Pessimism Meets Invariance: Provably Efficient Offline Mean-Field Multi-Agent RL*

Minshuo Chen[1] **Yan Li**[1] **Ethan Wang**[1] **Zhuoran Yang**[2] **Zhaoran Wang**[3] **Tuo Zhao**[1]
[1]Georgia Tech [2]University of California, Berkeley [3]Northwestern University
{mchen393, tourzhao}@gatech.edu

Abstract

Mean-Field Multi-Agent Reinforcement Learning (MF-MARL) is attractive in the applications involving a large population of homogeneous agents, as it exploits the permutation invariance of agents and avoids the curse of many agents. Most existing results only focus on online settings, in which agents can interact with the environment during training. In some applications such as social welfare optimization, however, the interaction during training can be prohibitive or even unethical in the societal systems. To bridge such a gap, we propose a SAFARI (peSsimistic meAn-Field vAlue iteRatIon) algorithm for off-line MF-MARL, which only requires a handful of pre-collected experience data. Theoretically, under a weak coverage assumption that the experience dataset contains enough information about the optimal policy, we prove that for an episodic mean-field MDP with a horizon H and N training trajectories, SAFARI attains a sub-optimality gap of $\mathcal{O}(H^2 d_{\text{eff}}/\sqrt{N})$, where d_{eff} is the effective dimension of the function class for parameterizing the value function, but independent on the number of agents. Numerical experiments are provided.

1 Introduction

Significant progress has been made towards multi-agent reinforcement learning (MARL) for many prominent sequential decision making problems, such as social welfare optimization (Leibo et al., 2017), fleet control of autonomous vehicles (Shalev-Shwartz et al., 2016) and playing multiplayer online battle arena (MOBA) games (Berner et al., 2019). As the joint state and action space scales exponentially with the number of agents, however, MARL becomes computationally expensive. One remedy is the mean-field regime when an extremely large number of homogenous agents are involved, e.g., social welfare optimization. The effect of each agent on the overall multi-agent system can become infinitesimal, and therefore all agents can be considered interchangeable/indistinguishable (Yang et al., 2018; Carmona et al., 2019; Li et al., 2021). Accordingly, the interaction among agents can be captured by some mean-field quantity such as the empirical distribution of states, and therefore each agent only needs to find the best response to the so-called "mean-field state", which avoids the curse of many agents.

Most existing results on mean-field MARL (MF-MARL) are for the online setting (Yang et al., 2018; Zhang et al., 2019), where the agents can interact with the environment during training. However, such interaction during training can be prohibitive for some important applications (Leibo et al., 2017; Mandel et al., 2014; Jaques et al., 2019; Levine et al., 2020). Taking social welfare optimization as an example, repeatedly conducting social experiments on human being can be unaffordable or even unethical in the societal systems. Therefore, we can only consider the offline settings, i.e., we learn the optimal policy based on some pre-collected experience data (Levine et al., 2020). Unfortunately,

*Extension to online setting is provided in a longer technical report version, which is available upon request.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

existing offline reinforcement learning (RL) algorithms and theories all focus on the single agent settings, and no algorithms and theories have been developed for MARL under the offline settings, regardless of the mean-field regime or not.

To bridge such a critical gap, we propose the first pessimistic algorithm – named SAFARI (peSsimistic meAn-Field vAlue iteRatIon) for mean-field MARL, which can provably achieve sample efficiency under the offline setting. Our proposed algorithm contains two important components: (1) To incorporate the permutation invariance of the homogenous agents, we adopt a RKHS (Reproducing Kernel Hilbert Space) mean-embedding approach for approximating value functions, which avoids the exponential blowup of the agents' state and action spaces; (2) We develop an uncertainty quantifier, and integrate it into the value iteration procedure as the penalty function. Such a penalty function can effectively screen the "spuriously correlated trajectories", i.e., which possibly happen to appear in the experience data, but are actually unrelated to the optimal policy, but by chance induce large cumulative rewards and hence may potentially mislead the learned policy.

Theoretically, we establish a data-dependent upper bound on the suboptimality of SAFARI for MF-MARL without the stringent assumptions on the sufficient coverage of the experience data (e.g., finite concentrability coefficients (Chen and Jiang, 2019) or uniformly lower bounded densities of visitation measure (Yin et al., 2020)). More specifically, we only assume that the experience data of N training trajectories contains enough information about the optimal policy. Then we prove that for an episodic MF-MARL problem with a horizon H, SAFARI attains a sub-optimality gap of $\mathcal{O}(H^2 d_{\text{eff}}/\sqrt{N})$, where d_{eff} is the effective dimension of the function class (RKHS) for parameterizing the value function and independent on the number of agents. In addition to the offline settings, our SAFARI algorithm can also be extended to MF-MARL under the online setting (OMPPO algorithm), which is of independent interest. Details are provided in a longer technical report version, which is available upon request.

The rest of this paper is organized as follows: Section 2 reviews related work on mean-field multi-agent reinforcement learning and offline reinforcement learning for the single agent settings; Section 3 introduces our problem setup of the mean-field MARL regime; Section 4 introduces our proposed SAFARI algorithm; Section 5 establishes the theoretical guarantees for SAFARI; Section 6 presents numerical experiments on the multi-agent particle cooperative navigation scenario; Section 7 draws a brief conclusion.

2 Related Work

• **Mean-Field MARL**. Existing literature has proposed various mean-field approximation approaches to model the population behavior of the agents for MARL with a large number, even infinitely many homogenous agents. Yang et al. (2017) investigate a mean-field game with deterministic linear state transitions, and reformulate it as a mean-field MDP, where the mean-field state lies in finite-dimensional probability simplex. Yang et al. (2018) propose a mean-field approximation approach over actions, which approximates the interaction between any given agent and the population by the interaction between the agent's action and the averaged actions of its neighboring agents. Such an averaging approach over the local actions, however, is only applicable when a sparse graph over agents is given, which requires extensive prior knowledge. Carmona et al. (2019) investigate a mean-field MDP from the perspective of mean-field control. As the mean-field state lies in a probability simplex and continuous in nature, they propose to discretize the joint state-action space such that conventional RL algorithms can be applied. Wang et al. (2020) investigate a mean-field MDP motivated by permutation invariance. They require a central controller managing the actions of all the agents, and therefore is restricted to handling the curse of many agents from the exponential blowup of joint state space. More recently, Li et al. (2021) investigate a similar mean-field MDP, which allows agents to make their own local actions without resorting to a centralized controller. All these methods focus on the online settings. In comparison, our proposed SAFARI algorithm and theory focus on the offline settings.

• **RL for Mean-Field Game.** Our work is also related to the literature that studies RL methods for mean-field games (Huang et al., 2003; Lasry and Lions, 2006a,b; Huang et al., 2007). Such a game can be viewed as the infinite-agent limit of general-sum Markov game with homogeneous agents, and the aggregated effect of the other agents is also summarized as a mean-field state. In contrast to mean-field MARL, the solution concept of mean-field game is the Nash equilibrium, which corresponds to a pair of a local policy π^* of the representative agent and a mean-field state d^* satisfying the following two properties: (i) when the mean-field state is set to d^*, π^* is the optimal policy of the representative agent; and (ii) when all agents adopt π^*, the resulting mean-field state is d^*. Recently, there are many recent works developing RL methods for solving mean-field games. See, e.g., Guo et al. (2019, 2020b,a); Fu et al. (2019); Anahtarcı et al. (2019); Anahtarci et al. (2020); Anahtarcı et al. (2020); Perrin et al. (2020); Elie et al. (2020); uz Zaman et al. (2020); Cui and Koeppl (2021) and the references therein. Most of these methods adopts a double-loop structure, where the inner loop finds the optimal local policy given the current mean field state and the outer loop updates the mean-field states. Moreover, these works often assume the data distribution is well-explored with either a generative model (Azar et al., 2012) or bounded concentrability coefficients (Munos, 2007). Our mean-field MARL problem is similar to the inner-loop problem of finding the optimal local policy in mean-field games. In contrast to these existing works, our algorithm and theory can be applied to datasets that are possibly not well-explored. Moreover, as mean-field MARL and mean-field games are different models, our work is not directly comparable to these works.

• **Offline Single-Agent RL**. Our work is also closely related to the literature on offline single-agent RL, which often focuses on either policy evaluation or policy optimization. In particular, in policy evaluation, the goal is to estimate the value function of a target policy, whereas in policy optimization, we aims to learn the optimal policy, which can be achieved via estimating the optimal value function. For both these tasks, in the offline setting, due to the lack of continuing exploration (Szepesvári, 2010), the distribution shift (Levine et al., 2020) is a fundamental challenge. That is, the trajectories in the dataset and those induced by the target policy or the optimal policy might have diverse distributions. Such a challenge is further exacerbated when function approximators are adopted to represent the desired value functions. To overcome such a challenge, most of the existing theoretical works imposes certain well-exploration assumptions on the dataset. Some of commonly made assumptions include uniformly lower bounded visitation measure of the behavior policy, uniformly upper bounded importance sampling ratio, and bounded concentrability coefficients. See, e.g., Antos et al. (2007, 2008); Munos and Szepesvári (2008); Farahmand et al. (2010, 2016); Scherrer et al. (2015); Jiang and Li (2016); Thomas and Brunskill (2016); Farajtabar et al. (2018); Liu et al. (2018); Xie et al. (2019); Nachum et al. (2019a,b); Tang et al. (2019); Zhang et al. (2020b); Chen and Jiang (2019); Kallus and Uehara (2019, 2020); Jiang and Huang (2020); Uehara et al. (2020); Duan et al. (2020); Yin and Wang (2020); Yin et al. (2020); Nachum and Dai (2020); Yang et al. (2020a); Fu et al. (2020b); Fan et al. (2020); Xie and Jiang (2020a,b); Liao et al. (2020); Zhang et al. (2020a); Ren et al. (2021) and the references therein.

However, in practice, such assumptions on the dataset often fail to hold (Fujimoto et al., 2019; Agarwal et al., 2020; Fu et al., 2020a; Gulcehre et al., 2020). In light of this, there is a line of recent works that proposes various pessimism-based offline single-agent RL algorithms with empirical evidence or theoretical guarantees (Yu et al., 2020; Kidambi et al., 2020; Kumar et al., 2020; Liu et al., 2020b; Buckman et al., 2020; Jin et al., 2020b; Xiao et al., 2021). In particular, Liu et al. (2020b) propose a regularized variant of fitted Q-iteration (Antos et al., 2007, 2008; Munos and Szepesvári, 2008), which is shown to attain the optimal policy within a restricted policy class without assuming the dataset is well-explored. Moreover, with an arbitrary dataset, Buckman et al. (2020); Jin et al. (2020b); Xiao et al. (2021) identify the critical role of pessimism in achieving offline sample efficiency. Among these works, our work is particularly related to Jin et al. (2020b), which develops a pessimistic variant of the value iteration algorithm with finite-dimensional linear function approximation. In comparison, our SAFARI algorithm extends such an algorithm to mean-field MARL and we propose to employ RKHS mean embedding for handling the difference between finite-agent empirical mean-field state and its infinite-agent counterpart. Moreover, our algorithm and analysis involve infinite-dimensional RKHS, which strictly generalizes those in Jin et al. (2020b).

Notation: Given a space $\mathcal{X}$, we denote $\mathcal{M}(\mathcal{X})$ as the collection of probability distributions supported on $\mathcal{X}$. Let $u, v, w \in \mathcal{H}$ be elements in a Hilbert space, we denote $\langle u, v\rangle$ as the inner product, and $u \otimes v$ as the outer product satisfying $(u \otimes v)w = u \langle v, w\rangle$. For a scalar a, we denote $\{a\}^+ = \max\{0, a\}$. We use $\mathcal{O}(\cdot)$ to hide absolute constants and log factors.

3 Mean-Field Multi-Agent RL

We consider a Multi-Agent Reinforcement Learning (MARL) problem with $m+1$ agents and time horizon H. For the i-th agent (also known as the Representative Agent (RA)), at step h, we denote $s_{i,h} \in \mathcal{S}$ and $a_{i,h} \in \mathcal{A}$ as its state and action, respectively. We assume $\mathcal{S}$ and $\mathcal{A}$ are compact.

Different from single agent RL problem, the transition kernel, reward function, and policy of a representative agent in MARL depend not only on its individual state, but the states of m other agents. Furthermore, we assume that the interaction of the representative agent to the other agents is permutation invariant, i.e., the influence of all the other agents is modeled using the empirical distribution of states $\widehat{d}_{s,h} = \frac{1}{m}\sum_{j\neq i}^{m} \delta_{s_{j,h}} \in \mathcal{M}(\mathcal{S})$. To this end, we define the transition kernel $p_h : \mathcal{S} \times \mathcal{M}(\mathcal{S}) \times \mathcal{A} \mapsto \mathcal{M}(\mathcal{S})$, the (deterministic) reward function $r_h : \mathcal{S} \times \mathcal{M}(\mathcal{S}) \times \mathcal{A} \mapsto \mathbb{R}$, and the policy $\pi_h : \mathcal{S} \times \mathcal{M}(\mathcal{S}) \mapsto \mathcal{M}(\mathcal{A})$ all depending on a "meta state" denoted as $\widehat{\omega}_h = (s_{i,h}, \widehat{d}_{s,h}) \in \mathcal{S} \times \mathcal{M}(\mathcal{S})$. For simplicity, we denote $\Omega = \mathcal{S} \times \mathcal{M}(\mathcal{S})$ as the meta state space.

Remark 1. The empirical distribution of states $\widehat{d}_{s,h}$ is naturally permutation invariant and evolves according to the transition kernel p_h and policy π_h. To see this, suppose each agent takes the same policy π_h at step h. Then at step $h+1$, the state $s_{h+1,j}$ of the j-th agent is sampled from the distribution $p_h(\cdot \mid s_{h,j} \times \widehat{d}_{s,h}, a_{h,j})$, where $a_{h,j}$ is determined by policy $\pi_h(\cdot \mid s_{h,j} \times \widehat{d}_{s,h})$. Collecting m states $s_{h+1,j}$ for $j \neq i$ induces the empirical distribution of states $\widehat{d}_{s,h+1}$.

We now define several important notions in MARL. Given a policy π, the value function $V_h^\pi : \Omega \mapsto \mathbb{R}$ at step $h \leq H$ for a representative agent is

$$V_h^\pi(\omega) = \mathbb{E}_\pi\left[\sum_{i=h}^{H} r_i(\omega_i, a_i) \mid \omega_h = \omega\right], \tag{1}$$

where $\mathbb{E}_\pi$ denotes the expectation over the randomness in trajectories induced by policy π. The action-value function (Q-function) $Q_h^\pi : \Omega \times \mathcal{A} \mapsto \mathbb{R}$ is defined as

$$Q_h^\pi(\omega, a) = \mathbb{E}_\pi\left[\sum_{i=h}^{H} r_i(\omega_i, a_i) \mid \omega_h = \omega, a_h = a\right].$$

By definition, V_h^π and Q_h^π are related via $V_h^\pi(\omega) = \int_{\mathcal{A}} Q_h^\pi(\omega, a)\pi(a|\omega)da \overset{\triangle}{=} \langle Q_h^\pi, \pi\rangle_{\mathcal{A}}$. Next, we define the Bellman operator and conditional transition operator. At each step $h \leq H$, the Bellman operator denoted as $\mathbb{B}_h$ is

$$\begin{aligned}(\mathbb{B}_h g)(\omega, a) &= \mathbb{E}\left[r_h(\omega_h, a_h) + g(\omega_{h+1}) \mid \omega_h = \omega, a_h = a\right] \\ &= r_h(\omega, a) + (\mathbb{P}_h g)(\omega, a),\end{aligned} \tag{2}$$

where g is a function defined on Ω, and $\mathbb{P}_h$ is referred to as the conditional transition operator.

Mean-Field MARL As the number of agents goes to infinity, the empirical distribution of states $\widehat{d}_s$ converges to a (continuous) limit d_s. Then the mean-field MARL problem for a representative agent is defined as a tuple $(\Omega, \mathcal{A}, H, P, r)$, where Ω and $\mathcal{A}$ are the meta state space and action space, respectively, H is the horizon, $P = \{p_h\}_{h=1}^H : \Omega \times \mathcal{A} \mapsto \mathcal{M}(\mathcal{S})$ is the transition kernel, and $r = \{r_h\}_{h=1}^H$ is the reward function defined on $\Omega \times \mathcal{A}$. Following Remark 1, the transition of d_s is also determined by $P = \{p_h\}_{h=1}^H$.

To tackle the infinite-dimensional joint distribution of states, we embed the meta state-action space $\Omega \times \mathcal{A}$ into a reproducing kernel Hilbert space (RKHS). Specifically, denote $\Xi = \mathcal{S} \times \mathcal{S} \times \mathcal{A}$ and let $K : \Xi \times \Xi \mapsto \mathbb{R}$ be a symmetric positive kernel. The corresponding feature mapping of kernel k is denoted as ψ, which verifies $\langle \psi(\cdot), \psi(\cdot)\rangle = K(\cdot, \cdot)$ can be infinite dimensional. For any $(\omega, a) \in \Omega \times \mathcal{A}$, we define mean embedding as

$$\mu(\omega, a) = \mathbb{E}_{s' \sim d_s}[\psi(s, s', a)]. \tag{3}$$

Based on the embedding, we parameterize the reward r_h and Markov transition p_h as linear functionals of $\mu(\omega, a)$ in RKHS $\mathcal{H}_K$ induced by kernel K, i.e.,

$$r_h(\omega, a) = \langle \mu(\omega, a), \theta_h\rangle, \quad p_h(\omega' \mid \omega, a) = \langle \mu(\omega, a), v_h(\omega')\rangle, \tag{4}$$

where θ_h, v_h are understood as "weights" and have bounded Hilbert norm (see Assumption 3). Such a parameterization encodes a rich family of functions, once the kernel is universal (Wang et al., 2020). By the definition of Q-function and value function, we can show that the Bellman operator can also be parameterized in $\mathcal{H}_K$.

Proposition 1. Suppose the reward function r_h and the transition kernel p_h is parameterized in $\mathcal{H}_K$ by (4) for $h = 1, \ldots, H$. Then for any $g : \Omega \mapsto \mathbb{R}$, the Bellman operator $(\mathbb{B}_h g)$ and conditional transition operator $(\mathbb{P}_h g)$ defined in (2) can be written as

$$(\mathbb{B}_h g)(\omega, a) = \langle \mu(\omega, a), w_g \rangle \,, \quad (\mathbb{P}_h g)(\omega, a) = \langle \mu(\omega, a), w_g + \theta_h \rangle \,,$$

where w_g depends on the function g.

The proof is provided in Appendix C.1, which follows from pure algebraic manipulation. From the perspective of policy learning in mean-field MARL, Proposition 1 motivates us to estimate the Bellman operator $\mathbb{B}_h$ in $\mathcal{H}_K$, and then optimize the estimated Q-function to obtain a policy. We introduce the detailed learning procedure in Section 4 (Algorithm 1).

4 Offline Pessimistic Value Iteration

In this section, we introduce our dataset and learning algorithm. We collect multiple trajectories of a representative agent in a mean-field MARL problem. Here the mean-field state distribution d_s is prohibitive to trace. Instead, we only independently observe the states of a finite number of agents. Accordingly, the batched dataset $\mathcal{D}_{N,H}$ consists of N trajectories of length H, within which the n-th sequence is $\tau_n = \left\{(s_h^n \in \mathcal{S}^{m+1}, a_h^n \in \mathcal{A}, r_h^n \in \mathbb{R})\right\}_{h=1}^H$. Without loss of generality, we assume $s_{h,0}$ is the state of the representative agent, and the reward function is bounded by 1, i.e., $|r_h(\omega, a)| \leq 1$ for any $\omega \in \Omega, a \in \mathcal{A}$. The collected trajectories are generated by some unknown behavior policy.

Recall $\widehat{d}_{s_h^n} = \frac{1}{m} \sum_{j=1}^m s_{h,j}^n$ is the empirical state distribution induced by s_h^n. (We slightly alter the notation to emphasize the empirical distribution is generated by the collection of m states $s_{h,1:m}^n$, while in the previous context, we use a general purpose notation $\widehat{d}_{s,h}$.) We denote $\widehat{\omega}_h^n = s_{h,0}^n \times \widehat{d}_{s_h^n}$, and compute the empirical mean embedding of $(\widehat{\omega}_h^n, a_h^n)$ as

$$\mu(\widehat{\omega}_h^n, a_h^n) = \mathbb{E}_{s' \sim \widehat{d}_{s_h^n}}[\psi(s_{h,0}^n, s', a_h^n)] = \tfrac{1}{m} \textstyle\sum_{j=1}^m \psi(s_{h,0}^n, s_{h,j}^n, a_h^n).$$

Under mild conditions, the empirical mean embedding $\mu(\widehat{\omega}_h^n, a_h^n)$ concentrates around the infinite agent mean embedding $\mu(\omega_h^n, a_h^n)$ defined in (3), where ω_h^n is the infinite agent meta state. See a detailed error quantification in Lemma 3.

