right)^{-1} \left(\overline{\mathbf{\Psi}}_h^t\right)^\top \overline{\mathbf{y}}_h^t \in \mathcal{H},$$

which proves item (a) of our results. The same argument holds for $\underline{\boldsymbol{\theta}}_h^t$.

Algorithm 2 `KernelCCE-VTR+`

Input: bonus parameter $\lambda_1, \lambda_2 > 0$.
for episode $t = 1, 2, \ldots, T$ **do**
    Receive initial state x_1^t
    for step $h = H, H-1, \ldots, 1$ **do**
        Estimate $\overline{R}_h^t, \underline{R}_h^t$ as in Eq. (B.2)
        Calculate $\overline{Q}_h^t(\cdot), \underline{Q}_h^t(\cdot)$ as in Eq. (6.2)
        For each x, let $\sigma_h^t(x) = \texttt{FIND\_CCE}(\overline{Q}_h^t, \underline{Q}_h^t, x)$
        Let $\overline{V}_h^t(x) = \mathbb{E}_{(a,b)\sim\sigma_h^t(x)}\overline{Q}_h^t(x,a,b)$ and $\underline{V}_h^t(x) = \mathbb{E}_{(a,b)\sim\sigma_h^t(x)}\underline{Q}_h^t(x,a,b)$
    end for
    for step $h = 1, 2, \ldots, T$ **do**
        Sample $(a_h^t, b_h^t) \sim \sigma_h^t(x_h^t)$
        P_1 takes action a_h^t, P_2 takes action b_h^t
        Observe next state x_{h+1}^t
    end for
end for

(b) By definition of $\overline{w}_h^t$, $\overline{\boldsymbol{k}}_h^t$ and $\overline{\mathbf{K}}_h^t$, we have

$$
\begin{aligned}
\overline{w}_h^t(z) &= \lambda^{-1/2} \cdot \left[\boldsymbol{k}_{\overline{V}_{h+1}^t, \overline{V}_{h+1}^t}(z,z) - \overline{\boldsymbol{k}}_h^t(z)^\top \left(\overline{\mathbf{K}}_h^t + \lambda \mathbf{I}\right)^{-1} \overline{\boldsymbol{k}}_h^t(z)\right]^{1/2} \\
&= \lambda^{-1/2} \cdot \left[\boldsymbol{k}_{\overline{V}_{h+1}^t, \overline{V}_{h+1}^t}(z,z) - \boldsymbol{\phi}_{\overline{V}_{h+1}^t}^\top(z) \left(\overline{\boldsymbol{\Psi}}_h^t\right)^\top \left(\overline{\mathbf{K}}_h^t + \lambda \mathbf{I}\right)^{-1} \left(\overline{\boldsymbol{\Psi}}_h^t\right) \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)\right]^{1/2} \\
&= \lambda^{-1/2} \cdot \left[\boldsymbol{k}_{\overline{V}_{h+1}^t, \overline{V}_{h+1}^t}(z,z) - \boldsymbol{\phi}_{\overline{V}_{h+1}^t}^\top(z) \left(\overline{\boldsymbol{\Lambda}}_h^t\right)^{-1} \left(\overline{\boldsymbol{\Psi}}_h^t\right)^\top \left(\overline{\boldsymbol{\Psi}}_h^t\right) \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)\right]^{1/2} \\
&= \lambda^{-1/2} \cdot \left[\boldsymbol{\phi}_{\overline{V}_{h+1}^t}^\top \left(\overline{\boldsymbol{\Lambda}}_h^t\right)^{-1} \left(\overline{\boldsymbol{\Lambda}}_h^t\right) \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z) - \boldsymbol{\phi}_{\overline{V}_{h+1}^t}^\top \left(\overline{\boldsymbol{\Lambda}}_h^t\right)^{-1} \left(\overline{\boldsymbol{\Psi}}_h^t\right)^\top \left(\overline{\boldsymbol{\Psi}}_h^t\right) \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)\right]^{1/2} \\
&= \left[\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z) (\overline{\boldsymbol{\Lambda}}_h^t)^{-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)\right]^{1/2}.
\end{aligned}
$$

This concludes the proof of item (b) of our result. The same argument holds for $\underline{w}_h^t(z)$.

(c) Noting that from the definition of $\overline{\boldsymbol{\Lambda}}_h^t$ in Eq. (A.1),

$$
\begin{aligned}
\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z) = \left(\overline{\boldsymbol{\Lambda}}_h^t\right)^{-1} \left(\overline{\boldsymbol{\Lambda}}_h^t\right) \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z) &= \left(\overline{\boldsymbol{\Lambda}}_h^t\right)^{-1} \left(\lambda \mathbf{I}_{\mathcal{H}} + \left(\overline{\boldsymbol{\Psi}}_h^t\right)^\top \left(\overline{\boldsymbol{\Psi}}_h^t\right)\right) \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z) \\
&= \left(\overline{\boldsymbol{\Lambda}}_h^t\right)^{-1} \left(\overline{\boldsymbol{\Psi}}_h^t\right)^\top \left(\overline{\boldsymbol{\Psi}}_h^t\right) \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z) + \lambda \cdot \left(\overline{\boldsymbol{\Lambda}}_h^t\right)^{-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z).
\end{aligned}
$$

Applying the results in the proof of item (a) on $\left(\overline{\boldsymbol{\Lambda}}_h^t\right)^{-1} \left(\overline{\boldsymbol{\Psi}}_h^t\right)^\top$, we have that

$$
\begin{aligned}
\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z) &= \left(\overline{\boldsymbol{\Psi}}_h^t\right)^\top \left[\overline{\mathbf{K}}_h^t + \lambda \mathbf{I}\right]^{-1} \left(\overline{\boldsymbol{\Psi}}_h^t\right) \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z) + \lambda \cdot \left(\overline{\boldsymbol{\Lambda}}_h^t\right)^{-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z) \qquad \text{(A.2)} \\
&= \left(\overline{\boldsymbol{\Psi}}_h^t\right)^\top \left[\overline{\mathbf{K}}_h^t + \lambda \mathbf{I}\right]^{-1} \overline{\boldsymbol{k}}_h^t(z) + \lambda \cdot \left(\overline{\boldsymbol{\Lambda}}_h^t\right)^{-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z),
\end{aligned}
$$

which concludes the proof of item (c).

□

B Details of `KernelCCE-VTR+`

In this section, we propose more details for the algorithm `KernelCCE-VTR+`. We consider the following ridge regression problem with each term weighed by its estimated variance:

$$\overline{\boldsymbol{\theta}}_{h,1}^{t} = \min_{\boldsymbol{\theta}\in\mathcal{H}} \sum_{\tau=1}^{t-1} \left[\overline{V}_{h+1}^{\tau}(x_{h+1}^{\tau}) - \left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^{\tau}}(z_h^{\tau}), \boldsymbol{\theta} \right\rangle_{\mathcal{H}} \right]^2 / \left(\overline{R}_h^{\tau} \right)^2 + \lambda_1 \left\|\boldsymbol{\theta}\right\|_{\mathcal{H}}^2 ,$$

$$\underline{\boldsymbol{\theta}}_{h,1}^{t} = \min_{\boldsymbol{\theta}\in\mathcal{H}} \sum_{\tau=1}^{t-1} \left[\underline{V}_{h+1}^{\tau}(x_{h+1}^{\tau}) - \left\langle \boldsymbol{\phi}_{\underline{V}_{h+1}^{\tau}}(z_h^{\tau}), \boldsymbol{\theta} \right\rangle_{\mathcal{H}} \right]^2 / \left(\underline{R}_h^{\tau} \right)^2 + \lambda_1 \left\|\boldsymbol{\theta}\right\|_{\mathcal{H}}^2 .$$

Here we use $\overline{R}_h^{\tau}, \underline{R}_h^{\tau}$ to denote upper bounds on the conditional variance of $\overline{V}_{h+1}^{\tau}(x_{h+1}^{\tau})$ and $\underline{V}_{h+1}^{\tau}(x_{h+1}^{\tau})$ respectively, which we will specify in later subsections. Next we define the necessary quantities in estimating the regret bound. Similarly as in previous sections, we define

$$\overline{\boldsymbol{\Psi}}_{h,1}^{t} := \left(\boldsymbol{\phi}_{\overline{V}_{h+1}^{1}}(z_h^1)/\overline{R}_h^1, \dots \boldsymbol{\phi}_{\overline{V}_{h+1}^{t-1}}(z_h^{t-1})/\overline{R}_h^{t-1} \right)^{\top} \in \mathcal{H}^{t-1}, \quad \text{and}$$

$$\underline{\boldsymbol{\Psi}}_{h,1}^{t} := \left(\boldsymbol{\phi}_{\underline{V}_{h+1}^{1}}(z_h^1)/\underline{R}_h^1, \dots \boldsymbol{\phi}_{\underline{V}_{h+1}^{t-1}}(z_h^{t-1})/\underline{R}_h^{t-1} \right)^{\top} \in \mathcal{H}^{t-1}.$$

The gram matrix $\overline{\mathbf{K}}_{h,1}^{t}$, vector-valued function $\overline{\boldsymbol{k}}_{h,1}^{t}$ and the confidence region centered at $\overline{\boldsymbol{\theta}}_{h,1}^{t}$ in the RKHS $\mathcal{H}$ are defined accordingly by replacing $\overline{\boldsymbol{\Psi}}_{h}^{t}, \underline{\boldsymbol{\Psi}}_{h}^{t}$ by $\overline{\boldsymbol{\Psi}}_{h,1}^{t}, \underline{\boldsymbol{\Psi}}_{h,1}^{t}$ respectively. The optimistic (pessimistic version can be defined accordingly) estimates of the action value function have the following closed form solution:

$$\overline{Q}_h^t(z) = \Pi_{[-H,H]}[r_h(z) + \overline{\boldsymbol{k}}_{h,1}^{t}(z)^{\top}(\overline{\mathbf{K}}_{h,1}^{t} + \lambda\mathbf{I})^{-1}\overline{\boldsymbol{y}}_{h,1}^{t} + \beta_t \cdot \overline{w}_{h,1}^{t}(z)], \tag{B.1}$$

where

$$\overline{\boldsymbol{y}}_{h,1}^{t} := \left[\overline{V}_{h+1}^{1}(x_h^1)/\overline{R}_h^1, \dots \overline{V}_{h+1}^{t-1}(x_h^{t-1})/\overline{R}_h^{t-1} \right]^{\top},$$

and

$$\overline{w}_{h,1}^{t}(z) = \lambda_1^{-1/2} \cdot \left[\boldsymbol{k}_{\overline{V}_{h+1}^{t}, \overline{V}_{h+1}^{t}}(z,z) - \overline{\boldsymbol{k}}_{h,1}^{t}(z)^{\top} \left(\overline{\mathbf{K}}_{h,1}^{t} + \lambda_1 \cdot \mathbf{I} \right)^{-1} \overline{\boldsymbol{k}}_{h,1}^{t}(z) \right]^{1/2}.$$

The full version of the algorithm is presented formally in Algorithm 2.

B.1 Variance Estimator

In order to determine the values of $\overline{R}_h^{\tau}, \underline{R}_h^{\tau}$, we note that we can solve a ridge regression problem for estimating the expected square of the value function:

$$\overline{\boldsymbol{\theta}}_{h,2}^{t} = \min_{\boldsymbol{\theta}\in\mathcal{H}} \sum_{\tau=1}^{t-1} \left[\left(\overline{V}_{h+1}^{\tau}(x_{h+1}^{\tau}) \right)^2 - \left\langle \boldsymbol{\phi}_{(\overline{V}_{h+1}^{\tau})^2}(z_h^{\tau}), \boldsymbol{\theta} \right\rangle_{\mathcal{H}} \right]^2 + \lambda_2 \left\|\boldsymbol{\theta}\right\|_{\mathcal{H}}^2 ,$$

$$\underline{\boldsymbol{\theta}}_{h,2}^{t} = \min_{\boldsymbol{\theta}\in\mathcal{H}} \sum_{\tau=1}^{t-1} \left[\left(\underline{V}_{h+1}^{\tau}(x_{h+1}^{\tau}) \right)^2 - \left\langle \boldsymbol{\phi}_{(\underline{V}_{h+1}^{\tau})^2}(z_h^{\tau}), \boldsymbol{\theta} \right\rangle_{\mathcal{H}} \right]^2 + \lambda_2 \left\|\boldsymbol{\theta}\right\|_{\mathcal{H}}^2 .$$

By defining

$$\overline{\boldsymbol{\Psi}}_{h,2}^{t} := \left(\boldsymbol{\phi}_{(\overline{V}_{h+1}^{1})^2}(z_h^1), \dots \boldsymbol{\phi}_{(\overline{V}_{h+1}^{t-1})^2}(z_h^{t-1}) \right)^{\top} \in \mathcal{H}^{t-1},$$

$$\underline{\boldsymbol{\Psi}}_{h,2}^{t} := \left(\boldsymbol{\phi}_{(\underline{V}_{h+1}^{1})^2}(z_h^1), \dots \boldsymbol{\phi}_{(\underline{V}_{h+1}^{t-1})^2}(z_h^{t-1}) \right)^{\top} \in \mathcal{H}^{t-1},$$

we can define the gram matrix $\overline{\mathbf{K}}_{h,2}^t$, vector-valued function $\overline{\boldsymbol{k}}_{h,2}^t$, and

$$\overline{w}_{h,2}^t(z) = \lambda_2^{-1/2} \cdot \left[\boldsymbol{k}_{(\overline{V}_{h+1}^t)^2,(\overline{V}_{h+1}^t)^2}(z,z) - \overline{\boldsymbol{k}}_{h,2}^t(z)^\top \left(\overline{\mathbf{K}}_{h,2}^t + \lambda_2 \cdot \mathbf{I}\right)^{-1} \overline{\boldsymbol{k}}_{h,2}^t(z)\right]^{1/2}.$$

The variance estimator is thus defined as:

$$\begin{aligned}\mathbb{V}^{\text{est}}\overline{V}_{h+1}^t(z_h^t) &:= \left\langle \boldsymbol{\phi}_{(\overline{V}_{h+1}^t)^2}(z_h^t), \overline{\boldsymbol{\theta}}_{h,2}^t \right\rangle_{\mathcal{H}} - \left(\left\langle \boldsymbol{\phi}_{(\overline{V}_{h+1}^t)^2}(z_h^t), \overline{\boldsymbol{\theta}}_{h,1}^t \right\rangle_{\mathcal{H}}\right)^2 \\ &\approx \mathbb{P}_h \left(\underline{V}_{h+1}^t(x_{h+1}^t)\right)^2 - \left(\mathbb{P}_h \underline{V}_{h+1}^t(x_{h+1}^t)\right)^2,\end{aligned}$$

and

$$\begin{aligned}\left(\overline{R}_h^t\right)^2 &:= \max\{\mathbb{V}^{\text{est}}\overline{V}_{h+1}^t(z_h^t) + \overline{E}_h^t, (\alpha_t)^2\}, \\ \overline{E}_h^t &:= \min\left\{H^2, \beta_t^{(2)}\overline{w}_{h,2}^t\right\} + \min\left\{H^2, 2H\beta_t^{(1)}\overline{w}_{h,1}^t\right\}. \end{aligned} \tag{B.2}$$

Up till now, we have finished the definition of the variance estimator for the upper-value estimator. The lower-value estimator can be defined in a similar fashion, and we omit the details.

B.2 Main Results

In this section, we provide theoretical results for the regret bound under the weighted setting described above. First we propose a key lemma which suggests that our constructed $\overline{\boldsymbol{\theta}}_{h,1}^t$ and $\overline{\boldsymbol{\theta}}_{h,2}^t$ are good estimates to $\boldsymbol{\theta}_h^*$ with high probability.

Lemma 12. Assuming that for any $h \in [H]$, $\|\boldsymbol{\theta}_h^*\|_{\mathcal{H}} \leq B$. Let $\alpha_t, \beta_t^{(1)}, \beta_t^{(2)}$ satisfy $\alpha_t = \alpha$,

$$\beta_t^{(1)} = (16H/\alpha)\sqrt{\Gamma_{\mathbf{K}}(T, \lambda_1\alpha^2)}\sqrt{\log(4t^2H/\delta)} + (8H/\alpha)\log(4t^2H/\delta) + \sqrt{\lambda_1} \cdot B, \tag{B.3}$$

$$\beta_t^{(2)} = 16H^2\sqrt{\Gamma_{\mathbf{K}}(T, \lambda_2/H^2)}\sqrt{\log(4t^2H/\delta)} + 8H^2\log(4t^2H/\delta) + \sqrt{\lambda_2} \cdot B, \tag{B.4}$$

then for any $\delta > 0$, there exists an event $\mathcal{E}$ satisfying $\mathbb{P}(\mathcal{E}) \geq 1 - 2\delta$ such that on $\mathcal{E}$ the following holds for any $(t, h) \in [T] \times [H]$ and any $(x, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{A}$:

$$\left|\left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z_h^t), \boldsymbol{\theta}_h^* - \overline{\boldsymbol{\theta}}_{h,1}^t \right\rangle_{\mathcal{H}}\right| \leq \beta_t^{(1)} \cdot \overline{w}_{h,1}^k(z_h^k),$$

and

$$\left|\left\langle \boldsymbol{\phi}_{(\overline{V}_{h+1}^t)^2}(z_h^t), \boldsymbol{\theta}_h^* - \overline{\boldsymbol{\theta}}_{h,2}^t \right\rangle_{\mathcal{H}}\right| \leq \beta_t^{(2)} \cdot \overline{w}_{h,2}^k(z_h^k).$$

We now propose our main theorem, which is the formal version of Theorem 7 and suggests that the regret bound of Algorithm 2 is upper bounded by $\widetilde{O}\left(\Gamma_{\mathbf{K}}(T, \lambda)H^2\sqrt{T}\right)$.

