Breaking the Curse of Many Agents: Provable
Mean Embedding Q-Iteration for Mean-Field Reinforcement Learning

Lingxiao Wang[1] Zhuoran Yang[2] Zhaoran Wang[1]

Abstract

Multi-agent reinforcement learning (MARL) achieves significant empirical successes. However, MARL suffers from the curse of many agents. In this paper, we exploit the symmetry of agents in MARL. In the most generic form, we study a mean-field MARL problem. Such a mean-field MARL is defined on mean-field states, which are distributions that are supported on continuous space. Based on the mean embedding of the distributions, we propose MF-FQI algorithm, which solves the mean-field MARL and establishes a non-asymptotic analysis for MF-FQI algorithm. We highlight that MF-FQI algorithm enjoys a "blessing of many agents" property in the sense that a larger number of observed agents improves the performance of MF-FQI algorithm.[1]

1. Introduction

Reinforcement learning (RL) (Sutton & Barto, 2018) searches for the optimal policy for sequential decision making through interacting with environments and learning from experiences. Multi-agent reinforcement learning (MARL) (Bu et al., 2008) generalizes RL to multi-agent systems. For competitive tasks such as zero-sum game and general-sum game, various MARL algorithms (Hu & Wellman, 2003; Littman, 1994; Wang & Sandholm, 2003) are proposed in search for the Nash equilibrium (Nash, 1951). Meanwhile, for cooperative tasks, MARL searches for the optimal policy that maximizes the social welfare (Ng, 1975), i.e., the expected total reward obtained by all agents (Claus & Boutilier,

---

[1]Department of Industrial Engineering and Management Sciences, Northwestern University, USA [2]Department of Operations Research and Financial Engineering, Princeton University, USA. Correspondence to: Lingxiao Wang <lingxiaowang2022@u.northwestern.edu>.

[1]See https://arxiv.org/abs/2006.11917 for the full version.

1998; Džeroski et al., 2001; Guestrin et al., 2002; Kar et al., 2013; Lauer & Riedmiller, 2000; Panait & Luke, 2005; Tan, 1993; Wang & Sandholm, 2003; Zambaldi et al., 2018). Combined with the breakthrough in deep learning, MARL achieves significant empirical successes in both settings, e.g., autonomous driving (Shalev-Shwartz et al., 2016), Go (Silver et al., 2016; 2017), esports (OpenAI, 2018; Vinyals et al., 2019), and robotics (Yang & Gu, 2004).

Despite its empirical successes, MARL remains challenging in the "many-agent" setting, as the capacity of state-action space grows exponentially in the number of agents, which hinders the learning of value function and policy due to the curse of dimensionality. Such a challenge is named as the "curse of many agents". One way to break such a curse is through mean-field approximation, which exploits the symmetry of homogeneous agents and summarizes them as a population. In the most general form, such a population is represented by a distribution over the state space of individual agents, while the reward and transition are parametrized as functionals of distributions (Acciaio et al., 2018). Although mean-field MARL demonstrates remarkable efficiency in applications such as large-scale fleet management (Lin et al., 2018) and ridesharing order dispaching (Li et al., 2019), its theoretical analysis remains scarce. In particular, despite significant progress (Guo et al., 2019; Jiachen et al., 2017; Jiang & Lu, 2018; Yang et al., 2018), we still lack a principled model-free algorithm that allows individual agents to have continuous states, which requires approximating nonlinear functionals of infinite-dimensional mean-field states, e.g., value function and policy.

In this paper, we study mean-field MARL in the collaborative setting, where the mean-field states are distributions over a continuous space $\mathcal{S}$. Here $\mathcal{S}$ denotes the state space of individual agents. In particular, we consider the setting with a centralized controller, which has a finite action space $\mathcal{A}$. Such a setting is extensively studied in the analysis of societal-scale systems (Gomes et al., 2015; Guéant et al., 2011; Moll et al., 2019). As a simplified example, the central bank or the central government decides whether to raise the interest rate or reduce the fiscal budget, respectively, both with the goal of maximizing social welfare. In such an example, the action space only contains two actions. How-

ever, the action taken by the centralized controller affects the dynamics of billions of individuals. Such a setting can be viewed as centralized mean-field control (Carmona et al., 2013; Fornasier & Solombrino, 2014; Huang et al., 2012) with an infinite number of homogeneous agents, which faces two challenges: (i) learning the value function and policy is intractable as they are functionals of distributions, which are infinite dimensional as $\mathcal{S}$ is continuous, and (ii) the mean-field state is only accessible through the observation of a finite number of agents, which only provides partial information. To tackle these challenges, we resort to the mean embedding of mean-field states into a reproducing kernel Hilbert space (RKHS) (Gretton et al., 2007; Smola et al., 2007; Sriperumbudur et al., 2010), which allows us to parametrize value functions as nonlinear functionals over the RKHS. Based on value function approximation, we propose the mean-field fitted Q-iteration algorithm (MF-FQI), which provably attains the optimal value function at a linear rate of convergence. In particular, we show that MF-FQI breaks the curse of many agents in the sense that its computational complexity only scales linearly in the number of observed agents, and moreover, the statistical accuracy enjoys a "blessing of many agents", that is, a larger number of observed agents improves the statistical accuracy. Moreover, we characterize the phase transition in the statistical accuracy in terms of the batch size in fitted Q-iteration and the number of observed agents.

Our Contribution. Our contribution is three-fold: (i) We propose the first model-free mean-field MARL algorithm, namely MF-FQI, that allows for continuous support with provable guarantees. (ii) We prove that MF-FQI breaks the curse of many agents by establishing its nonasymptotic computational and statistical rates of convergence. (iii) We motivate a principled framework for exploiting the invariance in MARL, e.g., exchangeability, via mean embedding.

Related Works. Our work is related to mean-field games and mean-field control. The study of mean-field games focuses on the search of the Nash equilibrium (Carmona & Delarue, 2018; Guéant et al., 2011; Huang et al., 2003; 2007; Lasry & Lions, 2006a;b; 2007), whereas the goal of mean-field control is to optimally control a McKean-Vlasov process (Acciaio et al., 2018; Andersson & Djehiche, 2011; Bensoussan et al., 2013; Buckdahn et al., 2009; 2011; Carmona et al., 2015; Meyer-Brandis et al., 2012). Most of these works focus on the continuous-time setting and require the knowledge of the transition model. In contrast, we consider the model-free and discrete-time setting.

Our work falls in the study of mean-field MARL, which generalizes finite-agent MARL by incorporating the notion of mean fields. Previous works investigate MARL in both cooperative and competitive settings (Guo et al., 2019; Jiachen et al., 2017; Jiang & Lu, 2018; Lin et al., 2019; Suttle et al.,

2019; Yang et al., 2018; Zhang et al., 2018a;b; 2019a). See (Zhang et al., 2019b) for the survey. In (Jiachen et al., 2017), a similar setting of mean-field MARL is studied, where the mean-field states are supported on a discrete space and the transition is linear in the state and action. In contrast, our work is model-free and allows for continuous support. In (Guo et al., 2019; Yang et al., 2018), mean-field MARL algorithms are proposed in the competitive setting, which have provable guarantees when the support is discrete. In comparison, we consider the cooperative setting with continuous support and establish nonasymptotic guarantees.

Our work exploits the exchangeability of agents via mean embedding. See e.g., (Fukumizu et al., 2008; Gretton et al., 2009; 2012; Smola et al., 2007; Sriperumbudur et al., 2010; Tolstikhin et al., 2017) and references therein for the study of mean embedding. Our work is closely related to various statistical models that exploit invariance, such as set kernels (Gärtner et al., 2002; Haussler, 1999; Kondor & Jebara, 2003) and deep sets (Zaheer et al., 2017). We refer to (Bloem-Reddy & Teh, 2019) for a detailed survey on learning with invariance.

Notations. For a topological space X, we denote by $\mathcal{C}_B(X)$ the set of bounded and continuous real functions on X. We denote by $\mathcal{M}(X)$ the space of all the probability distributions supported on X. For $x \in X$, we denote by $\delta_x \in \mathcal{M}(X)$ the point mass at x. For a real-valued function f defined on X, we denote by $\|f\|_{p,\nu}$ the $L_p(\nu)$ norm for $p \geq 1$, where $\nu \in \mathcal{M}(X)$. We write $\|f\|_\nu = \|f\|_{2,\nu}$ for notational simplicity.

2. Mean-Field MARL via Mean Embedding

In this section, we first motivate mean-field MARL by an example of N-player control with invariance. We then introduce the problem setup of mean-field MARL and the mean embedding of distributions. Finally, we propose the MF-FQI algorithm, which solves mean-field MARL based on the mean embedding.