Pessimistic Value Iteration Our goal is to learn an optimal policy to be deployed for all the agents based on the experience data of the representative agent. The idea is to estimate the Q-function at each time step in the RKHS $\mathcal{H}_K$, and then optimize the Q-function to obtain an optimal policy. In more detail, at step $h \leq H$, we estimate Bellman operator by optimizing the empirical mean squared Bellman error

$$(\widehat{\mathbb{B}}_h \widehat{V}_{h+1}) = \underset{f}{\operatorname{argmin}} \sum_{n=1}^N \Big(f(\mu(\widehat{\omega}_h^n, a_h^n)) - r_h^n - \widehat{V}_{h+1}(\widehat{\omega}_{h+1}^n) \Big)^2 + \lambda \left\| f \right\|_{\mathcal{H}}^2 , \tag{5}$$

where $\lambda \geq 1$ controls the regularization strength, $\widehat{V}$ is the estimated value function, and $\|\cdot\|_{\mathcal{H}}$ denotes the Hilbert norm.

The solution to (5) can be written in a closed form. For notational simplicity, we define

$$K((\omega, a), \cdot) = \mathbb{E}_{s' \sim d_s}[K((s, s', a), \cdot)] \quad \text{with} \quad \omega = s \times d_s.$$

Then we denote the Gram matrix $K_h \in \mathbb{R}^{N \times N}$ as

$$[K_h]_{\ell,\ell'} = K((\widehat{\omega}_h^\ell, a_h^\ell), (\widehat{\omega}_h^{\ell'}, a_h^{\ell'})) \triangleq \mathbb{E}_{s_1 \sim \widehat{d}_h^\ell, s_2 \sim \widehat{d}_h^{\ell'}} \langle \psi(s_{h,0}^\ell, s_1, a_h^\ell), \psi(s_{h,0}^{\ell'}, s_2, a_h^{\ell'}) \rangle$$

for $\ell, \ell' = 1, \ldots, N$. Meanwhile, for any (ω, a), we denote feature vector $\phi_h(\omega, a) = \left[K((\widehat{\omega}_h^1, a_h^1), (\omega, a)), \ldots, K((\widehat{\omega}_h^N, a_h^N), (\omega, a))\right]^\top \in \mathbb{R}^N$. Then the estimated Bellman operator

$\widehat{\mathbb{B}}_h \widehat{V}_{h+1}$ can be written as

$$
\begin{aligned}
&(\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(\omega, a) = \phi_h(\omega, a)^\top \widehat{\alpha}_h \\
\text{with} \quad & \widehat{\alpha}_h = (K_h + \lambda I)^{-1} [r_h^1 + \widehat{V}_{h+1}(\widehat{\omega}_{h+1}^1), \ldots, r_h^N + \widehat{V}_{h+1}(\widehat{\omega}_{h+1}^N)]^\top,
\end{aligned} \tag{6}
$$

We summarize the proposed SAFARI algorithm in Algorithm 1.

Algorithm 1 Pessimistic Mean-Field Value Iteration (SAFARI)

Input: Dataset $\mathcal{D}_{N,H}$, coefficient β, regularization coefficient λ.
Initialize: Set $\widehat{V}_{H+1} = 0$.
for $h = H, H-1, \ldots, 1$ **do**
  Compute $\Lambda_h = K_h + \lambda I$.
  Estimate $\widetilde{Q}_h(\omega, a) \triangleq (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(\omega, a) = \phi_h(\omega, a)^\top \widehat{\alpha}_h$ as in (6).
  Set $\Gamma_h(\omega, a) = \beta \cdot \lambda^{-1/2} \left(K((\omega, a), (\omega, a)) - \phi_h(\omega, a)^\top \Lambda_h^{-1} \phi_h(\omega, a)\right)^{1/2}$.
  Let $\widehat{Q}_h(\omega, a) = \min\{\widetilde{Q}_h(\omega, a) - \Gamma_h(\omega, a), H - h + 1\}^+$.
  Optimal policy $\widehat{\pi}_h = \operatorname{argmax}_\pi \langle \widehat{Q}_h(\omega, \cdot), \pi(\cdot \mid \omega) \rangle_{\mathcal{A}}$.
  Set $\widehat{V}_h(\omega) = \langle \widehat{Q}_h(\omega, \cdot), \widehat{\pi}_h(\cdot \mid \omega) \rangle_{\mathcal{A}}$.
end for
Output: Estimated Q-function $\widehat{Q}_h$, value function $\widehat{V}_h$, and optimal policy $\widehat{\pi}_h$ for $h = 1, \ldots, H$.

The quantity Γ_h quantifies the uncertainty in estimating the Bellman operator $\mathbb{B}_h \widehat{V}_{h+1}$ using kernel ridge regression. We subtract Γ_h for estimating the Bellman operator to account for the spurious correlation in the experience data (see Technical Overview following Theorem 1 for a detailed explanation). We truncate $\widehat{Q}_h$ at $H - h + 1$, since the reward function is bounded by 1.

5 Suboptimality of Policy Learned by SAFARI

We investigate the performance of the optimal policy $\widehat{\pi}$ learned by Algorithm 1. Before we proceed, we state the following assumptions.

Assumption 1 (Boundedness of Kernel)**.** Kernel $K(\cdot, \cdot)$ is bounded, i.e., without loss of generality, we assume $\sup_{\xi \in \Xi} |K(\xi, \xi)| \leq 1$.

By Cauchy-Schwarz inequality, Assumption 1 implies for any $\xi_1, \xi_2 \in \Xi$, $K(\xi_1, \xi_2) \leq \sqrt{K(\xi_1, \xi_1) K(\xi_2, \xi_2)} \leq 1$. Such an assumption holds for a rich family of commonly used kernels, e.g., RBF kernel and Laplacian kernel, and is a standard assumption in literature (Caponnetto and De Vito, 2007; Muandet et al., 2012).

The second assumption characterizes the spectrum of kernel K. We first introduce the integral operator induced by kernel K. Let $f : \Xi \mapsto \mathbb{R}$ be a square-integrable function. Then we define the integral operator $\mathcal{T}_K$ as

$$
(\mathcal{T}_K f)(\xi) = \int K(\xi, x) f(x) dx \quad \text{for} \quad \xi \in \Xi.
$$

By Mercer's theorem (Hearst et al., 1998), $\mathcal{T}_K$ has corresponding positive eigenvalues σ_i and eigenfunctions ν_i. Then the kernel K admits a decomposition

$$
K(\xi_1, \xi_2) = \sum_{i=1}^{\infty} \sigma_i \nu_i(\xi_1) \nu_i(\xi_2).
$$

Assumption 2 (Spectrum of Kernel)**.** The eigenvalue σ_i satisfies one of the following three conditions:

1. *(Finite Spectrum).* There exists a positive integer γ, such that $\sigma_i = 0$ for all $i > \gamma$.

2. *(Exponential Decay).* There exist positive constants C_1, C_2 and exponent $\gamma > 0$ such that $\sigma_i \leq C_1 \exp(-C_2 i^\gamma)$.

3. *(Polynomial Decay)*. There exists a positive constant C and exponent $\gamma \geq 3 + \mathcal{O}(\frac{1}{d})$ such that $\sigma_i \leq C i^{-\gamma}$, where d is the dimension of $\mathcal{S} \times \mathcal{S} \times \mathcal{A}$.

Furthermore, in *(Exponential Decay)* and *(Polynomial Decay)*, we assume the eigenfunction ν_i is uniformly bounded, i.e., $\sup_i \|\nu_i\|_\infty \leq 1$.

As we will show in our theory, the decay rate of the spectrum significantly influences the performance of the proposed SAFARI algorithm. We give examples to better interpret the three categories above. In *(Finite Spectrum)* case, by (4), the reward function and transition kernel is a linear function of a finite dimensional feature map. Such a parameterization is satisfied by linear MDP (Yang and Wang, 2019; Jin et al., 2020a). In *(Exponential Decay)* and *(Polynomial Decay)* cases, the feature map is infinite dimensional. For example, RBF kernel belongs to *(Exponential Decay)* case, while Laplacian kernel and neural tangent kernel belong to *(Polynomial Decay)* case. We assume $\gamma \geq 3 + \mathcal{O}(1/d)$ in *(Polynomial Decay)* for technical simplicity, yet it is not restrictive: Laplacian kernel and neural tangent kernel both have a polynomial decay rate of $\gamma = d$ (Bietti and Bach, 2020).

The last assumption imposes some regularity on the reward function and transition probabilities.

Assumption 3 (Boundedness)**.** The weights θ_h and v_h in reward function r_h and Markov transition kernel p_h are bounded for any $h = 1, \ldots, H$, respectively, i.e., $\|\theta_h\|_{\mathcal{H}} \leq 1$ and $\int_\Omega \|v_h(x)\|_{\mathcal{H}}\, dx \leq \sqrt{d_{\text{eff}}}$, where $d_{\text{eff}} = \sup_{K_h} \log\det(I + K_h/\lambda)$ is the effective dimension of $\mathcal{H}_K$ with supremum over all Gram matrix $K_h \in \mathbb{R}^{N \times N}$.

The effective dimension describes the complexity of $\mathcal{H}_K$ for parameterizing the MDP (Yang et al., 2020b), whose scale is closely related to the spectrum of kernel K. In the special case of K having a γ-finite spectrum as in Assumption 2, we have $d_{\text{eff}} = \mathcal{O}(\gamma)$, which resembles the dimensionality of a finite dimensional Euclidean space.

We measure the pointwise suboptimality of the learned policy $\widehat{\pi}$. We define the global optimal policy by the recursion,

$$\pi_h^* = \operatorname*{argmax}_\pi \ \langle Q_h^*, \pi \rangle_{\mathcal{A}}, \quad \text{with} \quad Q_h^* = \mathbb{B}_h V_{h+1}^*, V_h^* = \langle Q_h^*, \pi_h^* \rangle_{\mathcal{A}}, \text{ and } V_{H+1}^* = 0.$$

Then the suboptimality of $\widehat{\pi}$ is given as

$$\texttt{SubOpt}(\widehat{\pi}; \omega) = V_1^{\pi^*}(\omega) - V_1^{\widehat{\pi}}(\omega).$$

Our main result is provided in the following theorem, which upper bounds $\texttt{SubOpt}(\widehat{\pi}; \omega)$.

Theorem 1. Suppose Assumption 1 – 3 hold. For any $\delta \in (0, 1)$, let $\widehat{\pi}_h$ be the policy returned by Algorithm 1 with

$$m \geq \log(2/\delta), \quad \lambda = 1, \quad \beta = \begin{cases} c \max\{d, \gamma\} H \sqrt{\log(\max\{d, \gamma\} HN/\delta)} & \textit{(Finite Spectrum)} \\ cH \sqrt{d \left(\log(HN/\delta)\right)^{1+2/\gamma}} & \textit{(Exponential Decay)} \\ cN^{\frac{d+1}{d+\gamma}} H \sqrt{d \log(HN/\delta)} & \textit{(Polynomial Decay)} \end{cases},$$

where d is the dimension of $\Xi = \mathcal{S} \times \mathcal{S} \times \mathcal{A}$ and c is some constant depending on C, C_1, C_2 and Lebesgue measure of Ξ. Then for any meta state ω, with probability at least $1 - \delta$ over the randomness of the dataset $\mathcal{D}_{N,H}$, we have

$$\texttt{SubOpt}(\widehat{\pi}; \omega) \leq 2 \sum_{h=1}^H \mathbb{E}_{\pi^*}[\Gamma_h(\omega_h, a_h) \mid \omega_1 = \omega].$$

Theorem 1 indicates that the suboptimality of learned policy depends on the uncertainty quantifier Γ_h. The scale of Γ_h depends on how well the collected data explore the state-action space. Moreover, from a Bayesian learning perspective, Γ_h measures the eliminated uncertainty in estimating the Bellman operator given dataset $\mathcal{D}_{N,H}$ (Jin et al., 2020b). To better understand the convergence of `SubOpt`, we specialize Theorem 1 under a weak data coverage assumption.

Assumption 4 (Weak Coverage)**.** Suppose the dataset is collected under some behavior policy $\bar{\pi}$ such that there exists a constant $c_{\min} > 0$ satisfying

$$\inf_{\|f\|_{\mathcal{H}}=1} \langle f, \mathbb{E}_{\bar{\pi}}[\mu(\omega_h, a_h) \otimes \mu(\omega_h, a_h)] f \rangle \geq c_{\min} \quad \text{for any} \quad h = 1, \ldots, H.$$

Recall that μ is the mean embedding in $\mathcal{H}_K$.

Assumption 4 says that the operator $\mathbb{E}_{\bar{\pi}}[\mu(\omega_h, a_h) \otimes \mu(\omega_h, a_h)]$ is positive definite. Intuitively, this requires that the collected data relatively well spread over the state-action space. We present the following Corollary providing a concrete convergence rate of `SubOpt`.

Corollary 1. Under the setting in Theorem 1, we additionally assume Assumption 4 holds. Then for $N \geq \Omega(\log(d_{\text{eff}} H/\delta))$ sufficiently large, with probability $1 - \delta$, we have

$$\texttt{SubOpt}(\widehat{\pi}; \omega) = \mathcal{O}\left(H^2 d_{\text{eff}} \sqrt{\frac{\log(d_{\text{eff}} HN/\delta)}{N}}\right).$$

Here d_{eff} is the effective dimension of RKHS $\mathcal{H}_K$, which takes value

$$d_{\text{eff}} = \begin{cases} \max\{d, \gamma\} \log N & \textit{(Finite Spectrum)} \\ d(\log N)^{1+1/\gamma} & \textit{(Exponential Decay)} \\ dN^{\frac{d+1}{d+\gamma}} \log N & \textit{(Polynomial Decay)} \end{cases}.$$

Impact of Kernel Spectrum The spectrum of kernel K significantly influences the performance of the learned policy. In *(Finite Spectrum)* case, the effective dimension scales linearly with dimension d and γ, and `SubOpt` converges at a rate of $\mathcal{O}(H^2 \max\{d, \gamma\}/\sqrt{N})$, which recovers the result of Corollary 4.5 in Jin et al. (2020b) on linear MDP. In *(Exponential Decay)* case, the convergence rate is $\mathcal{O}(H^2 d(\log N)^{1+1/\gamma}/\sqrt{N})$, which is similar to *(Finite Spectrum)* case with additional logarithmic dependence on N. However, in *(Polynomial Decay)* case, the convergence rate is considerably slower, and relies heavily on the decay rate γ. Consider, for instance, Laplacian kernel and NTK, whose spectrum decays with $\gamma = d$. Then `SubOpt` converges at a rate of $\mathcal{O}(H^2 dN^{-\frac{1}{2d}} \log N)$, which suffers from the curse of dimensionality without further assumptions on data.

No Curse of Many Agents The convergence of `SubOpt` does not suffer from the curse of many agents. In particular, both Theorem 1 and Corollary 1 only impose a mild requirement on the number m of neighboring agents to be sampled. This is due to the permutation invariance in mean-field MARL, since the interactive influence of neighboring agents are captured by the distribution of states.

Technical Overview We briefly discuss the proof of Theorem 1 and Corollary 1. The full proof is deferred to Appendix A and B. We first decompose `SubOpt` into three terms (see Lemma 1):

$$\texttt{SubOpt}(\pi; \omega) = \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3.$$

Here $\mathcal{E}_1 = \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}[\widehat{Q}_h(\omega_h, a_h) - (\mathbb{B}_h \widehat{V}_{h+1})(\omega_h, a_h) \mid \omega_1 = \omega]$ reflects the uncertainty in estimating the Bellman operator. Note that the evaluating trajectory is generated by the learned policy $\widehat{\pi}$, which has spurious correlation with the estimated Bellman operator; $\mathcal{E}_2 = \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[(\mathbb{B}_h \widehat{V}_{h+1})(\omega_h, a_h) - \widehat{Q}_h(\omega_h, a_h) \mid \omega_1 = \omega]$ is the estimation error of Bellman operator again, yet it is evaluated by a trajectory generated by π^*. Compared to $\mathcal{E}_1$, $\mathcal{E}_2$ does not suffer from the spurious correlation between the learned policy and the estimated Bellman operator. Lastly, $\mathcal{E}_3 = \sum_{h=1}^{H} \mathbb{E}_{\pi}[\langle \widehat{Q}_h(\omega_h, \cdot), \pi_h^*(\cdot \mid \omega_h) - \widehat{\pi}_h(\cdot \mid \omega_h) \rangle_{\mathcal{A}} \mid \omega_1 = \omega]$ is the optimization error. By the optimality of $\widehat{\pi}$, we immediately have $\mathcal{E}_3 \leq 0$.

In order to tackle $\mathcal{E}_1$ and $\mathcal{E}_2$, we properly choose Γ_h so that the event $E = \{|\mathbb{B}_h \widehat{V}_{h+1} - \widehat{\mathbb{B}}_h \widehat{V}_{h+1}| \leq \Gamma_h\}$ happens with high probability. In fact, Γ_h is understood as the uncertainty quantifier of estimating $\mathbb{B}_h \widehat{V}_{h+1}$ with high confidence $1 - \delta$. Then we can show $\mathcal{E}_1 \leq 0$ conditioned on event E, meanwhile $\mathcal{E}_2 \leq 2 \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\Gamma_h(\omega_h, a_h) \mid \omega_1 = \omega]$. To this end, we reduce the upper bound of `SubOpt` to bounding the uncertainty quantifier Γ_h, which allows us to leverage statistical tools. In particular, Γ_h consists of two types of statistical error: 1) covariate concentration error on mean embedding, i.e., finite agent empirical embedding $\mu(\widehat{\omega}_h^n, a_h^n)$ concentration error with respect to population counterpart $\mu(\omega_h^n, a_h^n)$; 2) regression error in Bellman operator estimation. We bound 1) by concentration of empirical means in Hilbert spaces (see Lemma 3). In bounding 2), we exploit the closed form solution of kernel ridge regression and concentration of self-normalizing processes (see Lemma 5).

6 Numerical Experiment

We perform experiments on the multi-agent particle environment (MPE, Lowe et al. (2017)), a popular benchmark used in prior work (Mordatch and Abbeel, 2018; Liu et al., 2020a). Here, we consider the *cooperative navigation* scenario, where N agents must spread to cooperatively cover N landmarks across the map. Each agent is able to observe information about the k closest landmarks and agents, and receives a global reward $r = -\sum_{i=1}^{N} \min_{j\in[N]} \|y_i - x_j\|_2$, where $x_i \in \mathbb{R}^2$ and $y_i \in \mathbb{R}^2$ are agent and landmark positions, respectively. Implementation of all environments follows from the official codebase of Liu et al. (2020a). See hyperparameter choices and more details in Appendix E. Sample code is also available at `https://github.com/wange011/offline-pessimistic`.

Data Generation To receive optimal reward on *cooperative navigation*, individual agent must learn to coordinate their behaviors to each cover a different landmark. As a result, we generate data for the offline setting by training a MARL policy and collecting experience data after convergence. We use *counterfactual multi-agent policy gradients* (COMA) to address the problem of *credit assignment* by learning a joint critic that marginalizes out an individual agent's action with a counterfactual baseline (Foerster et al., 2018). This, in turn, allows the agent-level policies to learn sufficient coordination by evaluating their individual impact on the team reward. Both the policy and critic networks are implemented as traditional MLPs, with 64 and 512 nodes in a single hidden layer, respectively, and we use parameter sharing for policy networks. To sanity check the performance of COMA, we train the *individual actor-critic* (IAC) algorithm (Konda and Tsitsiklis, 2000), which applies the policy gradient to train independent actor-critics. Given the lack of an in-built coordination mechanism, IAC is expected to perform suboptimally on multi-agent settings.