Theorem 13. Assuming that for any $h \in [H]$, $\|\boldsymbol{\theta}_h^*\|_{\mathcal{H}} \leq B$. Let $\lambda = 1/B^2$, $d_{\text{eff}} = \Gamma_{\mathbf{K}}(T, \lambda)$, $\lambda_1 = d_{\text{eff}}/(B^2H^2)$, $\lambda_2 = H^2/B^2$, and taking $\beta_t, \beta_t^{(1)}, \beta_t^{(2)}$ as in Eq. (E.5), (B.3) and (B.4), then with probability at least $1 - \delta$, the following holds that:

$$\begin{aligned}\text{Regret}(T) &:= \sum_{t=1}^T V_1^{*,\nu^t}(x_h^t) - V_1^{\pi^t,*}(x_h^t) \\ &\leq \widetilde{O}\left(d_{\text{eff}}^2H^3 + \sqrt{d_{\text{eff}}H^4 + d_{\text{eff}}^2H^3}\sqrt{T} + \left(d_{\text{eff}}^7H^8 + d_{\text{eff}}^4H^9\right)^{1/4}T^{1/4}\right).\end{aligned}$$

Proofs of Lemma 12 and Theorem 13 are deferred to Section E.

Algorithm 3 NeuralCCE-VTR

```
Input: bonus parameter β_t > 0.
for episode t = 1, 2, ..., T do
    Receive initial state x_1^t
    for step h = H, H − 1, ..., 1 do
        Solve the optimization problem (C.1)
        Calculate \overline{Q}_h^t(·), \underline{Q}_h^t(·) as in Eq. (C.4)
        For each x, let σ_h^t(x) = FIND_CCE(\overline{Q}_h^t, \underline{Q}_h^t, x)
        Let \overline{V}_h^t(x_h^t) = E_{(a,b)∼σ_h^t(x_h^t)} \overline{Q}_h^t(x_h^t, a, b) and \underline{V}_h^t(x_h^t) = E_{(a,b)∼σ_h^t(x_h^t)} \underline{Q}_h^t(x_h^t, a, b)
    end for
    for step h = 1, 2, ..., T do
        Sample (a_h^t, b_h^t) ∼ σ_h^t(x_h^t)
        P_1 takes action a_h^t, P_2 takes action b_h^t
        Observe next state x_{h+1}^t
    end for
end for
```

C Neural Network (NN) Function Approximation

C.1 Neural Function Approximation

In this subsection, we show that neural network function approximation can be treated as a special case of kernel function approximation with misspecification. We denote $z := (x, a, b)$ as a vector in $\mathbb{R}^d$ that satisfies $\|z\| = 1$ and represent the parameters of a L-Layer fully connected neural network f by $\boldsymbol{\theta} := \left[\operatorname{vec}(\mathbf{W}_1)^\top, \operatorname{vec}(\mathbf{W}_2)^\top, \ldots, \operatorname{vec}(\mathbf{W}_L)^\top\right]^\top$, where $\mathbf{W}_1 \in \mathbb{R}^{m\times d}$, $\mathbf{W}_l \in \mathbb{R}^{m\times m}$ for $2 \le l \le L-1$ and $\mathbf{W}_L \in \mathbb{R}^{1\times m}$. The neural network $f(z;\boldsymbol{\theta})$ with parameter set $\boldsymbol{\theta}$ can be defined as:

$$f(z;\boldsymbol{\theta}) = \sqrt{m}\mathbf{W}_L G\left(\cdots G\left(\mathbf{W}_2 G\left(\mathbf{W}_1 z\right)\right)\right),$$

where $G(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is an activation function. For $1 \le l \le L-1$, $\mathbf{W}_l$ is initialized as $\mathbf{W}_l = (\mathbf{W}, \mathbf{0}; \mathbf{0}, \mathbf{W})$, where each entry of $\mathbf{W}$ is generated independently from normal distribution $N(0, 4/m)$; $\mathbf{W}_L$ is initialized as $\mathbf{W}_L = (\boldsymbol{w}^\top, -\boldsymbol{w}^\top)$, where each entry of $\boldsymbol{w}$ is generated independently from $N(0, 2/m)$. Given the initialized parameter $\boldsymbol{\theta}^{(0)}$, we choose the feature map as the gradient of f at $\boldsymbol{\theta}^{(0)}$:

$$\boldsymbol{\phi}(z) = \nabla_{\boldsymbol{\theta}} f(z;\boldsymbol{\theta}^{(0)})/\sqrt{m}.$$

Then we define the weighted kernel function $\boldsymbol{k}_{V_1,V_2}(\cdot,\cdot)$ in Definition 1 with $\boldsymbol{\phi}(z)$. Similarly, we define the effective dimension $\Gamma_{\mathbf{K}}(T, \lambda)$ with respect to the kernel function $\boldsymbol{k}_{V_1,V_2}(\cdot,\cdot)$, in the same fashion of Definition 2. Our assumption is that for $\forall h \in [H]$ our transition probability $\mathbb{P}_h$ can be modeled by the neural network with parameter $\boldsymbol{\theta}_h^*$ satisfying $\left\|\boldsymbol{\theta}_h^* - \boldsymbol{\theta}^{(0)}\right\|_2 \le B$:

$$\mathbb{P}_h(x'|z) = f(x', z; \boldsymbol{\theta}_h^*).$$

Now we show the details of our Algorithm 3. Similarily as in Eq. (4.1), we solve penalized ridge regression problem for the min-player and the max-player respectively

$$\overline{\boldsymbol{\theta}}_h^t = \min_{\boldsymbol{\theta}\in\mathbb{R}^P} \sum_{\tau=1}^{t-1} \left[\overline{V}_{h+1}^\tau(x_{h+1}^\tau) - f_{\overline{V}_{h+1}^\tau}(z_h^\tau;\boldsymbol{\theta})\right]^2 + \lambda \cdot \left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}\right\|^2,$$

$$\underline{\boldsymbol{\theta}}_h^t = \min_{\boldsymbol{\theta}\in\mathbb{R}^P} \sum_{\tau=1}^{t-1} \left[\underline{V}_{h+1}^\tau(x_{h+1}^\tau) - f_{\underline{V}_{h+1}^\tau}(z_h^\tau;\boldsymbol{\theta})\right]^2 + \lambda \cdot \left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}\right\|^2, \tag{C.1}$$

where $p = md + m^2(L-2) + m$ is the dimension of the parameter space, and $f_{\overline{V}_{h+1}^\tau}, f_{\underline{V}_{h+1}^\tau}$ are defined similarly as $\boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}$ as follows:

$$f_{\overline{V}_{h+1}^\tau}(z;\boldsymbol{\theta}) = \sum_{s'\in\mathcal{S}} \overline{V}_{h+1}^\tau(s') f(s', z;\boldsymbol{\theta}), \qquad f_{\underline{V}_{h+1}^\tau}(z;\boldsymbol{\theta}) = \sum_{s'\in\mathcal{S}} \underline{V}_{h+1}^\tau(s') f(s', z;\boldsymbol{\theta}).$$

For given $\overline{\boldsymbol{\theta}}_h^t, \underline{\boldsymbol{\theta}}_h^t$, we define

$$\overline{\boldsymbol{\Psi}}_h^t := \left(\boldsymbol{\phi}_{\overline{V}_{h+1}^1}(z_h^1; \overline{\boldsymbol{\theta}}_h^2), \dots \boldsymbol{\phi}_{\overline{V}_{h+1}^{t-1}}(z_h^{t-1}; \overline{\boldsymbol{\theta}}_h^t)\right)^\top,$$

$$\underline{\boldsymbol{\Psi}}_h^t := \left(\boldsymbol{\phi}_{\underline{V}_{h+1}^1}(z_h^1; \underline{\boldsymbol{\theta}}_h^2), \dots \boldsymbol{\phi}_{\underline{V}_{h+1}^{t-1}}(z_h^{t-1}; \underline{\boldsymbol{\theta}}_h^t)\right)^\top. \tag{C.2}$$

Furthermore,

$$\overline{\boldsymbol{\Lambda}}_h^t := \lambda \mathbf{I} + (\overline{\boldsymbol{\Psi}}_h^t)^\top \overline{\boldsymbol{\Psi}}_h^t, \qquad \underline{\boldsymbol{\Lambda}}_h^t := \lambda \mathbf{I} + (\underline{\boldsymbol{\Psi}}_h^t)^\top \underline{\boldsymbol{\Psi}}_h^t,$$

and

$$\overline{w}_h^t(z) := \left[\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z; \overline{\boldsymbol{\theta}}_h^t)^\top (\overline{\boldsymbol{\Lambda}}_h^t)^{-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z; \overline{\boldsymbol{\theta}}_h^t)\right]^{1/2},$$

$$\underline{w}_h^t(z) := \left[\boldsymbol{\phi}_{\underline{V}_{h+1}^t}(z; \underline{\boldsymbol{\theta}}_h^t)^\top (\underline{\boldsymbol{\Lambda}}_h^t)^{-1} \boldsymbol{\phi}_{\underline{V}_{h+1}^t}(z; \underline{\boldsymbol{\theta}}_h^t)\right]^{1/2}. \tag{C.3}$$

Using the $\overline{\boldsymbol{\Lambda}}_h^t, \underline{\boldsymbol{\Lambda}}_h^t, \overline{w}_h^t, \underline{w}_h^t$, we estimate the optimal value functions as

$$\overline{Q}_h^t(z) = \Pi_{[-H,H]}\{r_h(z) + f_{\overline{V}_{h+1}^t}(z; \overline{\boldsymbol{\theta}}_h^t) + \beta \cdot \overline{w}_h^t(z)\},$$

$$\underline{Q}_h^t(z) = \Pi_{[-H,H]}\{r_h(z) + f_{\underline{V}_{h+1}^t}(z; \underline{\boldsymbol{\theta}}_h^t) - \beta \cdot \underline{w}_h^t(z)\}. \tag{C.4}$$

Combining with the procedures of finding a CCE, we present the full version of our algorithm as in Algorithm 3.

We have the following result on the neural network at initialization.

Lemma 14. There exist constants $C_i > 0$ such that for any $\delta \in (0,1)$, if B satisfies that

$$B \geq C_1 m^{-1} L^{-3/2} \max\{\log^{-3/2} m, \log^{3/2}(|\mathcal{Z}|HL^2/\delta)\},$$
$$B \leq C_2 L^{-6} (\log m)^{-3/2},$$

then with probability at least $1-\delta$, we have for all $z \in \mathcal{Z}$, $h \in [H]$ and $V_h : \mathcal{S} \to [-1,1]$,

$$|\mathbb{P}_h V_h(z) - \langle \boldsymbol{\phi}_{V_h}(z), \boldsymbol{\theta}_h^* - \boldsymbol{\theta}^{(0)} \rangle| \leq C_3 |\mathcal{S}| B^{4/3} m^{-1/6} L^3 \sqrt{\log m},$$

and

$$\|\boldsymbol{\phi}_{V_h}(z)\|_2 \leq C := C_4 |\mathcal{S}| \sqrt{L}.$$

Lemma 14, whose proof is deferred to Appendix H, suggests that in the NN approximation setting, Assumption 9 for the misspecified kernel approximation setting is satisfied with $\iota_{\mathrm{mis}} = C_3|\mathcal{S}|B^{4/3}m^{-1/6}L^3\sqrt{\log m}$ and with probability at least $1-\delta$. The misspecified error is sufficiently small when m is large. We note that the definition of $\boldsymbol{\phi}(z)$ in the NN setting does not match the boundedness assumption in Section 3.2. We balance the scale of $\boldsymbol{\phi}(z)$ by the constant C in Lemma 14 which goes into the choice of $\lambda = C^2(1+1/T)$. With these at hand, we are ready to present our main result for NN approximation:

Theorem 15 (NN approximation)**.** Let C be the constant in Lemma 14. Assuming that for any $h \in [H]$, $\left\|\boldsymbol{\theta}_h^* - \boldsymbol{\theta}^{(0)}\right\|_2 \leq B$. Set $\lambda = C^2(1+1/T)$ in the `KernelCCE-VTR` Algorithm. For any $\delta > 0$ and any β_t satisfying

$$\left(\frac{\beta_t}{H}\right)^2 \geq 2\Gamma_{\mathbf{K}}(T,\lambda) + 3 + 6 \cdot \log\left(\frac{1}{\delta}\right) + 3\lambda\left(\frac{B}{H}\right)^2 + 3 \cdot C^2 \cdot B^{8/3} \cdot m^{-1/12} \cdot t \cdot \log m,$$

there exists a global constant $c > 0$ such that with probability at least $1-2\delta$, we have

$$\mathrm{Regret}(T) \leq c\left(\beta_T H \sqrt{T \cdot \Gamma_{\mathbf{K}}(T,\lambda)} + 1 + B^{4/3} H^2 T m^{-1/6} \sqrt{\log m}\right).$$

Theorem 15 suggests that when we use an overparameterized deep neural network ($m \gg 1$) to approximate the transition dynamic, `KernelCCE-VTR` achieves an $\widetilde{O}(\Gamma_{\mathbf{K}}(T,\lambda)H^2\sqrt{T})$ regret, which is of the same order as that in Theorem 4. We defer the proof of Theorem 15 to Appendix G.

D Proof of Results for `KernelCCE-VTR`

In this subsection, we provide the proof of our main Theorem 4 on RKHS.

D.1 Proof of Theorem 4

We recall that the duality gap is defined as $\sum_{t=1}^{T} V_1^{*,\nu^t}(x_1^t) - V_1^{\pi^t,*}(x_1^t)$. As can be seen in our algorithm, we maintain an optimistic estimate of $V_h^{*,v^t}(\cdot)$ as $\overline{V}_{h+1}^t(\cdot)$ and a pessimistic estimate of $V_h^{\pi^t,*}(\cdot)$ as $\underline{V}_{h+1}^t(\cdot)$. Hence the term $\overline{V}_{h+1}^t(x_h^t) - \underline{V}_{h+1}^t(x_h^t)$ is approximately the upper bound of the duality gap. We write the decomposition formally as below:

$$V_h^{*,\nu^t}(x_h^t) - V_h^{\pi^t,*}(x_h^t) = \underbrace{\overline{V}_h^t(x_h^t) - \underline{V}_h^t(x_h^t)}_{\text{I}} - \underbrace{\Big(V_h^{\pi^t,*}(x_h^t) - \underline{V}_h^t(x_h^t)\Big)}_{\text{II}} - \underbrace{\Big(\overline{V}_h^t(x_h^t) - V_h^{*,\nu^t}(x_h^t)\Big)}_{\text{III}}. \tag{D.1}$$

We use $\overline{\delta}_h^t$ to denote an important quantity $\left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z_h^t), \boldsymbol{\theta}_h^* - \overline{\boldsymbol{\theta}}_h^t \right\rangle_{\mathcal{H}}$ (and $\left\langle \boldsymbol{\phi}_{\underline{V}_{h+1}^t}(z_h^t), \boldsymbol{\theta}_h^* - \underline{\boldsymbol{\theta}}_h^t \right\rangle_{\mathcal{H}}$) in estimating the duality gap. In the rest of the proof we aim to show that all of the above three terms can be bounded by a quantity related to $\overline{\delta}_h^t$ ($\underline{\delta}_h^t$) and a stochastic random variable that forms a martingale difference sequence when considering for all $h \in [H], t \in [T]$.

For bounding term I, we first define two sequences of zero mean variables:

$$\begin{aligned} \gamma_h^t &:= \overline{Q}_h^t(x_h^t, a_h^t, b_h^t) - \underline{Q}_h^t(x_h^t, a_h^t, b_h^t) - \mathbb{E}_{(a,b)}\left[\overline{Q}_h^t(x, a, b) - \underline{Q}_h^t(x, a, b)\right], \\ \xi_h^t &:= \Big(\mathbb{P}_h(\overline{V}_{h+1}^t - \underline{V}_{h+1}^t)\Big)(x_h^t, a_h^t, b_h^t) - \Big(\overline{V}_{h+1}^t(x_{h+1}^t) - \underline{V}_{h+1}^t(x_{h+1}^t)\Big), \end{aligned} \tag{D.2}$$

where γ_h^t depicts the stochastic error with respect to the policy and ξ_h^t depicts the stochastic error with respect to the transition. We refer the readers to the proof of Lemma 16 for detailed explanations on these two error term. Given the above definition, we have the following Lemma 16.