2.1. Exchangeability in MARL

We first consider an N-player control problem with a centralized controller in discrete time. Such a setting is extensively studied in the analysis of societal-scale systems (Gomes et al., 2015; Guéant et al., 2011; Moll et al., 2019), such as the example of central bank or central government in §1. At each time step t, the central controller takes an action $a_t \in \mathcal{A}$ based on the current joint state $s_t = (s_{1,t}, \ldots, s_{N,t})$, where $s_{i,t} \in \mathcal{S}$ is the state of the i-th agent at time t. The immediate reward r_t follows a distribution that depends on the current state $s_t \in \mathcal{S}^N$ and action $a_t \in \mathcal{A}$. The transition of the joint state follows a distribution, which is determined by the current state $s_t \in \mathcal{S}^N$ and action $a_t \in \mathcal{A}$. In summary,

it holds that

$$r_t \sim r(s_t, a_t), \quad S_{t+1} \sim P(\cdot \mid s_t, a_t). \quad (2.1)$$

The process defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r)$ is a Markov decision process (MDP). We define a policy $\pi : \mathcal{S}^N \mapsto \mathcal{M}(\mathcal{A})$ as a mapping that maps a joint state $s \in \mathcal{S}^N$ to a probability distribution $\pi(\cdot \mid s)$ over $\mathcal{A}$. We define the value function corresponding to the policy π as

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \cdot r(S_t, A_t) \,\Big|\, S_0 = s\right], \quad (2.2)$$

where $A_t \sim \pi(\cdot|S_t)$, and $S_{t+1} \sim \mathbb{P}(\cdot \mid S_t, A_t)$, and $\gamma \in (0, 1)$ is the discount factor. Similarly, we define the action-value function corresponding to the policy π as follows,

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \cdot r(S_t, A_t) \,\Big|\, S_0 = s, A_0 = a\right], \quad (2.3)$$

where $A_t \sim \pi(\cdot \mid S_t)$, and $S_{t+1} \sim \mathbb{P}(\cdot \mid S_t, A_t)$. We define the Bellman operator T^π as follows,

$$T^\pi Q(s, a) = \mathbb{E}\big[r(s, a) + \gamma \cdot Q(S', A')\big], \quad (2.4)$$

where $S' \sim P(\cdot \mid s, a)$ and $A' \sim \pi(\cdot \mid s)$. Our goal is to find the optimal policy that maximizes the expected total reward as follows,

$$Q^*(s, a) = \sup_\pi Q^\pi(s, a), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \quad (2.5)$$

We denote by π^* the optimal solution of (2.5). It can be shown that $Q^* = Q^{\pi^*}$, and the following Bellman optimality equation holds,

$$Q^*(s, a) = TQ^*(s, a) = \mathbb{E}\Big[r(s, a) + \gamma \cdot \max_{a \in \mathcal{A}} Q^*(S', a)\Big], \quad (2.6)$$

where $S' \sim P(\cdot \mid s, a)$. Here we call T the Bellman optimality operator.

Curse of Many Agents: The learning of the optimal action-value function Q^* under the N-player setting suffers from the curse of many agents. More specifically, as N increases, the capacity of the joint state space $\mathcal{S}^N$ grows exponentially in N and incurs intractability in the learning of the action-value function Q. To address such a curse, we exploit the exchangeability of the MDP in (2.1). More specifically, we assume that the MDP is exchangeable in the sense that

$$r(s_t, a_t) \overset{\mathrm{d}}{=} r\big(\sigma(s_t), a_t\big), \quad P(s_{t+1} \mid s_t, a_t) = P\big(\sigma(s_{t+1}) \mid \sigma(s_t), a_t\big), \quad (2.7)$$

which holds for any $s_t, s_{t+1} \in \mathcal{S}^N$, $a_t \in \mathcal{A}$, and $\sigma \in \mathbb{S}_N$. Here σ is a block-wise permutation of the vector $s_t \in \mathcal{S}^N$, and $\mathbb{S}_N$ is the permutation group of order N. Under the exchangeability defined in (2.7), the following proposition shows that the optimal policy is invariant to permutations of the joint state.

Proposition 2.1 (Invariance of Q^*). If (2.7) holds, then it holds for any $\sigma \in \mathbb{S}_N$, $s \in \mathcal{S}^N$, and $a \in \mathcal{A}$ that

$$Q^*(s, a) = Q^*\big(\sigma(s), a\big). \quad (2.8)$$

Moreover, it holds for any $\sigma \in \mathbb{S}_n$, $s \in \mathcal{S}^N$, and $a \in \mathcal{A}$ that $\pi^*(a \mid s) \overset{\mathrm{d}}{=} \pi^*(a \mid \sigma(s))$.

Proof. See §B.1 for a detailed proof. □

Meanwhile, the following proposition proves that the action-value function Q^π is invariant to permutations of the joint states.

Proposition 2.2 (Invariant Representation). Let π be invariant to permutations such that $\pi(\cdot \mid s) \overset{\mathrm{d}}{=} \pi(\cdot \mid \sigma(s))$. If (2.7) holds, then it holds for some $g : \mathcal{M}(\mathcal{S}) \times \mathcal{A} \mapsto \mathbb{R}$ that

$$Q^\pi(s, a) = g(M_s, a). \quad (2.9)$$

Here $M_s(\cdot)$ is the empirical measure supported on the set $\{s_i\}_{i \in [N]}$ corresponding to $s = (s_1, \ldots, s_N)$, which takes the form of

$$M_s = \frac{1}{N} \sum_{i=1}^N \delta_{s_i}, \quad (2.10)$$

where recall that δ_{s_i} is the point mass at $s_i \in \mathcal{S}$ for all $i \in [N]$.

Proof. See §B.2 for a detailed proof. □

By Propositions 2.1 and 2.2, the optimal action-value function Q^* is related to the joint state through the empirical state distribution M_s defined in (2.10). When the number of agents N goes to infinity, the empirical state distribution M_s converges to a limiting continuous distribution. To capture such a limiting dynamics of infinite agents with exchangeability, we define an MDP with $\mathcal{M}(\mathcal{S})$, the space of probability measures supported on $\mathcal{S}$, as the mean-field state space as follows.

Mean-Field MARL. We define the discounted mean-field MDP by the tuple $(\mathcal{M}(\mathcal{S}), \mathcal{A}, P, r, \gamma)$. Here $\mathcal{A}$ is the action space and $\mathcal{M}(\mathcal{S})$ is the mean-field state space, which is the space of all the distributions supported on $\mathcal{S}$. Given a mean-field state $p_s \in \mathcal{M}(\mathcal{S})$ and an action $a \in \mathcal{A}$, the immediate reward follows the distribution $r(a, p_s)$, where $r : \mathcal{A} \times \mathcal{M}(\mathcal{S}) \mapsto \mathcal{M}(\mathbb{R})$. The Markov kernel $P(\cdot \mid \cdot)$ maps the

action-state pair (a, p_s) to a distribution on $\mathcal{M}(\mathcal{S})$, which is the distribution of the mean-field state after transition from the action-state pair (a, p_s). For a policy $\pi(\cdot \,|\, p_s)$ that maps from $\mathcal{M}(\mathcal{S})$ to $\mathcal{M}(\mathcal{A})$, we define the action-value function as

$$Q^\pi(a, p_s) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \cdot r(P_{s,t}, A_t) \,\bigg|\, P_{s,0} = p_s, A_0 = a\right], \tag{2.11}$$

where $A_t \sim \pi(\cdot \,|\, P_{s,t})$ and $P_{s,t+1} \sim P(\cdot \,|\, a, p_{s,t})$. Correspondingly, we define the Bellman evaluation operator T^π as follows,

$$T^\pi Q(a, p_s) = \mathbb{E}\big[r(a, p_s) + \gamma \cdot Q(A', P_{s'})\big], \tag{2.12}$$

where $P_{s'} \sim P(\cdot \,|\, a, p_s)$ and $A' \sim \pi(\cdot \,|\, a, p_s)$. We define the optimal action-value function as $Q^* = \sup_\pi Q^\pi$. The Bellman optimality equation then takes the form of

$$Q^*(a, p_s) = TQ^*(a, p_s) = \mathbb{E}\Big[r(a, p_s) + \gamma \cdot \max_{a \in \mathcal{A}} Q^*(a, P_s')\Big], \tag{2.13}$$

where $P_{s'} \sim P(\cdot \,|\, a, p_s)$. We write $\Omega = \mathcal{S} \times \mathcal{A}$ and define the space of state-action configurations $\widetilde{\mathcal{M}}(\Omega)$ as follows,

$$\widetilde{\mathcal{M}}(\Omega) = \big\{\omega_{a,p_s} = \delta_a \times p_s : a \in \mathcal{A}, p_s \in \mathcal{M}(\mathcal{S})\big\}. \tag{2.14}$$

Here we denote by $\delta_a \in \mathcal{M}(\mathcal{A})$ the point mass at action $a \in \mathcal{A}$ and by $\delta_a \times p_s$ the product measure on $\Omega = \mathcal{S} \times \mathcal{A}$ induced by δ_a and p_s. See §A for the definition of a topological structure that allows us to define distributions on $\widetilde{\mathcal{M}}(\Omega)$. Note that the transition kernel $P(\cdot | a, p_s)$ equivalently defines a Markov kernel from $\widetilde{\mathcal{M}}(\Omega)$ to $\mathcal{M}(\mathcal{S})$. With a slight abuse of notations, we denote by $P(\cdot | \omega_{a,p_s})$ such a Markov kernel and do not distinguish between them. Similarly, we denote by $r(\omega)$ and $Q(\omega)$ the immediate reward and action-value function defined on the state-action configuration $\omega \in \widetilde{\mathcal{M}}(\Omega)$, respectively. We assume that the action set $\mathcal{A}$ is finite, and the immediate reward is upper bounded by a positive absolute constant $R_{\max}$. It then holds that the action-value functions are upper bounded by $Q_{\max} = R_{\max}/(1 - \gamma)$.