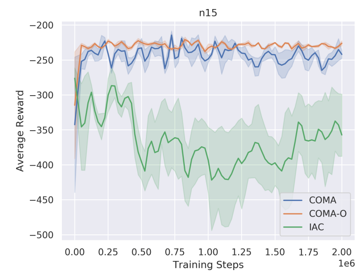
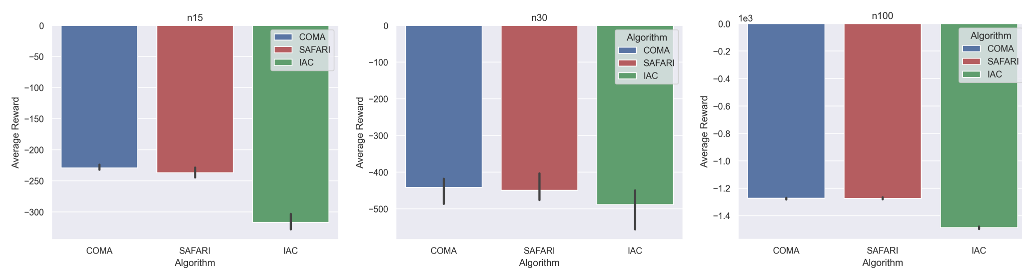


Figure 1: Training reward on the 15 agent environment.

As all agents take the same action in the mean-field MARL formulation, COMA produces experiences by selecting the action that corresponds with the plurality vote (mode) of individual agent policy outputs. However, to demonstrate that this does not greatly inhibit convergence behavior, we train IAC and the original COMA implementation, labeled COMA-O, without this restriction. As demonstrated in Figure 1, with error bar computed over 3 independent random seeds, COMA-O performs the best. It is worth noting that COMA receives slightly lower rewards yet still performing significantly better than IAC with the same number of learnable parameters.



(a) 15 agents (b) 30 agents (c) 100 agents

Figure 2: Average reward after training. COMA and IAC are evaluated off loaded pre-trained models.

In Figure 2, we implement our SAFARI algorithm with varying number of agents on $n = 500$ sample episodes of experience data. We evaluate the performance over a horizon $H = 50$ on 3 different random seeds. We observe that SAFARI is able to perform comparably to COMA in settings with $m = 15$, 30, and 100 agents. Due to mean-field permutation invariance, we see that the performance gap between SAFARI and COMA does not widen as the number of agents increases, a behavior that is normally expected given the exponential growth of the joint state-action space.

7 Conclusion

This paper proposes a SAFARI (Pessimistic Mean-Field Value Iteration) algorithm in offline mean-field MARL. We prove a suboptimality bound $\mathcal{O}(H^2 d_{\text{eff}}/\sqrt{N})$, and provide concrete rate of convergence under a weak data coverage assumption. The suboptimality bound is free of the curse of many agents due to the permutation invariance in mean-field formulation. We also extend to the online setting in a longer technical report version.

Acknowledgment

Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their supports. Zhuoran Yang acknowledges Simons Institute (Theory of Reinforcement Learning).

References

ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Improved algorithms for linear stochastic bandits. In *NIPS*, vol. 11.

AGARWAL, R., SCHUURMANS, D. and NOROUZI, M. (2020). An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*.

ALTUN, Y. and SMOLA, A. (2006). Unifying divergence minimization and statistical inference via convex duality. In *International Conference on Computational Learning Theory*. Springer.

ANAHTARCI, B., KARIKSIZ, C. D. and SALDI, N. (2019). Fitted q-learning in mean-field games. *arXiv preprint arXiv:1912.13309*.

ANAHTARCI, B., KARIKSIZ, C. D. and SALDI, N. (2020). Q-learning in regularized mean-field games. *arXiv preprint arXiv:2003.12151*.

ANAHTARCI, B., KARIKSIZ, C. D. and SALDI, N. (2020). Value iteration algorithm for mean-field games. *Systems & Control Letters*, **143** 104744.

ANTOS, A., SZEPESVÁRI, C. and MUNOS, R. (2007). Fitted Q-iteration in continuous action-space MDPs. In *Advances in Neural Information Processing Systems*.

ANTOS, A., SZEPESVÁRI, C. and MUNOS, R. (2008). Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, **71** 89–129.

AZAR, M. G., MUNOS, R. and KAPPEN, B. (2012). On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*.

BERNER, C., BROCKMAN, G., CHAN, B., CHEUNG, V., DĘBIAK, P., DENNISON, C., FARHI, D., FISCHER, Q., HASHME, S., HESSE, C. ET AL. (2019). Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*.

BIETTI, A. and BACH, F. (2020). Deep equals shallow for relu networks in kernel regimes. *arXiv preprint arXiv:2009.14397*.

BUCKMAN, J., GELADA, C. and BELLEMARE, M. G. (2020). The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*.

CAI, Q., YANG, Z., JIN, C. and WANG, Z. (2020). Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*. PMLR.

CAPONNETTO, A. and DE VITO, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, **7** 331–368.

CARMONA, R., LAURIÈRE, M. and TAN, Z. (2019). Model-free mean-field reinforcement learning: mean-field mdp and mean-field q-learning. *arXiv preprint arXiv:1910.12802*.

CHEN, J. and JIANG, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*. PMLR.

CHOWDHURY, S. R. and GOPALAN, A. (2017). On kernelized multi-armed bandits. In *International Conference on Machine Learning*. PMLR.

CUI, K. and KOEPPL, H. (2021). Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

DUAN, Y., JIA, Z. and WANG, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*.

ELIE, R., PÉROLAT, J., LAURIÈRE, M., GEIST, M. and PIETQUIN, O. (2020). On the convergence of model free learning in mean field games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34.

FAN, J., WANG, Z., XIE, Y. and YANG, Z. (2020). A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control*.

FARAHMAND, A.-M., GHAVAMZADEH, M., SZEPESVÁRI, C. and MANNOR, S. (2016). Regularized policy iteration with nonparametric function spaces. *Journal of Machine Learning Research*, **17** 4809–4874.

FARAHMAND, A.-M., SZEPESVÁRI, C. and MUNOS, R. (2010). Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*.

FARAJTABAR, M., CHOW, Y. and GHAVAMZADEH, M. (2018). More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*.

FOERSTER, J., FARQUHAR, G., AFOURAS, T., NARDELLI, N. and WHITESON, S. (2018). Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32.

FU, J., KUMAR, A., NACHUM, O., TUCKER, G. and LEVINE, S. (2020a). D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*.

FU, Z., YANG, Z., CHEN, Y. and WANG, Z. (2019). Actor-critic provably finds nash equilibria of linear-quadratic mean-field games. In *International Conference on Learning Representations*.

FU, Z., YANG, Z. and WANG, Z. (2020b). Single-timescale actor-critic provably finds globally optimal policy. *arXiv preprint arXiv:2008.00483*.

FUJIMOTO, S., MEGER, D. and PRECUP, D. (2019). Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*.

GULCEHRE, C., WANG, Z., NOVIKOV, A., PAINE, T. L., COLMENAREJO, S. G., ZOLNA, K., AGARWAL, R., MEREL, J., MANKOWITZ, D., PADURARU, C. ET AL. (2020). RL Unplugged: Benchmarks for offline reinforcement learning. *arXiv preprint arXiv:2006.13888*.

GUO, X., HU, A., XU, R. and ZHANG, J. (2019). Learning mean-field games. *arXiv preprint arXiv:1901.09585*.

GUO, X., HU, A., XU, R. and ZHANG, J. (2020a). A general framework for learning mean-field games. *arXiv preprint arXiv:2003.06069*.

GUO, X., XU, R. and ZARIPHOPOULOU, T. (2020b). Entropy regularization for mean field games with learning. *arXiv preprint arXiv:2010.00145*.

HEARST, M. A., DUMAIS, S. T., OSUNA, E., PLATT, J. and SCHOLKOPF, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, **13** 18–28.

HUANG, M., CAINES, P. E. and MALHAMÉ, R. P. (2003). Individual and mass behaviour in large population stochastic wireless power control problems: centralized and nash equilibrium solutions. In *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, vol. 1. IEEE.

HUANG, M., CAINES, P. E. and MALHAMÉ, R. P. (2007). Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized $varepsilon$-Nash equilibria. *IEEE transactions on automatic control*, **52** 1560–1571.

JAQUES, N., GHANDEHARIOUN, A., SHEN, J. H., FERGUSON, C., LAPEDRIZA, A., JONES, N., GU, S. and PICARD, R. (2019). Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*.

JIANG, N. and HUANG, J. (2020). Minimax value interval for off-policy evaluation and policy optimization. In *Advances in Neural Information Processing Systems*.

JIANG, N. and LI, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*.

JIN, C., YANG, Z., WANG, Z. and JORDAN, M. I. (2020a). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*.

JIN, Y., YANG, Z. and WANG, Z. (2020b). Is pessimism provably efficient for offline RL? *arXiv preprint arXiv:2012.15085*.

KALLUS, N. and UEHARA, M. (2019). Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *arXiv preprint arXiv:1909.05850*.

KALLUS, N. and UEHARA, M. (2020). Doubly robust off-policy value and gradient estimation for deterministic policies. *arXiv preprint arXiv:2006.03900*.

KIDAMBI, R., RAJESWARAN, A., NETRAPALLI, P. and JOACHIMS, T. (2020). MOReL: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*.

KONDA, V. R. and TSITSIKLIS, J. N. (2000). Actor-critic algorithms. In *Advances in neural information processing systems*.

KUMAR, A., ZHOU, A., TUCKER, G. and LEVINE, S. (2020). Conservative Q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*.

LASRY, J.-M. and LIONS, P.-L. (2006a). Jeux à champ moyen. i–le cas stationnaire. *Comptes Rendus Mathématique*, **343** 619–625.

LASRY, J.-M. and LIONS, P.-L. (2006b). Jeux à champ moyen. ii–Horizon fini et contrôle optimal. *Comptes Rendus Mathématique*, **343** 679–684.

LEIBO, J. Z., ZAMBALDI, V., LANCTOT, M., MARECKI, J. and GRAEPEL, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*.

LEVINE, S., KUMAR, A., TUCKER, G. and FU, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

LI, Y., WANG, L., YANG, J., WANG, E., WANG, Z., ZHAO, T. and ZHA, H. (2021). Permutation invariant policy optimization for mean-field multi-agent reinforcement learning: A principled approach. *arXiv preprint arXiv:2105.08268*.

LIAO, P., QI, Z. and MURPHY, S. (2020). Batch policy learning in average reward Markov decision processes. *arXiv preprint arXiv:2007.11771*.

LIU, I.-J., YEH, R. A. and SCHWING, A. G. (2020a). Pic: permutation invariant critic for multi-agent deep reinforcement learning. In *Conference on Robot Learning*. PMLR.

LIU, Q., LI, L., TANG, Z. and ZHOU, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*.

LIU, Y., SWAMINATHAN, A., AGARWAL, A. and BRUNSKILL, E. (2020b). Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*.

LOWE, R., WU, Y., TAMAR, A., HARB, J., ABBEEL, P. and MORDATCH, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275*.

MANDEL, T., LIU, Y.-E., LEVINE, S., BRUNSKILL, E. and POPOVIC, Z. (2014). Offline policy evaluation across representations with applications to educational games. In *AAMAS*.

MORDATCH, I. and ABBEEL, P. (2018). Emergence of grounded compositional language in multi-agent populations. In *Thirty-second AAAI conference on artificial intelligence*.

MUANDET, K., FUKUMIZU, K., DINUZZO, F. and SCHÖLKOPF, B. (2012). Learning from distributions via support measure machines. *arXiv preprint arXiv:1202.6504*.

MUNOS, R. (2007). Performance bounds in ℓ_p-norm for approximate value iteration. *SIAM journal on control and optimization*, **46** 541–561.

MUNOS, R. and SZEPESVÁRI, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, **9** 815–857.

NACHUM, O., CHOW, Y., DAI, B. and LI, L. (2019a). DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*.

NACHUM, O. and DAI, B. (2020). Reinforcement learning via Fenchel-Rockafellar duality. *arXiv preprint arXiv:2001.01866*.

NACHUM, O., DAI, B., KOSTRIKOV, I., CHOW, Y., LI, L. and SCHUURMANS, D. (2019b). AlgaeDICE: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*.

PERRIN, S., PÉROLAT, J., LAURIÈRE, M., GEIST, M., ELIE, R. and PIETQUIN, O. (2020). Fictitious play for mean field games: Continuous time analysis and applications. *arXiv preprint arXiv:2007.03458*.

REN, T., LI, J., DAI, B., DU, S. S. and SANGHAVI, S. (2021). Nearly horizon-free offline reinforcement learning. *arXiv preprint arXiv:2103.14077*.

SCHERRER, B., GHAVAMZADEH, M., GABILLON, V., LESNER, B. and GEIST, M. (2015). Approximate modified policy iteration and its application to the game of Tetris. *Journal of Machine Learning Research*, **16** 1629–1676.

SHALEV-SHWARTZ, S., SHAMMAH, S. and SHASHUA, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*.

SRINIVAS, N., KRAUSE, A., KAKADE, S. M. and SEEGER, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

SZEPESVÁRI, C. (2010). *Algorithms for reinforcement learning*. Morgan & Claypool.

TANG, Z., FENG, Y., LI, L., ZHOU, D. and LIU, Q. (2019). Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*.

THOMAS, P. and BRUNSKILL, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*.

UEHARA, M., HUANG, J. and JIANG, N. (2020). Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*.

UZ ZAMAN, M. A., ZHANG, K., MIEHLING, E. and BAÇAR, T. (2020). Reinforcement learning in non-stationary discrete-time linear-quadratic mean-field games. In *IEEE Conference on Decision and Control (CDC)*. IEEE.

WANG, L., YANG, Z. and WANG, Z. (2020). Breaking the curse of many agents: Provable mean embedding q-iteration for mean-field reinforcement learning. In *International Conference on Machine Learning*. PMLR.

XIAO, C., WU, Y., LATTIMORE, T., DAI, B., MEI, J., LI, L., SZEPESVARI, C. and SCHUURMANS, D. (2021). On the optimality of batch policy optimization algorithms. *arXiv preprint arXiv:2104.02293*.

XIE, T. and JIANG, N. (2020a). Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*.

XIE, T. and JIANG, N. (2020b). $Q^\star$-approximation schemes for batch reinforcement learning: A theoretical comparison. *arXiv preprint arXiv:2003.03924*.

XIE, T., MA, Y. and WANG, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*.

YANG, J., YE, X., TRIVEDI, R., XU, H. and ZHA, H. (2017). Learning deep mean field games for modeling large population behavior. *arXiv preprint arXiv:1711.03156*.

YANG, L. and WANG, M. (2019). Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*.

YANG, M., NACHUM, O., DAI, B., LI, L. and SCHUURMANS, D. (2020a). Off-policy evaluation via the regularized Lagrangian. In *Advances in Neural Information Processing Systems*.

YANG, Y., LUO, R., LI, M., ZHOU, M., ZHANG, W. and WANG, J. (2018). Mean field multi-agent reinforcement learning. In *International Conference on Machine Learning*. PMLR.

YANG, Z., JIN, C., WANG, Z., WANG, M. and JORDAN, M. I. (2020b). On function approximation in reinforcement learning: Optimism in the face of large state spaces.

YIN, M., BAI, Y. and WANG, Y.-X. (2020). Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning. *arXiv preprint arXiv:2007.03760*.

YIN, M. and WANG, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*.

YU, T., THOMAS, G., YU, L., ERMON, S., ZOU, J., LEVINE, S., FINN, C. and MA, T. (2020). MOPO: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*.

ZHANG, J., KOPPEL, A., BEDI, A. S., SZEPESVÁRI, C. and WANG, M. (2020a). Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*.

ZHANG, K., YANG, Z. and BAŞAR, T. (2019). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*.

ZHANG, R., DAI, B., LI, L. and SCHUURMANS, D. (2020b). GenDICE: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*.

Supplementary Material for "Pessimism Meets Invariance: Provably Efficient Offline Mean-Field Multi-Agent RL"

A Proof of Theorem 1 — Offline Pessimistic Policy Learning

Throughout the proofs, we adopt the following convention on inner product and outer product in Hilbert space $\mathcal{H}_K$. Let $V = [v_1, \ldots, v_d]^\top$ and $U = [u_1, \ldots, u_d]$ be collections of elements in $\mathcal{H}_K$ ($v_i, u_j \in \mathcal{H}$). For any $w \in \mathcal{H}_K$, we denote $\langle V, w \rangle = [\langle v_1, w\rangle, \ldots, \langle v_d, w\rangle]^\top \in \mathbb{R}^d$, meanwhile, $\left\langle V^\top, w \right\rangle = \langle V, w \rangle^\top \in \mathbb{R}^d$ is a row vector. We also denote $\left\langle V^\top, U \right\rangle = \sum_{i=1}^d \langle v_i, u_i \rangle \in \mathbb{R}$, while $\left\langle V, U^\top \right\rangle \in \mathbb{R}^{d \times d}$ is a matrix. For outer product, we similarly denote $V^\top \otimes U = \sum_{i=1}^d v_i \otimes u_i$ and $V \otimes U^\top = [v_i \otimes u_j]_{i,j}$ as operators. Such a convention coincides with the standard vector algebra in finite dimensional spaces.

Proof. The full proof consists of four steps. In each step, we require several technical lemmas, whose proofs are deferred to Appendix C.

Step 1: Suboptimality Decomposition. We decompose `SubOpt` into three terms.

Lemma 1. Given a policy $\pi = \{\pi_h\}_{h=1}^H$ and Q-function $\{Q_h\}_{h=1}^H$ with $V_h = \langle Q_h, \pi_h \rangle_{\mathcal{A}}$, for any meta state ω, `SubOpt` can be decomposed into three terms,

$$\begin{aligned}
\texttt{SubOpt}(\pi; \omega) = & \underbrace{\sum_{h=1}^H \mathbb{E}_\pi \left[Q_h(\omega_h, a_h) - (\mathbb{B}_h V_{h+1})(\omega_h, a_h) \,\middle|\, \omega_1 = \omega \right]}_{\mathcal{E}_1} \\
& + \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*} \left[(\mathbb{B}_h V_{h+1})(\omega_h, a_h) - Q_h(\omega_h, a_h) \,\middle|\, \omega_1 = \omega \right]}_{\mathcal{E}_2} \\
& + \underbrace{\sum_{h=1}^H \mathbb{E}_\pi \left[\langle Q_h(\omega_h, \cdot), \pi_h^*(\cdot \mid \omega_h) - \pi_h(\cdot \mid \omega_h) \rangle_{\mathcal{A}} \,\middle|\, \omega_1 = \omega \right]}_{\mathcal{E}_3}.
\end{aligned} \tag{7}$$

The proof is provided in Appendix C.2. We instantiate π, Q_h, and V_h in Lemma 1 to $\widehat{\pi}_h$, $\widehat{Q}_h$, and $\widehat{V}_h$ returned by Algorithm 1. The optimality of $\widehat{\pi}_h$, i.e., $\widehat{\pi}_h = \operatorname{argmax}_\pi \langle \widehat{Q}_h, \pi_h \rangle_{\mathcal{A}}$ implies that the third term $\mathcal{E}_3$ in (7) is non-positive. Therefore, $\texttt{SubOpt}(\widehat{\pi}; \omega)$ admits the upper bound

$$\texttt{SubOpt}(\widehat{\pi}; \omega) \leq \mathcal{E}_1 + \mathcal{E}_2. \tag{8}$$

Step 2: Pessimism Correction and Simplified Suboptimality Upper Bound. We further simplify (8) by assuming the following concentration condition:

$$\left| (\mathbb{B}_h \widehat{V}_{h+1})(\omega, a) - (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(\omega, a) \right| \leq \Gamma_h(\omega, a) \quad \text{for any} \quad (\omega, a) \in \Xi. \tag{9}$$

We will show in **Step 4** that choosing β and $\lambda = 1$ as in Theorem 1, Γ_h computed in Algorithm 1 verifies condition (9) with probability at least $1 - \delta$. Conditioned on (9), we can show term $\mathcal{E}_1$ is negative and term $\mathcal{E}_2$ is bounded by $2\Gamma_h$.