Lemma 16. Under the settings of Lemma 3, we have the following recursive bound for $\forall h \in [H]$:

$$\begin{aligned} &\overline{V}_h^t(x_h^t) - \underline{V}_h^t(x_h^t) \\ &\le \overline{V}_{h+1}^t(x_{h+1}^t) - \underline{V}_{h+1}^t(x_{h+1}^t) + 2\beta_t \min\{1, \overline{w}_h^t(x_h^t)\} + 2\beta_t \min\{1, \underline{w}_h^t(x_h^t)\} + \xi_h^t + \gamma_h^t. \end{aligned}$$

Proof of Lemma 16. By the update rule of Algorithm 3, we have the following relation:

$$\overline{V}_h^t(x_h^t) - \underline{V}_h^t(x_h^t) = \mathbb{E}_{(a,b)\sim\sigma_h^t(x_h^t)}\left[\overline{Q}_h^t(x_h^t, a, b) - \underline{Q}_h^t(x_h^t, a, b)\right]. \tag{D.3}$$

We note that the RHS of Eq. (D.3) is an expectation over the CCE distribution $\sigma_h^t(x_h^t)$, which can be decomposed into one sample from the distribution plus a noise term as follows:

$$\overline{V}_h^t(x_h^t) - \underline{V}_h^t(x_h^t) = \overline{Q}_h^t(x_h^t, a_h^t, b_h^t) - \underline{Q}_h^t(x_h^t, a_h^t, b_h^t) + \gamma_h^t, \tag{D.4}$$

where

$$\gamma_h^t := \overline{Q}_h^t(x_h^t, a_h^t, b_h^t) - \underline{Q}_h^t(x_h^t, a_h^t, b_h^t) - \mathbb{E}_{(a,b)}\left[\overline{Q}_h^t(x, a, b) - \underline{Q}_h^t(x, a, b)\right].$$

Furthermore, for bounding the difference between the upper confidence Q estimation and the lower confidence Q estimation, we have

$$\begin{aligned} &\overline{Q}_h^t(z_h^t) - \underline{Q}_h^t(z_h^t) \\ &\le \left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z_h^t), \overline{\boldsymbol{\theta}}_h^t \right\rangle_{\mathcal{H}} - \left\langle \boldsymbol{\phi}_{\underline{V}_{h+1}^t}(z_h^t), \underline{\boldsymbol{\theta}}_h^t \right\rangle_{\mathcal{H}} + \beta_t \overline{w}_h^t(z_h^t) + \beta_t \underline{w}_h^t(z_h^t) \\ &= \left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z_h^t), \overline{\boldsymbol{\theta}}_h^t \right\rangle_{\mathcal{H}} - \left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z_h^t), \boldsymbol{\theta}_h^* \right\rangle_{\mathcal{H}} + \left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z_h^t), \boldsymbol{\theta}_h^* \right\rangle_{\mathcal{H}} - \left\langle \boldsymbol{\phi}_{\underline{V}_{h+1}^t}(z_h^t), \boldsymbol{\theta}_h^* \right\rangle_{\mathcal{H}} \\ &\quad + \left\langle \boldsymbol{\phi}_{\underline{V}_{h+1}^t}(z_h^t), \boldsymbol{\theta}_h^* \right\rangle_{\mathcal{H}} - \left\langle \boldsymbol{\phi}_{\underline{V}_{h+1}^t}(z_h^t), \underline{\boldsymbol{\theta}}_h^t \right\rangle_{\mathcal{H}} + \beta_t \overline{w}_h^t(z_h^t) + \beta_t \underline{w}_h^t(z_h^t) \end{aligned}$$

$$= \left\langle \boldsymbol{\phi}_{\overline{V}^t_{h+1}}(z^t_h), \overline{\boldsymbol{\theta}}^t_h - \boldsymbol{\theta}^*_h \right\rangle_{\mathcal{H}} + \left(\mathbb{P}_h(\overline{V}^t_{h+1} - \underline{V}^t_{h+1})\right)(z^t_h) + \left\langle \boldsymbol{\phi}_{\underline{V}^t_{h+1}}(z^t_h), \boldsymbol{\theta}^*_h - \underline{\boldsymbol{\theta}}^t_h \right\rangle_{\mathcal{H}}$$
$$+ \beta_t \overline{w}^t_h(z^t_h) + \beta_t \underline{w}^t_h(z^t_h).$$

By utilizing Lemma 3, we further arrive at:

$$\overline{Q}^t_h(z^t_h) - \underline{Q}^t_h(z^t_h) \leq \left(\mathbb{P}_h(\overline{V}^t_{h+1} - \underline{V}^t_{h+1})\right)(z^t_h) + 2\beta_t \overline{w}^t_h(z^t_h) + 2\beta_t \underline{w}^t_h(z^t_h),$$

where again by extracting the sequence

$$\xi^t_h := \left(\mathbb{P}_h(\overline{V}^t_{h+1} - \underline{V}^t_{h+1})\right)(x^t_h, a^t_h, b^t_h) - \left(\overline{V}^t_{h+1}(x^t_{h+1}) - \underline{V}^t_{h+1}(x^t_{h+1})\right),$$

we have

$$\overline{Q}^t_h(z^t_h) - \underline{Q}^t_h(z^t_h) \leq \overline{V}^t_{h+1}(x^t_{h+1}) - \underline{V}^t_{h+1}(x^t_{h+1}) + 2\beta \overline{w}^t_h(z^t_h) + 2\beta \underline{w}^t_h(z^t_h) + \xi^t_h. \tag{D.5}$$

Combining Eq. (D.4) and (D.5) concludes the following recursive bound:

$$\overline{V}^t_h(x^t_h) - \underline{V}^t_h(x^t_h) \leq \overline{V}^t_{h+1}(x^t_{h+1}) - \underline{V}^t_{h+1}(x^t_{h+1}) + 2\beta_t \overline{w}^t_h(z^t_h) + 2\beta_t \underline{w}^t_h(z^t_h) + \xi^t_h + \gamma^t_h$$

Moreover, due to the fact that $\overline{V}^t_h(x^t_h) - \underline{V}^t_h(x^t_h) \leq 2H$, we rewrite the above inequality into:

$$\begin{aligned}
&\overline{V}^t_h(x^t_h) - \underline{V}^t_h(x^t_h) \\
&\leq \min\left\{2H, \overline{V}^t_{h+1}(x^t_{h+1}) - \underline{V}^t_{h+1}(x^t_{h+1}) + 2\beta_t \overline{w}^t_h(z^t_h) + 2\beta_t \underline{w}^t_h(z^t_h) + \xi^t_h + \gamma^t_h\right\} \\
&\leq \min\left\{2H, 2\beta_t \overline{w}^t_h(z^t_h) + 2\beta_t \underline{w}^t_h(z^t_h)\right\} + \overline{V}^t_{h+1}(x^t_{h+1}) - \underline{V}^t_{h+1}(x^t_{h+1}) + \xi^t_h + \gamma^t_h \\
&\leq 2\beta_t \min\{1, \overline{w}^t_h(z^t_h)\} + 2\beta_t \min\{1, \underline{w}^t_h(z^t_h)\} + \overline{V}^t_{h+1}(x^t_{h+1}) - \underline{V}^t_{h+1}(x^t_{h+1}) + \xi^t_h + \gamma^t_h,
\end{aligned}$$

where the last inequality is due to the choice of β satisfying $\beta/H \geq 1$. This completes the proof of Lemma 16. □

For bounding II and III, we use induction to prove that III ≥ 0 for every h, that is,

$$\overline{V}^t_h(x^t_h) - V^{*,\nu^t}_h(x^t_h) \geq 0. \tag{D.6}$$

Then the same statement will also hold for II due to the symmetric property. The statement holds for $h = H+1$, where III $= 0$ (since $\overline{V}^t_{H+1} = V^{*,\nu^t}_{H+1} = 0$ by definition). Suppose the statement holds for $h+1$. Let $(a,b) \in \mathcal{A}_1 \times \mathcal{A}_2$ and $z := (x^t_h, a, b)$. If $\overline{Q}^t_h(z) \geq H$, then by definition, III ≥ 0. Suppose $\overline{Q}^t_h(z) < H$, then by definition of $\overline{Q}^t_h(z)$, we have

$$\begin{aligned}
\overline{Q}^t_h(z) - Q^{*,\nu^t}_h(z) &= \left\langle \boldsymbol{\phi}_{\overline{V}^t_{h+1}}(z), \overline{\boldsymbol{\theta}}^t_h - \boldsymbol{\theta}^*_h \right\rangle_{\mathcal{H}} + \left(\mathbb{P}_h(\overline{V}^t_{h+1} - V^{*,\nu^t}_{h+1})\right)(z) + \beta_t \overline{w}^t_h(z) \\
&\geq -\beta_t \overline{w}^t_h(z) + \beta_t \overline{w}^t_h(z) = 0,
\end{aligned} \tag{D.7}$$

where the first inequality holds due to the statement holds for $h+1$, which leads to $\overline{V}^t_{h+1} - V^{*,\nu^t}_{h+1} \geq 0$, and Lemma 3 that gives a bound for $\left\langle \boldsymbol{\phi}_{\overline{V}^t_{h+1}}(z), \overline{\boldsymbol{\theta}}^t_h - \boldsymbol{\theta}^*_h \right\rangle_{\mathcal{H}}$. Next, we have

$$\begin{aligned}
\overline{V}^t_h(x^t_h) - V^{*,\nu^t}_h(x^t_h) &= \mathbb{E}_{(a,b)\sim\sigma^t_h(x^t_h)} \overline{Q}^t_h(x^t_h, a, b) - \mathbb{E}_{a\sim \mathrm{br}(\nu^t_h), b\sim\nu^t_h} Q^{*,\nu^t}_h(x^t_h, a, b) \\
&\geq \mathbb{E}_{a\sim \mathrm{br}(\nu^t_h), b\sim\nu^t_h} \overline{Q}^t_h(x^t_h, a, b) - \mathbb{E}_{a\sim \mathrm{br}(\nu^t_h), b\sim\nu^t_h} Q^{*,\nu^t}_h(x^t_h, a, b) \\
&= \mathbb{E}_{a\sim \mathrm{br}(\nu^t_h), b\sim\nu^t_h} \left[\overline{Q}^t_h(x^t_h, a, b) - Q^{*,\nu^t}_h(x^t_h, a, b)\right] \geq 0,
\end{aligned}$$

where $\nu^t_h := \mathcal{P}_2 \sigma^t_h$ and $\pi^t_h := \mathcal{P}_1 \sigma^t_h$ is the projection of σ^t_h on the first and second coordinate respectively and br is the best response policy of a given distribution. The last inequality holds due to (D.7). Therefore, the statement holds for h, which suggests that the induction holds.

Combining with Eq. (D.1), we arrive at a bound in terms of $\overline{w}^t_h(z^t_h), \underline{w}^t_h(z^t_h)$ and the martingale difference sequences:

$$V^{*,\nu^t}_1(x^t_h) - V^{\pi^t,*}_1(x^t_h) \leq \sum_{h=1}^{H} \left(2\beta_t \min\{1, \overline{w}^t_h(z^t_h)\} + 2\beta_t \min\{1, \underline{w}^t_h(z^t_h)\} + \xi^t_h + \gamma^t_h\right), \tag{D.8}$$

where ν^t is the policy that operates according to ν_h^t at time h and π^t is the sequence of π_h^t accordingly. The rest of the proof follows by bounding $\sum_{t=1}^T \sum_{h=1}^H \min\{1, \overline{w}_h^t(z_h^t)\}, \sum_{t=1}^T \sum_{h=1}^H \min\{1, \underline{w}_h^t(z_h^t)\}$ and the martingale difference sequences.

The bound of $\sum_{t=1}^T \sum_{h=1}^H \min\{1, \overline{w}_h^t(z_h^t)\}$ and $\sum_{t=1}^T \sum_{h=1}^H \min\{1, \underline{w}_h^t(z_h^t)\}$ comes directly from the following lemma 17 which can be simply derived from Lemma 11 in Abbasi-Yadkori et al. [2011] and is an analogue of Lemma E.3 of Yang et al. [2020]:

Lemma 17 (Lemma E.3 of Yang et al. [2020])**.** For any sequence $\{\mathbf{x}_t\}_{t\geq 1}$ taking values on the RKHS $\mathcal{H}$ satisfying $\forall t$, $\|\mathbf{x}_t\|_{\mathcal{H}} \leq L$. Let $I_{\mathcal{H}}$ be the identity operator on $\mathcal{H}$ and $\mathbf{\Lambda}_0 := \lambda \mathbf{I}_{\mathcal{H}}$ the multiplication operator by λ. Furthermore, if we let $\mathbf{\Lambda}_t := \mathbf{\Lambda}_0 + \sum_{i=1}^t \mathbf{x}_i \mathbf{x}_i^\top$ be a positive definite operator from $\mathcal{H}$ to $\mathcal{H}$ and $\mathbf{K}_t \in \mathbb{R}^{t\times t}$ the gram matrix of $\mathcal{H}$ obtained from $\{\mathbf{x}_t\}_{t\geq 1}$. Then the following holds for $\forall t > 0$:

$$\sum_{i=1}^t \min\{1, \mathbf{x}_i^\top \mathbf{\Lambda}_{t-1}^{-1} \mathbf{x}_i\} \leq 2\text{logdet}(\mathbf{I} + \mathbf{K}_t/\lambda).$$

We recall that by Lemma 11, $\overline{w}_h^t = \left[\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top (\overline{\mathbf{\Lambda}}_h^t)^{-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)\right]^{1/2}$, and the same holds for $\underline{w}_h^t$. Let $\mathbf{x}_t = \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z_h^t)$ and $\mathbf{\Lambda}_0 = \lambda \mathbf{I}$ in Lemma 17 and by applying the Cauchy-Schwarz inequality, we have that

$$\sum_{h=1}^H \sum_{t=1}^T \beta_t \min\{1, \overline{w}_h^t(z_h^t)\}, \quad \sum_{h=1}^H \sum_{t=1}^T \beta_t \min\{1, \underline{w}_h^t(z_h^t)\} \leq 2\beta H \cdot \sqrt{T}\sqrt{\Gamma_{\mathbf{K}}(T,\lambda)}, \tag{D.9}$$

where $\beta_t = \beta$ takes values as in Lemma 3 for each t and $\Gamma_{\mathbf{K}}(T, \lambda)$ is defined as the supremum over all V's and z's as in Definition 2. For the martingale difference sequence $\xi_h^t + \gamma_h^t$, as $|\xi_h^t + \gamma_h^t| \leq 4H$, we bound it by Azuma-Hoeffding which gives us with probability at least $1 - \delta$:

$$\sum_{t=1}^T \sum_{h=1}^H \xi_h^t + \gamma_h^t \leq \mathcal{O}\left(H\sqrt{TH} \cdot \log(1/\delta)\right).$$

Combining the above inequality with the bound in (D.9) and (D.8) concludes our proof of Theorem 4.

D.2 Proof of Corollary 6

Due to the selection of t_0, we have

$$V_1^{*,\nu^{t_0}}(x_1) - V_1^{\pi^{t_0},*}(x_1) \leq \overline{V}_1^{t_0}(x_1) - \underline{V}_1^{t_0}(x_1) \leq \frac{1}{T}\sum_{t=1}^T \overline{V}_1^t(x_1) - \underline{V}_1^t(x_1) \leq \sqrt{\frac{\beta^2 H^2 \Gamma_{\mathbf{K}}(T,\lambda)}{T}}, \tag{D.10}$$

where the first inequality holds due to (D.6) and its counterpart for $\underline{V}_1^t$, the second one holds due to the selection of t_0. From (D.10) we can see that by selecting T as what our statement suggests, the ϵ-approximate NE can be guaranteed.

E Proof of Results for `KernelCCE-VTR+`

In this section we give the proof of results in Appendix B. One of the key results of this paper is the following Bernstein self-normalized concentration inequality:

Theorem 18 (Bernstein inequality for vector-valued martingales)**.** Let $\{\mathcal{G}_t\}_{t=1}^\infty$ be a filtration, $\{\mathbf{x}_t, \eta_{t+1}\}_{t\geq 1}$ be a stochastic process so that $\mathbf{x}_t \in \mathbb{R}^d$ is $\mathcal{G}_t$-measurable and $\eta_{t+1} \in \mathbb{R}$ is $\mathcal{G}_{t+1}$-measurable. Fix $R, L, \sigma, \lambda > 0$, $\boldsymbol{\mu}^* \in \mathbb{R}^d$. For $t \geq 1$ we observe $\langle \boldsymbol{\mu}^*, \mathbf{x}_t\rangle + \eta_{t+1}$ and suppose that $\eta_{t+1}, \mathbf{x}_t$ also satisfy

$$|\eta_{t+1}| \leq R, \quad \mathbb{E}[\eta_{t+1}|\mathcal{G}_t] = 0, \quad \mathbb{E}[\eta_{t+1}^2|\mathcal{G}_t] \leq \sigma^2, \quad \|\mathbf{x}_t\|_2 \leq L.$$

Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ we have

$$\left\| \sum_{i=1}^{t} \mathbf{x}_i \eta_{i+1} \right\|_{\mathbf{Z}_t^{-1}} \leq \beta_t, \quad \forall t > 0, \tag{E.1}$$

where for each $t \geq 1$, $\mathbf{Z}_t = \lambda \mathbf{I} + \sum_{i=1}^{t} \mathbf{x}_i \mathbf{x}_i^\top$, and

$$\beta_t = 8\sigma \sqrt{\log \det(\mathbf{I} + \mathbf{K}_t/\lambda) \log(4t^2/\delta)} + 4R \log(4t^2/\delta), \qquad [\mathbf{K}_t]_{i,j} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle_{\mathcal{H}}.$$

Proof. The proof can be derived by following the proof of Theorem 2 in Zhou et al. [2021a]. We only need to replace Lemma 12 in Zhou et al. [2021a] with Lemma 17, then the remaining of the proof goes through the same as Zhou et al. [2021a]. ☐

We first give the proof of Lemma 12.