In a multi-agent environment with infinite homogeneous agents and continuous state space $\mathcal{S}$, we cannot access the mean-field state p_s directly. Instead, we assume that we observe the states of N agents that follows the mean-field state p_s. In what follows, we construct an algorithm that solves the mean-field MARL via such a finite observation for each mean-field state. In the sequel, we denote by $\widehat{Q}_\kappa^\lambda$ and π_κ the outputs of MF-FQI.

We highlight that the mean-field MARL setting faces two challenges: (i) learning the value function and policy is intractable as they are functionals of distributions, which are infinite dimensional as $\mathcal{S}$ is continuous, and (ii) the mean-field state is only accessible through the observation of a finite number of agents, which only provides partial information. In what follows, we tackle these challenges via mean embedding.

2.2. Mean Embedding

To learn the optimal action-value function Q^* defined on $\widetilde{\mathcal{M}}(\Omega)$, which is a space of distributions, we introduce mean embedding, which embeds the space of distributions to a reproducing kernel Hilbert space (RKHS). We denote by $\mathcal{H}(k)$ the RKHS with reproducing kernel $k : \Omega \times \Omega \mapsto \mathbb{R}$. For any state-action configuration $\omega \in \widetilde{\mathcal{M}}(\Omega)$, the mean embedding $\mu_\omega(\cdot)$ of ω into the RKHS $\mathcal{H}(k)$ is defined as follows (Gretton et al., 2007; Smola et al., 2007; Sriperumbudur et al., 2010),

$$\mu_\omega(x) = \int_\Omega k(x, t)\, \mathrm{d}\omega(t) \in \mathcal{H}(k). \tag{2.15}$$

Let $\mathcal{X} = \{\mu_\omega : \omega \in \widetilde{\mathcal{M}}(\Omega)\} \subseteq \mathcal{H}(k)$. To tackle challenge (i), we introduce another reproducing kernel $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. Such a kernel generates an RKHS $\mathcal{H}(K)$ that include functions defined on $\mathcal{X}$. Our idea is then to approximate Q^* using functions in $\mathcal{H}(K)$. Note that upon a proper selection of kernel $K(\cdot, \cdot)$, the corresponding RKHS $\mathcal{H}(K)$ captures a rich family of functions defined on $\mathcal{X}$. As an example, for universal kernels such as the radial basis function (RBF) kernel, the corresponding RKHS is dense in $\mathcal{C}(\mathcal{X})$. To further regulate the behavior of mean embedding and RKHS, we introduce the following regularity conditions on kernels $k(\cdot, \cdot)$ and $K(\cdot, \cdot)$.

Assumption 2.3 (Regularity Condition of Kernels). We assume that the kernel $k(\cdot, \cdot)$ and $K(\cdot, \cdot)$ are continuous and bounded as follows,

$$k(u, u) \leq \varrho, \quad \forall u \in \Omega, \qquad K(\mu_\omega, \mu_\omega) \leq \varsigma, \quad \forall \omega \in \widetilde{\mathcal{M}}(\Omega), \tag{2.16}$$

where ϱ and ς are positive absolute constants. We assume that $k(\cdot, \cdot)$ is universal and the mean embeddings $\mu_\omega(\cdot)$ are continuous for any $\omega \in \widetilde{\mathcal{M}}(\Omega)$. We further assume that $K(\cdot, \mu_\omega)$ is Hölder continuous such that for any $x, y \in \widetilde{\mathcal{M}}(\Omega)$, it holds that

$$\|K(\cdot, \mu_x) - K(\cdot, \mu_y)\|_{\mathcal{H}(K)} \leq L \cdot \|\mu_x - \mu_y\|_{\mathcal{H}(k)}^h, \tag{2.17}$$

where L and h are positive absolute constants.

The assumption on the boundedness of the kernels in (2.16) is a standard assumption in the learning with kernel embedding (Caponnetto & De Vito, 2007; Lin et al., 2017; Muandet et al., 2012; Szabó et al., 2015). The universality assumption on $k(\cdot, \cdot)$ ensures that each mean embedding uniquely

characterizes a distribution (Gretton et al., 2007; 2012). The continuous assumption on the embeddings μ_ω is a mild regularity condition, which holds if the kernel $k(\cdot, \cdot)$ is universal and the domain Ω is compact (Sriperumbudur et al., 2010). Meanwhile, the Hölder continuity of $K(\cdot, \cdot)$ in Assumption 2.3 is a mild regularity condition. Such an assumption holds for a rich family of commonly used reproducing kernels, such as the linear kernel $K(\mu_x, \mu_y) = \langle \mu_x, \mu_y \rangle_{\mathcal{H}(k)}$ and the RBF kernel $K(\mu_x, \mu_y) = \exp(-\|\mu_x - \mu_y\|^2_{\mathcal{H}(k)}/\sigma^2)$.

We highlight that the Hölder continuity of $K(\cdot, \cdot)$ allows for an approximation of mean embedding μ_p based on the empirical approximation $\widehat{p}$ of the distribution $p \in \widetilde{\mathcal{M}}(\Omega)$ with finite observations, which further allows for an approximation of the action-value function with finite observations, thus tackles the challenge (ii). For an empirical approximation of state-action configuration $\omega_{a,\widehat{p}_s} = \delta_a \times \widehat{p}_s$, where $\widehat{p}_s$ is the empirical distribution of the observed states $\{s_i\}_{i \in N}$, the mean embedding takes the following form,

$$\mu_{\omega_{a,\widehat{p}_s}}(\cdot) = \frac{1}{N} \sum_{i=1}^{N} k(\cdot, (a, s_i)), \qquad (2.18)$$

which is invariant to the permutation of states $\{s_i\}_{i \in [N]}$. Such an invariance is also exploited by the neural-network based approach named deep sets (Zaheer et al., 2017). In what follows, we connect our mean embedding approach to the invariant deep reinforcement learning under the framework of overparametrized two-layer neural networks (Arora et al., 2019; Jacot et al., 2018; Neyshabur et al., 2018; Zhang et al., 2016).

Connection to Invariant Deep Reinforcement Learning. In what follows, we assume that $(a, s_i) \in \mathbb{R}^d$ and write $x_i = (a, s_i)$. We define the feature mappings $\{\phi_j(\cdot)\}_{j \in [m]}$ and $\{\Phi_\ell(\cdot)\}_{\ell \in [M]}$ as follows,

$$\phi_j(x) = \frac{1}{\sqrt{m}} \cdot b_j \cdot \mathbb{1}\{w_j^\top x > 0\} \cdot x,$$

$$\Phi_r(q) = \frac{1}{\sqrt{\ell}} \cdot b_r' \cdot \mathbb{1}\{W_r'^\top q > 0\} \cdot q, \qquad (2.19)$$

where $b_j, b_r' \sim \text{Unif}\{-1, 1\}$, $w_j \sim N(0, I_{d\ell}/(d\ell))$, and $W_r' \sim N(0, I_\ell/\ell)$. Correspondingly, we define the kernels $k_m(x, y) = \sum_{j=1}^{m} \phi_j(x)^\top \phi_j(y)$ and $K_\ell(p, q) = \sum_{r=1}^{\ell} \Phi_r(p)^\top \Phi_r(q)$.

Note that the mean embedding of a point mass δ_{x_i} by the kernel k_m is an array $\mu_i = [\phi_1(x_i)^\top, \ldots, \phi_m(x_i)^\top]^\top \in \mathbb{R}^{md}$. For a mean embedding $\mu \in \mathbb{R}^{md}$, we consider the parametrization of action-value functions $Q(\mu) = \widetilde{Q}(D^\top \mu)$, where $\widetilde{Q} \in \mathcal{H}(K_\ell)$ and $D = [D_1, \ldots, D_m] \in \mathbb{R}^{md \times \ell}$ with $D_j \in \mathbb{R}^{d \times \ell}$ ($j \in [m]$). Let μ_p be the mean embedding of the empirical measure supported on $\{x_i\}_{i \in [N]}$.