Lemma 2. In the setup of Theorem 1, let $\widehat{Q}_h$ and Γ_h be computed as in Algorithm 1. Conditioned on (9), for any (ω, a), the following sandwich inequality holds true,

$$0 \leq (\mathbb{B}_h \widehat{V}_{h+1})(\omega, a) - \widehat{Q}_h(\omega, a) \leq 2\Gamma_h(\omega, a), \quad \text{for} \quad h = 1, \ldots, H.$$

The proof is provided in Appendix C.3. Lemma 2 immediately implies that term $\mathcal{E}_1$ in (7) is non-positive, and $\mathcal{E}_2$ is upper bounded by $2\Gamma_h$. As a result, we simplify (8) as

$$\texttt{SubOpt}(\widehat{\pi};\omega) \le 2\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\Gamma_h(\omega_h,a_h)\mid \omega_1=\omega\right]. \tag{10}$$

Step 3: Establishing Condition (9). Recall that we require $\Gamma_h(\omega,a)$ satisfying

$$\left|(\mathbb{B}_h\widehat{V}_{h+1})(\omega,a)-(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(\omega,a)\right| \le \Gamma_h(\omega,a) \quad \text{for any} \quad (\omega,a)\in\Xi,$$

with probability at least $1-\delta$. It suffices to characterize the concentration of $\widehat{\mathbb{B}}_h\widehat{V}_{h+1}$ to $\mathbb{B}_h\widehat{V}_{h+1}$. By Proposition 1, we can write $\mathbb{B}_h\widehat{V}_{h+1}=\langle\mu(\omega,a),\alpha_h\rangle$ for some weight α_h. We also denote $\widehat{r}_h^n = r_h(\widehat{\omega}_h^n,a_h^n)$. We then decompose $(\mathbb{B}_h\widehat{V}_{h+1})(\omega,a)-(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(\omega,a)$ into three terms,

$$\begin{aligned}
&(\mathbb{B}_h\widehat{V}_{h+1})(\omega,a)-(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(\omega,a)\\
&=\langle\mu(\omega,a),\alpha_h\rangle-\phi_h(\omega,a)^\top\widehat{\alpha}_h\\
&=\langle\mu(\omega,a),\alpha_h\rangle-\phi_h(\omega,a)^\top\Lambda_h^{-1}\left[\widehat{r}_h^1+\widehat{V}_{h+1}(\widehat{\omega}_{h+1}^1),\ldots,\widehat{r}_h^N+\widehat{V}_{h+1}(\widehat{\omega}_{h+1}^N)\right]^\top\\
&\quad+\underbrace{\phi_h(\omega,a)^\top\Lambda_h^{-1}\left[\widehat{r}_h^1-r_h^1,\ldots,\widehat{r}_h^N-r_h^N\right]^\top}_{(A)}\\
&=(A)+\underbrace{\langle\mu(\omega,a),\alpha_h\rangle-\phi_h(\omega,a)^\top\Lambda_h^{-1}\left[(\mathbb{B}_h\widehat{V}_{h+1})(\widehat{\omega}_n^1,a_h^1),\ldots,(\mathbb{B}_h\widehat{V}_{h+1})(\widehat{\omega}_n^N,a_h^N)\right]^\top}_{(B)}\\
&\quad-\underbrace{\phi_h(\omega,a)^\top\Lambda_h^{-1}\left[\widehat{r}_h^1+\widehat{V}_{h+1}(\widehat{\omega}_{h+1}^1)-\mathbb{B}_h\widehat{V}_{h+1}(\widehat{\omega}_h^1,a_h^1),\ldots,\widehat{r}_h^N+\widehat{V}_{h+1}(\widehat{\omega}_{h+1}^N)-\mathbb{B}_h\widehat{V}_{h+1}(\widehat{\omega}_h^N,a_h^N)\right]^\top}_{(C)}.
\end{aligned}$$

Consequently, we have

$$\left|(\mathbb{B}_h\widehat{V}_{h+1})(\omega,a)-(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(\omega,a)\right| \le |(A)|+|(B)|+|(C)|.$$

Intuitively, term (A) measures the error induced by the empirical estimation of the mean-field distribution of states. Term (B) corresponds to the bias of kernel ridge regression, and Term (C) is the statistical error. We tackle these terms separately in the sequel.

• **Bounding Term** (A). We show reward $\widehat{r}_h^n$ concentrates around r_h^n, by establishing the concentration of empirical mean embedding to its population counterpart.

Lemma 3. Let $\widehat{\omega}_m$ be the empirical mean embedding corresponding to m agents i.i.d. sampled from infinite-agent state distribution. Given any $\delta_A>0$, with probability at least $1-\delta_A$, for any $a\in\mathcal{A}$, we have

$$\|\mu(\widehat{\omega}_m,a)-\mu(\omega,a)\|_{\mathcal{H}_K} \le \sqrt{\frac{2}{m}}+\sqrt{\frac{2\log(1/\delta_A)}{m}}.$$

The proof is provided in Appendix C.4. Combining Assumption 3 and Lemma 3, with probability $1-\delta_A$, it holds

$$\sup_n |\widehat{r}_h^n-r_h^n| \le \|\theta_h\|_{\mathcal{H}_K}\left(\sqrt{\frac{2}{m}}+\sqrt{\frac{2\log(1/\delta_A)}{m}}\right) \le \left(\sqrt{\frac{2}{m}}+\sqrt{\frac{2\log(1/\delta_A)}{m}}\right). \tag{11}$$

We are now ready to prove the following upper bound on (A).

Lemma 4. Suppose Assumption 1 and 2 hold. With probability $1-\delta_A$, it holds

$$|(A)| \le 2\left(\sqrt{\frac{1}{m}}+\sqrt{\frac{\log(1/\delta_A)}{m}}\right)\sqrt{\log\det(I+K_h/\lambda)}\left\|\Sigma_h^{-1/2}\mu(\omega,a)\right\|_{\mathcal{H}}.$$

The proof is provided in Appendix C.5. We note that $\log\det(I + K_h/\lambda)$ is known as the effective dimension of an RKHS (Yang et al., 2020b). The scale of $\log\det(I + K_h/\lambda)$ closely ties to the spectrum of kernel K. See Lemma 10 for an upper bound on $\log\det(I + K_h/\lambda)$. It is enough to set $\delta_A = \delta/2$, which yields

$$|(A)| \leq 2\left(\sqrt{\frac{1}{m}} + \sqrt{\frac{\log(2/\delta)}{m}}\right)\sqrt{\log\det(I + K_h/\lambda)}\left\|\Sigma_h^{-1/2}\mu(\omega, a)\right\|_{\mathcal{H}} \tag{12}$$

with probability $1 - \delta/2$.

• **Bounding Term** (B). We derive a useful decomposition of the mean embedding μ to simplify term (B). We let $\Phi_h = [\mu_1, \ldots, \mu_N]^\top$ for any fixed collection of $\mu_1, \ldots, \mu_N$. Then we define the regularized covariance operator as

$$\Sigma_h = \lambda I_{\mathcal{H}_K} + \Phi_h^\top \otimes \Phi_h,$$

where $I_{\mathcal{H}_K}$ is the identity operator on $\mathcal{H}_K$, and $\Phi_h^\top \otimes \Phi_h = \sum_{n=1}^N \mu_1 \otimes \mu_N$. The operator Σ_h has eigenvalues lower bounded by λ. Therefore, its inverse operator Σ_h^{-1} is well-defined. Now we check the identity

$$\begin{aligned}\Sigma_h^{-1}\Phi^\top &= \left(\lambda I_{\mathcal{H}_K} + \Phi_h^\top \otimes \Phi_h\right)^{-1}\Phi_h^\top \\ &\overset{(i)}{=} \Phi_h^\top\left(\lambda I + \left\langle \Phi_h, \Phi_h^\top\right\rangle\right)^{-1} \\ &= \Phi_h^\top\left(\lambda I + K_h\right)^{-1},\end{aligned} \tag{13}$$

where $\left\langle \Phi_h, \Phi_h^\top\right\rangle = [\langle \mu_\ell, \mu_{\ell'}\rangle]_{\ell,\ell'} \in \mathbb{R}^{N\times N}$, and equality (i) follows from

$$\left(\lambda I_{\mathcal{H}_K} + \Phi_h^\top \otimes \Phi_h\right)\Phi_h^\top = \lambda\Phi_h^\top + (\Phi_h^\top \otimes \Phi_h)\Phi_h^\top = \lambda\Phi_h^\top + \Phi_h^\top\left\langle \Phi_h, \Phi_h^\top\right\rangle = \Phi_h^\top\left(\lambda I + \left\langle \Phi_h, \Phi_h^\top\right\rangle\right)$$

which implies $\Phi_h^\top\left(\lambda I + \left\langle \Phi_h, \Phi_h^\top\right\rangle\right)^{-1} = \left(\lambda I_{\mathcal{H}_K} + \Phi_h^\top \otimes \Phi_h\right)^{-1}\Phi_h^\top$. We are ready to decompose mean embedding μ as

$$\begin{aligned}\mu(\cdot) &= \Sigma_h^{-1}\Sigma_h\mu(\cdot) \\ &= \Sigma_h^{-1}\left(\lambda I_{\mathcal{H}_K} + \Phi_h^\top \otimes \Phi_h\right)\mu(\cdot) \\ &= \lambda\Sigma_h^{-1}\mu(\cdot) + \Sigma_h^{-1}\Phi_h^\top \otimes \Phi_h\mu(\cdot) \\ &\overset{(i)}{=} \lambda\Sigma_h^{-1}\mu(\cdot) + \left(\Phi_h^\top\left(\lambda I + K_h\right)^{-1} \otimes \Phi_h\right)\mu(\cdot) \\ &\overset{(ii)}{=} \lambda\Sigma_h^{-1}\mu(\cdot) + \Phi_h^\top\left(\lambda I + K_h\right)^{-1}\langle \Phi_h, \mu(\cdot)\rangle \\ &= \lambda\Sigma_h^{-1}\mu(\cdot) + \Phi_h^\top\left(\lambda I + K_h\right)^{-1}\phi_h(\cdot) \\ &= \lambda\Sigma_h^{-1}\mu(\cdot) + \Phi_h^\top\Lambda_h^{-1}\phi_h(\cdot),\end{aligned} \tag{14}$$

where step (i) follows from (13), and step (ii) uses the definition of outer product. We use (14) to simplify (B) and derive an upper bound. We overload Φ_h by replacing fixed collection $\mu_1, \ldots, \mu_N$ with $\mu(\widehat{\omega}_h^1, a_h^1), \ldots, \mu(\widehat{\omega}_h^N, a_h^N)$. By substituting $\mathbb{B}_h\widehat{V}_{h+1}(\omega, a) = \langle \mu(\omega, a), \alpha_h\rangle$ into (B), we have

$$\begin{aligned}(B) &= \langle \mu(\omega, a), \alpha_h\rangle - \phi_h(\omega, a)^\top\Lambda_h^{-1}\left[\left\langle \mu(\widehat{\omega}_h^1, a_h^1), \alpha_h\right\rangle, \ldots, \left\langle \mu(\widehat{\omega}_h^N, a_h^N), \alpha_h\right\rangle\right]^\top \\ &\overset{(i)}{=} \left\langle \lambda\Sigma_h^{-1}\mu(\omega, a) + \Phi_h^\top\Lambda_h^{-1}\phi_h(\omega, a), \alpha_h\right\rangle \\ &\quad - \phi_h(\omega, a)^\top\Lambda_h^{-1}\left[\left\langle \mu(\widehat{\omega}_h^1, a_h^1), \alpha_h\right\rangle, \ldots, \left\langle \mu(\widehat{\omega}_h^N, a_h^N), \alpha_h\right\rangle\right]^\top \\ &= \lambda\left\langle \Sigma_h^{-1}\mu(\omega, a), \alpha_h\right\rangle.\end{aligned}$$

By Cauchy-Schwarz inequality, we have

$$(B) \leq \lambda\left\|\Sigma_h^{-1/2}\mu(\omega, a)\right\|_{\mathcal{H}}\left\|\Sigma_h^{-1/2}\alpha_h\right\|_{\mathcal{H}} \overset{(i)}{\leq} \sqrt{\lambda}\,\|\alpha_h\|_{\mathcal{H}}\left\|\Sigma_h^{-1/2}\mu(\omega, a)\right\|_{\mathcal{H}},$$

where inequality (i) follows from the operator norm of Σ_h^{-1} being upper bounded by λ^{-1}. To finish bounding term (B), we derive an upper bound on $\|\alpha_h\|_{\mathcal{H}}$. By Proposition 1, we have

$$\alpha_h = \int_\Omega \widehat{V}_{h+1}(x)v_h(x)dx + \theta_h.$$

The estimated value function satisfies $|\widehat{V}_{h+1}| \leq H-h$. Therefore, by the triangle inequality, we deduce

$$
\begin{aligned}
\|\alpha_h\|_{\mathcal{H}} &\leq \left\| \int_\Omega \widehat{V}_{h+1}(x) v_h(x) dx \right\|_{\mathcal{H}} + \|\theta_h\|_{\mathcal{H}} \\
&\leq (H-h) \int_\Omega \|v_h(x)\|_{\mathcal{H}} \, dx + \|\theta_h\|_{\mathcal{H}} \\
&\overset{(i)}{\leq} (H-h)\sqrt{d_{\text{eff}}} + 1 \\
&\leq H\sqrt{d_{\text{eff}}},
\end{aligned}
$$

where inequality (i) holds due to Assumption 3. Consequently, we derive

$$
|(B)| \leq \sqrt{\lambda} H \sqrt{d_{\text{eff}}} \left\| \Sigma_h^{-1/2} \mu(\omega, a) \right\|_{\mathcal{H}}. \tag{15}
$$

• **Bounding Term** (C). We use (13) to write

$$
\phi_h(\omega, a)^\top \Lambda_h^{-1} = \left\langle \mu(\omega, a), \Phi_h^\top \right\rangle \Lambda_h^{-1} = \left\langle \mu(\omega, a), \Phi_h^\top \Lambda_h^{-1} \right\rangle = \left\langle \mu(\omega, a), \Sigma_h^{-1} \Phi_h^\top \right\rangle. \tag{16}
$$

Denote $\Delta^n(\widehat{V}_{h+1}) = \widehat{r}_h^n + \widehat{V}_{h+1}(\widehat{\omega}_{h+1}^n) - \mathbb{B}_h \widehat{V}_{h+1}(\widehat{\omega}_h^n, a_h^n)$. Then term (C) can be rewrite as

$$
\begin{aligned}
(C) &= \left\langle \mu(\omega, a), \Sigma_h^{-1} \Phi_h^\top \right\rangle \left[\Delta^1(\widehat{V}_{h+1}), \ldots, \Delta^n(\widehat{V}_{h+1}) \right]^\top \\
&= \left\langle \mu(\omega, a), \Sigma_h^{-1} \sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n) \Delta^n(\widehat{V}_{h+1}) \right\rangle \\
&\overset{(i)}{\leq} \left\| \Sigma_h^{-1/2} \mu(\omega, a) \right\|_{\mathcal{H}} \underbrace{\left\| \Sigma_h^{-1/2} \sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n) \Delta^n(\widehat{V}_{h+1}) \right\|_{\mathcal{H}}}_{(\star)},
\end{aligned}
$$

where (i) invokes Cauchy-Schwarz inequality. We construct a space of functions that contains $\widehat{V}_h$ to decouple the dependence between the data $\mathcal{D}_{N,H}$ and $\widehat{V}_h$ in $(\star)$. Specifically, we define $\mathcal{V}_h$ consisting of functions in the form of

$$
\begin{aligned}
&\mathcal{V}_h(R, B, \lambda) = \\
&\Big\{ V_h(\omega) : V_h(\omega) = \max_{a \in \mathcal{A}} \left[\min \left\{ \langle \mu(\omega, a), \theta \rangle - \beta \sqrt{\langle \mu(\omega, a), \Sigma \cdot \mu(\omega, a) \rangle}, H-h+1 \right\}^+ \right], \\
&\qquad \|\theta\|_{\mathcal{H}} \leq R, \beta \in [0, B], \lambda^{-1} I_{\mathcal{H}_K} \succeq \Sigma \succeq 0 \Big\}.
\end{aligned} \tag{17}
$$

Note that when taking $\Sigma = \Sigma_h^{-1}$ in $\beta\sqrt{\langle \mu(\omega, a), \Sigma \cdot \mu(\omega, a) \rangle}$, it becomes an equivalent form of $\Gamma_h(\omega, a)$. To see this, we take inner product on both sides of (14) with ψ, and derive

$$
\begin{aligned}
K((\omega, a), (\omega, a)) &= \langle \mu(\omega, a), \mu(\omega, a) \rangle \\
&= \lambda \left\langle \mu(\omega, a), \Sigma_h^{-1} \mu(\omega, a) \right\rangle + \left\langle \mu(\omega, a), \Phi^\top \right\rangle \Lambda_h^{-1} \phi_h(\omega, a) \\
&= \lambda \left\langle \mu(\omega, a), \Sigma_h^{-1} \mu(\omega, a) \right\rangle + \phi_h(\omega, a)^\top \Lambda_h^{-1} \phi_h(\omega, a).
\end{aligned}
$$

By rearranging terms, we deduce

$$
\lambda \left\langle \mu(\omega, a), \Sigma_h^{-1} \mu(\omega, a) \right\rangle = K((\omega, a), (\omega, a)) - \phi_h(\omega, a)^\top \Lambda_h^{-1} \phi_h(\omega, a),
$$

and further $\Gamma_h(\omega, a) = \beta\sqrt{\left\langle \mu(\omega, a), \Sigma_h^{-1} \mu(\omega, a) \right\rangle}$. As a result, we have $\widehat{V}_h \in \mathcal{V}_h(R, B, \lambda)$ for properly chosen B and R, which are determined in **Step 4**.

We discretize $\mathcal{V}_{h+1}(R, B, \lambda)$ with respect to the ℓ_∞ norm, and find the closest element to replace $\widehat{V}_{h+1}$. In more detail, for any $\epsilon > 0$, we denote $\{V_{h+1,j}\}_{j=1}^{\mathcal{N}(\epsilon, \mathcal{V}_{h+1}(R,B,\lambda), \|\cdot\|_\infty)}$ as an ϵ-covering of

$\mathcal{V}_{h+1}(R,B,\lambda)$, where $\mathcal{N}(\epsilon, \mathcal{V}_{h+1}(R,B,\lambda), \|\cdot\|_\infty)$ is known as the covering number. By definition, there exists an index J, such that $\left\|\widehat{V}_{h+1} - V_{h+1,J}\right\|_\infty \le \epsilon$. By the triangle inequality, we have

$$
\begin{aligned}
(\star) &= \left\|\Sigma_h^{-1/2} \sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\left(\Delta^n(\widehat{V}_{h+1}) - \Delta^n(V_{h+1,J}) + \Delta^n(V_{h+1,J})\right)\right\|_{\mathcal{H}} \\
&\le \left\|\Sigma_h^{-1/2} \sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\left(\Delta^n(\widehat{V}_{h+1}) - \Delta^n(V_{h+1,J})\right)\right\|_{\mathcal{H}} + \left\|\Sigma_h^{-1/2} \sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\Delta^n(V_{h+1,J})\right\|_{\mathcal{H}}.
\end{aligned}
$$

The first term above can be bounded by

$$
\begin{aligned}
&\left\|\Sigma_h^{-1/2} \sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\left(\Delta^n(\widehat{V}_{h+1}) - \Delta^n(V_{h+1,J})\right)\right\|_{\mathcal{H}} \\
&\overset{(i)}{=} \left\|\Sigma_h^{-1/2} \sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\left(\widehat{V}_{h+1}(\widehat{\omega}_{h+1}^n) - \mathbb{B}_h\widehat{V}_{h+1}(\widehat{\omega}_h^n, a_h^n) - \widehat{V}_{h+1,J}(\widehat{\omega}_{h+1}^n) + \mathbb{B}_h\widehat{V}_{h+1,J}(\widehat{\omega}_h^n, a_h^n)\right)\right\|_{\mathcal{H}} \\
&\overset{(ii)}{\le} 2\epsilon\left\|\Sigma_h^{-1/2} \sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\right\|_{\mathcal{H}} \\
&\overset{(iii)}{\le} 2\epsilon\lambda^{-1/2} N \max_{\ell,\ell'} \left|K((\widehat{\omega}_h^\ell, a_h^\ell), (\widehat{\omega}_h^{\ell'}, a_h^{\ell'}))\right| \\
&\le 2\epsilon\lambda^{-1/2}N,
\end{aligned}
$$

where equality (i) uses the definition of Δ^n, inequality (ii) follows from the definition of $\widehat{V}_{h+1,J}$, and inequality (iii) holds since $\Sigma_h \succeq \lambda I_{\mathcal{H}}$. Consequently, we bound $(\star)$ as

$$
(\star) \le \sup_{j \le \mathcal{N}(\epsilon, \mathcal{V}_{h+1}(R,B,\lambda), \|\cdot\|_\infty)} \left\|\Sigma_h^{-1/2} \sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\Delta^n(V_{h+1,j})\right\|_{\mathcal{H}} + 2\epsilon\lambda^{-1/2}N. \tag{18}
$$

A crucial observation is that $V_{h+1,j}$ no longer coupled with the data $\mathcal{D}_{N,H}$, which allows us to derive uniform concentration on the first term above. We need the following lemma.