E.1 Proof of Lemma 12

Proof. We only provide the proof of the max-player, and results of the min-player can be derived similarily. We recall that we define $\overline{\mathbf{y}}_{h,1}^t$ to be the vector of regression targets

$$\left(\overline{V}_{h+1}^1(x_{h+1}^1)/\overline{R}_h^1, \ldots, \overline{V}_{h+1}^{t-1}(x_{h+1}^{t-1})/\overline{R}_h^{t-1} \right)^\top \in \mathbb{R}^{t-1}.$$

Furthermore by Lemma 11 in Section A, we know that $\overline{\boldsymbol{\theta}}_{h,1}^t = \left(\overline{\boldsymbol{\Psi}}_{h,1}^t\right)^\top \left[\overline{\mathbf{K}}_{h,1}^t + \lambda_1 \cdot \mathbf{I}\right]^{-1} \overline{\mathbf{y}}_{h,1}^t$ and $\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z) = \left(\overline{\boldsymbol{\Psi}}_{h,1}^t\right)^\top (\overline{\mathbf{K}}_{h,1}^t + \lambda_1 \cdot \mathbf{I})^{-1} \overline{\boldsymbol{k}}_{h,1}^t(z) + \lambda_1 \cdot (\overline{\boldsymbol{\Lambda}}_{h,1}^t)^{-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)$, which enable us to bound the difference $\left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z), \overline{\boldsymbol{\theta}}_{h,1}^t - \boldsymbol{\theta}_{h,1}^* \right\rangle_{\mathcal{H}}$ as follows:

$$\begin{aligned}
\left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z), \overline{\boldsymbol{\theta}}_{h,1}^t - \boldsymbol{\theta}_h^* \right\rangle_{\mathcal{H}} &= \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top \left(\overline{\boldsymbol{\Psi}}_{h,1}^t\right)^\top \left[\overline{\mathbf{K}}_{h,1}^t + \lambda_1 \cdot \mathbf{I}\right]^{-1} \overline{\mathbf{y}}_{h,1}^t \\
&\quad - \left(\overline{\boldsymbol{\theta}}_h^*\right)^\top \left[\left(\overline{\boldsymbol{\Psi}}_{h,1}^t\right)^\top \left[\overline{\mathbf{K}}_{h,1}^t + \lambda_1 \cdot \mathbf{I}\right]^{-1} \overline{\boldsymbol{k}}_{h,1}^t(z) + \lambda_1 \cdot (\overline{\boldsymbol{\Lambda}}_{h,1}^t)^{-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)\right] \\
&= \underbrace{(\overline{\boldsymbol{k}}_{h,1}^t)^\top \left[\overline{\mathbf{K}}_{h,1}^t + \lambda_1 \cdot \mathbf{I}\right]^{-1} \left[\overline{\mathbf{y}}_{h,1}^t - \overline{\boldsymbol{\Psi}}_{h,1}^t \boldsymbol{\theta}_h^*\right]}_{\mathrm{I}_1} - \underbrace{\lambda_1 \cdot \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top (\overline{\boldsymbol{\Lambda}}_{h,1}^t)^{-1} \boldsymbol{\theta}_h^*}_{\mathrm{I}_2}.
\end{aligned}$$

For bounding I_2, we apply the Cauchy-Schwarz inequality and have

$$\begin{aligned}
\lambda_1 \cdot \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top (\overline{\boldsymbol{\Lambda}}_h^t)^{-1} \boldsymbol{\theta}_h^* &\leq \left\| \lambda_1 \cdot \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top (\overline{\boldsymbol{\Lambda}}_{h,1}^t)^{-1} \right\|_{\mathcal{H}} \cdot \|\boldsymbol{\theta}_h^*\|_{\mathcal{H}} \\
&\overset{(a)}{\leq} B \cdot \sqrt{\lambda_1 \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top (\overline{\boldsymbol{\Lambda}}_{h,1}^t)^{-1} \lambda_1 (\overline{\boldsymbol{\Lambda}}_{h,1}^t)^{-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)} \overset{(b)}{\leq} \sqrt{\lambda_1} B \cdot \overline{w}_{h,1}^t(z),
\end{aligned}$$

where (a) is due to the assumption that $\|\boldsymbol{\theta}_h^*\|_{\mathcal{H}} \leq B$, and (b) is by the definition of $\overline{w}_h^t$ and the fact that $\left(\overline{\boldsymbol{\Lambda}}_{h,1}^t\right)^{-1}$ is a self-adjoint mapping on the RKHS $\mathcal{H}$ satisfying $\left\| \left(\overline{\boldsymbol{\Lambda}}_{h,1}^t\right)^{-1} \right\|_{op} \leq \frac{1}{\lambda_1}$.

For bounding I_1, we observe the following equality:

$$\begin{aligned}
&(\overline{\boldsymbol{k}}_{h,1}^t)^\top \left[\overline{\mathbf{K}}_{h,1}^t + \lambda_1 \cdot \mathbf{I}\right]^{-1} \left[\overline{\mathbf{y}}_{h,1}^t - \overline{\boldsymbol{\Psi}}_{h,1}^t \boldsymbol{\theta}_h^*\right] \\
&= \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top (\overline{\boldsymbol{\Lambda}}_{h,1}^t)^{-1} (\overline{\boldsymbol{\Psi}}_{h,1}^t)^\top \left[\overline{\mathbf{y}}_{h,1}^t - \boldsymbol{\Psi}_{h,1}^t \boldsymbol{\theta}_h^*\right] \\
&= \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top (\overline{\boldsymbol{\Lambda}}_{h,1}^t)^{-1} \sum_{\tau=1}^{t-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}(z_h^\tau) \left[\overline{V}_{h+1}^\tau(x_{h+1}^\tau) - (\mathbb{P}_h \overline{V}_{h+1}^\tau)(z_h^\tau)\right] / \left(\overline{R}_h^\tau\right)^2.
\end{aligned}$$

Again by applying the Cauchy-Schwarz inequality, we bound the RHS of the above equality as

$$|\mathrm{I}_1| \le \left\|\boldsymbol{\phi}_{\overline{V}^t_{h+1}}(z)\right\|_{\left(\overline{\boldsymbol{\Lambda}}^t_{h,1}\right)^{-1}} \cdot \left\|\sum_{\tau=1}^{t-1} \boldsymbol{\phi}_{\overline{V}^\tau_{h+1}}(z^\tau_h)\left[\overline{V}^\tau_{h+1}(x^\tau_{h+1}) - (\mathbb{P}_h\overline{V}^\tau_{h+1})(z^\tau_h)\right] / \left(\overline{R}^\tau_h\right)^2\right\|_{\left(\overline{\boldsymbol{\Lambda}}^t_{h,1}\right)^{-1}}.$$

We note that for the given h considered in Lemma 12. If we define $\{\mathcal{F}_t\}_{t\ge 0}$ as the σ-algebra generated by all data before iteration $t-1$ along with all data before time h at iteration t, then

$$\eta_{\tau+1} := \left(\overline{V}^\tau_{h+1}(x^\tau_{h+1}) - (\mathbb{P}_h\overline{V}^\tau_{h+1})(z^\tau_h)\right)/\overline{R}^\tau_h \in \mathcal{F}_{\tau+1} \tag{E.2}$$

is a mean zero random variable with respect to filtration $\mathcal{F}_\tau$. By the choice of $\overline{R}_h$ such that $\overline{R}_h \ge \alpha_t$, we can bound the absolute value of $\eta_{\tau+1}$ by $\left|\left(\overline{V}^\tau_{h+1}(x^\tau_{h+1}) - (\mathbb{P}_h\overline{V}^\tau_{h+1})(z^\tau_h)\right)/\overline{R}^\tau_h\right| \le 2H/\alpha_\tau$. We take $\eta_{\tau+1}$ as in Eq. (E.2) in Theorem 18 and $\{\mathbf{x}_t\}_{t\ge 1}$ is $\left\{\boldsymbol{\phi}_{\overline{V}^t_{h+1}}(z^t_h)/\overline{R}^t_h\right\}_{t\ge 1} \in \mathcal{F}_\tau$. Then by directly utilizing Theorem 18, we have that the following inequality holds with probability at least $1-\delta/H$,

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{t-1} \boldsymbol{\phi}_{\overline{V}^\tau_{h+1}}(z^\tau_h)\left[\overline{V}^t_{h+1}(x^\tau_{h+1}) - (\mathbb{P}_h\overline{V}^t_{h+1})(z^\tau_h)\right]/\overline{R}^\tau_h\right\|^2_{\left(\overline{\boldsymbol{\Lambda}}^t_{h,1}\right)^{-1}}\\
&\le 16H/\alpha\sqrt{\log\det(\mathbf{I}+\mathbf{K}^{(1)}_t/\lambda_1)\log(4t^2H/\delta)} + 8H/\alpha\log(4t^2H/\delta)\\
&\le 16H/\alpha\sqrt{\log\det(\mathbf{I}+\mathbf{K}_t/\left(\lambda_1(\alpha_t)^2\right))\log(4t^2H/\delta)} + 8H/\alpha\log(4t^2H/\delta)\\
&\le 16H/\alpha\sqrt{\Gamma_{\mathbf{K}}(T,\lambda_1(\alpha_t)^2)}\sqrt{\log(4t^2H/\delta)} + 8H/\alpha\log(4t^2H/\delta),
\end{aligned}$$

where $\mathbf{K}^{(1)}_t$ is the gram matrix for $\{\mathbf{x}_\tau\}_{\tau\in[t-1]} = \left\{\boldsymbol{\phi}_{\overline{V}^t_{h+1}}(z^\tau_h)/\overline{R}^\tau_h\right\}_{\tau\in[t-1]}$, $\mathbf{K}_t$ is the gram matrix for $\left\{\boldsymbol{\phi}_{\overline{V}^t_{h+1}}(z^\tau_h)\right\}_{\tau\in[t-1]}$.

On the other hand, when estimating $\mathbb{P}_h\left(\overline{V}^t_{h+1}\right)^2$, we have the following result regarding $\overline{\boldsymbol{\theta}}^t_{h,2}$ holds with probability at least $1-\delta/H$:

$$\left|\left\langle \boldsymbol{\phi}_{(\overline{V}^t_{h+1})^2}(z), \overline{\boldsymbol{\theta}}^t_{h,2} - \boldsymbol{\theta}^*_h\right\rangle_{\mathcal{H}}\right| \le 16H^2\sqrt{\Gamma_{\mathbf{K}}(T,\lambda_2/H^2)}\sqrt{\log(4t^2H/\delta)} + 8H^2\log(4t^2H/\delta) + \sqrt{\lambda_2}\cdot B.$$

By letting

$$\beta^{(1)}_t = 16H/\alpha\sqrt{\Gamma_{\mathbf{K}}(T,\lambda_1(\alpha_t)^2)}\sqrt{\log(4t^2H/\delta)} + 8H/\alpha\log(4t^2H/\delta) + \sqrt{\lambda_1}\cdot B,$$

and

$$\beta^{(2)}_t = 16H^2\sqrt{\Gamma_{\mathbf{K}}(T,\lambda_2/H^2)}\sqrt{\log(4t^2H/\delta)} + 8H^2\log(4t^2H/\delta) + \sqrt{\lambda_2}\cdot B,$$

we have that from the above theoretical derivation and by taking union bounds over $h\in[H]$,

$$\left|\boldsymbol{\phi}_{\overline{V}^\tau_{h+1}}(z)^\top(\boldsymbol{\theta}^*_h - \overline{\boldsymbol{\theta}}^t_{h,1})\right| \le \beta^{(1)}_t\cdot\overline{w}^t_{h,1}(z),$$

and

$$\left|\boldsymbol{\phi}_{(\overline{V}^\tau_{h+1})^2}(z)^\top(\boldsymbol{\theta}^*_h - \overline{\boldsymbol{\theta}}^t_{h,2})\right| \le \beta^{(2)}_t\cdot\overline{w}^t_{h,2}(z), \tag{E.3}$$

with probability at least $1-2\delta$. This concludes our proof. □

From Lemma 12, we can prove that $\overline{R}^t_h$ is an upper bound of the actual variance of $\overline{V}^t_{h+1}$.

Lemma 19. Following the setting of Lemma 12 and assume that event $\mathcal{E}$ occurs, the following holds for any $(t,h)\in T\times H$:

$$\left|\mathbb{V}^{\text{est}}\overline{V}^t_{h+1}(z^t_h) - \mathbb{V}\overline{V}^t_{h+1}(z^t_h)\right| \le \min\left\{H^2, \beta^{(2)}_t\overline{w}^t_{h,2}\right\} + \min\left\{H^2, 2H\beta^{(1)}_t\overline{w}^t_{h,1}\right\}.$$

Proof. By the triangle inequality we have that

$$
\begin{aligned}
&|\mathbb{V}^{\text{est}}\overline{V}^t_{h+1}(z^t_h) - \mathbb{V}\overline{V}^t_{h+1}(z^t_h)| \\
&\leq \left|\langle \boldsymbol{\phi}_{(\overline{V}^t_{h+1})^2}(z^t_h), \boldsymbol{\theta}^*_h\rangle_{\mathcal{H}} - \left[\langle \boldsymbol{\phi}_{(\overline{V}^t_{h+1})^2}(z^t_h), \overline{\boldsymbol{\theta}}^t_{h,2}\rangle_{\mathcal{H}}\right]_{[0,H^2]}\right| \\
&\quad + \left|(\langle \boldsymbol{\phi}_{\overline{V}^t_{h+1}}(z^t_h), \boldsymbol{\theta}^*_h\rangle_{\mathcal{H}})^2 - \left[\langle \boldsymbol{\phi}_{\overline{V}^t_{h+1}}(z^t_h), \overline{\boldsymbol{\theta}}^t_{h,1}\rangle_{\mathcal{H}}\right]^2_{[-H,H]}\right| \\
&\leq \min\left\{H^2, \left|\langle \boldsymbol{\phi}_{(\overline{V}^t_{h+1})^2}(z^t_h), \boldsymbol{\theta}^*_h - \overline{\boldsymbol{\theta}}^t_{h,2}\rangle_{\mathcal{H}}\right|\right\} + \min\left\{H^2, 2H\left|\langle \boldsymbol{\phi}_{\overline{V}^t_{h+1}}(z^t_h), \boldsymbol{\theta}^*_h - \overline{\boldsymbol{\theta}}^t_{h,1}\rangle_{\mathcal{H}}\right|\right\} \\
&\leq \min\left\{H^2, \beta^{(2)}_t \overline{w}^t_{h,2}\right\} + \min\left\{H^2, 2H\beta^{(1)}_t \overline{w}^t_{h,1}\right\},
\end{aligned}
\tag{E.4}
$$

where the last inequality directly comes from Lemma 12. □

Lemma 20 (Fine-tuned bound)**.** Assuming that for any $h \in [H]$, $\|\boldsymbol{\theta}^*_h\|_{\mathcal{H}} \leq B$. Let β_t satisfy

$$
\beta_t \geq 16\sqrt{\Gamma_{\mathbf{K}}(T, \lambda_1(\alpha_t)^2)}\sqrt{\log(4t^2H/\delta)} + 8H/\alpha \log(4t^2H/\delta) + \sqrt{\lambda_1} \cdot B. \tag{E.5}
$$

Then on the event defined in Lemma 12, there exists an event $\mathcal{E}_1$ such that the following holds with probability at least $1 - \delta$ for any $(t, h) \in [T] \times [H]$ and any $(x, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{A}$:

$$
\left|\langle \boldsymbol{\phi}_{\overline{V}^t_{h_1}}(z^t_h), \boldsymbol{\theta}^*_h - \overline{\boldsymbol{\theta}}^t_{h,1}\rangle_{\mathcal{H}}\right| \leq \beta_t \cdot \overline{w}^k_{h,1}(z^k_h).
$$