It then holds for some $\{\alpha_r\}_{r \in [\ell]} \subseteq \mathbb{R}^d$ that

$$Q(\mu_p) = \sum_{r=1}^{\ell} \alpha_r^\top \Phi_r(D^\top \mu_p)$$

$$= \frac{1}{\sqrt{\ell}} \sum_{r=1}^{\ell} b_r' \cdot \mathbb{1}\{W_r'^\top \rho > 0\} \cdot \alpha_r^\top (D^\top \mu_p)$$

$$= f(D^\top \mu_p). \qquad (2.20)$$

Note that if α_r is sufficiently close to W_r', then $Q(\mu_p)$ is close to $\widetilde{f}(D^\top \mu_p)$, where

$$\widetilde{f}(\rho) = \frac{1}{\sqrt{\ell}} \sum_{r=1}^{\ell} b_r' \cdot \mathbb{1}\{\alpha_r^\top \rho > 0\} \cdot \alpha_r^\top(\rho), \quad \forall \rho \in \mathbb{R}^\ell. \qquad (2.21)$$

Here $\widetilde{f}(\cdot)$ is a two-layer neural network with parameters b_r', α_r ($r \in [\ell]$) and ReLU activation function. Meanwhile, it holds that

$$D^\top \mu_p = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sqrt{m}} \sum_{j=1}^{m} b_j \cdot \mathbb{1}\{w_j^\top x_i > 0\} \cdot D_j^\top x_i$$

$$= \frac{1}{N} \sum_{i=1}^{N} \psi(x_i). \qquad (2.22)$$

Similarly, if D_j is close to w_j, then $\psi(x_i)$ is close to $\widetilde{\psi}(x_i)$ defined as follows,

$$\widetilde{\psi}(x_i) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} b_j \cdot \mathbb{1}\{D_j^\top x_i > 0\} \cdot D_j^\top x_i, \quad \forall x_i \in \mathbb{R}^d. \qquad (2.23)$$

Here $\widetilde{\phi}(x_i)$ is a two-layer neural network with input x_i, parameters b_j' ($j \in [m]$) and D, and ReLU activation function. Note that for the functions f and ψ defined in (2.20) and (2.22) to approximately take the form of neural networks, we requires the parameters α and D to be sufficiently close to W and w, respectively. Such a requirement is formally characterized by the study of overparametrized neural networks. More specifically, if the widths m and ℓ of neural networks are sufficiently large (which depends on the deviation of parameters α and D from their respective initializations W and w), then the functions f and ψ well approximates the neural networks $\widetilde{f}$ and $\widetilde{\psi}$ defined in (2.21) and (2.23), respectively.

In conclusion, under the mean embedding with the feature mappings defined in (2.19), the parameterization of action-value function takes the form of $Q(\mu_p) = f(1/N \cdot \sum_{i=1}^{N} \psi(x_i))$, where f and ψ are approximations of two-layer neural networks $\widetilde{f}$ and $\widetilde{\psi}$, respectively. Hence, the action-value function $Q(\mu_p)$ approximately takes the form of deep sets (Zaheer et al., 2017) with $\{x_i\}_{i \in [N]}$ as the set input.

2.3. Mean-Field Fitted Q-Iteration

In what follows, we establish a value-based algorithm that solves mean-field MARL problem in §2.1 based on fitted Q-iteration (Ernst et al., 2005). More specifically, we propose an algorithm that learns the optimal action-value function Q^* by the sample $\{\delta_{a_i} \times p_{i,s}\}_{i \in [n]}$ that follows a sampling distribution ν over the space of state-action configurations $\widetilde{\mathcal{M}}(\Omega)$. For each state-action configuration $\delta_{a_i} \times p_{i,s}$, the mean-field state $p_{i,s} \in \mathcal{M}(\mathcal{S})$ is available to us through the observed states $\{s_{i,j}\}_{j \in [N]}$, which are sampled independently from the mean-field state $p_{i,s}$. We further observe the immediate reward r_i and the states $\{s'_{i,j}\}_{j \in [N]}$ that are independently sampled from the mean-field state $p_{i,s'}$. Here $p_{i,s'} \sim P(\cdot \,|\, a_i, p_{i,s})$ is the mean-field state after transition from the state-action configuration $(a_i, p_{i,s})$. Given the batch of data $\{(\{s_{i,j}\}_{j \in [N]}, a_i, r_i, \{s'_{i,j}\}_{j \in [N]})\}_{i \in [n]}$, the mean-field fitted Q-iteration (MF-FQI) sequentially computes

$$\widehat{y}_{i,k} = r_i + \max_{a \in \mathcal{A}} \widehat{Q}_k^\lambda(\mu_{\omega_{a,\widehat{p}_{i,s'}}}) \tag{2.24}$$

at the k-th iteration. Here $\mu_{\omega_{a,\widehat{p}_{i,s'}}}$ is the mean embedding of the distribution $\omega_{a,\widehat{p}_{i,s'}} = \delta_a \times \widehat{p}_{i,s'}$, $\widehat{p}_{i,s'}$ is the empirical distribution supported on the set $\{s'_{i,j}\}_{j \in [N]}$, and Q_k^λ is the approximation of the optimal action-value function at the k-th iteration of MF-FQI. Upon computing $\{\widehat{y}_{i,k}\}_{i \in [n]}$ according to (2.24), MF-FQI then updates the approximation of the optimal action-value function in the RKHS $\mathcal{H}(K)$ by solving the following optimization problem,

$$Q_{k+1}^\lambda = \operatorname*{argmin}_{f \in \mathcal{H}(K)} \frac{1}{n} \sum_{i=1}^n \big(f(\mu_{\widehat{\omega}_i}) - \widehat{y}_{i,k}\big)^2 + \lambda \cdot \|f\|_{\mathcal{H}(K)}^2,$$

$$\widehat{Q}_{k+1}^\lambda = \min\{Q_{k+1}^\lambda, Q_{\max}\}, \tag{2.25}$$

where $\widehat{\omega}_i = \delta_a \times \widehat{p}_{i,s}$ and $\widehat{p}_{i,s}$ is the empirical distribution supported on the set $\{s_{i,j}\}_{j \in [N]}$. We summarize MF-FQI defined by (2.24) and (2.25) in Algorithm 1. We highlight that MF-FQI has a linear computational complexity in terms of the number of observed agents N. Therefore, MF-FQI is computationally tractable even for a large number of observed agents N.

3. Main Results

In this section, we establish the theoretical guarantee of MF-FQI defined in Algorithm 1. In the sequel, we denote by $\widehat{Q}_\kappa^\lambda$ and π_κ the outputs of MF-FQI. Our goal is to establish an upper bound for $\|Q^* - Q^{\pi_\kappa}\|_{1,\mu}$, where μ is the measurement distribution over $\widetilde{\mathcal{M}}(\Omega)$.

We first introduce the definition of concentration coefficients. For a policy π_1, we define $E^{\pi_1}\nu$ as the distribution of $\Lambda_1 = \delta_{A_1} \times P_{1,s}$, where $P_{1,s} \sim P(\cdot \,|\, \nu)$ and

---

Algorithm 1 Mean-Field Fitted Q-Iteration (MF-FQI)

1: **Input:** Batch of data $\{(\{s_{i,j}\}_{j \in [N]}, a_i, r_i, \{s'_{i,j}\}_{j \in [N]})\}_{i \in [n]}$, reproducing kernels $k(\cdot, \cdot)$ and $K(\cdot, \cdot)$, number of iterations κ, parameter λ, initial action-value function $\widehat{Q}_0^\lambda$.

2: For all $i \in [n]$, compute mean embeddings $\mu_{\widehat{\omega}_i}$ and $\mu_{a,\widehat{p}_{i,s'}}$ for all $a \in \mathcal{A}$ as follows,

$$\mu_{\widehat{\omega}_i}(\cdot) = \frac{1}{N} \sum_{j=1}^N k\big(\cdot, (a_i, s_{i,j})\big),$$

$$\mu_{a,\widehat{p}_{i,s'}}(\cdot) = \frac{1}{N} \sum_{j=1}^N k\big(\cdot, (a, s'_{i,j})\big), \quad \forall a \in \mathcal{A}.$$

3: **for** $k = 0, 1, \ldots, \kappa - 1$ **do**

4:     Compute $\widehat{y}_{i,k} = r_i + \max_{a \in \mathcal{A}} \widehat{Q}_k^\lambda(\mu_{a,\widehat{p}_{i,s'}})$ for all $i \in [n]$.

5:     Update the action-value function $\widehat{Q}_{k+1}^\lambda$ by

$$Q_{k+1}^\lambda = \operatorname*{argmin}_{f \in \mathcal{H}(K)} \frac{1}{n} \sum_{i=1}^n \big(f(\mu_{\widehat{\omega}_i}) - \widehat{y}_i\big)^2 + \lambda \cdot \|f\|_{\mathcal{H}(K)}^2,$$

$$\widehat{Q}_{k+1}^\lambda = \min\{Q_{k+1}^\lambda, Q_{\max}\}.$$

6: **end for**

7: **Output:** $\{\widehat{Q}_i^\lambda\}_{i \in [\kappa]}$.

$A_1 \sim \pi_1(\cdot \,|\, P_{1,s})$. Similarly, for policies $\{\pi_i\}_{i \in [\ell]}$, we recursively define $E^{\pi_\ell} \circ E^{\pi_{\ell-1}} \circ \ldots \circ E^{\pi_1}\nu$ as the distribution of $\Lambda_\ell = \delta_{A_\ell} \times P_{\ell,s}$, where $P_{\ell,s} \sim P(\cdot \,|\, E^{\pi_{\ell-1}} \circ \ldots \circ E^{\pi_1}\nu)$ and $A_\ell \sim \pi_\ell(\cdot \,|\, P_{\ell,s})$. In what follows, we define the concentration coefficients that measures the difference between the sampling distribution ν and the measurement distribution μ on $\widetilde{\mathcal{M}}(\Omega)$.