Lemma 5 (Restatement of Lemma B.2 in Jin et al. (2020b))**.** For any $h \le H$, let V_{h+1} be a given value function. Then, for any $\delta_C \in (0,1)$, we have

$$
\mathbb{P}\left(\left\|\Sigma_h^{-1/2} \sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\Delta^n(V_{h+1})\right\|_{\mathcal{H}}^2 > H^2\left(2\log\frac{1}{\delta_C} + \log\det(\lambda I + K_h)\right)\right) \le \delta_C.
$$

The proof is provided in Appendix C.6. Taking union bound over the covering of $\mathcal{V}_{h+1}$, we immediately have

$$
\begin{aligned}
\mathbb{P}\left(\sup_j \left\|\Sigma_h^{-1/2} \sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\Delta^n(V_{h+1,j})\right\|_{\mathcal{H}}^2 > H^2\left(2\log\frac{1}{\delta_C} + \log\det(\lambda I + K_h)\right)\right) \\
\le \delta_C \mathcal{N}(\epsilon, \mathcal{V}_{h+1}(R,B,\lambda), \|\cdot\|_\infty).
\end{aligned}
$$

We choose $\delta_C = \mathcal{N}^{-1}(\epsilon, \mathcal{V}_{h+1}(R,B,\lambda), \|\cdot\|_\infty) \cdot \delta/2$ such that

$$
\begin{aligned}
&\sup_j \left\|\Sigma_h^{-1/2} \sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\Delta^n(V_{h+1,j})\right\|_{\mathcal{H}}^2 \\
&\qquad \le H^2\left(2\log\frac{2\mathcal{N}(\epsilon, \mathcal{V}_{h+1}(R,B,\lambda), \|\cdot\|_\infty)}{\delta} + \log\det(\lambda I + K_h)\right)
\end{aligned} \tag{19}
$$

holds with probability $1 - \delta/2$. The remaining step is to bound the covering number $\mathcal{N}(\epsilon, \mathcal{V}_{h+1}(R,B,\lambda), \|\cdot\|_\infty)$.

Lemma 6. Suppose Assumption 1 and 2 hold. Recall the definition of $\mathcal{V}_h$ in (17). For any $h = 1, \ldots, H$, it holds

$$
\begin{aligned}
&\log \mathcal{N}(\epsilon, \mathcal{V}_h(R, B, \lambda), \|\cdot\|_\infty) \\
&\leq \begin{cases} \gamma \log(1 + 6R/\epsilon) + \gamma^2 \log(1 + 6B\sqrt{\gamma}/\epsilon), & (\textit{Finite Spectrum}) \\ C_3 \left(\log \frac{3R}{\epsilon} + C_4\right)^{\frac{1+\gamma}{\gamma}} + C_6 \left(\log \frac{3B}{\epsilon} + \log\log \frac{3B}{\epsilon} + C_7\right)^{\frac{2+\gamma}{\gamma}}, & (\textit{Exponential Decay}) \\ C_5 \left(\frac{3R}{\epsilon}\right)^{\frac{2}{\gamma-1}} \left(\log \frac{12R}{\epsilon} + 1\right) + C_8 \left(\frac{3B}{\epsilon}\right)^{\frac{4}{\gamma-1}} \left(\log \frac{3B}{\epsilon} + C_9\right), & (\textit{Polynomial Decay}) \end{cases} \\
&\quad + \log\left(1 + \frac{3B}{\epsilon\sqrt{\lambda}}\right),
\end{aligned}
$$

where constants C_i depend on C, C_1, C_2, λ, and γ in *(Exponential Decay)* and *(Polynomial Decay)*, for $i = 3, \ldots, 10$.

The proof is provided in Appendix C.7. Combining (18), (19) and Lemma 6, we obtain

$$
\begin{aligned}
|(C)| \leq &\left(H\sqrt{2\log \mathcal{N}(\epsilon, \mathcal{V}_{h+1}(R, B, \lambda), \|\cdot\|_\infty) + 2\log 2/\delta + \log\det(\lambda I + K_h)} + 2\epsilon\lambda^{-1/2}N\right) \\
&\cdot \left\|\Sigma_h^{-1/2}\mu(\omega, a)\right\|_{\mathcal{H}}.
\end{aligned} \tag{20}
$$

with probability $1 - \delta/2$.

Step 4: Completing the Proof. We choose proper ϵ, R, B, λ, and verify condition (9) to finish the proof. We set $\lambda = 1$, and determine R first. By parameterizing $f(\mu(\omega, a)) = \langle \mu(\omega, a), w\rangle$ in (5), we can solve the kernel ridge regression and obtain a closed form solution at step h as

$$
\widehat{w}_h = \Sigma_h^{-1}\Phi_h^\top [r_h^1 + \widehat{V}_{h+1}(\widehat{\omega}_{h+1}^1), \ldots, r_h^N + \widehat{V}_{h+1}(\widehat{\omega}_{h+1}^N)]^\top.
$$

Recall that by the definition of $\mathcal{V}_h$ in (17), R can be chosen as an upper bound on $\|\widehat{w}_h\|_{\mathcal{H}}$, which is provided in the following Lemma.

Lemma 7. For any $\lambda \geq 1$, we have

$$
\|\widehat{w}_h\|_{\mathcal{H}} \leq H\lambda^{-1/2}\sqrt{\log\det(I + K_h/\lambda)}.
$$

The proof is provided in Appendix C.8. As a result, we choose $R = H\lambda^{-1/2}\sqrt{\log\det(I + K_h/\lambda)}$. We choose B according to the spectrum of kernel K.

Case 1: *(Finite Spectrum)*. We set $B = c\gamma H\sqrt{\log(\max\{d, \gamma\}HN/\delta)}$ for some sufficiently large absolute constant c, which is exactly the choice of β. We first simplify the upper bound on $|(C)|$ in (20). By only keeping the dominating terms in the covering number and effective dimension $\log\det(I + K_h/\lambda)$ (see Lemma 10), for small $\epsilon \in (0, 1)$, we have

$$
\begin{aligned}
&2\log \mathcal{N}(\epsilon, \mathcal{V}_{h+1}(R, B, \lambda), \|\cdot\|_\infty) + 2\log 2/\delta + \log\det(\lambda I + K_h) \\
&\leq 4\gamma^2 \log\left(1 + \frac{6\max\{R, B\}\sqrt{\gamma}}{\epsilon}\right) + 2\log 2/\delta + C_{\text{eff-FS}} \cdot \max\{d, \gamma\}\log N \\
&\leq 4\gamma^2 \log\left(\frac{6cH\gamma^{3/2}\sqrt{\log(\max\{d, \gamma\}HN/\delta)}}{\epsilon\delta}\right) + C_{\text{eff-FS}} \cdot \max\{d, \gamma\}\log N,
\end{aligned}
$$

where the last inequality is valid when constant c is large. Setting $\epsilon = \frac{\gamma H\sqrt{\log(\max\{d, \gamma\}HN/\delta)}}{2N}$, we derive

$$
\begin{aligned}
&H\sqrt{2\log \mathcal{N}(\epsilon, \mathcal{V}_{h+1}(R, B, \lambda), \|\cdot\|_\infty) + 2\log 2/\delta + \log\det(\lambda I + K_h)} + 2\epsilon\lambda^{-1/2}N \\
&\overset{(i)}{\leq} 2\gamma H\sqrt{\log\left(12cHN\sqrt{\gamma}/\delta\right)} + H\sqrt{C_{\text{eff-FS}} \cdot \max\{d, \gamma\}\log N} + \gamma H\sqrt{\log(\max\{d, \gamma\}HN/\delta)} \\
&\overset{(ii)}{\leq} \frac{c}{2}\gamma H\sqrt{\log(\max\{d, \gamma\}HN/\delta)},
\end{aligned}
$$

where inequality (i) follows from $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and inequality (ii) is always valid when c is properly chosen. Combining (12), (15), and (20), we have

$$
\begin{aligned}
&\left|(\mathbb{B}_h \widehat{V}_{h+1})(\omega, a) - (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(\omega, a)\right| \\
&\le |(A)| + |(B)| + |(C)| \\
&\le \left(\left(\sqrt{\frac{1}{m}} + \sqrt{\frac{\log(2/\delta)}{m}} + H \right) \sqrt{\max\{d, \gamma\} \log N} + \frac{c}{2} \gamma H \sqrt{\log(\max\{d, \gamma\} HN/\delta)} \right) \\
&\quad \cdot \left\| \Sigma_h^{-1/2} \mu(\omega, a) \right\|_{\mathcal{H}} \\
&\overset{(i)}{\le} c\gamma H \sqrt{\log(\max\{d, \gamma\} HN/\delta)} \left\| \Sigma_h^{-1/2} \mu(\omega, a) \right\|_{\mathcal{H}} \\
&= \Gamma_h(\omega, a),
\end{aligned}
$$

where inequality (i) holds, since $m \ge \log(2/\delta)$ and c is sufficiently large. As a consequence, condition (9) holds true, and we bound $\texttt{SubOpt}(\widehat{\pi}; \omega)$ by

$$
\texttt{SubOpt}(\widehat{\pi}; \omega) \le 2 \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[\Gamma_h(\omega_h, a_h) \mid \omega_1 = \omega \right].
$$

Case 2: *(Exponential Decay)*. We set $B = cH\sqrt{d\left(\log(HN/\delta)\right)^{1+2/\gamma}}$ for a sufficiently large constant c. Utilizing similar analysis in **Case 1**, for small $\epsilon \in (0, 1)$ and sufficiently large c, we simplify the bound for $|(C)|$:

$$
\begin{aligned}
&2 \log \mathcal{N}(\epsilon, \mathcal{V}_{h+1}(R, B, \lambda), \|\cdot\|_\infty) + 2 \log 2/\delta + \log\det(\lambda I + K_h) \\
&\le (C_3 + C_6) \log \left(\frac{6 \max\{R, B\}}{\epsilon} + (C_4 + C_7) \right)^{\frac{2+\gamma}{\gamma}} + 2 \log 2/\delta + C_{\text{eff-ED}} \cdot d (\log N)^{\frac{1+\gamma}{\gamma}} \\
&\le (C_3 + C_6) d \log \left(\frac{12cH\sqrt{d(\log(HN/\delta))^{1+2/\gamma}}}{\epsilon\delta} \right)^{\frac{2+\gamma}{\gamma}} + C_{\text{eff-ED}} \cdot d (\log N)^{\frac{1+\gamma}{\gamma}}.
\end{aligned}
$$

Choosing $\epsilon = \frac{H\sqrt{d(\log(HN/\delta)^{1+2/\gamma}}}{2N}$, we derive

$$
\begin{aligned}
&H \sqrt{2 \log \mathcal{N}(\epsilon, \mathcal{V}_{h+1}(R, B, \lambda), \|\cdot\|_\infty) + 2 \log 2/\delta + \log\det(\lambda I + K_h)} + 2\epsilon\lambda^{-1/2} N \\
&\le \sqrt{2(C_3 + C_6)} H \sqrt{d \log(24cN/\delta)} + H \sqrt{C_{\text{eff-ED}} \cdot d (\log N)^{1+1/\gamma}} + H \sqrt{d(\log(HN/\delta))^{1+2/\gamma}} \\
&\le \frac{c}{2} H \sqrt{d(\log(HN/\delta))^{1+2/\gamma}}.
\end{aligned}
$$

Lastly, combining (12), (15), and (20), we have

$$
\begin{aligned}
&\left|(\mathbb{B}_h \widehat{V}_{h+1})(\omega, a) - (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(\omega, a)\right| \\
&\le |(A)| + |(B)| + |(C)| \\
&\le \left(\left(\sqrt{\frac{1}{m}} + \sqrt{\frac{\log(2/\delta)}{m}} + H \right) \sqrt{C_{\text{eff-ED}} \cdot d (\log N)^{1+1/\gamma}} + \frac{c}{2} H \sqrt{d(\log(HN/\delta))^{1+2/\gamma}} \right) \\
&\quad \cdot \left\| \Sigma_h^{-1/2} \mu(\omega, a) \right\|_{\mathcal{H}} \\
&\le cH \sqrt{d(\log(HN/\delta))^{1+2/\gamma}} \left\| \Sigma_h^{-1/2} \mu(\omega, a) \right\|_{\mathcal{H}} \\
&= \Gamma_h(\omega, a).
\end{aligned}
$$

As a consequence, condition (9) holds true.

Case 3: *(Polynomial Decay)*. We set $B = cN^{\frac{d+1}{d+\gamma}} H \sqrt{d \log(HN/\delta)}$ for a sufficiently large constant c. For small $\epsilon \in (0,1)$ and sufficiently large c, we begin with simplifying the bound for $|(C)|$:

$$\begin{aligned}
&2 \log \mathcal{N}(\epsilon, \mathcal{V}_{h+1}(R,B,\lambda), \|\cdot\|_\infty) + 2\log 2/\delta + \log\det(\lambda I + K_h) \\
&\le (C_5 + C_8) \left(\frac{3\max\{R,B\}}{\epsilon}\right)^{4/(\gamma-1)} \log\left(\frac{12\max\{R,B\}}{\epsilon} + C_9 + 1\right) + 2\log 2/\delta \\
&\quad + C_{\text{eff-PD}} \cdot N^{\frac{d+1}{d+\gamma}} d \log N \\
&\le (C_5 + C_8) \left(\frac{6cN^{\frac{d+1}{d+\gamma}} H \sqrt{d\log(HN/\delta)}}{\epsilon}\right)^{\frac{4}{\gamma-1}} + C_{\text{eff-PD}} \cdot N^{\frac{d+1}{d+\gamma}} d\log N.
\end{aligned}$$

Choosing $\epsilon = \frac{1}{2} N^{-1+\frac{d+1}{d+\gamma}} H \sqrt{d\log(HN/\delta)}$, we derive

$$\begin{aligned}
&H\sqrt{2 \log \mathcal{N}(\epsilon, \mathcal{V}_{h+1}(R,B,\lambda), \|\cdot\|_\infty) + 2\log 2/\delta + \log\det(\lambda I + K_h)} + 2\epsilon\lambda^{-1/2} N \\
&\le \sqrt{2(C_5+C_8)} H \left(12cN\right)^{\frac{2}{\gamma-1}} + H\sqrt{C_{\text{eff-PD}} \cdot N^{\frac{d+1}{d+\gamma}} d\log N} + N^{\frac{d+1}{d+\gamma}} H\sqrt{d\log(HN/\delta)} \\
&\overset{(i)}{\le} \frac{c}{2} H N^{\frac{d+1}{d+\gamma}} H \sqrt{d\log(HN/\delta)},
\end{aligned}$$

where in (i), we have $\frac{2}{\gamma-1} \le \frac{d+1}{d+\gamma}$ for $\gamma \ge 3 + 4/(d-1)$. Lastly, combining (12), (15), and (20), we have

$$\begin{aligned}
&\left|(\mathbb{B}_h \widehat{V}_{h+1})(\omega,a) - (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(\omega,a)\right| \\
&\le |(A)| + |(B)| + |(C)| \\
&\le \left(\left(\sqrt{\frac{1}{m}} + \sqrt{\frac{\log(2/\delta)}{m}} + H\right)\sqrt{C_{\text{eff-PD}} \cdot N^{\frac{d+1}{d+\gamma}} d\log N} + \frac{c}{2} H N^{\frac{d+1}{d+\gamma}} H\sqrt{d\log(HN/\delta)}\right) \\
&\quad \cdot \left\|\Sigma_h^{-1/2} \mu(\omega,a)\right\|_{\mathcal{H}} \\
&\le cHN^{\frac{d+1}{d+\gamma}} H\sqrt{d\log(HN/\delta)} \left\|\Sigma_h^{-1/2}\mu(\omega,a)\right\|_{\mathcal{H}} \\
&= \Gamma_h(\omega,a).
\end{aligned}$$

As a consequence, condition (9) holds true. Therefore, we complete the proof. □

B Proof of Corollary 1

Proof. Given Theorem 1, we only need to show the convergence of Γ_h under Assumption 4. In particular, we show that $\|\Sigma_h^{-1/2}\mu(\omega,a)\|_{\mathcal{H}}$ converges at a rate of $1/\sqrt{N}$. We denote $A_n = \mu(\omega_h^n, a_h^n) \otimes \mu(\omega_h^n, a_h^n) - \mathbb{E}_{\bar{\pi}}[\mu(\omega_h, a_h) \otimes \mu(\omega_h, a_h)]$, which verifies $\mathbb{E}_{\bar{\pi}}[A_n] = 0$. We denote $Z = \sum_{n=1}^N A_n$, and distinguish three cases according to the spectrum of K. (In the following proof, we denote $\|\cdot\|_{\text{op}}$ as the operator norm.)

Case 1: *(Finite Spectrum)*. Using the same argument in Lemma 9, we represent $\mu(\omega_h^n, a) \otimes \mu(\omega_h^n, a)$ as a $\gamma \times \gamma$ matrix W_h^n, since kernel K has a γ-finite spectrum. We also denote $W_h = \mathbb{E}_{\bar{\pi}}[\mu(\omega_h, a_h) \otimes \mu(\omega_h, a_h)]$. We keep the notation A_n and Z unchanged, yet overload the two with $A_n = W_h^n - W_h$ and $Z = \sum_{n=1}^N A_n$.

By the boundedness of kernel, we deduce that the matrix operator norm $\|W_h\|_{\text{op}}$ is bounded by 1. Meanwhile, the operator norm of A_n is bounded by $\|A_n\|_{\text{op}} \le \|W_h^n\|_{\text{op}} + \|W_h\|_{\text{op}} \le 2$, using the triangle inequality. Furthermore, we have

$$\left\|\mathbb{E}_{\bar{\pi}}[ZZ^\top]\right\|_{\text{op}} = N\left\|A_n A_n^\top\right\|_{\text{op}} \le N\left\|A_n\right\|_{\text{op}}\left\|A_n^\top\right\|_{\text{op}} \le 4N,$$

and similarly

$$\left\|\mathbb{E}_{\bar{\pi}}[Z^\top Z]\right\|_{\text{op}} = N\left\|A_n^\top A_n\right\|_{\text{op}} \le 4N.$$

Applying matrix Bernstein inequality, for any $t > 0$, we have

$$\mathbb{P}\left(\|Z\|_{\text{op}} \geq t\right) \leq 2\gamma \exp\left(-\frac{t^2/2}{4N + 2t/3}\right).$$

Taking $t = \sqrt{10N \log(4\gamma H/\delta)}$, for sufficiently large $N \geq 5\log(4\gamma H/\delta)$, we obtain

$$\|Z\|_{\text{op}} \leq \sqrt{10N \log(4\gamma H/\delta)} \quad \text{holds with probability} \quad 1 - \delta/2H.$$

Case 2: *(Exponential Decay)*. We truncate the spectrum of kernel K at M, where the positive integer M will be determined later. Denote $\bar{\mu}$ as the truncated version of μ. Using the truncation error in Lemma 9 with $\lambda = 1$, we bound

$$\begin{aligned} e_M &\triangleq \|\mu(\omega, a) \otimes \mu(\omega, a) - \bar{\mu}(\omega, a) \otimes \bar{\mu}(\omega, a)\|_{\text{op}} \\ &\leq \begin{cases} 2\sqrt{\frac{C_1}{C_2}\exp(-C_2 M^\gamma)} & \gamma \geq 1 \\ 2\sqrt{\frac{C_1 M^{1-\gamma}}{C_2 \gamma}\exp(-C_2 M^\gamma)} & \gamma \in (0, 1) \end{cases}. \end{aligned}$$

We let $\bar{A}_n = \bar{\mu}(\omega_h^n, a_h^n) \otimes \bar{\mu}(\omega_h^n, a_h^n) - \mathbb{E}_{\bar{\pi}}[\bar{\mu}(\omega_h, a_h) \otimes \bar{\mu}(\omega_h, a_h)]$ and $\bar{Z} = \sum_{n=1}^N \bar{A}_n$. Then we can derive

$$\begin{aligned} \mathbb{P}\left(\|Z\|_{\text{op}} \geq t\right) &\leq \mathbb{P}\left(\|\bar{Z}\|_{\text{op}} \geq t - 2Ne_M\right) \\ &\leq 2M \exp\left(-\frac{\left(t - 2Ne_M\right)^2/2}{4N + 2\left(t - 2Ne_M\right)/3}\right). \end{aligned}$$