Proof. From the definition of $\overline{R}_h$ and $\overline{E}^t_h$, we know that

$$
\overline{R}^t_h \geq \mathbb{V}^{\text{est}}\overline{V}^t_{h+1}(z^t_h) + \min\left\{H^2, \beta^{(2)}_t \overline{w}^t_{h,2}\right\} + \min\left\{H^2, 2H\beta^{(1)}_t \overline{w}^t_{h,1}\right\}.
$$

Combining with the result in Lemma 19 where we bound the absolute difference between the estimated variance and the true variance in Eq. (E.4), we have that on the event $\mathcal{E}$ defined in Lemma 12:

$$
\overline{R}^t_h \geq \mathbb{V}\overline{V}^t_{h+1}(z^t_h).
$$

We derive a fine-tuned bound on the variance of η_{t+1} defined in Eq. (E.2) that on event $\mathcal{E}$:

$$
\mathbb{E}\left[\eta^2_{t+1} \mid \mathcal{F}_t\right] = \mathbb{V}\overline{V}^t_{h+1}(z^t_h) / \left(\overline{R}^t_h\right)^2 \leq 1.
$$

The rest of the proof follows by a direct application of Theorem 18. □

Lemma 21. On the event $\mathcal{E} \cap \mathcal{E}_1$, there exists an event $\mathcal{E}_2$ such that $\mathcal{E}_2$ holds with probability at least $1 - \delta$, we have

$$
\begin{aligned}
\sum_{t=1}^T \sum_{h=1}^H \left(\overline{R}^t_h\right)^2 \leq{} & HT\alpha^2 + 3(H^2T + H^3\log(1/\delta)) + 4H\sum_{t=1}^T \sum_{h=1}^H \mathbb{P}_h[\overline{V}^t_{h+1} - V^{\boldsymbol{\mu}^t}_{h+1}] \\
& + 2\beta^{(2)}_T\sqrt{TH} \cdot \sqrt{2H\Gamma_{\mathbf{K}}(T, \lambda_2/H^2)} \\
& + 7\beta^{(1)}_T H^2\sqrt{TH} \cdot \sqrt{2H\Gamma_{\mathbf{K}}(T, \lambda_1/\alpha^2)}.
\end{aligned}
$$

Proof. First by considering the definition of $\overline{R}_h$, we know that

$$
\begin{aligned}
&\sum_{t=1}^T \sum_{h=1}^H \left(\overline{R}^t_h\right)^2 \\
&\leq \sum_{t=1}^T \sum_{h=1}^H \left(\alpha^2_t + \mathbb{V}^{\text{est}}\overline{V}^t_{h+1}(z^t_h) + \overline{E}^t_h\right) \\
&= \sum_{t=1}^T \sum_{h=1}^H \left(\alpha^2_t + \mathbb{V}^{\text{est}}\overline{V}^t_{h+1}(z^t_h) + \min\left\{H^2, \beta^{(2)}_t \overline{w}^t_{h,2}\right\} + \min\left\{H^2, 2H\beta^{(1)}_t \overline{w}^t_{h,1}\right\}\right)
\end{aligned}
$$

$$\leq HT\alpha^2 + \sum_{t=1}^{T}\sum_{h=1}^{H}\left(\mathbb{V}\overline{V}_{h+1}^{t}(z_h^t) + 2\min\left\{H^2, \beta_t^{(2)}\overline{w}_{h,2}^t\right\} + 2\min\left\{H^2, 2H\beta_t^{(1)}\overline{w}_{h,1}^t\right\}\right)$$

$$\leq HT\alpha^2 + \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H}\left[\mathbb{V}\overline{V}_{h+1}^{t}(z_h^t) - \mathbb{V}\overline{V}_{h+1}^{\pi_t}(z_h^t)\right]}_{\text{I}} + \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{V}\overline{V}_{h+1}^{\pi_t}(z_h^t)}_{\text{II}}$$

$$+ \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H}2\min\left\{H^2, \beta_t^{(2)}\overline{w}_{h,2}^t\right\} + \sum_{t=1}^{T}\sum_{h=1}^{H}2\min\left\{H^2, 2H\beta_t^{(1)}\overline{w}_{h,1}^t\right\}}_{\text{III}}.$$

The rest of the proof for bounding $\text{I}, \text{II}, \text{III}$ goes the same as in the proof of Lemma A.6 in Chen et al. [2022], except that we replace Lemma B.4 in Chen et al. [2022] with Lemma 17. ☐

Proof of Theorem 13. The first part of the proof follows almost the same as in the proof of Theorem 4 by replacing $\overline{w}_h^t$ with $\overline{w}_{h,1}^t$ and $\underline{w}_h^t$ with $\underline{w}_{h,1}^t$, except that now we have $\beta\overline{R}_h^t \geq 2H$ so that we have (E.6) instead of (D.8).

$$V_1^{*,\nu^t}(x_1^t) - V_1^{\pi^t,*}(x_1^t)$$
$$\leq \sum_{h=1}^{H}\left(4\beta_t\overline{R}_h^t\min\{1, \overline{w}_{h,1}^t(z_h^t)/\overline{R}_h^t\} + 4\beta_t\underline{R}_h^t\min\{1, \underline{w}_{h,1}^t(z_h^t)/\overline{R}_h^t\} + \xi_h^t + \gamma_h^t\right), \tag{E.6}$$

Similarly, for any $1 \leq h' \leq H$, we have

$$\overline{V}_{h'}^t(x_{h'}^t) - \underline{V}_{h'}^t(x_{h'}^t)$$
$$\leq \sum_{h=1}^{H}\left(2\beta_t\overline{R}_h^t\min\{1, \overline{w}_{h,1}^t(z_h^t)/\overline{R}_h^t\} + 2\beta_t\underline{R}_h^t\min\{1, \underline{w}_{h,1}^t(z_h^t)/\overline{R}_h^t\} + \xi_h^t + \gamma_h^t\right).$$

Applying Azuma-Hoeffding inequality onto (E.7), we have with probability at least $1-\delta$,

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{P}_h[\overline{V}_{h+1}^t - \underline{V}_{h+1}^t](x_h^t, a_h^t, b_h^t)$$
$$\leq \sum_{t=1}^{T}\sum_{h=1}^{H}\left(2\beta_t\overline{R}_h^t\min\{1, \overline{w}_{h,1}^t(z_h^t)/\overline{R}_h^t\} + 2\beta_t\underline{R}_h^t\min\{1, \underline{w}_{h,1}^t(z_h^t)/\overline{R}_h^t\} + \xi_h^t + \gamma_h^t\right) + \sum_{t=1}^{T}\sum_{h=1}^{H}\xi_h^t. \tag{E.7}$$

Then we estimate the two summation terms $\sum_{t=1}^{T}\sum_{h=1}^{H}\overline{R}_h^t\min\{1, \overline{w}_{h,1}^t(z_h^t)/\overline{R}_h^t\}$ and $\sum_{t=1}^{T}\sum_{h=1}^{H}\overline{R}_h^t\min\{1, \underline{w}_{h,1}^t(z_h^t)/\overline{R}_h^t\}$ separately. By definitions in Lemma 11, Section A, we know that

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\overline{R}_h^t\min\{1, \overline{w}_{h,1}^t(z_h^t)/\overline{R}_h^t\}$$
$$= \sum_{t=1}^{T}\sum_{h=1}^{H}\overline{R}_h^t\min\left\{1, \left[\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top\overline{\boldsymbol{\Lambda}}_{h,1}^t\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)\right]^{1/2}/\overline{R}_h^t\right\}$$
$$\leq \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\overline{R}_h^t\right)^2}\sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\min\left\{1, \left[\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top\overline{\boldsymbol{\Lambda}}_{h,1}^t\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)\right]/\overline{R}_h^t\right\}}$$
$$\leq \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\overline{R}_h^t\right)^2}\cdot\sqrt{2H\Gamma_{\mathbf{K}}(T, \lambda_1\alpha^2)}.$$

Similarly,

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\underline{R}_h^t \min\{1, \underline{w}_{h,1}^t(z_h^t)/\underline{R}_h^t\} \le \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\underline{R}_h^t\right)^2}\cdot\sqrt{2H\Gamma_{\mathbf{K}}(T,\lambda_1\alpha^2)}.$$

By Lemma 21, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\overline{R}_h^t\right)^2+\left(\underline{R}_h^t\right)^2\\
&= O\bigg(HT\alpha^2+H^2T+H^3\log(1/\delta)+H\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{P}_h[\overline{V}_{h+1}^t-\underline{V}_{h+1}^t]\\
&\quad+\beta_T^{(2)}\sqrt{TH}\sqrt{H\Gamma_{\mathbf{K}}(T,\lambda_2/H^2)}+\beta_t^{(1)}H^2\sqrt{TH}\sqrt{H\Gamma_{\mathbf{K}}(T,\lambda_1/\alpha^2)}\bigg)\\
&\le O\bigg(HT\alpha^2+H^2T+H^3\log(1/\delta))\\
&\quad+H^2\beta_t\sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\overline{R}_h^t\right)^2+\left(\underline{R}_h^t\right)^2}\cdot\sqrt{H\cdot\Gamma_{\mathbf{K}}(T,\lambda_1\alpha^2)}+H^3\sqrt{HT\log(H/\delta)}\\
&\quad+\beta_t^{(2)}\sqrt{TH}\sqrt{H\Gamma_{\mathbf{K}}(T,\lambda_2/H^2)}+\beta_t^{(1)}H^2\sqrt{TH}\sqrt{H\Gamma_{\mathbf{K}}(T,\lambda_1\alpha^2)}\bigg),
\end{aligned} \tag{E.8}$$

where the inequality holds due to Cauchy-Schwarz inequality. Next, by taking

$$\alpha = H/\sqrt{\Gamma_{\mathbf{K}}(T,1/B^2)},\ \lambda_2 = H^2/B^2,\ \lambda_1 = 1/(\alpha^2B^2),$$

we have

$$\beta_t^{(1)} = 16H/\alpha\sqrt{\Gamma_{\mathbf{K}}(T,\lambda_1(\alpha)^2)}\sqrt{\log(4t^2H/\delta)}+8H/\alpha\log(4t^2H/\delta)+\sqrt{\lambda_1}\cdot B = \widetilde{O}(\Gamma_{\mathbf{K}}(T,H^2/B^2)),$$

$$\beta_t^{(2)} = 16H^2\sqrt{\Gamma_{\mathbf{K}}(T,\lambda_2/H^2)}\sqrt{\log(4t^2H/\delta)}+8H^2\log(4t^2H/\delta)+\sqrt{\lambda_2}\cdot B = \widetilde{O}(H^2),$$

$$\beta_t = 16\sqrt{\Gamma_{\mathbf{K}}(T,\lambda_1(\alpha)^2)}\sqrt{\log(4t^2H/\delta)}+8H/\alpha\log(4t^2H/\delta)+\sqrt{\lambda_1}\cdot B = \widetilde{O}(\sqrt{\Gamma_{\mathbf{K}}(T,H^2/B^2)}).$$

For simplicity, let $d_{\text{eff}} := \Gamma_{\mathbf{K}}(T,H^2/B^2)$, then by (E.8) we have

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\overline{R}_h^t\right)^2+\left(\underline{R}_h^t\right)^2\\
&\le \widetilde{O}\left(\sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\overline{R}_h^t\right)^2+\left(\underline{R}_h^t\right)^2}H^{5/2}d_{\text{eff}}+H^3d_{\text{eff}}^{3/2}T^{1/2}+H^{7/2}T^{1/2}+H^3T/d_{\text{eff}}+H^2T\right).
\end{aligned}$$

With the fact that $x \le a\sqrt{x}+b$ leads to $x = O(a^2+b)$, we have

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\overline{R}_h^t\right)^2+\left(\underline{R}_h^t\right)^2 = \widetilde{O}(d_{\text{eff}}^2H^5+H^3d_{\text{eff}}^{3/2}T^{1/2}+H^{7/2}T^{1/2}+H^3T/d_{\text{eff}}+H^2T). \tag{E.9}$$

Finally, substituting (E.9) into (E.6) and bound the summation of ξ_h^t, γ_h^t by Azuma-Hoeffding inequality, we have

$$\begin{aligned}
&\sum_{t=1}^{T}V_1^{*,\nu^t}(x_h^t)-V_1^{\pi^t,*}(x_h^t)\\
&\le \widetilde{O}\left(\beta_t\sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\overline{R}_h^t\right)^2+\left(\underline{R}_h^t\right)^2}\cdot\sqrt{H\cdot d_{\text{eff}}}+H\sqrt{2HT\log(H/\delta)}\right)\\
&= \widetilde{O}\left(d_{\text{eff}}^2H^3+d_{\text{eff}}^{1.75}H^2T^{0.25}+d_{\text{eff}}H^{2.25}T^{0.25}+\sqrt{d_{\text{eff}}}H^2\sqrt{T}+d_{\text{eff}}H^{1.5}\sqrt{T}\right)\\
&= \widetilde{O}\left(d_{\text{eff}}^2H^3+\sqrt{d_{\text{eff}}H^4+d_{\text{eff}}^2H^3}\sqrt{T}+\left(d_{\text{eff}}^7H^8+d_{\text{eff}}^4H^9\right)^{1/4}T^{1/4}\right).
\end{aligned}$$

This completes the proof of the theorem. □

F Proof of Results for `KernelCCE-VTR` with Misspecification

In this section we prove Theorem 10.

Lemma 22. Assuming that for any $h \in [H]$, $\|\boldsymbol{\theta}_h^*\|_{\mathcal{H}} \le B$. Let $\lambda = 1 + 1/T$ and β_t satisfies

$$\left(\frac{\beta_t}{H}\right)^2 \ge 3\Gamma_{\mathbf{K}}(T,\lambda) + 3 + 6 \cdot \log\left(\frac{1}{\delta}\right) + 3\lambda\left(\frac{B}{H}\right)^2 + 3\iota_{\text{mis}}^2 t. \tag{F.1}$$

Then for any $\delta > 0$, with probability at least $1-\delta$ the following holds for any $(t,h) \in [T] \times [H]$ and any $z \in \mathcal{Z}$:

$$\left|\left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z), \overline{\boldsymbol{\theta}}_h^t \right\rangle_{\mathcal{H}} - \mathbb{P}_h \overline{V}_{h+1}^t(z)\right| \le \beta_t \cdot \overline{w}_h^t(z) + H \cdot \iota_{\text{mis}},$$
$$\left|\left\langle \boldsymbol{\phi}_{\underline{V}_{h+1}^t}(z), \underline{\boldsymbol{\theta}}_h^t \right\rangle_{\mathcal{H}} - \mathbb{P}_h \underline{V}_{h+1}^t(z)\right| \le \beta_t \cdot \underline{w}_h^t(z) + H \cdot \iota_{\text{mis}}.$$

The proof of Theorem 10 shares similar techniques with the proof of Theorem 4, except that we need Lemma 22 instead of Lemma 3. We detail the whole proof for completeness.

We recall that the duality gap is defined as $\sum_{t=1}^T V_1^{*,\nu^t}(x_1^t) - V_1^{\pi^t,*}(x_1^t)$. As can be seen in our Algorithm, we maintain an optimistic estimate of $V_h^{*,v^t}(\cdot)$ as $\overline{V}_{h+1}^t(\cdot)$ and a pessimistic estimate of $V_h^{\pi^t,*}(\cdot)$ as $\underline{V}_{h+1}^t(\cdot)$. Hence the term $\overline{V}_{h+1}^t(x_h^t) - \underline{V}_{h+1}^t(x_h^t)$ is approximately the upper bound of the duality gap. We write the decomposition formally as below:

$$V_h^{*,\nu^t}(x_h^t) - V_h^{\pi^t,*}(x_h^t) = \underbrace{\overline{V}_h^t(x_h^t) - \underline{V}_h^t(x_h^t)}_{\text{I}} - \underbrace{\left(V_h^{\pi^t,*}(x_h^t) - \underline{V}_h^t(x_h^t)\right)}_{\text{II}} - \underbrace{\left(\overline{V}_h^t(x_h^t) - V_h^{*,\nu^t}(x_h^t)\right)}_{\text{III}}. \tag{F.2}$$

We use $\overline{\delta}_h^t$ to denote an important quantity $\left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z_h^t), \boldsymbol{\theta}_h^* - \overline{\boldsymbol{\theta}}_h^t \right\rangle_{\mathcal{H}}$ (and $\left\langle \boldsymbol{\phi}_{\underline{V}_{h+1}^t}(z_h^t), \boldsymbol{\theta}_h^* - \underline{\boldsymbol{\theta}}_h^t \right\rangle_{\mathcal{H}}$) in estimating the duality gap. In the rest of the proof we aim to show that all of the above three terms can be bounded by a quantity related to $\overline{\delta}_h^t$ ($\underline{\delta}_h^t$) and a stochastic random variable that forms a martingale difference sequence when considering for all $h \in [H], t \in [T]$.