Assumption 3.1. (Concentration Coefficients) Let ν be the sampling distribution on $\widetilde{\mathcal{M}}(\Omega)$. Let μ be the measurement distribution on $\widetilde{\mathcal{M}}(\Omega)$. We assume that for any policies $\{\pi_i\}_{i \in [\ell]}$, the distribution $E^{\pi_\ell} \circ E^{\pi_{\ell-1}} \circ \ldots \circ E^{\pi_1}\nu$ is absolutely continuous with respect to μ. We define the ℓ-th concentration coefficients between ν and μ as follows,

$$\phi(\ell; \mu, \nu) \tag{3.1}$$

$$= \sup_{\pi_1, \ldots, \pi_\ell} \left(\mathbb{E}_\mu\left[\left(\frac{\mathrm{d}E^{\pi_\ell} \circ E^{\pi_{\ell-1}} \circ \ldots \circ E^{\pi_1}\nu}{\mathrm{d}\mu} \right)^2 \right] \right)^{1/2}.$$

We assume that $\phi(\ell; \mu, \nu) < +\infty$ for any $\ell \in \mathbb{N}$. We further assume that there exist a positive absolute constant $\Phi(\mu, \nu)$ such that

$$\sum_{\ell=1}^\infty \gamma^{\ell-1} \cdot \ell \cdot \phi(\ell; \mu, \nu) \le \Phi(\mu, \nu)/(1-\gamma)^2. \tag{3.2}$$

Assumption 3.2 is a standard assumption in the theoretical analysis of reinforcement learning (Antos et al., 2008; Chen & Jiang, 2019; Farahmand et al., 2010; 2016; Lazaric et al., 2016; Munos & Szepesvári, 2008; Scherrer, 2013; Scherrer et al., 2015; Szepesvári & Munos, 2005; Yang et al., 2019). Under Assumption 3.1, the following theorem upper bounds the error $\|Q^* - Q^{\pi_\kappa}\|_{1,\mu}$ of MF-FQI.

Theorem 3.2 (Error Propagation). Let $\{\widehat{Q}_i^\lambda\}_{i \in [\kappa]}$ be the output of Algorithm 1. Let π_κ be the greedy policy corresponding to $\widehat{Q}_\kappa^\lambda$. Under Assumption 3.1, it holds that

$$\|Q^* - Q^{\pi_\kappa}\|_{1,\mu} \leq \underbrace{\frac{2\gamma \cdot \Phi(\mu, \nu)}{(1-\gamma)^2} \cdot \max_{i \in [\kappa]} \|\widehat{Q}_i^\lambda - T\widehat{Q}_{i-1}^\lambda\|_\nu}_{(a)}$$
$$+ \underbrace{\frac{4\gamma^{\kappa+1} \cdot Q_{\max}}{1-\gamma}}_{(b)}. \quad (3.3)$$

Proof. See §B.3 for a detailed proof. $\square$

Following from Theorem 3.2, the error of MF-FQI is upper bounded by the sum of the two terms on the right-hand side of (3.3). Here term (b) characterizes the algorithmic error that hinges on the number of iterations κ. Meanwhile, term (a) characterizes the one-step approximation error that hinges on the approximation $\widehat{Q}_i^\lambda$ of $T\widehat{Q}_{i-1}^\lambda$. In the sequel, we upper bound the one-step approximation error characterized by term (a). To this end, we first impose the following regularity condition on the Bellman optimality operator T and the RKHS $\mathcal{H}(K)$.

Assumption 3.3 (Regularity Condition of T and $\mathcal{H}(K)$). We define the integral operator $\mathcal{C}$ as follows,

$$\mathcal{C}f(x) = \int_{\widetilde{\mathcal{M}}(\Omega)} K(x, \mu_\omega) f(\mu_\omega) d\nu(\omega). \quad (3.4)$$

We assume that the eigenvalues $\{t_n\}_{n \in \mathbb{N}}$ of $\mathcal{C}$ is bounded such that $\alpha \leq n^b t_n \leq \beta$ for all $n \in \mathbb{N}$, where α, β, and $b > 1$ are positive absolute constants. We further assume that for any output $Q^\lambda \in \mathcal{H}(K)$ of the regression problem defined in (2.25), it holds for some $g \in \mathcal{H}(K)$ that

$$Q_{\mathcal{H},T} = \mathcal{C}^{(c-1)/2} g, \quad \|g\|_{\mathcal{H}(K)} \leq R. \quad (3.5)$$

Here $R > 0$ and $c \in [1, 2]$ are absolute constants, and $Q_{\mathcal{H},T}$ is defined as follows,

$$\widehat{Q}^\lambda = \min\{Q^\lambda, Q_{\max}\},$$
$$Q_{\mathcal{H},T} = \Pi_{\mathcal{H}(K)}(T\widehat{Q}^\lambda) = \underset{f \in \mathcal{H}(K)}{\operatorname{argmin}} \|f - T\widehat{Q}^\lambda\|_\nu, \quad (3.6)$$

where we denote by $\Pi_{\mathcal{H}(K)}$ the projection onto $\mathcal{H}(K)$ with respect to the norm $\|\cdot\|_\nu$.

Assumption 3.3 is a mild regularity assumption on the RKHS $\mathcal{H}(K)$ and the Bellman optimality operator T. Similar assumptions arises in the analysis of kernel ridge regression (Caponnetto & De Vito, 2007; Lin et al., 2017; Szabó et al., 2015). The parameters b and c in Assumption 3.3 define a prior space $\mathcal{P}(b, c)$ in the context of kernel ridge regression (Caponnetto & De Vito, 2007). Intuitively, the parameter c controls the smoothness of $Q_{\mathcal{H},T}$ defined in (3.6), and the parameter b controls the size of $\mathcal{H}(K)$. Under Assumption 3.3, the following theorem characterizes the one-step approximation error of MF-FQI defined in Algorithm 1.

Theorem 3.4 (One-step Approximation Error). Let η, τ be two constants such that $0 < \eta + \tau < 1$. Let $C(\eta) = 32 \log^2(6/\eta)$. Under Assumptions 2.3 and 3.3, for

$$N \geq 2\varrho \cdot (1 + \sqrt{\log(|\mathcal{A}| \cdot n/2\tau)})^2 \cdot (64L^2\varsigma^2/\lambda^2)^{1/h},$$
$$n \geq \frac{2C(\eta)\varsigma\beta b}{(b-1)\lambda^{1+1/b}}, \quad \lambda \leq \|\mathcal{C}\|_{\mathcal{H}(K)}, \quad (3.7)$$

it holds with probability at least $1 - \eta - \tau$ that

$$\|\widehat{Q}_k^\lambda - T\widehat{Q}_{k-1}^\lambda\|_\nu^2 \leq \mathcal{G}_1 + \mathcal{G}_2 + \psi_T^2, \quad \forall k \in [\kappa], \quad (3.8)$$

where

$$\mathcal{G}_1 = \left(\frac{8L^2 Q_{\max}^2 (1 + \sqrt{\log(|\mathcal{A}| \cdot n/2\tau)})^{2h} \cdot (2\varrho)^h}{\lambda \cdot N^h} \right)$$
$$\cdot \left(1 + \frac{5\varsigma^2}{\lambda^2} \right),$$
$$\mathcal{G}_2 = C(\eta) \cdot \left(R\lambda^c + \frac{\varsigma^2 R}{\lambda^{2-c}n^2} + \frac{\varsigma R\lambda^{c-1}}{4n} \right.$$
$$\left. + \frac{\varsigma M^2}{\lambda n^2} + \frac{\Sigma^2\beta b}{(b-1)n\lambda^{1/b}} \right), \quad (3.9)$$

and the term ψ_T is defined as follows,

$$\psi_T = \sup_{k \in [\kappa]} \|T\widehat{Q}_k^\lambda - \Pi_{\mathcal{H}(K)}(TQ_k^\lambda)\|_\nu. \quad (3.10)$$

Here M and Σ are positive absolute constants, the parameters ς, ϱ, h are defined in Assumption 2.3, and the parameters $\mathcal{C}$, b, c are defined in Assumption 3.3, and $\Pi_{\mathcal{H}(K)}$ the projection onto $\mathcal{H}(K)$ with respect to the norm $\|\cdot\|_\nu$.

Proof. See §B.4 for a detailed proof. $\square$

Following from Theorem 3.4, the one-step approximation error is upper bounded by the sum of the three terms, $\mathcal{G}_1$, $\mathcal{G}_2$, and ψ_T, on the right-hand side of (3.8). Here the term $\mathcal{G}_1$ characterizes the error by approximating the mean-field state via N observed agents, the term $\mathcal{G}_2$ characterizes the error by estimating TQ with the batch of size n, and the term ψ_T characterizes the error in of approximating TQ by

functions from the RKHS $\mathcal{H}(K)$. If we further assume that $TQ \in \mathcal{H}(K)$ for all $Q \in \mathcal{H}(K)$ with $Q \leq Q_{\max}$, then term ψ_T vanishes and $Q^* \in \mathcal{H}(K)$ is the unique fixed point of T in $\mathcal{H}(K)$. Combining the error propagation in Theorem 3.2 and the one-step approximation error in Theorem 3.4, we obtain the following theorem that upper bounds the error of MF-FQI defined in Algorithm 1.