We choose $M = \frac{c_0}{C_2}(\log \frac{C_1 N}{C_2})^{1/\gamma}$ for some absolute constant $c_0 \leq 2$ and $t = 8\sqrt{N \log(4Hd_{\text{eff}}/\delta)}$, which gives rise to

$$t - 2Ne_M = 8\sqrt{N \log(4Hd_{\text{eff}}/\delta)} - 4\sqrt{N} \leq \sqrt{10N \log(4Hd_{\text{eff}}/\delta)}.$$

For sufficiently large $N \geq 5\log(4Hd_{\text{eff}}/\delta)$, we obtain

$$\|Z\|_{\text{op}} \leq 8\sqrt{N \log(4Hd_{\text{eff}}/\delta)} \quad \text{holds with probability} \quad 1 - \delta/2H.$$

Case 3: *(Polynomial Decay)*. We consider truncating the spectrum at M again. Using the truncation error of polynomial decay case in Lemma 9, we have

$$\begin{aligned} \mathbb{P}\left(\|Z\|_{\text{op}} \geq t\right) &\leq \mathbb{P}\left(\|\bar{Z}\|_{\text{op}} \geq t - 4N\sqrt{(\gamma-1)^{-1}CM^{-\gamma+1}}\right) \\ &\leq 2M \exp\left(-\frac{\left(t - 4N\sqrt{(\gamma-1)^{-1}CM^{-\gamma+1}}\right)^2/2}{4N + 2\left(t - 4N\sqrt{(\gamma-1)^{-1}CM^{-\gamma+1}}\right)/3}\right). \end{aligned}$$

We choose $M = \left(\frac{CN}{\gamma-1}\right)^{\frac{1}{\gamma-1}}$ and $t = 8\sqrt{N \log(4Hd_{\text{eff}}/\delta)}$, which gives rise to

$$t - 4N\sqrt{(\gamma-1)^{-1}CM^{-\gamma+1}} = 8\sqrt{N \log(4Hd_{\text{eff}}/\delta)} - 4\sqrt{N} \leq \sqrt{10N \log(4Hd_{\text{eff}}/\delta)}.$$

Further, when $\gamma \geq 2 + 1/d$, we have $\frac{d+1}{d+\gamma} > \frac{1}{\gamma-1}$, which implies $M \leq d_{\text{eff}}$. Thus, for $N \geq 5\log(4Hd_{\text{eff}}/\delta)$, we have

$$\|Z\|_{\text{op}} \leq 8\sqrt{N \log(4Hd_{\text{eff}}/\delta)} \quad \text{holds with probability} \quad 1 - \delta/2H.$$

We rewrite Z using the covariance operator Σ_h as

$$\begin{aligned} Z = \Sigma_h - \lambda I_{\mathcal{H}_K} - N\mathbb{E}_{\bar{\pi}}[\mu(\omega_h, a_h) \otimes \mu(\omega_h, a_h)] \\ + \underbrace{\sum_{n=1}^N \left\{\mu(\widehat{\omega}_h^n, a_h^n) \otimes \mu(\widehat{\omega}_h^n, a_h^n) - \mu(\omega_h^n, a_h^n) \otimes \mu(\omega_h^n, a_h^n)\right\}}_{\mathcal{E}}. \end{aligned}$$

We bound the operator norm of $\mathcal{E}$ by

$$\begin{aligned}\|\mathcal{E}\|_{\text{op}} &= \sup_{\|f\|_{\mathcal{H}}\leq 1} \|\mathcal{E}f\|_{\mathcal{H}} \\ &\leq N \left\|\mu(\widehat{\omega}_h^n, a_h^n)\left\langle\mu(\widehat{\omega}_h^n, a_h^n), f\right\rangle - \mu(\omega_h^n, a_h^n)\left\langle\mu(\omega_h^n, a_h^n), f\right\rangle\right\|_{\mathcal{H}} \\ &\leq 2N \left\|\mu(\widehat{\omega}_h^n, a_h^n) - \mu(\omega_h, a_h)\right\|_{\mathcal{H}}.\end{aligned}$$

By Lemma 3, we have $\|\mu(\widehat{\omega}_h^n, a_h^n) - \mu(\omega_h, a_h)\|_{\mathcal{H}} \leq \sqrt{\frac{2}{m}} + \sqrt{\frac{2\log(4H/\delta)}{m}}$ with probability $1-\delta/4H$. Thus, for sufficiently large $m \geq 32N^2\log(4H/\delta)$, $\|\mathcal{E}\|_{\text{op}} \leq \sqrt{N}$ with probability $1-\delta/4H$. Therefore, with probability $1-\frac{3}{4}\delta$, we deduce that

$$\begin{aligned}\left\|\frac{1}{N}(\Sigma_h - \lambda I) - \mathbb{E}_{\bar{\pi}}[\mu(\omega_h, a_h)\otimes\mu(\omega_h, a_h)]\right\|_{\text{op}} &\leq \frac{1}{N}\|Z\|_{\text{op}} + \frac{1}{N}\|\mathcal{E}\|_{\text{op}} \\ &\leq 8\sqrt{\frac{\log(4d_{\text{eff}}H/\delta)}{N}} + \sqrt{\frac{1}{N}}\end{aligned}$$

holds simultaneously for all $h = 1, \ldots, H$. For sufficiently large $N \geq \frac{1024}{c_{\min}^2}\log(4d_{\text{eff}}H/\delta)$, we have

$$\frac{1}{N}\Sigma_h \succeq \mathbb{E}_{\bar{\pi}}[\mu(\omega_h, a_h)\otimes\mu(\omega_h, a_h)] - \left(8\sqrt{\frac{\log(4d_{\text{eff}}H/\delta)}{N}} + \sqrt{\frac{1}{N}}\right)\cdot I_{\mathcal{H}_K} \succeq \frac{c_{\min}}{2}\cdot I_{\mathcal{H}_K}.$$

This further implies

$$\left\|\Sigma_h^{-1/2}\mu(\omega, a)\right\|_{\mathcal{H}} \leq \|\mu(\omega, a)\|_{\mathcal{H}}\left\|\Sigma_h^{-1}\right\|_{\text{op}} \leq \frac{2}{c_{\min}\sqrt{N}}.$$

Combining with Theorem 1 and taking $\delta = \delta/4$ therein, we have

$$\begin{aligned}\texttt{SubOpt}(\widehat{\pi};\omega) &\leq 2\sum_{h=1}^{H}[\Gamma_h(\omega_h, a_h) \mid \omega_1 = \omega] \\ &\leq 2\beta\sum_{h=1}^{H}\left[\frac{2}{c_{\min}}N^{-1/2} \mid \omega_1 = \omega\right] \\ &= \mathcal{O}\left(H^2 d_{\text{eff}}\sqrt{\frac{\log(d_{\text{eff}}HN/\delta)}{N}}\right),\end{aligned}$$

where in the last inequality, we substitute into the choice of β and note that by Lemma 10, $\beta = \mathcal{O}(Hd_{\text{eff}}\sqrt{\log(d_{\text{eff}}HN/\delta)})$. The proof is complete. □

C Proofs of Supporting Lemmas for Theorem 1

We provide proofs of technical lemmas for establishing our main results.

C.1 Proof of Proposition 1

Proof. We first show $\mathbb{B}_h g$ and $\mathbb{P}_h g$ can be parameterized in $\mathcal{H}_K$. Using the definition in (2), we derive

$$\begin{aligned}(\mathbb{P}_h g)(\omega, a) &= \mathbb{E}\left[g(\omega_{h+1}) \mid \omega_h = \omega, a_h = a\right] \\ &= \int_\Omega g(x)p_h(x \mid \omega, a)dx \\ &= \int_\Omega g(x)\left\langle\mu(\omega, a), v_h(x)\right\rangle dx \\ &= \left\langle\mu(\omega, a), \int_\Omega g(x)v_h(x)dx\right\rangle \in \mathcal{H}_{\mathcal{K}}.\end{aligned}$$

Similarly, for $\mathbb{B}_h g$, we have

$$
\begin{aligned}
(\mathbb{B}_h g)(\omega, a) &= (\mathbb{P}_h g)(\omega, a) + r_h(\omega, a) \\
&= \left\langle \mu(\omega, a), \int_\Omega g(x) v_h(x) dx \right\rangle + \langle \mu(\omega, a), \theta_h \rangle \\
&= \left\langle \mu(\omega, a), \int_\Omega g(x) v_h(x) dx + \theta_h \right\rangle \in \mathcal{H}_K.
\end{aligned}
$$

The proof is complete. □

C.2 Proof of Lemma 1

Proof. We write $\texttt{SubOpt}(\widehat{\pi}; \omega)$ as

$$
\texttt{SubOpt}(\pi; \omega) = V_1^{\pi^*}(\omega) - V_1(\omega) + V_1(\omega) - V_1^{\pi}(\omega), \tag{21}
$$

where $V_h = \langle Q_h, \pi \rangle_{\mathcal{A}}$. Note the difference between V_1 and V_1^{π}, where in the former, each V_h is computed from a given Q-function Q_h. By the extended value difference in Section B.1 of Cai et al. (2020) (see also Lemma A.1 in Jin et al. (2020b)), we have

$$
\begin{aligned}
V_1(\omega) - V_1^{\pi}(\omega) &= \mathbb{E}_\pi \left[\langle Q_1(\omega_1, \cdot), \pi(\cdot \mid \omega_1) \rangle_{\mathcal{A}} - \langle Q_1^{\pi}(\omega_1, \cdot), \pi(\cdot \mid \omega_1) \rangle_{\mathcal{A}} \mid \omega_1 = \omega \right] \\
&= \mathbb{E}_\pi \Big[\langle Q_1(\omega_1, \cdot), \pi(\cdot \mid \omega_1) \rangle_{\mathcal{A}} - \langle (\mathbb{B}_1 V_2)(\omega_1, \cdot), \pi(\cdot \mid \omega_1) \rangle_{\mathcal{A}} \\
&\quad + \langle (\mathbb{B}_1 V_2)(\omega_1, \cdot), \pi(\cdot \mid \omega_1) \rangle_{\mathcal{A}} - \langle Q_1^{\pi}(\omega, \cdot), \pi(\cdot \mid \omega) \rangle_{\mathcal{A}} \,\Big|\, \omega_1 = \omega \Big] \\
&= \mathbb{E}_\pi \left[Q_1(\omega_1, a_1) - (\mathbb{B}_1 V_2)(\omega_1, a_1) \mid \omega_1 = \omega \right] + \mathbb{E}_\pi \left[V_2(\omega_2) - V_2^{\pi}(\omega_2) \mid \omega_1 = \omega \right] \\
&= \cdots \\
&= \sum_{h=1}^{H} \mathbb{E}_\pi \left[Q_h(\omega_h, a_h) - (\mathbb{B}_h V_{h+1})(\omega_h, a_h) \mid \omega_1 = \omega \right].
\end{aligned} \tag{22}
$$

Analogously, we derive

$$
\begin{aligned}
V_1^{\pi^*}(\omega) - V_1(\omega) = &\sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[\langle Q_h(\omega_h, \cdot), \pi_h(\cdot \mid \omega_h) - \pi^*(\cdot \mid \omega_h) \rangle \,\Big|\, \omega_1 = \omega \right] \\
&+ \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[Q_h(\omega_h, a_h) - (\mathbb{B}_h V_{h+1})(\omega_h, a_h) \,\Big|\, \omega_1 = \omega \right].
\end{aligned} \tag{23}
$$

Substituting (22) and (23) into (21), we obtain the desired decomposition in Lemma 1. □

C.3 Proof of Lemma 2

Proof. We first prove the left inequality, i.e.,

$$
0 \le (\mathbb{B}_h \widehat{V}_{h+1})(\omega, a) - \widehat{Q}_h(\omega, a).
$$

Conditioned on the event $\left|(\mathbb{B}_h \widehat{V}_{h+1})(\omega, a) - (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(\omega, a)\right| \le \Gamma_h(\omega, a)$, we have

$$
(\mathbb{B}_h \widehat{V}_{h+1})(\omega, a) - \widehat{Q}_h(\omega, a) \overset{(i)}{\ge} (\mathbb{B}_h \widehat{V}_{h+1})(\omega, a) - \widetilde{Q}_h(\omega, a) + \Gamma(\omega, a) \ge 0,
$$

where inequality (i) follows from $\widehat{Q}$ is a bounded truncated version of $\widetilde{Q}_h$. Therefore, the left inequality holds for any (ω, a). Next, we show the right inequality, i.e.,

$$
(\mathbb{B}_h \widehat{V}_{h+1})(\omega, a) - \widehat{Q}_h(\omega, a) \le 2\Gamma_h(\omega, a).
$$

Observe $\widehat{V}_{h+1} \le H - h$ in Algorithm 1. Combining with $|r_h| \le 1$, we have

$$
(\mathbb{B}_h \widehat{V}_{h+1})(\omega, a) \le H - h + 1.
$$

This implies $(\widehat{\mathbb{B}}_h \widehat{V}_{h+1}) \leq H - h + 1 - \Gamma_h$. Therefore, we deduce

$$\widehat{Q}_h \geq \widetilde{Q}_h - \Gamma_h = (\widehat{\mathbb{B}}_h \widehat{V}_{h+1}) - \Gamma_h.$$

Since we have proved $(\mathbb{B}_h \widehat{V}_{h+1})(\omega, a) - \widehat{Q}_h(\omega, a) \geq 0$, we derive

$$(\mathbb{B}_h \widehat{V}_{h+1})(\omega, a) - \widehat{Q}_h(\omega, a) \leq (\mathbb{B}_h \widehat{V}_{h+1})(\omega, a) - (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(\omega, a) + \Gamma_h(\omega, a) \leq 2\Gamma_h(\omega, a).$$

The proof is complete. $\square$

C.4 Proof of Lemma 3

Proof. We denote

$$g(\widehat{\omega}_m, a) = \|\mu(\widehat{\omega}_m, a) - \mu(\omega, a)\|_{\mathcal{H}}.$$

Consider two meta states $\widehat{\omega}_m = s_0 \times \widehat{d}_s$ and $\widehat{\omega}' = s_0 \times \widehat{d}_s'$ with only the k-th agent ($k \geq 1$) having distinct states s_k and $(s_k)'$, respectively. We bound the difference in function value:

$$\begin{aligned}
|g(\widehat{\omega}_m, a) - g(\widehat{\omega}_m', a)| &= \|\mu(\widehat{\omega}_m, a) - \mu(\omega, a)\|_{\mathcal{H}_K} - \|\mu(\widehat{\omega}_m', a) - \mu(\omega, a)\|_{\mathcal{H}} \\
&\leq \|\mu(\widehat{\omega}_m, a) - \mu(\widehat{\omega}_m', a)\|_{\mathcal{H}} \\
&\leq \frac{1}{m} \|\psi(s_0, s_k, a) - \psi(s_0, s_k', a)\|_{\mathcal{H}} \\
&\leq \frac{2}{m}.
\end{aligned}$$

By Mcdiarmid's inequality, for any $\delta_0 > 0$, we have

$$g(\widehat{\omega}_m, a) \leq \mathbb{E}[g(\widehat{\omega}_m, a)] + \delta_0 \quad \text{with probability at least} \quad 1 - \exp\left(-\frac{\delta_0^2 m}{2}\right). \tag{24}$$

It remains to bound $\mathbb{E}[g(\widehat{\omega}_m, a)]$. Some algebraic manipulation gives rise to

$$\begin{aligned}
\mathbb{E}[g(\widehat{\omega}_m, a)] &\overset{(i)}{\leq} \sqrt{\mathbb{E}\left[\|\mu(\widehat{\omega}_m, a) - \mu(\omega, a)\|_{\mathcal{H}}^2\right]} \\
&= \sqrt{\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} \psi(s_0, s_i, a) - \mu(\omega, a)\right]} \\
&= \sqrt{\frac{1}{m}\mathbb{E}_{s \sim d_s, s' \sim d_s'}\left[k((s_0, s, a), (s_0, s, a)) - k((s_0, s, a), (s_0, s', a))\right]} \\
&\leq \sqrt{2/m},
\end{aligned} \tag{25}$$

where inequality (i) follows from Jensen's inequality. (A similar computation appears in Theorem 15 of Altun and Smola (2006).) Substituting (25) into the right-hand side of (24), with probability at least $1 - \exp\left(-\frac{\delta_0^2 m}{2}\right)$, we have

$$\|\mu(\widehat{\omega}_m, a) - \mu(\omega, a)\|_{\mathcal{H}} \leq \sqrt{2/m} + \delta_0.$$

Taking $\delta_0 = \sqrt{\frac{2\log(1/\delta_A)}{m}}$, we deduce

$$\|\mu(\widehat{\omega}_m, a) - \mu(\omega, a)\|_{\mathcal{H}} \leq \sqrt{\frac{2}{m}} + \sqrt{\frac{2\log(1/\delta_A)}{m}} \quad \text{with probability at least} \quad 1 - \delta_A.$$

The proof is complete. $\square$

C.5 Proof of Lemma 4

Proof. Using identity (16), we bound (A) as

$$
\begin{aligned}
|(A)| &= \left|\phi_h(\omega,a)^\top \Lambda_h^{-1}\left[\widehat{r}_h^1 - r_h^1, \ldots, \widehat{r}_h^N - r_h^N\right]^\top\right| \\
&= \left|\left\langle \mu(\omega,a), \Sigma_h^{-1}\Phi_h^\top\left[\widehat{r}_h^1 - r_h^1, \ldots, \widehat{r}_h^N - r_h^N\right]^\top\right\rangle\right| \\
&\overset{(i)}{\leq} \|\theta_h\|_{\mathcal{H}}\left(\sqrt{\frac{2}{m}} + \sqrt{\frac{2\log(1/\delta_A)}{m}}\right)\sum_{n=1}^N \left|\left\langle \mu(\omega,a), \Sigma_h^{-1}\mu(\widehat{\omega}_h^n, a_h^n)\right\rangle\right| \\
&\overset{(ii)}{\leq} \left(\sqrt{\frac{2}{m}} + \sqrt{\frac{2\log(1/\delta_A)}{m}}\right)\left\|\Sigma_h^{-1/2}\mu(\omega,a)\right\|_{\mathcal{H}}\sqrt{\sum_{n=1}^N \left\langle \mu(\widehat{\omega}_h^n, a_h^n), \Sigma_h^{-1}\mu(\widehat{\omega}_h^n, a_h^n)\right\rangle}, \quad (26)
\end{aligned}
$$

where inequality (i) invokes (11) and holds with probability $1-\delta_A$, inequality (ii) follows from Cauchy-Schwarz inequality. It remains to bound $\sum_{n=1}^N \left\langle \psi_h(\mu_{\widehat{\varsigma}_h^n}), \Sigma_h^{-1}\psi_h(\mu_{\widehat{\varsigma}_h^n})\right\rangle_{\mathcal{H}}$. By Lemma 11 in Abbasi-Yadkori et al. (2011) (see also Lemma E.3 in Yang et al. (2020b)), we have

$$
\sum_{n=1}^N \left\langle \mu(\widehat{\omega}_h^n, a_h^n), \Sigma_h^{-1}\mu(\widehat{\omega}_h^n, a_h^n)\right\rangle_{\mathcal{H}} \leq 2\log\det\left(I + K_h/\lambda\right). \quad (27)
$$

Substituting (27) into (26), we obtain

$$
|(A)| \leq 2\left(\sqrt{\frac{1}{m}} + \sqrt{\frac{\log(1/\delta_A)}{m}}\right)\sqrt{\log\det(I + K_h/\lambda)}\left\|\Sigma_h^{-1/2}\mu(\omega,a)\right\|_{\mathcal{H}}
$$

with probability $1-\delta_A$. □

C.6 Proof of Lemma 5

Proof. The proof is based on concentration of measure in self-normalizing sequences. Let $\mathcal{F}_{h,\tau}$ be the σ-algebra generated by data $\{(s_h^n, a_h^n, r_h^n)\}_{n=1}^\tau$. By the Markov property and definition of Bellman operator, we have $\mathbb{E}[\Delta^\tau(V_{h+1}) \mid \mathcal{F}_{h,\tau-1}] = 0$. Moreover, $\Delta^n(V_{h+1}) \leq H$, since the reward function is bounded by 1 in Assumption 3. By Theorem 1 in Chowdhury and Gopalan (2017), with probability at least $1-\delta$, for any $\eta > 0$, we have

$$
\Delta(V_{h+1})^\top((K_h + \eta I)^{-1} + I)^{-1}\Delta(V_{h+1}) \leq 2H^2\log\det((1+\eta)I + K_h) + 2H^2\log\frac{1}{\delta}, \quad (28)
$$

where $\Delta(V_{h+1}) = [\Delta^1(V_{h+1}), \ldots, \Delta^N(V_{h+1})]^\top$. In the remaining of the proof, we take $\eta = \lambda - 1$ and show

$$
\left\|\Sigma_h^{-1/2}\sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\Delta^n(V_{h+1})\right\|_{\mathcal{H}}^2 \leq \Delta(V_{h+1})^\top((K_h + (\lambda-1)I)^{-1} + I)^{-1}\Delta(V_{h+1}).
$$

As a consequence, (28) is an upper bound of $\left\|\Sigma_h^{-1/2}\sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\Delta^n(V_{h+1})\right\|_{\mathcal{H}}^2$, which is the desired result.