For bounding term I, we first define two sequences of zero mean variables:

$$\gamma_h^t := \overline{Q}_h^t(x_h^t, a_h^t, b_h^t) - \underline{Q}_h^t(x_h^t, a_h^t, b_h^t) - \mathbb{E}_{(a,b)}\left[\overline{Q}_h^t(x,a,b) - \underline{Q}_h^t(x,a,b)\right],$$
$$\xi_h^t := \left(\mathbb{P}_h(\overline{V}_{h+1}^t - \underline{V}_{h+1}^t)\right)(x_h^t, a_h^t, b_h^t) - \left(\overline{V}_{h+1}^t(x_{h+1}^t) - \underline{V}_{h+1}^t(x_{h+1}^t)\right), \tag{F.3}$$

where γ_h^t depicts the stochastic error with respect to the policy and ξ_h^t depicts the stochastic error with respect to the transition. We refer the readers to the proof of Lemma 23 for detailed explanations on these two error term. Given the above definition, we have the following Lemma 23.

Lemma 23. Under the settings of Lemma 22, we have the following recursive bound for $\forall h \in [H]$:

$$\overline{V}_h^t(x_h^t) - \underline{V}_h^t(x_h^t)$$
$$\le \overline{V}_{h+1}^t(x_{h+1}^t) - \underline{V}_{h+1}^t(x_{h+1}^t) + 2\beta_t \min\{1, \overline{w}_h^t(x_h^t)\} + 2\beta_t \min\{1, \underline{w}_h^t(x_h^t)\} + 2H \cdot \iota_{\text{mis}} + \xi_h^t + \gamma_h^t. \tag{F.4}$$

Proof of Lemma 23. Follows the same derivative as in Lemma 16 as it still holds that $\beta_t \ge H$ in Lemma 22 □

For bounding II and III, we observe that the gap between two consecutive points of the values of term II can be decomposed as

$$\begin{aligned}
&\left(\overline{V}^t_h(x^t_h) - V^{*,\nu^t_h}_h(x^t_h)\right) - \left(\overline{V}^t_{h+1}(x^t_{h+1}) - V^{*,\nu^t_h}_{h+1}(x^t_{h+1})\right) \\
&= \underbrace{\left(\overline{V}^t_h(x^t_h) - V^{*,\nu^t_h}_h(x^t_h)\right) - \left(\overline{Q}^t_h(z^t_h) - Q^{*,\nu^t_h}_h(z^t_h)\right)}_{\mathrm{I}_1} \\
&\quad + \underbrace{\left(\overline{Q}^t_h(z^t_h) - Q^{*,\nu^t_h}_h(z^t_h)\right) - \left(\mathbb{P}_h(\overline{V}^t_{h+1} - V^{*,\nu^t_h}_{h+1})\right)(z^t_h)}_{\mathrm{I}_2} \\
&\quad + \underbrace{\left(\mathbb{P}_h(\overline{V}^t_{h+1} - V^{*,\nu^t_h}_{h+1})\right)(z^t_h) - \left(\overline{V}^t_{h+1}(x^t_{h+1}) - V^{*,\nu^t_h}_{h+1}(x^t_{h+1})\right)}_{\mathrm{I}_3}.
\end{aligned} \tag{F.5}$$

In the two-player game setting II and III are symmetric, so the terms $\mathrm{I}_1, \mathrm{I}_2, \mathrm{I}_3$ has their correspondence denoted as $\mathrm{I}'_1, \mathrm{I}'_2, \mathrm{I}'_3$ separately. Utilizing the bound in Lemma 3, we can bound $|\mathrm{I}_2|$ and $|\mathrm{I}'_2|$ by $2\beta_t \min\{1, \overline{w}^t_h(z^t_h)\} + H \cdot \iota_{\mathrm{mis}}$ and $2\beta_t \min\{1, \underline{w}^t_h(z^t_h)\} + H \cdot \iota_{\mathrm{mis}}$ respectively via similar techniques as in the Proof of Lemma 16 and the fact that $\left|\overline{Q}^t_h(z^t_h) - Q^{*,\nu^t_h}_h(z^t_h)\right| \le 2H$, $\left|\underline{Q}^t_h(z^t_h) - Q^{\pi^t_h,*}_h(z^t_h)\right| \le 2H$ as well as the following inequalities:

$$\begin{aligned}
&\left|\left(\overline{Q}^t_h(z^t_h) - Q^{*,\nu^t}_h(z^t_h)\right) - \left(\mathbb{P}_h(\overline{V}^t_{h+1} - V^{*,\nu^t}_{h+1})\right)(z^t_h)\right| \le 2\beta_t \overline{w}^t_h(z^t_h) + H \cdot \iota_{\mathrm{mis}}, \\
&\left|\left(\underline{Q}^t_h(z^t_h) - Q^{\pi^t,*}_h(z^t_h)\right) - \left(\mathbb{P}_h(\underline{V}^t_{h+1} - V^{\pi^t,*}_{h+1})\right)(z^t_h)\right| \le 2\beta_t \underline{w}^t_h(z^t_h) + H \cdot \iota_{\mathrm{mis}}.
\end{aligned} \tag{F.6}$$

For I_3 and I'_3, we note that both terms are stochastic noises with mean zero, where the stochasticity lies in the transition probability. We denote them as $\alpha^1_{h,t}$ and $\alpha^2_{h,t}$ respectively.

Finally for bounding I_1 and I'_1, we utilize the properties of the CCE. (For simplicity we only prove the bound for I_1, it is trivial to generalize to the bound for I_1)

$$\begin{aligned}
\overline{V}^t_h(x^t_h) - V^{*,\nu^t_h}_h(x^t_h) &= \mathbb{E}_{(a,b)\sim\sigma^t_h(x^t_h)} \overline{Q}^t_h(x^t_h, a, b) - \mathbb{E}_{a\sim \mathrm{br}(\nu^t_h), b\sim\nu^t_h} Q^{*,\nu^t_h}_h(x^t_h, a, b), \\
&\ge \mathbb{E}_{a\sim \mathrm{br}(\nu^t_h), b\sim\nu^t_h} \overline{Q}^t_h(x^t_h, a, b) - \mathbb{E}_{a\sim \mathrm{br}(\nu^t_h), b\sim\nu^t_h} Q^{*,\nu^t_h}_h(x^t_h, a, b) \\
&= \mathbb{E}_{a\sim \mathrm{br}(\nu^t_h), b\sim\nu^t_h} \left[\overline{Q}^t_h(x^t_h, a, b) - Q^{*,\nu^t_h}_h(x^t_h, a, b)\right],
\end{aligned}$$

where $\nu^t_h := \mathcal{P}_2\sigma^t_h$ and $\pi^t_h := \mathcal{P}_1\sigma^t_h$ is the projection of σ^t_h on the first and second coordinate respectively and br is the best response policy of a given distribution. Defining

$$\begin{aligned}
\zeta^1_{h,t} &:= \mathbb{E}_{a\sim \mathrm{br}(\nu^t_h), b\sim\nu^t_h} \left[\overline{Q}^t_h(x^t_h, a, b) - Q^{*,\nu^t_h}_h(x^t_h, a, b)\right] - \left[\overline{Q}^t_h(x^t_h, a^t_h, b^t_h) - Q^{*,\nu^t_h}_h(x^t_h, a^t_h, b^t_h)\right], \\
\zeta^2_{h,t} &:= \mathbb{E}_{a\sim \pi^t_h, b\sim \mathrm{br}(\pi^t_h)} \left[Q^{\pi^t_h,*}_h(x^t_h, a, b) - \underline{Q}^t_h(x^t_h, a, b)\right] - \left[Q^{\pi^t_h,*}_h(x^t_h, a^t_h, b^t_h) - \underline{Q}^t_h(x^t_h, a^t_h, b^t_h)\right],
\end{aligned}$$

we are at the conclusion that

$$\begin{aligned}
\overline{V}^t_h(x^t_h) - V^{*,\nu^t_h}_h(x^t_h) &\ge \overline{Q}^t_h(x^t_h, a^t_h, b^t_h) - Q^{*,\nu^t_h}_h(x^t_h, a^t_h, b^t_h) + \zeta^1_{h,t}, \\
V^{\pi^t_h,*}_h(x^t_h) - \underline{V}^t_h(x^t_h) &\ge Q^{\pi^t_h,*}_h(x^t_h, a^t_h, b^t_h) - \underline{Q}^t_h(x^t_h, a^t_h, b^t_h) + \zeta^2_{h,t}.
\end{aligned}$$

Bringing this lower bound result together with the previous absolute bound (F.6) into Eq. (F.5) and its counterpart for the min-player, we have

$$\begin{aligned}
&\left(V^{*,\nu^t_h}_h(x^t_h) - \overline{V}^t_h(x^t_h)\right) - \left(V^{*,\nu^t_h}_{h+1}(x^t_{h+1}) - \overline{V}^t_{h+1}(x^t_{h+1})\right) \\
&\qquad + \left(\underline{V}^t_h(x^t_h) - V^{\pi^t_h,*}_h(x^t_h)\right) - \left(\underline{V}^t_{h+1}(x^t_{h+1}) - V^{\pi^t_h,*}_{h+1}(x^t_{h+1})\right) \\
&\ge \alpha^1_{h,t} + \alpha^2_{h,t} + \zeta^1_{h,t} + \zeta^2_{h,t} - 2\beta_t \min\{1, \overline{w}^t_h(z^t_h)\} - 2\beta_t \min\{1, \underline{w}^t_h(z^t_h)\} - 2H \cdot \iota_{\mathrm{mis}}.
\end{aligned} \tag{F.7}$$

Combining with Eq. (F.2), we arrive at a bound in terms of $\overline{w}_h^t(z_h^t), \underline{w}_h^t(z_h^t)$ and the martingale difference sequences:

$$\sum_{h=1}^{H} V_h^{*,\nu^t}(x_h^t) - V_h^{\pi^t,*}(x_h^t)$$

$$\leq \sum_{h=1}^{H} \left(4\beta_t \min\{1, \overline{w}_h^t(x_h^t)\} + 4\beta_t \min\{1, \underline{w}_h^t(x_h^t)\} + 4H \cdot \iota_{\mathrm{mis}} + \xi_h^t + \gamma_h^t + \alpha_{h,t}^1 + \alpha_{h,t}^2 + \zeta_{h,t}^1 + \zeta_{h,t}^2\right),$$

where ν^t is the policy that operates according to ν_h^t at time h and π^t is the sequence of π_h^t accordingly. The rest of the proof follows by bounding $\sum_{t=1}^T \sum_{h=1}^H \min\{1, \overline{w}_h^t(z_h^t)\}, \sum_{t=1}^T \sum_{h=1}^H \min\{1, \underline{w}_h^t(z_h^t)\}$ and the martingale difference sequences. Following the same techniques as in the proof of Theorem 4, we again apply Lemma 17 and Cauchy-Schwarz inequality, together with the Azuma-Hoeffding for bounded martingale differences, we arrive at our final result.

G Proof of Results for `KernelCCE-VTR` with Neural Approximation

In this section, we proof our result for the neural network approximation.

Proof of Theorem 15. Let C be defined in Lemma 14. We define a rescaled version of $\boldsymbol{\phi}$ as $\widetilde{\boldsymbol{\phi}}/C$, then we know that for any bounded value functions $V_h(\cdot) : \mathcal{S} \mapsto [-1, 1]$,

$$|\mathbb{P}_h V_h(z) - \langle \widetilde{\boldsymbol{\phi}}_{V_h}(z), C(\boldsymbol{\theta}_h^* - \boldsymbol{\theta}^{(0)})\rangle| \leq C_1 |\mathcal{S}| B^{4/3} m^{-1/6} L^3 \sqrt{\log m}, \qquad \|\widetilde{\boldsymbol{\phi}}_{V_h}(z)\|_2 \leq 1.$$

Defining $\widetilde{\boldsymbol{\theta}}_h^* := C\boldsymbol{\theta}_h^*$ and $\widetilde{\boldsymbol{\theta}}^{(0)} := C\boldsymbol{\theta}^{(0)}$ and taking $\iota_{\mathrm{mis}} := C \cdot B^{4/3} \cdot m^{-1/6} \cdot \sqrt{\log m}$, by Theorem 10 we know that for $\widetilde{\lambda} = 1 + \frac{1}{T}$, any $\delta > 0$ and any β satisfying

$$\left(\frac{\beta}{H}\right)^2 \geq 2\Gamma_{\widetilde{\mathbf{K}}}(T, \widetilde{\lambda}) + 3 + 6 \cdot \log\left(\frac{1}{\delta}\right) + 3\lambda \left(\frac{\widetilde{B}}{H}\right)^2 + 3C^2 \cdot B^{8/3} \cdot m^{-1/12} \cdot \log m \cdot t,$$

there exists a global constant $c > 0$ such that with probability at least $1 - \delta$, we have

$$\mathrm{Regret}(T) \leq c \left(\beta H \sqrt{T \cdot \Gamma_{\widetilde{K}}(T, \lambda)} + 1 + H^2 T \iota_{\mathrm{mis}}\right).$$

where $\widetilde{B} = C \cdot B$, and

$$\Gamma_{\widetilde{K}}(T, \lambda) := \sup_{(V_i)_i, (z_i)_i} \frac{1}{2} \log\det(\mathbf{I} + \widetilde{\mathbf{K}}(\{V_i\}_i, \{z_i\}_i)/\lambda), \tag{G.1}$$

where $\widetilde{\mathbf{K}} \in \mathbb{R}^{T \times T}$ is the matrix based on the kernel function k induced by the feature mapping $\widetilde{\boldsymbol{\phi}}$, where

$$\boldsymbol{k}_{V_1, V_2}(z_1, z_2) = \langle \widetilde{\boldsymbol{\phi}}_{V_1}(z_1), \widetilde{\boldsymbol{\phi}}_{V_2}(z_2)\rangle.$$

Rescaling gives

$$\Gamma_{\widetilde{\mathbf{K}}}(T, \widetilde{\lambda}) = \Gamma_{\mathbf{K}}(T, C^2 \widetilde{\lambda}).$$

Choosing $\lambda := C^2(1 + \frac{1}{T})$ completes our proof of Theorem 15.

□

H Proof of Auxillary Lemmas

In this section, we prove the essential lemmas in the proof of our main theorems. First of all we present a concentration bound for self-normalized processes in an RKHS $\mathcal{H}$, which is critical in determining the main term of the regret bound.

Theorem 24 (Self-Normalized Concentration Bounds for RKHS [Chowdhury and Gopalan, 2017, Yang et al., 2020])**.** Let $\{\mathbf{x}_t\}_{t\geq 1}$ be a discrete time stochastic process taking values in $\mathcal{Z}$, $\mathcal{H}$ is an RKHS with kernel $\mathbf{K}(\cdot,\cdot) : \mathcal{Z}\times\mathcal{Z}\mapsto\mathbb{R}$, $\{\mathcal{F}_t\}_{t\geq 0}$ is a given filtration. We assume that $\mathbf{x}_t$ is $\mathcal{F}_{t-1}$ measurable in the sense that for $\forall t\geq 1$, $\mathbf{x}_t\in\mathcal{F}_{t-1}$. Furthermore, $\{\epsilon_t\}_{t\geq 1}$ is a real-valued stochastic process with each ϵ_t $\mathcal{F}_t$ measurable and σ-sub-Gaussian. Define $\mathbf{K}_t\in\mathbb{R}^{(t-1)\times(t-1)}$ as the Gram matrix for data $\{\mathbf{x}_\tau\}_{\tau\in[t-1]}$ of the RKHS $\mathcal{H}$. Then for any $\lambda > 1$ and $\delta\in(0,1)$, with probability at least $1-\delta$, the following holds simultaneosly for all $t\geq 0$:

$$\|\varepsilon_{1:t-1}\|^2_{((\mathbf{K}_t+(\lambda-1)\cdot\mathbf{I})^{-1}+\mathbf{I})^{-1}} \leq 2\sigma^2\log\frac{\sqrt{\det(\lambda\mathbf{I}+\mathbf{K}_t)}}{\delta}.$$

Proof. See Lemma E.1 in Yang et al. [2020] and Theorem 1 in Chowdhury and Gopalan [2017] for the detailed proof. □