Theorem 3.5 (Theoretical Guarantee of MF-FQI). Let π_κ be the output of MF-FQI and $n = N^a$. Under the assumptions imposed by Theorem 3.2 and 3.4, for

$$\|\mathcal{C}\|_{\mathcal{H}(K)} \geq \max\big\{(a \cdot \log N/N)^{h/(c+3)}, 1/N^{ab/(bc+1)}\big\},$$

it holds with probability at least $1 - \eta - \tau$ that

$$\|Q^* - Q^{\pi_\kappa}\|_{1,\mu} \leq \frac{2\gamma \cdot \Phi(\mu, \nu)}{(1-\gamma)^2} \cdot \bigg(\underbrace{C \cdot \Xi}_{(i)} + \underbrace{\psi_T}_{(ii)}\bigg)$$
$$+ \underbrace{\frac{4\gamma^{\kappa+1} \cdot Q_{\max}}{1-\gamma}}_{(iii)}, \qquad (3.11)$$

where C is a positive absolute constant, ψ_T is defined in (3.10) of Theorem 3.4, and

$$\Xi = \max\bigg\{\bigg(\frac{\log(|\mathcal{A}| \cdot N/\tau)}{N}\bigg)^{\frac{hc}{2(c+3)}}, 1/N^{\frac{abc}{2(bc+1)}}\bigg\}.$$

Here the integral operator $\mathcal{C}$ is defined in (3.4) of Assumption 3.3, the parameter h is defined in Assumption 2.3, and parameters b, c are defined in Assumption 3.3.

Proof. See §B.5 for a detailed proof. $\square$

By Theorem 3.5, the approximation error of the action-value function attained by MF-FQI is characterized by the three terms on the right-hand side of (3.11). Here term (i) characterizes the statistical error, which is small for a sufficiently large number of observed agents N and batch size $n = N^a$. Term (ii) is the bias ψ_T defined in in (3.10) of Theorem 3.4, which vanishes if the Bellman optimality operator T is closed in the RKHS $\mathcal{H}(K)$. Term (iii) characterizes the algorithmic error, which is small for a sufficiently large number of iterations κ.

In the sequel, we assume that T is closed in $\mathcal{H}(K)$ and thus $\psi_T = 0$ for simplicity. Note that the algorithmic error characterized by term (ii) has a linear rate of convergence, which is negligible comparing with the statistical error characterized by term (i) if the iteration number is sufficiently large. More specifically, if it holds for some positive absolute constant C that

$$\kappa \geq C \cdot \max\bigg\{\frac{hc}{2(c+3) \cdot \log(1/\gamma)} \cdot \log\Big(\frac{N}{\log(|\mathcal{A}| \cdot N)}\Big),$$
$$\frac{abc}{2(bc+1) \cdot \log(1/\gamma)} \cdot \log N\bigg\}, \qquad (3.12)$$

then the dominating term on the right-hand side of (3.11) in Theorem 3.5 is term (i) that characterizes the statistical error.

Phase Transition. Note that MF-FQI involves two stage of sampling, where the first stage samples n mean-field state from the sampling distribution ν, and the second stage samples N states from each mean-field state. In what follows, we discuss the connection between the performance of MF-FQI and the sample complexity of the two-stage sampling involved. More specifically, we discuss the phase transition in the statistical error of MF-FQI when $a = \log n / \log N$ transits from zero to infinity. We categorize the phase transition into the following regimes in terms of a.

1. For $a > h \cdot (c + 1/b)/(c + 3)$, the rate of convergence of the statistical error takes the form of $\mathcal{O}((\log(|\mathcal{A}| \cdot N)/N)^{hc/(2c+6)})$. In this regime, increasing the number of observations N for each mean-field state improves the performance of MF-FQI, whereas increasing the batch size n of mean-field state cannot improve the performance of MF-FQI.

2. For $0 < a < h \cdot (c + 1/b)/(c + 3)$, the rate of convergence of the statistical error takes the form of $\mathcal{O}(1/n^{bc/(2bc+2)})$. In this regime, increasing batch size n of mean-field state improves the performance of MF-FQI, whereas increasing the number of observations N for each mean-field state cannot improve the performance of MF-FQI.

In conclusion, under regularity conditions, MF-FQI approximately achieves the optimal policy for a sufficiently large number of iteration κ, batch size n, and number of observations N. We highlight that MF-FQI enjoys a "blessing of many agents" property. More specifically, for a sufficiently large batch size n of the mean-field state, a larger number N of observed agents improves the learning of Q^*.

4. Limitation and Future Work

In this work, we propose MF-FQI, which tackles mean-field reinforcement learning problem with symmetric agents and a centralized controller. Such a setting is extensively studied in the analysis of societal-scale systems (Gomes et al., 2015; Guéant et al., 2011; Moll et al., 2019), such as the example of central bank or central government in §1. MF-FQI tackles the "curse of many agents" via mean embedding of the mean-field state for the (inexact) policy evaluation step, which approximately calculates the action-value function for the greedy policy. Based on the action-value function, we obtain a greedy policy, which corresponds to the policy improvement step. Such an approach is intractable when the action space also suffers from the "curse of many agents", as Q-learning requires taking the maximum over the action

space at each iteration, which can be combinatorially large if each agent takes its own action. However, the mean embedding technique is still applicable for the policy evaluation step even if each agent takes its own action, which can be coupled with other policy optimization methods, such as policy gradient (Sutton & Barto, 2018) and proximal policy optimization (Schulman et al., 2015; 2017). By replacing the greedy policy improvement step with other policy optimization methods, we are able to tackle the "curse of many agents" of both the state space and the action space, which is left as our future research.

References

Acciaio, B., Backhoff-Veraguas, J., and Carmona, R. Extended mean field control problems: stochastic maximum principle and transport perspective. *arXiv preprint arXiv:1802.05754*, 2018.

Altun, Y. and Smola, A. Unifying divergence minimization and statistical inference via convex duality. In *International Conference on Computational Learning Theory*, pp. 139–153. Springer, 2006.

Andersson, D. and Djehiche, B. A maximum principle for SDEs of mean-field type. *Applied Mathematics & Optimization*, 63(3):341–356, 2011.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

Bensoussan, A., Frehse, J., Yam, P., et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

Bloem-Reddy, B. and Teh, Y. W. Probabilistic symmetry and invariant neural networks. *arXiv preprint arXiv:1901.06082*, 2019.

Bu, L., Babu, R., De Schutter, B., et al. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Buckdahn, R., Djehiche, B., Li, J., Peng, S., et al. Mean-field backward stochastic differential equations: a limit approach. *The Annals of Probability*, 37(4):1524–1565, 2009.

Buckdahn, R., Djehiche, B., and Li, J. A general stochastic maximum principle for SDEs of mean-field type. *Applied Mathematics & Optimization*, 64(2):197–216, 2011.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Carmona, R. and Delarue, F. *Probabilistic Theory of Mean Field Games with Applications I-II*. Springer, 2018.

Carmona, R., Delarue, F., and Lachapelle, A. Control of McKean–Vlasov dynamics versus mean field games. *Mathematics and Financial Economics*, 7(2):131–166, 2013.

Carmona, R., Delarue, F., et al. Forward–backward stochastic differential equations and controlled McKean–Vlasov dynamics. *The Annals of Probability*, 43(5):2647–2700, 2015.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. *arXiv preprint arXiv:1905.00360*, 2019.

Claus, C. and Boutilier, C. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998:746–752, 1998.

Džeroski, S., De Raedt, L., and Driessens, K. Relational reinforcement learning. *Machine Learning*, 43(1-2):7–52, 2001.

Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(Apr):503–556, 2005.

Farahmand, A.-m., Szepesvári, C., and Munos, R. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, pp. 568–576, 2010.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. Regularized policy iteration with nonparametric function spaces. *Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Fornasier, M. and Solombrino, F. Mean-field optimal control. *ESAIM: Control, Optimisation and Calculus of Variations*, 20(4):1123–1152, 2014.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, pp. 489–496, 2008.

Gärtner, T., Flach, P. A., Kowalczyk, A., and Smola, A. J. Multi-instance kernels. In *International Conference on Machine Learning*, volume 2, pp. 7, 2002.

Gomes, D. A., Nurbekyan, L., and Pimentel, E. Economic models and mean-field games theory. In *30th Brazilian Mathematics Colloquium*, 2015.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pp. 513–520, 2007.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4): 5, 2009.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Guéant, O., Lasry, J.-M., and Lions, P.-L. Mean field games and applications. In *Paris-Princeton lectures on mathematical finance 2010*, pp. 205–266. Springer, 2011.

Guestrin, C., Lagoudakis, M., and Parr, R. Coordinated reinforcement learning. In *ICML*, volume 2, pp. 227–234. Citeseer, 2002.

Guo, X., Hu, A., Xu, R., and Zhang, J. Learning mean-field games. *arXiv preprint arXiv:1901.09585*, 2019.

Haussler, D. Convolution kernels on discrete structures. Technical report, Department of Computer Science, University of California, 1999.

Hu, J. and Wellman, M. P. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4(Nov):1039–1069, 2003.

Huang, M., Caines, P. E., and Malhamé, R. P. 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*, volume 1, pp. 98–103. IEEE, 2003.

Huang, M., Caines, P. E., and Malhamé, R. P. Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ε-Nash equilibria. *IEEE Transactions on Automatic Control*, 52(9):1560–1571, 2007.