Using vector notations, we rewrite $\Sigma_h^{-1/2}\sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\Delta^n(V_{h+1})$ as $\Sigma_h^{-1/2}\Phi_h^\top\Delta(V_{h+1})$. Then we derive

$$
\begin{aligned}
\left\|\Sigma_h^{-1/2}\sum_{n=1}^N \mu(\widehat{\omega}_h^n, a_h^n)\Delta^n(V_{h+1})\right\|_{\mathcal{H}}^2 &= \left\|\Sigma_h^{-1/2}\Phi_h^\top\Delta(V_{h+1})\right\|_{\mathcal{H}}^2 \\
&= \Delta(V_{h+1})^\top\left\langle \Phi_h, \Sigma_h^{-1}\Phi_h^\top\right\rangle\Delta(V_{h+1}) \\
&\overset{(i)}{=} \Delta(V_{h+1})^\top\left\langle \Phi_h, \Phi_h^\top\right\rangle(\lambda I + K_h)^{-1}\Delta(V_{h+1}) \\
&= \Delta(V_{h+1})^\top K_h(\lambda I + K_h)^{-1}\Delta(V_{h+1}),
\end{aligned}
$$

where equality (i) invokes the identity in (13). Now we need to show $\Delta(V_{h+1})^\top((K_h + (\lambda - 1)I)^{-1} + I)^{-1}\Delta(V_{h+1}) \geq \Delta(V_{h+1})^\top K_h(\lambda I + K_h)^{-1}\Delta(V_{h+1})$. Indeed, by the matrix inversion lemma, we have

$$\begin{aligned}
\Delta(V_{h+1})^\top((K_h + (\lambda - 1)I)^{-1} + I)^{-1}\Delta(V_{h+1}) &= \Delta(V_{h+1})^\top(I - (K_h + \lambda I)^{-1})\Delta(V_{h+1}) \\
&\geq \Delta(V_{h+1})^\top(I - \lambda(K_h + \lambda I)^{-1})\Delta(V_{h+1}) \\
&= \Delta(V_{h+1})^\top K_h(\lambda I + K_h)^{-1}\Delta(V_{h+1}).
\end{aligned}$$

The proof is complete. □

C.7 Proof of Lemma 6

Proof. We reduce the covering of $\mathcal{V}_h(R, B, \lambda)$ to Cartesian product of coverings on θ, β, and Σ. Specifically, let f_1, f_2 be two elements in $\mathcal{H}_h$. We denote

$$\begin{aligned}
f_1(\omega, a) &= \max_{a \in \mathcal{A}} \left[\min\left\{\langle \mu(\omega, a), \theta_1\rangle - \beta_1\sqrt{\langle \mu(\omega, a), \Sigma_1\mu(\omega, a)\rangle}, H - h + 1\right\}^+\right] \quad \text{and} \\
f_2(\omega, a) &= \max_{a \in \mathcal{A}} \left[\min\left\{\langle \mu(\omega, a), \theta_2\rangle - \beta_2\sqrt{\langle \mu(\omega, a), \Sigma_2\mu(\omega, a)\rangle}, H - h + 1\right\}^+\right].
\end{aligned}$$

We evaluate the difference between f_1 and f_2:

$$\begin{aligned}
&\|f_1 - f_2\|_\infty \\
&\overset{(i)}{\leq} \sup_{(\omega,a)} \left|\langle \mu(\omega, a), \theta_1 - \theta_2\rangle - \left(\beta_1\sqrt{\langle \mu(\omega, a), \Sigma_1^{-1}\mu(\omega, a)\rangle} - \beta_2\sqrt{\langle \mu(\omega, a), \Sigma_2^{-1}\mu(\omega, a)\rangle}\right)\right| \\
&\leq \sup_{(\omega,a)} |\langle \mu(\omega, a), \theta_1 - \theta_2\rangle| + \sup_{(\omega,a)} |\beta_1 - \beta_2|\sqrt{\langle \mu(\omega, a), \Sigma_1\mu(\omega, a)\rangle} \\
&\quad + \sup_{(\omega,a)} \beta_2\left|\sqrt{\langle \mu(\omega, a), \Sigma_1\mu(\omega, a)\rangle} - \sqrt{\langle \mu(\omega, a), \Sigma_2\mu(\omega, a)\rangle}\right| \\
&\overset{(ii)}{\leq} \|\theta_1 - \theta_2\|_{\mathcal{H}} + \lambda^{-1/2}|\beta_1 - \beta_2| \\
&\quad + B\sup_{(\omega,a)} \left|\sqrt{\langle \mu(\omega, a), \Sigma_1\mu(\omega, a)\rangle} - \sqrt{\langle \mu(\omega, a), \Sigma_2\mu(\omega, a)\rangle}\right| \\
&\leq \|\theta_1 - \theta_2\|_{\mathcal{H}} + \lambda^{-1/2}|\beta_1 - \beta_2| \\
&\quad + B\sup_{(\omega,a)} \sqrt{|\langle \mu(\omega, a), (\Sigma_1 - \Sigma_2)\mu(\omega, a)\rangle|},
\end{aligned} \tag{29}$$

where inequality (i) removes the truncation operation in f_1, f_2, and inequality (ii) follows from $\sigma_1 \succeq \lambda I_{\mathcal{H}_K}$. Decomposition (29) suggests that an ϵ-covering of $\mathcal{V}_h$ can be constructed from the Cartesian product of an $\epsilon/3$-covering on $\mathcal{F}_1 = \{\theta : \|\theta\|_{\mathcal{H}} \leq R\}$, an $\epsilon\sqrt{\lambda}/3$-covering on $\mathcal{F}_2 = \{\beta : 0 \leq \beta \leq B\}$, and an $\epsilon/(3B)$-covering on $\mathcal{F}_3 = \{\Sigma : \lambda^{-1}\mathcal{I}_{\mathcal{H}_K} \succeq \Sigma \succeq 0\}$. Correspondingly, the covering number of $\mathcal{V}_h$ is the product

$$\mathcal{N}(\epsilon, \mathcal{V}_h(R, B, \lambda), \|\cdot\|_\infty) = \mathcal{N}(\epsilon/3, \mathcal{F}_1, \|\cdot\|_\infty) \cdot \mathcal{N}(\epsilon\sqrt{\lambda}/3, \mathcal{F}_2, \|\cdot\|_\infty) \cdot \mathcal{N}(\epsilon/(3B), \mathcal{F}_3, \|\cdot\|_\infty).$$

By Lemma 8, we have

$$\log \mathcal{N}(\epsilon/3, \mathcal{F}_1, \|\cdot\|_\infty) \leq \begin{cases} \gamma \log(1 + 6R/\epsilon), & (\textit{Finite Spectrum}) \\ C_3\left(\log(3R/\epsilon) + C_4\right)^{1+1/\gamma}, & (\textit{Exponential Decay}) \\ C_5(3R/\epsilon)^{2/(\gamma-1)}\log(1 + 12R/\epsilon), & (\textit{Polynomial Decay}) \end{cases}.$$

A direct discretization yields

$$\log \mathcal{N}(\epsilon\sqrt{\lambda}/3, \mathcal{F}_2, \|\cdot\|_\infty) \leq \log\left(1 + \frac{3B}{\epsilon\sqrt{\lambda}}\right).$$

By Lemma 9, we have

$$
\begin{aligned}
&\log \mathcal{N}(\epsilon/(3B), \mathcal{F}_3, \|\cdot\|_\infty) \\
&\quad \le \begin{cases} \gamma^2 \log(1 + 6B\sqrt{\gamma}/(\lambda\epsilon)), & (\textit{Finite Spectrum}) \\ C_6 \left(\log(3B/\epsilon) + \log\log(3B/\epsilon) + C_7\right)^{1+2/\gamma}, & (\textit{Exponential Decay}) \\ C_8 (3B/\epsilon)^{4/(\gamma-1)} \left(\log(3B/\epsilon) + C_9\right), & (\textit{Polynomial Decay}) \end{cases}.
\end{aligned}
$$

Combining all these covering numbers, we deduce

$$
\begin{aligned}
&\log \mathcal{N}(\epsilon, \mathcal{V}_h(R, B, \lambda), \|\cdot\|_\infty) \\
&\le \begin{cases} \gamma \log(1 + 6R/\epsilon) + \gamma^2 \log(1 + 6B\sqrt{\gamma}/\epsilon), & (\textit{Finite Spectrum}) \\ C_3 \left(\log \frac{3R}{\epsilon} + C_4\right)^{\frac{1+\gamma}{\gamma}} + C_6 \left(\log \frac{3B}{\epsilon} + \log\log \frac{3B}{\epsilon} + C_7\right)^{\frac{\gamma+2}{\gamma}},, & (\textit{Exponential Decay}) \\ C_5 \left(\frac{3R}{\epsilon}\right)^{\frac{2}{\gamma-1}} \log\left(1 + \frac{12R}{\epsilon}\right) + C_8 \left(\frac{3B}{\epsilon}\right)^{\frac{4}{\gamma-1}} \left(\log \frac{3B}{\epsilon} + C_9\right), & (\textit{Polynomial Decay}) \end{cases} \\
&\quad + \log\left(1 + \frac{3B}{\epsilon\sqrt{\lambda}}\right).
\end{aligned}
$$

The proof is complete. □

C.8 Proof of Lemma 7

Proof. We observe $r_h^n + \widehat{V}_{h+1}$ uniformly bounded by H. For any f with $\|f\|_{\mathcal{H}} \le 1$, similar to the proof of Lemma 4, we have

$$
\begin{aligned}
\|\widehat{w}_h\|_{\mathcal{H}} &= \sup_{\|f\|_{\mathcal{H}} \le 1} \langle f, \widehat{w}_h \rangle \\
&= \sup_{\|f\|_{\mathcal{H}} \le 1} \left\langle f, \Sigma_h^{-1} \Phi_h^\top [r_h^1 + \widehat{V}_{h+1}(\widehat{\omega}_{h+1}^1), \ldots, r_h^N + \widehat{V}_{h+1}(\widehat{\omega}_{h+1}^N)]^\top \right\rangle \\
&\le \sup_{\|f\|_{\mathcal{H}} \le 1} H \lambda^{-1/2} \|f\|_{\mathcal{H}} \sqrt{\sum_{i=1}^N \left\langle \mu(\widehat{\omega}_h^n, a_h^n), \Sigma_h^{-1} \mu(\widehat{\omega}_h^n, a_h^n) \right\rangle} \\
&\le H \lambda^{-1/2} \sqrt{\log\det(I + K_h/\lambda)},
\end{aligned}
$$

where the second last inequality follows from Cauchy-Schwarz inequality and the last inequality invokes Lemma 11 in Abbasi-Yadkori et al. (2011). □

D Technical Results in RKHS

D.1 Covering Number of RKHS

Lemma 8 (Restatement of Lemma D.2 in Yang et al. (2020b))**.** Suppose Assumptions 1 and 2 hold. Let $\mathcal{H}_K(R) = \{f \in \mathcal{H}_K : \|f\|_{\mathcal{H}} \le R\}$ be a norm ball in the reproducing kernel Hilbert space $\mathcal{H}_K$. For any $\epsilon > 0$, the covering number $\mathcal{N}(\epsilon, \mathcal{H}_K(R), \|\cdot\|_\infty)$ is bounded by

$$
\log \mathcal{N}(\epsilon, \mathcal{H}_K(R), \|\cdot\|_\infty) \le \begin{cases} \gamma \log(1 + 2R/\epsilon), & (\textit{Finite Spectrum}) \\ C_3 \left(\log(R/\epsilon) + C_4\right)^{1+1/\gamma}, & (\textit{Exponential Decay}) \\ C_5 (R/\epsilon)^{2/(\gamma-1)} \log(1 + 4R/\epsilon), & (\textit{Polynomial Decay}) \end{cases},
$$

where C_3, C_4 are positive constants depending on $C_1, C_2, \log R$, and constant C_5 depends on $C, \log R, \gamma$.

Proof. The idea is to transform the covering of $\mathcal{H}_K(R)$ to a proper covering on a Euclidean ball, whose dimension is determined by the spectrum of kernel K.

Case 1: (*Finite Spectrum*). For any function $f \in \mathcal{H}_K(R)$, we have

$$
f = \sum_{i=1}^{\gamma} w_i \sqrt{\sigma_i} \nu_i \quad \text{with} \quad \sum_{i=1}^{\gamma} w_i^2 \le R^2.
$$

Consider two functions $f = \sum_{i=1}^{\gamma} w_i \sqrt{\sigma_i} \nu_i$ and $f' = \sum_{i=1}^{\gamma} w_i' \sqrt{\sigma_i} \nu_i$ satisfying $\sum_{i=1}^{\gamma} (w_i - w_i')^2 \leq \epsilon_0^2$. Then we have

$$\begin{aligned}
\|f - f'\|_\infty &= \left\| \sum_{i=1}^{\gamma} (w_i - w_i') \sqrt{\sigma_i} \nu_i \right\|_\infty \\
&\overset{(i)}{\leq} \sup_\mu \sqrt{\sum_{i=1}^{\gamma} \sigma_i \nu_i^2(\mu)} \sqrt{\sum_{i=1}^{\gamma} (w_i - w_i')^2} \\
&= \sqrt{\sup_\mu K(\mu, \mu) \cdot \sum_{i=1}^{\gamma} (w_i - w_i')^2} \\
&\leq \epsilon_0,
\end{aligned}$$

where inequality (i) uses the Cauchy-Schwarz inequality. The above inequality establishes the equivalence between an ϵ_0-covering on $\mathcal{B}^\gamma(R) = \{w \in \mathbb{R}^\gamma : \|w\|_2 \leq R\}$ and an ϵ_0-covering on $\mathcal{H}_K(R)$. By the volume ratio argument, we know

$$\mathcal{N}(\epsilon, \mathcal{B}^\gamma(R), \|\cdot\|_2) \leq (1 + 2R/\epsilon)^\gamma.$$

Therefore, we immediately obtain

$$\log \mathcal{N}(\epsilon, \mathcal{H}_K(R), \|\cdot\|_\infty) \leq \gamma \log(1 + 2R/\epsilon).$$

Case 2: (*Exponential Decay*). In both **Case 2** and **Case 3**, we properly truncate the spectrum of kernel K, which reduces to **Case 1**. Thanks to the specific spectrum decay, we are able to estimate the truncation error. Specifically, let M be a positive integer to be determined. We denote Π_M as the projection operator onto the eigenspace spanned by $\nu_1, \ldots, \nu_M$. Now for any $f \in \mathcal{H}_K(R)$ and $M \geq \left(\frac{1-\gamma}{C_2 \gamma}\right)^{1/\gamma}$ for $\gamma \in (0, 1)$, we bound

$$\begin{aligned}
\|f - \Pi_M f\|_\infty &= \left\| \sum_{i=M+1}^{\infty} w_i \sqrt{\sigma_i} \nu_i \right\|_\infty \\
&\leq \sqrt{\sum_{i=M+1}^{\infty} w_i^2} \sqrt{\sup_\mu \sum_{i=M+1}^{\infty} \sigma_i \nu_i^2(\mu)} \\
&\overset{(i)}{\leq} R \sqrt{\sum_{i=M+1}^{\infty} C_1 \exp\left(-C_2 i^\gamma\right)} \\
&\overset{(ii)}{\leq} R \sqrt{\int_M^\infty C_1 \exp\left(-C_2 x^\gamma\right) dx} \\
&\overset{(iii)}{\leq} \begin{cases} R\sqrt{\frac{C_1}{C_2} \exp(-C_2 M^\gamma)} & \gamma \geq 1 \\ R\sqrt{\frac{C_1 M^{1-\gamma}}{C_2 \gamma} \exp(-C_2 M^\gamma)} & \gamma \in (0, 1) \end{cases}.
\end{aligned} \tag{30}$$

where inequality (i) follows from the uniform upper bound on ν_i, inequality (ii) uses the monotonicity of exponential function, and inequality (iii) is valid due to the following computation. For $\gamma \geq 1$, we derive

$$\int_M^\infty C_1 \exp\left(-C_2 x^\gamma\right) dx \leq \int_M^\infty C_1 x^{\gamma - 1} \exp\left(-C_2 x^\gamma\right) dx = \frac{C_1}{C_2} \exp\left(-C_2 M^\gamma\right).$$

When $\gamma \in (0,1)$, and M sufficiently large, we can also bound the truncation error using a slightly more complicated argument. Indeed, using integration by parts, we have

$$\begin{aligned}
&\int_M^\infty C_1 \exp\left(-C_2 x^\gamma\right) dx \\
&= \int_{M^\gamma}^\infty C_1 \exp(-C_2 u) \frac{1}{\gamma} u^{1/\gamma - 1} du \\
&= \frac{C_1}{C_2\gamma} M^{1-\gamma} \exp(-C_2 M^\gamma) + \frac{C_1}{C_2} \int_{M^\gamma}^\infty \frac{1}{\gamma}\left(\frac{1}{\gamma} - 1\right) u^{1/\gamma - 2} \exp(-C_2 u) du \\
&\le \frac{C_1}{C_2\gamma} M^{1-\gamma} \exp(-C_2 M^\gamma) + \frac{C_1}{C_2 M^\gamma} \int_{M^\gamma}^\infty \frac{1}{\gamma}\left(\frac{1}{\gamma} - 1\right) u^{1/\gamma - 1} \exp(-C_2 u) du \\
&= \frac{C_1}{C_2\gamma} M^{1-\gamma} \exp(-C_2 M^\gamma) + \frac{1}{C_2 M^\gamma}\left(\frac{1}{\gamma} - 1\right) \int_M^\infty C_1 \exp\left(-C_2 x^\gamma\right) dx,
\end{aligned}$$

which implies, for $M \ge \left(\frac{1-\gamma}{C_2\gamma}\right)^{1/\gamma}$,

$$\begin{aligned}
\int_M^\infty C_1 \exp\left(-C_2 x^\gamma\right) dx &\le \frac{C_1 M^{1-\gamma}}{C_2\gamma - (1-\gamma) M^{-\gamma}} \exp(-C_2 M^\gamma) \\
&\le \frac{C_1 M^{1-\gamma}}{C_2\gamma} \exp(-C_2 M^\gamma).
\end{aligned}$$

To this end, we construct an ϵ_0-covering on $\Pi_M \mathcal{H}_K(R) = \{\Pi_M f : f \in \mathcal{H}_K(R)\}$, whose covering number is given in **Case 1**,

$$\mathcal{N}(\epsilon_0, \Pi_M \mathcal{H}_K(R), \|\cdot\|_\infty) \le (1 + 2R/\epsilon_0)^M.$$

For a given ϵ, and a proper absolute constant c_0, we choose M to be the smallest integer satisfying

$$R\sqrt{\int_M^\infty C_1 \exp\left(-C_2 x^\gamma\right) dx} \le \epsilon/2 \quad \Longrightarrow \quad M = c_0 \left\lceil \left(\frac{2}{C_2} \log \frac{2R}{\epsilon} + \frac{1}{C_2} \log \frac{C_1}{C_2}\right)^{1/\gamma} \right\rceil \tag{31}$$

and $\epsilon_0 = \epsilon/2$. We claim that the $\epsilon/2$-covering of $\Pi_M \mathcal{H}_K(R)$ with M chosen in (31) is also an ϵ-covering of $\mathcal{H}_K(R)$. To see this, for any given $f \in \mathcal{H}_K(R)$, there exists $\bar{f}_M$ in the covering of $\Pi_M \mathcal{H}_K(R)$ such that $\left\|\bar{f}_M - \Pi_M f\right\|_\infty \le \epsilon/2$. Then we have

$$\left\|f - \bar{f}_M\right\|_\infty \le \|f - \Pi_M f\|_\infty + \left\|\Pi_M f - \bar{f}_M\right\|_\infty \le \epsilon.$$

This implies the ϵ-covering number of $\mathcal{H}_K(R)$ is

$$\begin{aligned}
\log \mathcal{N}(\epsilon, \mathcal{H}_K(R), \|\cdot\|_\infty) &\le M \log(1 + 4R/\epsilon) \\
&= c_0 \left\lceil \left(\frac{2}{C_2} \log \frac{2R}{\epsilon} + \frac{1}{C_2} \log \frac{C_1}{C_2}\right)^{1/\gamma} \right\rceil \log(1 + 4R/\epsilon) \\
&= C_3 \left(\log(R/\epsilon) + C_4\right)^{1+1/\gamma},
\end{aligned}$$

where constants C_3, C_4 depend on C_1, C_2, R.