H.1 Proof of Lemma 3

Proof of Lemma 3. We only provide the proof of the max-player, and results of the min-player can be derived similarly. We recall the definition of $\overline{\mathbf{y}}_h^t$ is the vector of regression targets $\left(\overline{V}_{h+1}^1(x_{h+1}^1),\ldots,\overline{V}_{h+1}^{t-1}(x_{h+1}^{t-1})\right)^\top \in \mathbb{R}^{t-1}$. Furthermore by Lemma 11 in Section A, we know that $\overline{\theta}_h^t = \left(\overline{\mathbf{\Psi}}_h^t\right)^\top\left[\overline{\mathbf{K}}_h^t+\lambda\mathbf{I}\right]^{-1}\overline{\mathbf{y}}_h^t$ and $\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z) = \left(\overline{\mathbf{\Psi}}_h^t\right)^\top(\overline{\mathbf{K}}_h^t+\lambda\mathbf{I})^{-1}\overline{\boldsymbol{k}}_h^t(z)+\lambda\cdot(\overline{\mathbf{\Lambda}}_h^t)^{-1}\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)$, which enable us to bound the difference $\left\langle\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z),\overline{\boldsymbol{\theta}}_h^t-\boldsymbol{\theta}_h^*\right\rangle_{\mathcal{H}}$ as follows:

$$\begin{aligned}
\left\langle\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z),\overline{\boldsymbol{\theta}}_h^t-\boldsymbol{\theta}_h^*\right\rangle_{\mathcal{H}} &= \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top\left(\overline{\mathbf{\Psi}}_h^t\right)^\top\left[\overline{\mathbf{K}}_h^t+\lambda\mathbf{I}\right]^{-1}\overline{\mathbf{y}}_h^t \\
&\quad -\left(\overline{\boldsymbol{\theta}}_h^*\right)^\top\left[\left(\overline{\mathbf{\Psi}}_h^t\right)^\top\left[\overline{\mathbf{K}}_h^t+\lambda\mathbf{I}\right]^{-1}\overline{\boldsymbol{k}}_h^t(z)+\lambda\cdot(\overline{\mathbf{\Lambda}}_h^t)^{-1}\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)\right] \\
&= \underbrace{(\overline{\boldsymbol{k}}_h^t)^\top\left[\overline{\mathbf{K}}_h^t+\lambda\mathbf{I}\right]^{-1}\left[\overline{\mathbf{y}}_h^t-\overline{\mathbf{\Psi}}_h^t\boldsymbol{\theta}_h^*\right]}_{\mathrm{I}_1} - \underbrace{\lambda\cdot\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top(\overline{\mathbf{\Lambda}}_h^t)^{-1}\boldsymbol{\theta}_h^*}_{\mathrm{I}_2}.
\end{aligned}$$

For bounding I_2, we apply the Cauchy-Schwarz inequality and have

$$\begin{aligned}
\lambda\cdot\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top(\overline{\mathbf{\Lambda}}_h^t)^{-1}\boldsymbol{\theta}_h^* &\leq \left\|\lambda\cdot\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top(\overline{\mathbf{\Lambda}}_h^t)^{-1}\right\|_{\mathcal{H}}\cdot\|\boldsymbol{\theta}_h^*\|_{\mathcal{H}} \\
&\overset{(a)}{\leq} B\cdot\sqrt{\lambda\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top(\overline{\mathbf{\Lambda}}_h^t)^{-1}\lambda(\overline{\mathbf{\Lambda}}_h^t)^{-1}\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)} \overset{(b)}{\leq} \sqrt{\lambda}B\cdot\overline{w}_h^t(z),
\end{aligned}$$

where (a) is due to the assumption that $\|\boldsymbol{\theta}_h^*\|_{\mathcal{H}}\leq B$ and (b) is by the definition of $\overline{w}_h^t$ and the fact that $\left(\overline{\mathbf{\Lambda}}_h^t\right)^{-1}$ is a self-adjoint mapping on the RKHS $\mathcal{H}$ satisfying $\left\|\left(\overline{\mathbf{\Lambda}}_h^t\right)^{-1}\right\|_{op}\leq\frac{1}{\lambda}$.

For bounding I_1, we observe the following equality:

$$\begin{aligned}
(\overline{\boldsymbol{k}}_h^t)^\top\left[\overline{\mathbf{K}}_h^t+\lambda\mathbf{I}\right]^{-1}\left[\overline{\mathbf{y}}_h^t-\overline{\mathbf{\Psi}}_h^t\boldsymbol{\theta}_h^*\right] &= \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top(\overline{\mathbf{\Lambda}}_h^t)^{-1}(\overline{\mathbf{\Psi}}_h^t)^\top\left[\overline{\mathbf{y}}_h^t-\mathbf{\Psi}_h^t\boldsymbol{\theta}_h^*\right] \\
&= \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top(\overline{\mathbf{\Lambda}}_h^t)^{-1}\sum_{\tau=1}^{t-1}\boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}(z_h^\tau)\left[\overline{V}_{h+1}^\tau(x_{h+1}^\tau)-(\mathbb{P}_h\overline{V}_{h+1}^\tau)(z_h^\tau)\right].
\end{aligned}$$

Again by applying the Cauchy-Schwarz inequality, we bound the RHS of the above equality as

$$|\mathrm{I}_1| \leq \left\|\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)\right\|_{(\mathbf{\Lambda}_h^t)^{-1}}\cdot\left\|\sum_{\tau=1}^{t-1}\boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}(z_h^\tau)\left[\overline{V}_{h+1}^\tau(x_{h+1}^\tau)-(\mathbb{P}_h\overline{V}_{h+1}^\tau)(z_h^\tau)\right]\right\|_{(\mathbf{\Lambda}_h^t)^{-1}}.$$

We note that for a given h that we consider in Lemma 3. If we define $\{\mathcal{F}_t\}_{t\geq 0}$ as the σ-algebra generated by all data before iteration t and all data before step h at iteration $t+1$, $\overline{V}_{h+1}^\tau(x_{h+1}^\tau)-(\mathbb{P}_h\overline{V}_{h+1}^\tau)(z_h^\tau)\in\mathcal{F}_\tau$ is a mean zero random variable with respect to filtration

$\mathcal{F}_{\tau-1}$ with $\left|\overline{V}^{\tau}_{h+1}(x^{\tau}_{h+1}) - (\mathbb{P}_h \overline{V}^{\tau}_{h+1})(z^{\tau}_h)\right| \le H$. We take $\epsilon_\tau = \overline{V}^{\tau}_{h+1}(x^{\tau}_{h+1}) - (\mathbb{P}_h \overline{V}^{\tau}_{h+1})(z^{\tau}_h)$ in Theorem 24 and $\{\mathbf{x}_t\}_{t\ge 1}$ is $\left\{\phi_{\overline{V}^t_{h+1}}(z^t_h)\right\}_{t\ge1}$. Then by directly utilizing Theorem 24, we have that

$$\left\|\sum_{\tau=1}^{t-1} \phi_{\overline{V}^{\tau}_{h+1}}(z^{\tau}_h)\left[\overline{V}^{t}_{h+1}(x^{\tau}_{h+1}) - (\mathbb{P}_h \overline{V}^{t}_{h+1})(z^{\tau}_h)\right]\right\|^2_{(\mathbf{\Lambda}^t_h)^{-1}} = \left\|(\mathbf{\Psi}^t_h)^\top \epsilon^t_h\right\|^2_{(\mathbf{\Lambda}^t_h)^{-1}} = (\epsilon^t_h)^\top \mathbf{\Psi}^t_h \mathbf{\Lambda}_t^{-1} (\mathbf{\Psi}^t_h)^\top \epsilon^t_h.$$

By Lemma 11 and Theorem 24 again, we arrive at the following:

$$\begin{aligned}(\epsilon^t_h)^\top \mathbf{\Psi}^t_h \mathbf{\Lambda}_t^{-1} (\mathbf{\Psi}^t_h)^\top \epsilon^t_h &\le (\epsilon^t_h)^\top \mathbf{K}_t(\mathbf{K}_t + \lambda\mathbf{I})^{-1}\epsilon^t_h \le (\epsilon^t_h)^\top \left(\mathbf{K}_t + (\lambda-1)\cdot\mathbf{I}\right)(\mathbf{K}_t + \lambda\mathbf{I})^{-1}\epsilon^t_h \\ &= (\epsilon^t_h)^\top \left((\mathbf{K}_t + (\lambda-1)\cdot\mathbf{I})^{-1} + \mathbf{I}\right)^{-1}\epsilon^t_h \\ &\le H^2 \log\frac{\sqrt{\det(\lambda\mathbf{I} + \mathbf{K}_t)}}{\delta} \le 2H^2\cdot \text{logdet}(\lambda\mathbf{I} + \mathbf{K}_t) + 2H^2 \cdot \log\frac{1}{\delta}.\end{aligned}$$

By taking $\lambda = 1 + \frac{1}{T}$,

$$\begin{aligned}&\left|\phi(z)^\top(\boldsymbol{\theta}^*_h - \overline{\boldsymbol{\theta}}^t_h)\right| \\ &\le \left\{\left[H^2\cdot\text{logdet}\left[\lambda\mathbf{I} + \mathbf{K}_t\right] + 2H^2\cdot\log\left(\frac{1}{\delta}\right)\right]^{1/2} + \sqrt{\lambda}B\right\}\cdot b^t_h(z) \\ &\le \left\{\left[H^2\cdot\text{logdet}\left[\mathbf{I} + \mathbf{K}_t/\lambda\right] + (\lambda-1)tH^2 + 2H^2\cdot\log\left(\frac{1}{\delta}\right)\right]^{1/2} + \sqrt{\lambda}B\right\}\cdot b^t_h(z) \\ &\le \left\{\left[H^2\cdot\Gamma_{\mathbf{K}}(T,\lambda) + H^2 + 2H^2\cdot\log\left(\frac{1}{\delta}\right)\right]^{1/2} + \sqrt{\lambda}B\right\}\cdot b^t_h(z).\end{aligned}$$

Let $(\beta/H)^2 = 2\Gamma_{\mathbf{K}}(T,\lambda) + 2 + 4\cdot\log\left(\frac{1}{\delta}\right) + 2\lambda\left(\frac{B}{H}\right)^2$, we know from the above theoretical derivation that $\left|\phi(z)^\top(\boldsymbol{\theta}^*_h - \overline{\boldsymbol{\theta}}^t_h)\right| \le \beta\cdot b^t_h(z)$ with probability at least $1-\delta$, which concludes our proof. ☐

H.2 Proof of Lemma 22

Proof of Lemma 22. The first half of the proof of Lemma 22 follows exactly the same as in the proof of Lemma 3, we restate the proof for completeness and analyze the error terms brought by misspecification in the later half of the proof. We only provide the proof of the max-player, and results of the min-player can be derived similarily. We recall the definition of $\overline{\mathbf{y}}^t_h$ is the vector of regression targets $\left(\overline{V}^1_{h+1}(x^1_{h+1}), \ldots, \overline{V}^{t-1}_{h+1}(x^{t-1}_{h+1})\right)^\top \in \mathbb{R}^{t-1}$. Furthermore by Lemma 11 in Section A, we know that $\overline{\theta}^t_h = \left(\overline{\mathbf{\Psi}}^t_h\right)^\top\left[\overline{\mathbf{K}}^t_h + \lambda\mathbf{I}\right]^{-1}\overline{\mathbf{y}}^t_h$ and $\phi_{\overline{V}^t_{h+1}}(z) = \left(\overline{\mathbf{\Psi}}^t_h\right)^\top(\overline{\mathbf{K}}^t_h + \lambda\mathbf{I})^{-1}\overline{\boldsymbol{k}}^t_h(z) + \lambda\cdot(\overline{\mathbf{\Lambda}}^t_h)^{-1}\phi_{\overline{V}^t_{h+1}}(z)$, which enable us to bound the difference $\left\langle\phi_{\overline{V}^t_{h+1}}(z), \overline{\boldsymbol{\theta}}^t_h - \boldsymbol{\theta}^*_h\right\rangle_{\mathcal{H}}$ as follows:

$$\begin{aligned}\left\langle\phi_{\overline{V}^t_{h+1}}(z), \overline{\boldsymbol{\theta}}^t_h - \boldsymbol{\theta}^*_h\right\rangle_{\mathcal{H}} &= \phi_{\overline{V}^t_{h+1}}(z)^\top\left(\overline{\mathbf{\Psi}}^t_h\right)^\top\left[\overline{\mathbf{K}}^t_h + \lambda\mathbf{I}\right]^{-1}\overline{\mathbf{y}}^t_h \\ &\quad - \left(\overline{\boldsymbol{\theta}}^*_h\right)^\top\left[\left(\overline{\mathbf{\Psi}}^t_h\right)^\top\left[\overline{\mathbf{K}}^t_h + \lambda\mathbf{I}\right]^{-1}\overline{\boldsymbol{k}}^t_h(z) + \lambda\cdot(\overline{\mathbf{\Lambda}}^t_h)^{-1}\phi_{\overline{V}^t_{h+1}}(z)\right] \\ &= \underbrace{(\overline{\boldsymbol{k}}^t_h)^\top\left[\overline{\mathbf{K}}^t_h + \lambda\mathbf{I}\right]^{-1}\left[\overline{\mathbf{y}}^t_h - \overline{\mathbf{\Psi}}^t_h\boldsymbol{\theta}^*_h\right]}_{\mathrm{I}_1} - \underbrace{\lambda\cdot\phi_{\overline{V}^t_{h+1}}(z)^\top(\overline{\mathbf{\Lambda}}^t_h)^{-1}\boldsymbol{\theta}^*_h}_{\mathrm{I}_2}.\end{aligned}$$

For bounding I_2, we apply the Cauchy-Schwarz inequality and have

$$\begin{aligned}\lambda\cdot\phi_{\overline{V}^t_{h+1}}(z)^\top(\overline{\mathbf{\Lambda}}^t_h)^{-1}\boldsymbol{\theta}^*_h &\le \left\|\lambda\cdot\phi_{\overline{V}^t_{h+1}}(z)^\top(\overline{\mathbf{\Lambda}}^t_h)^{-1}\right\|_{\mathcal{H}}\cdot\|\boldsymbol{\theta}^*_h\|_{\mathcal{H}} \\ &\overset{(a)}{\le} B\cdot\sqrt{\lambda\phi_{\overline{V}^t_{h+1}}(z)^\top(\overline{\mathbf{\Lambda}}^t_h)^{-1}\lambda(\overline{\mathbf{\Lambda}}^t_h)^{-1}\phi_{\overline{V}^t_{h+1}}(z)} \overset{(b)}{\le} \sqrt{\lambda}B\cdot\overline{w}^t_h(z),\end{aligned}$$

where (a) is due to the assumption that $\|\boldsymbol{\theta}_h^*\|_{\mathcal{H}} \leq B$ and (b) is by the definition of $\overline{w}_h^t$ and the fact that $\left(\overline{\boldsymbol{\Lambda}}_h^t\right)^{-1}$ is a self-adjoint mapping on the RKHS $\mathcal{H}$ satisfying $\left\|\left(\overline{\boldsymbol{\Lambda}}_h^t\right)^{-1}\right\|_{op} \leq \frac{1}{\lambda}$.

For bounding I_1, we observe the following equality:

$$
\begin{aligned}
&(\overline{\boldsymbol{k}}_h^t)^\top \left[\overline{\mathbf{K}}_h^t + \lambda \mathbf{I}\right]^{-1} \left[\overline{\mathbf{y}}_h^t - \overline{\boldsymbol{\Psi}}_h^t \boldsymbol{\theta}_h^*\right] \\
&= \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top (\overline{\boldsymbol{\Lambda}}_h^t)^{-1} (\overline{\boldsymbol{\Psi}}_h^t)^\top \left[\overline{\mathbf{y}}_h^t - \boldsymbol{\Psi}_h^t \boldsymbol{\theta}_h^*\right] \\
&= \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top (\overline{\boldsymbol{\Lambda}}_h^t)^{-1} \sum_{\tau=1}^{t-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}(z_h^\tau) \left[\overline{V}_{h+1}^\tau(x_{h+1}^\tau) - (\mathbb{P}_h \overline{V}_{h+1}^\tau)(z_h^\tau)\right] \\
&\quad + \boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)^\top (\overline{\boldsymbol{\Lambda}}_h^t)^{-1} \sum_{\tau=1}^{t-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}(z_h^\tau) \left[(\mathbb{P}_h \overline{V}_{h+1}^\tau)(z_h^\tau) - \left\langle \boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}, \boldsymbol{\theta}_h^* \right\rangle_{\mathcal{H}}\right].
\end{aligned}
$$

Again by applying the Cauchy-Schwarz inequality, we bound the RHS of the above equality as

$$
\begin{aligned}
|\mathrm{I}_1| \leq & \left\|\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)\right\|_{(\boldsymbol{\Lambda}_h^t)^{-1}} \cdot \underbrace{\left\|\sum_{\tau=1}^{t-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}(z_h^\tau) \left[\overline{V}_{h+1}^\tau(x_{h+1}^\tau) - (\mathbb{P}_h \overline{V}_{h+1}^\tau)(z_h^\tau)\right]\right\|_{(\boldsymbol{\Lambda}_h^t)^{-1}}}_{A_1} \\
& + \left\|\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z)\right\|_{(\boldsymbol{\Lambda}_h^t)^{-1}} \cdot \underbrace{\left\|\sum_{\tau=1}^{t-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}(z_h^\tau) \left[(\mathbb{P}_h \overline{V}_{h+1}^\tau)(z_h^\tau) - \boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}^\top \boldsymbol{\theta}_h^*\right]\right\|_{(\boldsymbol{\Lambda}_h^t)^{-1}}}_{A_2}.
\end{aligned}
\tag{H.1}
$$