Huang, M., Caines, P. E., and Malhamé, R. P. Social optima in mean field LQG control: centralized and decentralized strategies. *IEEE Transactions on Automatic Control*, 57 (7):1736–1751, 2012.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Jiachen, Y., Xiaojing, Y., Rakshit, T., Huan, X., and Hongyuan, Z. Learning deep mean field games for modeling large population behavior. *arXiv preprint arXiv:1711.03156*, 2017.

Jiang, J. and Lu, Z. Learning attentional communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems*, pp. 7254–7264, 2018.

Kar, S., Moura, J. M., and Poor, H. V. $\mathcal{QD}$-learning: A collaborative distributed strategy for multi-agent reinforcement learning through $\mathrm{Consensus} + \mathrm{Innovations}$. *IEEE Transactions on Signal Processing*, 61(7):1848–1862, 2013.

Kondor, R. and Jebara, T. A kernel between sets of vectors. In *International Conference on Machine Learning*, pp. 361–368, 2003.

Lasry, J.-M. and Lions, P.-L. Jeux à champ moyen. i–le cas stationnaire. *Comptes Rendus Mathématique*, 343(9): 619–625, 2006a.

Lasry, J.-M. and Lions, P.-L. Jeux à champ moyen. ii–horizon fini et contrôle optimal. *Comptes Rendus Mathématique*, 343(10):679–684, 2006b.

Lasry, J.-M. and Lions, P.-L. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

Lauer, M. and Riedmiller, M. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *International Conference on Machine Learning*. Citeseer, 2000.

Lazaric, A., Ghavamzadeh, M., and Munos, R. Analysis of classification-based policy iteration algorithms. *Journal of Machine Learning Research*, 17(1):583–612, 2016.

Li, M., Jiao, Y., Yang, Y., Gong, Z., Wang, J., Wang, C., Wu, G., Ye, J., et al. Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning. *arXiv preprint arXiv:1901.11454*, 2019.

Lin, K., Zhao, R., Xu, Z., and Zhou, J. Efficient large-scale fleet management via multi-agent deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1774–1783. ACM, 2018.

Lin, S.-B., Guo, X., and Zhou, D.-X. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

Lin, Y., Zhang, K., Yang, Z., Wang, Z., Basar, T., Sandhu, R., and Liu, J. A communication-efficient multi-agent actor-critic algorithm for distributed reinforcement learning. *arXiv preprint arXiv:1907.03053*, 2019.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings*, pp. 157–163. Elsevier, 1994.

Meyer-Brandis, T., Øksendal, B., and Zhou, X. Y. A mean-field stochastic maximum principle via Malliavin calculus. *Stochastics An International Journal of Probability and Stochastic Processes*, 84(5-6):643–666, 2012.

Moll, B., Rachel, L., and Restrepo, P. Uneven growth: automation's impact on income and wealth inequality. *Manuscript, Princeton University*, 2019.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems*, pp. 10–18, 2012.

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

Nash, J. Non-cooperative games. *Annals of Mathematics*, pp. 286–295, 1951.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

Ng, Y.-K. Bentham or Bergson? Finite sensibility, utility functions and social welfare functions. *The Review of Economic Studies*, 42(4):545–569, 1975.

OpenAI. Openai Five. https://blog.openai.com/openai-five/, 2018.

Panait, L. and Luke, S. Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems*, 11(3):387–434, 2005.

Scherrer, B. On the performance bounds of some policy search dynamic programming algorithms. *arXiv preprint arXiv:1306.0539*, 2013.

Scherrer, B., Ghavamzadeh, M., Gabillon, V., Lesner, B., and Geist, M. Approximate modified policy iteration and its application to the game of Tetris. *Journal of Machine Learning Research*, 16:1629–1676, 2015.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354, 2017.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

Suttle, W., Yang, Z., Zhang, K., Wang, Z., Basar, T., and Liu, J. A multi-agent off-policy actor-critic algorithm for distributed reinforcement learning. *arXiv preprint arXiv:1903.06372*, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT press, 2018.

Szabó, Z., Gretton, A., Póczos, B., and Sriperumbudur, B. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*, pp. 948–957, 2015.

Szepesvári, C. and Munos, R. Finite time bounds for sampling based fitted value iteration. In *International Conference on Machine Learning*, 2005.

Tan, M. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *International Conference on Machine Learning*, pp. 330–337, 1993.

Tolstikhin, I., Sriperumbudur, B. K., and Muandet, K. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(1):3002–3048, 2017.

Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W. M., Dudzik, A., Huang, A., Georgiev, P., Powell, R., Ewalds, T., Horgan, D., Kroiss, M., Danihelka, I., Agapiou, J., Oh, J., Dalibard, V., Choi, D., Sifre, L., Sulsky, Y., Vezhnevets, S., Molloy, J., Cai, T., Budden, D., Paine, T., Gulcehre, C., Wang, Z., Pfaff, T., Pohlen, T., Wu, Y., Yogatama, D., Cohen, J., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Apps, C., Kavukcuoglu, K., Hassabis, D., and Silver,

D. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/, 2019.

Wang, X. and Sandholm, T. Reinforcement learning to play an optimal Nash equilibrium in team Markov games. In *Advances in Neural Information Processing Systems*, pp. 1603–1610, 2003.

Yang, E. and Gu, D. Multiagent reinforcement learning for multi-robot systems: A survey. Technical report, University of Essex, 2004.

Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. Mean field multi-agent reinforcement learning. *arXiv preprint arXiv:1802.05438*, 2018.

Yang, Z., Xie, Y., and Wang, Z. A theoretical analysis of deep Q-learning. *arXiv preprint arXiv:1901.00137*, 2019.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. In *Advances in Neural Information Processing Systems*, pp. 3391–3401, 2017.

Zambaldi, V., Raposo, D., Santoro, A., Bapst, V., Li, Y., Babuschkin, I., Tuyls, K., Reichert, D., Lillicrap, T., Lockhart, E., et al. Relational deep reinforcement learning. *arXiv preprint arXiv:1806.01830*, 2018.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Başar, T. Finite-sample analyses for fully decentralized multi-agent reinforcement learning. *arXiv preprint arXiv:1812.02783*, 2018a.

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Başar, T. Fully decentralized multi-agent reinforcement learning with networked agents. *arXiv preprint arXiv:1802.08757*, 2018b.

Zhang, K., Yang, Z., and Başar, T. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pp. 11598–11610, 2019a.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019b.

A. Topological Structures

We now establish a topological structure on the space $\widetilde{\mathcal{M}}(\Omega)$ adopted from (Szabó et al., 2015). Recall that we denote by $\Omega = \mathcal{S} \times \mathcal{A}$ the space of state-action pairs. We assume that Ω is a polish space, and denote by $\mathcal{B}(\Omega)$ the Borel σ-algebra of Ω. We denote by $\mathcal{M}_0(\mathcal{A})$ the space of all the point mass distributions on $\mathcal{A}$, and denote by $\mathcal{M}(\mathcal{S})$ the space of all the distributions on $\mathcal{S}$. We assume that both $\mathcal{M}_0(\mathcal{A})$ and $\mathcal{M}(\mathcal{S})$ are equipped with the weak topology such that for all $f \in \mathcal{C}_B(\mathcal{A})$ and $g \in \mathcal{C}_B(\mathcal{S})$, the mappings $p \mapsto \int f(x) \, dp(x)$ and $q \mapsto \int g(x) \, dq(x)$ are continuous for $p \in \mathcal{M}_0(\Omega)$ and $q \in \mathcal{M}(\Omega)$, respectively. Note that any $p \in \mathcal{M}_0(\mathcal{A})$ and $q \in \mathcal{M}(\mathcal{S})$ defines a product measure $\omega = p \times q \in \widetilde{\mathcal{M}}(\Omega)$ on $(\Omega, \mathcal{B}(\mathcal{A}) \otimes \mathcal{B}(\mathcal{S}))$. We endow the set $\widetilde{\mathcal{M}}(\Omega)$ with the product topology of corresponding weak topology defined on $\mathcal{M}_0(\mathcal{A})$ and $\mathcal{M}(\mathcal{S})$, which makes $\widetilde{\mathcal{M}}(\Omega)$ a Polish space.