Case 3: (*Polynomial Decay*). The idea is the same as in **Case 2**, except a different upper bound on the truncation error. Specifically, for polynomial decay spectrum and any $f \in \mathcal{H}_K(R)$, we have

$$\begin{aligned}
\|f - \Pi_M f\|_\infty &= \left\| \sum_{i=M+1}^\infty w_i \sqrt{\sigma_i} \nu_i \right\|_\infty \\
&\le R\sqrt{\int_M^\infty C x^{-\gamma} dx} \\
&\le R\sqrt{(\gamma - 1)^{-1} C M^{-\gamma+1}}.
\end{aligned} \tag{32}$$

By letting M be the smallest integer satisfying $R\sqrt{(\gamma-1)^{-1}CM^{-\gamma+1}} \leq \epsilon/2$, i.e.,

$$M = \left\lceil \left(\frac{4R^2C}{\epsilon^2(\gamma-1)}\right)^{\frac{1}{\gamma-1}} \right\rceil,$$

we have

$$\begin{aligned}\log \mathcal{N}(\epsilon, \mathcal{H}_K(R), \|\cdot\|_\infty) \leq M \log(1+4R/\epsilon) &= \left\lceil \left(\frac{4R^2C}{\epsilon^2(\gamma-1)}\right)^{\frac{1}{\gamma-1}} \right\rceil \log(1+4R/\epsilon) \\ &= C_5 \left(\frac{R}{\epsilon}\right)^{\frac{2}{\gamma-1}} \log(1+4R/\epsilon),\end{aligned}$$

where constant C_5 depends on C, R, γ. The proof is complete. $\square$

Lemma 9. Suppose Assumption 1 and 2 hold. Let $\mathcal{F}_K(\lambda) = \{\Sigma : \|\Sigma\|_{\mathrm{op}} \overset{\triangle}{=} \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \Sigma f\rangle_{\mathcal{H}} \leq \lambda^{-1}\}$ be a collection of operators of bounded operator norm defined on the RKHS $\mathcal{H}_K$. For any $\epsilon > 0$, the covering number $\mathcal{N}(\epsilon, \mathcal{F}_K(\lambda), \|\cdot\|_{\mathrm{op}})$ is bounded by

$$\log \mathcal{N}(\epsilon, \mathcal{F}_K(\lambda), \|\cdot\|_{\mathrm{op}}) \leq \begin{cases} \gamma^2 \log(1+2\sqrt{\gamma}/(\lambda\epsilon)), & (\mathit{Finite\ Spectrum}) \\ C_6\left(\log(1/\epsilon) + \log\log(1/\epsilon) + C_7\right)^{1+2/\gamma}, & (\mathit{Exponential\ Decay}) \\ C_8(1/\epsilon)^{4/(\gamma-1)}\left(\log(1/\epsilon) + C_9\right), & (\mathit{Polynomial\ Decay}) \end{cases},$$

where C_6, C_7 are positive constants depending on C_1, C_2, λ, and constants C_8, C_9 depend on C, λ, γ.

Proof. Similar to the proof of Lemma 8, the idea here is to transform the covering of $\mathcal{F}_K(\lambda)$ to a proper matrix covering, whose dimension is determined by the spectrum of kernel K.

Case 1: (*Finite Spectrum*). We show an equivalence between operator Σ and square matrix $M \in \mathbb{R}^{\gamma\times\gamma}$. For any unit norm eigenfunction $\sqrt{\sigma_i}\nu_i \in \mathcal{H}_K$, we denote

$$\Sigma(\sqrt{\sigma_i}\nu_i) = \sum_{j=1}^{\gamma} w_{ij}\sqrt{\sigma_j}\nu_j.$$

Let matrix $W_{ij} = w_{ij}$. Then for any $f \in \mathcal{H}_K$, we write $f = \sum_{j=1}^{\gamma} a_j\sqrt{\sigma_j}\nu_j$. Some algebra gives rise to

$$\langle f, \Sigma f\rangle_{\mathcal{H}} = a^\top W a \quad \text{with} \quad a = [a_1, \dots, a_\gamma]^\top.$$

This yields a one-to-one correspondence between $\mathcal{F}_K(\lambda)$ and $\mathcal{W}_\gamma(\lambda) = \{W \in \mathbb{R}^{\gamma\times\gamma} : \lambda^{-1}I \succeq W \succeq 0\}$. Therefore, it suffices to find the covering number of $\mathcal{W}_\gamma(\lambda)$. By vectorize a γ-by-γ matrix as a γ^2-dimensional vector, we obtain the covering number of $\mathcal{W}_\gamma(\lambda)$ using the volume ratio argument:

$$\log \mathcal{N}(\epsilon, \mathcal{W}_\gamma(\lambda), \|\cdot\|_{\mathrm{F}}) \leq \gamma^2 \log(1+2\sqrt{\gamma}/(\lambda\epsilon)),$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Accordingly, we have

$$\log \mathcal{N}(\epsilon, \mathcal{F}_K(\lambda), \|\cdot\|_{\mathrm{op}}) \leq \gamma^2 \log(1+2\sqrt{\gamma}/(\lambda\epsilon)).$$

Case 2: (*Exponential Decay*). We truncate the spectrum of kernel K again. Let M be a positive integer to be determined. We denote Π_M as the projection operator onto the eigenspace spanned by $\nu_1, \dots, \nu_M$. For any $f \in \mathcal{H}_K$ with $\|f\|_{\mathcal{H}} \leq 1$, the truncation error is bounded by (30):

$$\|f - \Pi_M f\|_\infty \leq \begin{cases} \sqrt{\frac{C_1}{C_2}\exp(-C_2M^\gamma)} & \gamma \geq 1 \\ \sqrt{\frac{C_1M^{1-\gamma}}{C_2\gamma}\exp(-C_2M^\gamma)} & \gamma \in (0,1) \end{cases}.$$

By the linearity, we have

$$\langle f, \Sigma f\rangle = \langle \Pi_M f, \Sigma(\Pi_M f)\rangle + \langle f - \Pi_M f, \Sigma(\Pi_M f)\rangle + \langle f, \Sigma(f - \Pi_M f)\rangle. \tag{33}$$

The last two terms in (33) can be bounded by the truncation error:

$$\langle f-\Pi_M f, \Sigma\Pi_M f\rangle + \langle f, \Sigma(f-\Pi_M f)\rangle \le 2\lambda^{-1}\begin{cases}\sqrt{\frac{C_1}{C_2}\exp(-C_2M^\gamma)} & \gamma\ge 1\\ \sqrt{\frac{C_1M^{1-\gamma}}{C_2\gamma}\exp(-C_2M^\gamma)} & \gamma\in(0,1)\end{cases}.$$

Given an $\epsilon>0$, by choosing M to be the smallest integer satisfying

$$2\lambda^{-1}\sqrt{\frac{C_1}{C_2}\exp(-C_2M^\gamma)} \le \epsilon/2 \quad\Longrightarrow\quad M = c_0\left\lceil\left(\frac{2}{C_2}\log\frac{4}{\lambda\epsilon}+\frac{1}{C_2}\log\frac{C_1}{C_2}\right)^{1/\gamma}\right\rceil, \tag{34}$$

for some absolute constant c_0, we only need to find an $\epsilon/2$-covering of Σ evaluated on $\Pi_M\mathcal{H}_K$. This reduces to **Case 1**, and the covering number is bounded by $(1+2\sqrt{M}/(\lambda\epsilon))^{M^2}$. With the choice of M in (34), we obtain

$$\begin{aligned}\log\mathcal{N}(\epsilon,\mathcal{F}_K(\lambda),\|\cdot\|_{\mathrm{op}}) &\le M^2\log(1+4\sqrt{M}/(\lambda\epsilon))\\ &= c_0\left\lceil\left(\frac{2}{C_2}\log\frac{4}{\lambda\epsilon}+\frac{1}{C_2}\log\frac{C_1}{C_2}\right)^{1/\gamma}\right\rceil^2\log(1+4\sqrt{M}/(\lambda\epsilon))\\ &= C_6\left(\log\left(\frac{\log 1/\epsilon}{\epsilon}\right)+C_7\right)^{1+2/\gamma},\end{aligned}$$

where constants C_6, C_7 depend on C_1, C_2, λ.

Case 3: (*Polynomial Decay*). The idea is the same as in **Case 2**. For polynomial decay spectrum and any $f\in\mathcal{H}_K$ with $\|f\|_{\mathcal{H}}\le 1$, by (32), the truncation error is bounded by

$$\|f-\Pi_M f\|_\infty \le \sqrt{(\gamma-1)^{-1}CM^{-\gamma+1}}.$$

By letting M be the smallest integer satisfying $2\sqrt{(\gamma-1)^{-1}CM^{-\gamma+1}}\le\epsilon/2$, i.e.,

$$M = \left\lceil\left(\frac{16C}{\epsilon^2(\gamma-1)}\right)^{\frac{1}{\gamma-1}}\right\rceil,$$

we have

$$\begin{aligned}\log\mathcal{N}(\epsilon,\mathcal{F}_K(\lambda),\|\cdot\|_{\mathrm{op}}) \le M^2\log(1+4\sqrt{M}/(\lambda\epsilon)) &= \left\lceil\left(\frac{16C}{\epsilon^2(\gamma-1)}\right)^{\frac{1}{\gamma-1}}\right\rceil^2\log(1+4\sqrt{M}/(\lambda\epsilon))\\ &= C_8\left(\frac{1}{\epsilon}\right)^{\frac{4}{\gamma-1}}(\log(1/\epsilon)+C_9),\end{aligned}$$

where constants C_8, C_9 depend on C, λ, γ. The proof is complete. □

D.2 Effective Dimension of RKHS

Lemma 10. Suppose that Assumptions 1 and 2 hold. Recall that the Gram matrix $[K_h]_{\ell,\ell'} = K((\widehat{\omega}_h^\ell, a_h^\ell),(\widehat{\omega}_h^{\ell'}, a_h^{\ell'}))$ defined on dataset $\mathcal{D}_{N,H}$. For any $h=1,\dots,H$ and fixed $\lambda>0$, we have

$$\log\det(I+K_h/\lambda) \le \begin{cases} C_{\text{eff-FS}}\cdot\max\{d,\gamma\}\log N & (\textit{Finite Spectrum})\\ C_{\text{eff-ED}}\cdot d(\log N)^{1+1/\gamma} & (\textit{Exponential Decay})\\ C_{\text{eff-PD}}\cdot dN^{\frac{d+1}{\gamma+d}}\log N & (\textit{Polynomial Decay})\end{cases},$$

where d is the dimension of mete state-action space $\mathcal{S}\times\mathcal{S}\times\mathcal{A}$, constant $C_{\text{eff-FS}}$ depends λ and Lebesgue measure of meta state-action space Ξ, constant $C_{\text{eff-ED}}$ depends on C_1, C_2, λ, and Lebesgue measure of meta state-action space Ξ, and constant $C_{\text{eff-PD}}$ depends on C, λ, γ, and Lebesgue measure of meta state-action space Ξ.

Proof. Each entry in the Gram matrix $K_h \in \mathbb{R}^{N\times N}$ can be written as $[K_h]_{\ell,\ell'} = \mathbb{E}_{s_1\sim\widehat{d}_h^{\ell}, s_2\sim\widehat{d}_h^{\ell'}}[K((s_{h,0}^{\ell}, s_1, a_h^{\ell}), (s_{h,0}^{\ell'}, s_2, a_h^{\ell'}))]$. Correspondingly, we define $\widetilde{K}_h \in \mathbb{R}^{N\times N}$ as $[\widetilde{K}_h]_{\ell,\ell'} = K((s_{h,0}^{\ell}, s_{h,1}^{\ell}, a_h^{\ell}), (s_{h,0}^{\ell'}, s_{h,1}^{\ell'}, a_h^{\ell'}))$, which can be viewed as only sampling a single agent from the mean-field state distribution. Observe that $\log\det$ is a concave function. By Jensen's inequality, we have

$$\log\det(I + K_h/\lambda) \le \mathbb{E}[\log\det(I + \widetilde{K}_h/\lambda)] \le \sup_{\widetilde{K}_h} \log\det(I + \widetilde{K}_h/\lambda),$$

where the expectation is taken over the empirical states of m agents. To bound $\log\det(I + K_h/\lambda)$, we only need to bound $\sup_{\widetilde{K}_h} \log\det(I + \widetilde{K}_h/\lambda)$. We introduce several notations. For fixed $\tau > 0$, we denote $C_\tau = 2\mu(\Xi)(2\tau + 1)$ with $\mu(\Xi)$ being the Lebesgue measure of Ξ and $n_\tau = N^\tau \log N$. Moreover, we denote d as the dimension of Ξ and $B_\sigma(N_\star) = \sum_{i=N_\star+1}^{\infty} \sigma_i$ as the tail spectrum of kernel K.

By Theorem 8 in Srinivas et al. (2009), for any positive integer $N_\star \le n_\tau$, we have

$$\sup_{\widetilde{K}_h} \log\det(I + \widetilde{K}_h/\lambda) \le N_\star \log(\lambda N^{1+\tau}) + C_\tau \lambda \log N (N^{\tau+1} B_\sigma(N_\star) + 1) + \mathcal{O}(N^{1-\tau/d}). \tag{35}$$

We choose $N_\star$ and τ according to the spectrum of kernel K.

Case 1: *(Finite Spectrum)*. We set $N_\star = \gamma$, which implies $B_\sigma(N_\star) = 0$, and $\tau = d$. Plugging into (35), we have

$$\sup_{\widetilde{K}_h} \log\det(I + \widetilde{K}_h/\lambda) \le \gamma \log(\lambda N^{1+\tau}) + C_\tau \lambda \log N + \mathcal{O}(1) = C_{\text{eff}-\text{FS}} \cdot \max\{d, \gamma\} \log N,$$

where $C_{\text{eff-FS}}$ depends on $\lambda, \mu(\Xi)$.

Case 2: *(Exponential Decay)*. We first bound the tail spectrum $B_\sigma(N_\star)$. By the computation in (30), we have

$$B_\sigma(N_\star) \le \begin{cases} \frac{C_1}{C_2}\exp(-C_2 N_\star^{\gamma}) & \gamma \ge 1 \\ \frac{C_1 N_\star^{1-\gamma}}{C_2 \gamma}\exp(-C_2 N_\star^{\gamma}) & \gamma \in (0,1) \end{cases}.$$

By setting $\tau = d$ again, substituting the upper bound of $B_\sigma(N_\star)$ into (35), and further choosing $N_\star = \frac{c_0}{C_2}(d+1)(\log N)^{1/\gamma}$, for some absolute constant c_0, we deduce

$$\begin{aligned}\sup_{\widetilde{K}_h} \log\det(I + \widetilde{K}_h/\lambda) &\le \frac{1}{C_2}(d+1)(\log N)^{1/\gamma} \log(\lambda N^{1+d}) + C_\tau \lambda \log N \left(1 + \frac{C_1}{C_2}\right) + \mathcal{O}(1) \\ &= C_{\text{eff-ED}} \cdot d(\log N)^{1+1/\gamma},\end{aligned}$$

where constant $C_{\text{eff-ED}}$ depends on $C_1, C_2, \lambda, \mu(\Xi)$.

Case 3: *(Polynomial Decay)*. Similar to **Case 2**, we bound $B_\sigma(N_\star)$ using (32) as

$$B_\sigma(N_\star) \le (\gamma - 1)^{-1} C N_\star^{-\gamma+1}.$$

Substituting the upper bound of $B_\sigma(N_\star)$ into (35), we obtain

$$\begin{aligned}&\sup_{\widetilde{K}_h} \log\det(I + \widetilde{K}_h/\lambda) \\ &\qquad \le N_\star \log(\lambda N^{1+\tau}) + C_\tau \lambda \log N \left(1 + N^{\tau+1}(\gamma-1)^{-1} C N_\star^{-\gamma+1}\right) + \mathcal{O}(N^{1-\tau/d}).\end{aligned}$$

Choosing $N_\star = N^{(\tau+1)/\gamma}$, we deduce

$$\begin{aligned}&\sup_{\widetilde{K}_h} \log\det(I + \widetilde{K}_h/\lambda) \\ &\qquad \le N^{(\tau+1)/\gamma} \log(\lambda N^{1+\tau}) + C_\tau \lambda \log N \left(1 + N^{(\tau+1)/\gamma}(\gamma-1)^{-1} C\right) + \mathcal{O}(N^{1-\tau/d}).\end{aligned}$$

Now we set $\tau = \frac{\gamma-1}{1+\gamma/d}$ so that $N^{(\tau+1)/\gamma} = N^{1-\tau/d}$, which further gives rise to

$$\begin{aligned}&\sup_{\widetilde{K}_h} \log\det(I + \widetilde{K}_h/\lambda)\\ &\qquad\qquad \leq N^{\frac{d+1}{d+\gamma}} \log(\lambda N^{\frac{(d+1)\gamma}{d+\gamma}}) + C_\tau \lambda \log N \left(1 + N^{\frac{d+1}{d+\gamma}} (\gamma-1)^{-1} C\right) + \mathcal{O}(N^{\frac{d+1}{d+\gamma}})\\ &\qquad\qquad = C_{\text{eff-PD}} \cdot dN^{\frac{d+1}{d+\gamma}} \log N,\end{aligned}$$

where constant $C_{\text{eff-PD}}$ depends on $C, \lambda, \gamma, \mu(\Xi)$. In this case, we need to verify $N_\star \leq n_\tau$, which implies $\gamma \geq 2 + 1/d$. □

E More Details on Experiment Implementation

We follow the implementation given by Liu et al. (2020a) for all experiments on this scenario. However, we fix the number of observable agents and landmarks, k, to 2, resulting in an observation space of 14. The environment then uses a discrete action space, corresponding to [NO-OP, LEFT, RIGHT, UP, DOWN], with movement in each direction being applied as a noisy force on the agent.

Note that we discretize the state space before episodes are used to train SAFARI by truncating continuous measurements to three significant figures. This allows us to sample fewer trajectories and shorten the length of the dataset, which in turn reduces the training time of the algorithm. As evidenced by Figure 2, performance of SAFARI is still comparable to the "expert" COMA even with a lower granularity in the representation of the state. We use the normal Gaussian kernel for the implementation here.

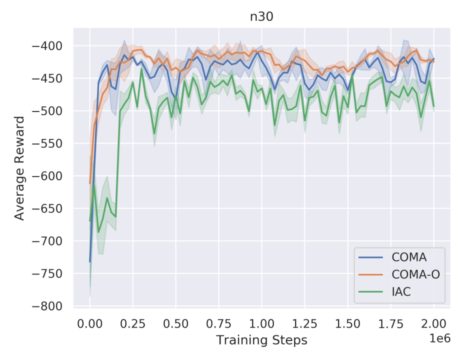


Figure 3: Training reward on the 30 agent environment.

We provide in Figure 3 the performance of COMA, COMA-O, and IAC on 20 evaluation episodes during training in the 30 agent environment. The same trend follows as with the 15 agent environment: COMA-O performs slight worse than COMA but still clearly better than IAC. The final convergence reward is also reflected in Figure 2. For the 100 agent case, we use the policies trained for 15 agents (a single network due to parameter sharing), as the exponential growth of the state-action space has greatly reduced the sample efficiency of COMA and significantly increased training time.

Hyperparameters All hyperparameters are tuned by logarithmic random search over the ranges given below:

Table 1: Algorithm Hyperparameters

Parameters	Value
COMA, VDN	
Discount γ	0.95
Exploration Rate	$(0.1, 0.3)$
Exploration Anneal	0.998
Policy learning rate	$(0.0001, 0.01)$
Critic learning rate	$(0.001, 0.01)$
Optimizer	Adam
SAFARI	
β	$(0.1, 1)$
λ	$(0.1, 0.99)$