For bounding A_1, we note that for a given h that we consider in Lemma 22. If we define $\{\mathcal{F}_t\}_{t\geq 0}$ as the σ-algebra generated by all data before iteration t and all data before step h at iteration $t+1$, $\overline{V}_{h+1}^\tau(x_{h+1}^\tau) - (\mathbb{P}_h \overline{V}_{h+1}^\tau)(z_h^\tau) \in \mathcal{F}_\tau$ is a mean zero random variable with respect to filtration $\mathcal{F}_{\tau-1}$ with $\left|\overline{V}_{h+1}^\tau(x_{h+1}^\tau) - (\mathbb{P}_h \overline{V}_{h+1}^\tau)(z_h^\tau)\right| \leq H$. We take $\epsilon_\tau = \overline{V}_{h+1}^\tau(x_{h+1}^\tau) - (\mathbb{P}_h \overline{V}_{h+1}^\tau)(z_h^\tau)$ in Theorem 24 and $\{\mathbf{x}_t\}_{t\geq 1}$ is $\left\{\boldsymbol{\phi}_{\overline{V}_{h+1}^t}(z_h^t)\right\}_{t\geq 1}$. Then by directly utilizing Theorem 24, we have that

$$
\begin{aligned}
&\left\|\sum_{\tau=1}^{t-1} \boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}(z_h^\tau) \left[\overline{V}_{h+1}^t(x_{h+1}^\tau) - (\mathbb{P}_h \overline{V}_{h+1}^t)(z_h^\tau)\right]\right\|_{(\boldsymbol{\Lambda}_h^t)^{-1}}^2 \\
&= \left\|(\boldsymbol{\Psi}_h^t)^\top \epsilon_h^t\right\|_{(\boldsymbol{\Lambda}_h^t)^{-1}}^2 = (\epsilon_h^t)^\top \boldsymbol{\Psi}_h^t \boldsymbol{\Lambda}_t^{-1} (\boldsymbol{\Psi}_h^t)^\top \epsilon_h^t.
\end{aligned}
$$

By Lemma 11 and Theorem 24 again, we arrive at the following:

$$
\begin{aligned}
(\epsilon_h^t)^\top \boldsymbol{\Psi}_h^t \boldsymbol{\Lambda}_t^{-1} (\boldsymbol{\Psi}_h^t)^\top \epsilon_h^t &\leq (\epsilon_h^t)^\top \mathbf{K}_t (\mathbf{K}_t + \lambda \mathbf{I})^{-1} \epsilon_h^t \leq (\epsilon_h^t)^\top \left(\mathbf{K}_t + (\lambda - 1) \cdot \mathbf{I}\right) (\mathbf{K}_t + \lambda \mathbf{I})^{-1} \epsilon_h^t \\
&= (\epsilon_h^t)^\top \left((\mathbf{K}_t + (\lambda - 1) \cdot \mathbf{I})^{-1} + \mathbf{I}\right)^{-1} \epsilon_h^t \\
&\leq 2H^2 \log \frac{\sqrt{\det(\lambda \mathbf{I} + \mathbf{K}_t)}}{\delta} \leq H^2 \cdot \operatorname{logdet}(\lambda \mathbf{I} + \mathbf{K}_t) + 2H^2 \cdot \log \frac{1}{\delta}.
\end{aligned}
$$

For bounding the term A_2 in Eq. (H.1), we apply the following lemma, which is the RKHS version of the Lemma 8 in Zanette et al. [2020]:

Lemma 25 (Lemma 8 in Zanette et al. [2020])**.** Let $\{\boldsymbol{a}_i\}_{i=1,\ldots,t}$ be any sequence of vectors in the RKHS $\mathcal{H}$ and $\{b_i\}_{i=1,\ldots,t}$ be any sequence of scalars such that $|b_i| \leq \epsilon \in \mathbb{R}^+$. For any $\lambda \geq 0$ and $t \in \mathbb{N}$ we have:

$$
\left\|\sum_{i=1}^t \boldsymbol{a}_i b_i\right\|_{\left[\sum_{i=1}^t \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \mathbf{I}\right]^{-1}} \leq t\epsilon^2.
$$

Proof of Lemma 25. By defining the feature matrix $\mathbf{A} := (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_t)$ and the vector $\boldsymbol{b} := (b_1, \ldots, b_t)^\top$, we have the following formulation

$$\sum_{t=1}^{t} \boldsymbol{a}_i b_i = \mathbf{A}\boldsymbol{b}, \qquad \left\| \sum_{i=1}^{t} \boldsymbol{a}_i b_i \right\|_{\left[\sum_{i=1}^{t} \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \mathbf{I}\right]^{-1}} = \|\mathbf{A}\boldsymbol{b}\|_{[\mathbf{A}\mathbf{A}^\top + \lambda \mathbf{I}_{\mathcal{H}}]^{-1}} = \boldsymbol{b}^\top \mathbf{A}^\top [\mathbf{A}\mathbf{A}^\top + \lambda \mathbf{I}_{\mathcal{H}}]^{-1} \mathbf{A}\boldsymbol{b}.$$

Via the same reasoning as in the proof of item (a) in Lemma 11,

$$\begin{aligned} \boldsymbol{b}^\top \mathbf{A}^\top [\mathbf{A}\mathbf{A}^\top + \lambda \mathbf{I}_{\mathcal{H}}]^{-1} \mathbf{A}\boldsymbol{b} &= \boldsymbol{b}^\top [\mathbf{A}^\top \mathbf{A} + \lambda \mathbf{I}_{\mathcal{H}}]^{-1} \mathbf{A}^\top \mathbf{A} \boldsymbol{b} \\ &= \boldsymbol{b}^\top \boldsymbol{b} - \lambda \cdot \boldsymbol{b}^\top [\mathbf{A}^\top \mathbf{A} + \lambda \mathbf{I}_{\mathcal{H}}]^{-1} \boldsymbol{b} \le \|\boldsymbol{b}\|^2 \le t\epsilon^2, \end{aligned}$$

which concludes our proof of Lemma 25. □

By letting $b_\tau = \left[(\mathbb{P}_h \overline{V}_{h+1}^\tau)(z_h^\tau) - \boldsymbol{\phi}_{\overline{V}_{h+1}^t}^\top \boldsymbol{\theta}_h^*\right]$ and $\boldsymbol{a}_\tau = \boldsymbol{\phi}_{\overline{V}_{h+1}^\tau}(z_h^\tau)$ and knowing that $|b_\tau| \le \iota_{\mathrm{mis}}$, we have the bound for item A_2 that $A_2 \le \iota_{\mathrm{mis}} \cdot \sqrt{t}$. Then by taking $\lambda = 1 + \frac{1}{T}$,

$$\begin{aligned} &\left| \boldsymbol{\phi}(z)^\top (\boldsymbol{\theta}_h^* - \overline{\boldsymbol{\theta}}_h^t) \right| \\ &\le \left\{ \left[H^2 \cdot \mathrm{logdet}\left[\lambda \mathbf{I} + \mathbf{K}_t\right] + 2H^2 \cdot \log\left(\frac{1}{\delta}\right) \right]^{1/2} + \sqrt{\lambda} B + H \cdot \iota_{\mathrm{mis}} \sqrt{t} \right\} \cdot b_h^t(z) && \text{(H.2)} \\ &\le \left\{ \left[H^2 \cdot \mathrm{logdet}\left[\mathbf{I} + \mathbf{K}_t/\lambda\right] + (\lambda - 1)tH^2 + 2H^2 \cdot \log\left(\frac{1}{\delta}\right) \right]^{1/2} + \sqrt{\lambda} B + H \cdot \iota_{\mathrm{mis}} \sqrt{t} \right\} \cdot b_h^t(z) && \text{(H.3)} \\ &\le \left\{ \left[H^2 \cdot \Gamma_{\mathbf{K}}(T, \lambda) + H^2 + 2H^2 \cdot \log\left(\frac{1}{\delta}\right) \right]^{1/2} + \sqrt{\lambda} B + H \cdot \iota_{\mathrm{mis}} \sqrt{t} \right\} \cdot b_h^t(z). && \text{(H.4)} \end{aligned}$$

Let $(\beta/H)^2 = 3\Gamma_{\mathbf{K}}(T, \lambda) + 3 + 6 \cdot \log\left(\frac{1}{\delta}\right) + 3\lambda\left(\frac{B}{H}\right)^2 + \iota_{\mathrm{mis}} t^2$, $\left| \boldsymbol{\phi}(z)^\top (\boldsymbol{\theta}_h^* - \overline{\boldsymbol{\theta}}_h^t) \right| \le \beta \cdot b_h^t(z)$ with probability at least $1 - \delta$. We present the last step of our proof in the following part. Equipped with Assumption 9, we are able to bound the distance between the optimal estimated expected value at time $h + 1$ with the true expected value as in the following Lemma 26, which concludes our proof:

Lemma 26. For any bounded value function $V(\cdot) : \mathcal{S} \mapsto [-1, 1]$ and any $z \in \mathcal{Z}$, there exists a $\boldsymbol{\theta}_h^* \in \mathcal{H}$ such that:

$$|\mathbb{P}_h V(z) - \langle \boldsymbol{\phi}_V(z), \boldsymbol{\theta}_h^* \rangle_{\mathcal{H}}| \le \iota_{\mathrm{mis}},$$

where $\boldsymbol{\phi}_V$ is defined in Section 3.2. The proof of Lemma 26 is a simply application of the definition of total variation distance.

□

H.3 Proof of Lemma 14

In this section we prove Lemma 14. We need the following lemmas.

Lemma 27 (Lemma 4.1 in Cao and Gu 2019, Zhou et al. 2020)**.** There exist constants $C_i > 0$ such that for any $\delta \in (0, 1)$, if B satisfies that

$$C_1 m^{-1} L^{-3/2} \max\{\log^{-3/2} m, \log^{3/2}(|\mathcal{Z}| L^2/\delta)\} \le B \le C_2 L^{-6} (\log m)^{-3/2},$$

then with probability at least $1 - \delta$, for all $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ satisfying $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in B(\boldsymbol{\theta}^{(0)}, B)$ and all $(s', z) \in \mathcal{S} \times \mathcal{Z}$, we have

$$|f(s', z; \boldsymbol{\theta}_1) - f(s', z; \boldsymbol{\theta}_2) - \langle \boldsymbol{\phi}(s', z; \boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle| \le C_3 B^{4/3} m^{-1/6} L^3 \sqrt{\log m}.$$

Lemma 28 (Lemma B.3 in Cao and Gu 2019, Zhou et al. 2020)**.** There exist constants $C_i > 0$ such that for any $\delta \in (0, 1)$, if B satisfies that

$$C_1 m^{-1} L^{-3/2} \max\{\log^{-3/2} m, \log^{3/2}(|\mathcal{Z}| L^2/\delta)\} \le B \le C_2 L^{-6} (\log m)^{-3/2},$$

then probability at least $1 - \delta$, for all $\boldsymbol{\theta}$ satisfying $\boldsymbol{\theta} \in B(\boldsymbol{\theta}^{(0)}, B)$ and all $(s', z) \in \mathcal{S} \times \mathcal{Z}$, we have $\|\boldsymbol{\phi}(s'|z)\|_2 \le C_3 \sqrt{L}$.

Proof of Lemma 14. By Lemma 28 we have

$$\|\boldsymbol{\phi}_V(z)\|_2 = \Big\|\sum_{s'} V(s')\boldsymbol{\phi}(s'|z)\Big\|_2 \le \sum_{s'} |V(s')|\|\boldsymbol{\phi}(s'|z)\|_2 \le C_1|\mathcal{S}|\sqrt{L}.$$

By the assumption of $\mathbb{P}_h(s'|z)$, we have $\boldsymbol{\theta}_h^* \in B(\boldsymbol{\theta}^{(0)}, B)$. Thus, by Lemma 27, we have with probability at least $1-\delta$, for all $s' \in \mathcal{S}, z \in \mathcal{Z}, h \in [H]$,

$$\begin{aligned}|\mathbb{P}_h(s'|z) - \langle \boldsymbol{\phi}(s', z; \boldsymbol{\theta}^{(0)}), \boldsymbol{\theta}_h^* - \boldsymbol{\theta}^{(0)}\rangle| &= |f(s', z; \boldsymbol{\theta}_h^*) - f(s', z; \boldsymbol{\theta}^{(0)}) - \langle \boldsymbol{\phi}(s', z; \boldsymbol{\theta}^{(0)}), \boldsymbol{\theta}_h^* - \boldsymbol{\theta}^{(0)}\rangle| \\ &\le C_2 B^{4/3} m^{-1/6} L^3 \sqrt{\log m},\end{aligned}$$

where the equality holds by the assumption of $\mathbb{P}_h$ and $f(s', z; \boldsymbol{\theta}^{(0)}) = 0$ guaranteed by the initialization scheme, the inequality holds due to Lemma 27. Therefore, for any value function $V : \mathcal{S} \to [-1, 1]$, we have

$$\begin{aligned}|\mathbb{P}_h V(z) - \langle \boldsymbol{\phi}_V(z), \boldsymbol{\theta}_h^* - \boldsymbol{\theta}^{(0)}\rangle| &= \Big|\sum_{s'} V(s')\mathbb{P}_h(s'|z) - \sum_{s'} V(s')\langle \boldsymbol{\phi}(s'|z), \boldsymbol{\theta}_h^* - \boldsymbol{\theta}^{(0)}\rangle\Big| \\ &\le \sum_{s'} |V(s')||\mathbb{P}_h(s'|z) - \langle \boldsymbol{\phi}(s'|z), \boldsymbol{\theta}_h^* - \boldsymbol{\theta}^{(0)}\rangle| \\ &\le C_2|\mathcal{S}|HB^{4/3}m^{-1/6}L^3\sqrt{\log m},\end{aligned}$$

where the second inequality we use the fact that $|V| \le 1$. □

I Implementation Details of FIND_CCE

Suppose that we have $Q_1, Q_2 \in \mathcal{S}\times\mathcal{A}\times\mathcal{A} \to \mathbb{R}$. Given a state $x \in \mathcal{S}$, let $P_1, P_2 \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ denote the matrices of Q values such that $[P_i]_{m,n} = Q_i(x, a_m, a_n)$ for $i = 1, 2$, where a_m, a_n denote the m-th and n-th actions of $\mathcal{A}$. Suppose the CCE of Q_1, Q_2 given x is denoted by a matrix $\sigma \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$, where $\sigma_{m,n}$ denotes the probability of selecting m-th and n-th actions. Then σ satisfies the following two groups of constraints:

- Since σ is a probability matrix, then we have

$$\forall 1 \le m, n \le |\mathcal{A}|,\ 0 \le \sigma_{m,n} \le 1, \tag{I.1}$$
$$\sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|} \sigma_{i,j} = 1. \tag{I.2}$$

- To satisfy (3.1), we have

$$\begin{aligned}&\forall 1 \le m \le |\mathcal{A}|,\ \sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|} \sigma_{i,j}[P_1]_{i,j} \ge \sum_{j=1}^{|\mathcal{A}|}[P_1]_{m,j}\sum_{i=1}^{|\mathcal{A}|}\sigma_{i,j}. \\ &\Leftrightarrow \forall 1 \le m \le |\mathcal{A}|,\ \sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|} \sigma_{i,j}([P_1]_{m,j} - [P_1]_{i,j}) \le 0 \end{aligned}\tag{I.3}$$

- To satisfy (3.2), we have

$$\begin{aligned}&\forall 1 \le n \le |\mathcal{A}|,\ \sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|} \sigma_{i,j}[P_2]_{i,j} \le \sum_{i=1}^{|\mathcal{A}|}[P_2]_{i,n}\sum_{j=1}^{|\mathcal{A}|}\sigma_{i,j}. \\ &\Leftrightarrow \forall 1 \le n \le |\mathcal{A}|,\ \sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|} \sigma_{i,j}([P_2]_{i,j} - [P_2]_{i,n}) \le 0. \end{aligned}\tag{I.4}$$

There are total $|\mathcal{A}|^2$ number of unknown variables ($\sigma_{m,n}$) with 1 equality constraint and $|\mathcal{A}|^2 + 2|\mathcal{A}|$ number of inequality constraints. The above linear system can be converted into a standard linear programming problem with $2|\mathcal{A}|^2$ number of variables $\sigma_{m,n}, \widehat{\sigma}_{m,n}, 1 \le m, n \le |\mathcal{A}|$, such that

$$\max_{\sigma_{m,n}, \widehat{\sigma}_{m,n}} \quad 0$$

$$
\begin{aligned}
\sigma_{m,n} &\geq 0 \\
\widehat{\sigma}_{m,n} &\geq 0 \\
\sigma_{m,n} + \widehat{\sigma}_{m,n} &\leq 1 \\
-\sigma_{m,n} - \widehat{\sigma}_{m,n} &\leq -1 \\
\sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|}\sigma_{i,j} &\leq 1 \\
-\sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|}\sigma_{i,j} &\leq -1 \\
\sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|}\sigma_{i,j}([P_1]_{m,j} - [P_1]_{i,j}) &\leq 0 \\
\sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|}\sigma_{i,j}([P_2]_{i,j} - [P_2]_{i,n}) &\leq 0
\end{aligned}
$$

The above linear system can be solved by Karmarkar's algorithm [Karmarkar, 1984] within $\widetilde{O}(|\mathcal{A}|^7)$ time complexity, or with Stochastic Central Path Method [Cohen et al., 2021] within $\widetilde{O}(|\mathcal{A}|^{2w})$ time complexity, where $w = 2.373...$ is the matrix multiplication constant.