B. Proof of Main Results

B.1. Proof of Proposition 2.1

We first show that if (2.7) holds, then $Q^*(s,a) = Q^*(\sigma(s),a)$ for any $s \in \mathcal{S}$ and $\sigma \in \mathbb{S}_n$, where $\mathbb{S}_n$ is the permutation group of order n. For any $\sigma \in \mathbb{S}_n$, we define the function $f(s,a) = Q^*(\sigma^{-1}(s),a)$. Therefore, it holds that

$$f\big(\sigma(s),a\big) = Q^*\big(\sigma^{-1} \circ \sigma(s),a\big) = Q^*(s,a).$$

Then, following from the Bellman optimality equation in (2.6), it holds that

$$f\big(\sigma(s),a\big) = \mathbb{E}\Big[r(s,a) + \gamma \cdot \max_{a \in \mathcal{A}} f\big(\sigma(S'),a\big)\Big], \tag{B.1}$$

where $S' \sim P(\cdot|s,a)$. We denote by $S'' \sim P(\cdot|\sigma(s),a)$. Following from (2.6), it holds that

$$\mathbb{P}(s' \mid s,a) = \mathbb{P}\big(\sigma(s') \mid \sigma(s),a\big). \tag{B.2}$$

Following from (B.2), it then holds that $S'' \stackrel{d}{=} \sigma(S')$. Meanwhile, it holds from (2.6) that $r(s,a) \stackrel{d}{=} r(\sigma(s),a)$. Therefore, following from (B.1), we obtain that

$$f\big(\sigma(s),a\big) = \mathbb{E}\Big[r\big(\sigma(s),a\big) + \gamma \cdot \max_{a \in \mathcal{A}} f\big(S'',a\big)\Big], \tag{B.3}$$

where $S'' \sim P(\cdot|\sigma(s),a)$. Note that (B.3) holds for any $\sigma \in \mathbb{S}_n$ and $s \in \mathcal{S}$. Therefore, the function $f(s,a)$ is an optimal action-value function. Following from the uniqueness of the optimal action-value function Q^* for a discounted MDP, it holds that

$$f(s,a) = Q^*(\sigma^{-1}(s),a) = Q^*(s,a). \tag{B.4}$$

Following from (B.4), it holds for all $\sigma \in \mathbb{S}_n$ that

$$\pi^*(a \mid s) = \operatorname*{argmax}_{a \in \mathcal{A}} Q^*(s,a) = \operatorname*{argmax}_{a \in \mathcal{A}} Q^*\big(\sigma(s),a\big) = \pi^*\big(a \mid \sigma(s)\big), \tag{B.5}$$

which conclude the proof of Proposition 2.1.

B.2. Proof of Proposition 2.2

Note that the policy $\pi(a|s)$ together with $\mathbb{P}(\cdot|S = s, A = a)$ defines a Markov process on (s,a) with transition dynamics

$$S_{t+1} \sim P(\cdot \mid S_t, A_t), \quad A_{t+1} \sim \pi(\cdot \mid S_{t+1}).$$

We denote by $R(S_0, A_0)$ the discounted total reward following the MDP with initial state S_0 and A_0, which takes the form

$$R(S_0, A_0) = \sum_{t=0}^{\infty} \gamma^t \cdot r(S_t, A_t). \tag{B.6}$$

Moreover, following from the definition of action-value function in (2.3), it holds that

$$Q(s,a) = \mathbb{E}\big[R(S,A) \,\big|\, S = s, A = a\big]. \tag{B.7}$$

Note that for any invariant policy $\pi(a \,|\, s) \overset{\mathrm{d}}{=} \pi(a \,|\, \sigma(s))$, R is invariant to the permutation of state S_0. With a slight abuse of notation, we denote by $R(S,a) = R(S,A) \,|\, A = a$. Then following from the definition of $R(S,A)$ in (B.6), we obtain that

$$R(S_0, a) \overset{d}{=} R\big(\sigma(S_0), a\big), \quad \sigma \in \mathbb{S}_n, \tag{B.8}$$

which holds for any $a \in \mathcal{A}$. Following from Theorem 12 in (Bloem-Reddy & Teh, 2019), it holds for some f_a that

$$R(S_0, a) = f_a\big(\eta, M(S_0)\big), \quad \eta \sim \mathrm{Unif}[0,1]. \tag{B.9}$$

Here M_{S_0} is the maximal invariant on S under the permutation group S_N, which takes the form of (Bloem-Reddy & Teh, 2019)

$$M_{S_0}(\cdot) = \frac{1}{N} \sum_{i=1}^{N} \delta_{S_{0,i}},$$

where $S_{0,i}$ is the i-th component of S_0, and δ_x is the point mass at x. Following from (B.7), we complete the proof of Proposition 2.2 by setting $g(M,a) = \mathbb{E}_\eta[f_a(\eta, M) \,|\, S_0 = s]$.

B.3. Proof of Theorem 3.2

The proof strategy is similar to that of Theorem 4.5 by (Yang et al., 2019). We denote by $\widehat{Q}_\kappa^\lambda$ the output of MF-FQI in the κ-th iteration. We further denote by

$$\epsilon_\kappa = T\widehat{Q}_{\kappa-1}^\lambda - \widehat{Q}_\kappa^\lambda.$$

The function ϵ_κ is then the one-step error that corresponds to the κ-th iteration. Recall that we denote by π_κ the greedy policy with respect to $\widehat{Q}_\kappa^\lambda$, and Q^{π_κ} the action-value function with respect to the policy π_κ. For each function Q and policy π, we define the bellman operator T^π and evaluation operator E^π as follows,

$$E^\pi Q(\delta_a \times p_s) = \mathbb{E}\big[Q(\delta_{A'} \times P_{s'}) \,\big|\, A' \sim \pi(\cdot \,|\, p_s), \, P_{s'} \sim P(\cdot \,|\, \delta_a \times p_s)\big],$$
$$T^\pi Q(\delta_a \times p_s) = \mathbb{E}\big[r(\delta_a \times p_s)\big] + \gamma \cdot E^\pi Q(\delta_a \times p_s). \tag{B.10}$$

Recall that we denote by Q^* and π^* the optimal action-value function and corresponding greedy policy. In what follows, we establish the error bound $\|Q^* - Q^{\pi_\kappa}\|_{1,\mu}$, where μ is a distribution over the space $\widetilde{M}(\Omega)$. The idea is to first establish a recursive relation between $Q^* - \widehat{Q}_\kappa^\lambda$, and then upper bound the error of $|Q^* - Q^{\pi_\kappa}|$ by $|Q^* - \widehat{Q}_\kappa^\lambda|$. We denote by π_Q the greedy policy with respect to Q. It then holds that

$$T^{\pi_Q}(\delta_a \times p_s) = \mathbb{E}\big[r(\delta_a \times p_s)\big] + \gamma \cdot \mathbb{E}\Big[\max_{a' \in \mathcal{A}} Q(\delta_{a'} \times P_{s'}) \,\Big|\, P_{s'} \sim P(\cdot \,|\, \delta_a \times p_s)\Big]$$
$$= TQ(\delta_a \times p_s). \tag{B.11}$$

Meanwhile, it holds for any policy π that

$$\mathbb{E}\Big[\max_{a' \in \mathcal{A}} Q(\delta_{a'} \times P_{s'}) \,\Big|\, P_{s'} \sim P(\cdot \,|\, \delta_a \times p_s)\Big] \geq E^\pi Q(\delta_a \times p_s). \tag{B.12}$$

Combining (B.11) and (B.12), we obtain the following identity,

$$T^{\pi_Q} Q(\omega) = TQ(\omega) \geq T^\pi Q(\omega), \tag{B.13}$$

which holds for any $\omega \in \widetilde{\mathcal{M}}(\Omega)$ and policy π. Note that $T^{\pi^*} Q^* = Q^*$. Therefore, following from (B.13), we obtain that

$$Q^* - \widehat{Q}_{\kappa+1}^\lambda = Q^* - T\widehat{Q}_k^\lambda + \epsilon_{\kappa+1} \leq T^{\pi^*} Q^* - T^{\pi^*} \widehat{Q}_\kappa^\lambda + \epsilon_{\kappa+1}$$
$$= \gamma \cdot E^{\pi^*}(Q^* - \widehat{Q}_\kappa^\lambda) + \epsilon_{\kappa+1}, \tag{B.14}$$

where the last equality follows from the definition of Bellman operator in (B.10). Similarly, it holds that

$$Q^* - \widehat{Q}^\lambda_{\kappa+1} = T^{\pi^*} Q^* - T\widehat{Q}^\lambda_\kappa + \epsilon_{\kappa+1} \geq T^{\pi_k} Q^* - T^{\pi_\kappa} \widehat{Q}^\lambda_\kappa + \epsilon_{\kappa+1}$$
$$= \gamma \cdot E^{\pi_\kappa}(Q^* - \widehat{Q}^\lambda_\kappa) + \epsilon_{\kappa+1}, \tag{B.15}$$

where recall that we denote by π_k the greedy policy with respect to $\widehat{Q}^\lambda_\kappa$. Following the definition of E^π in (B.10), it holds that the operator E^π is linear. Therefore, upon iteratively applying (B.14) and (B.15), it holds for any $K > \kappa$ that

$$Q^* - \widehat{Q}^\lambda_K \leq \gamma^{K-\kappa} \cdot (E^{\pi^*})^{K-\kappa}(Q^* - \widehat{Q}^\lambda_\kappa) + \sum_{i=\kappa}^{K-1} \gamma^{K-1-i} \cdot (E^{\pi^*})^{K-1-i}\epsilon_{i+1},$$

$$Q^* - \widehat{Q}^\lambda_K \geq \gamma^{K-\kappa} \cdot \left(\prod_{i=\kappa}^{K-1} E^{\pi_i}\right)(Q^* - \widehat{Q}^\lambda_\kappa) + \sum_{i=\kappa}^{K-1} \gamma^{K-1-i} \cdot \left(\prod_{j=i+1}^{K-1} E^{\pi_j}\right)\epsilon_{i+1}, \tag{B.16}$$

where we denote by $\prod_{j=i+1}^{K-1} E^{\pi_j}$ the composition of operators $E^{\pi_{K-1}} \circ E^{\pi_{K-2}} \circ \ldots \circ E^{\pi_{i+1}}$. We now bound the error $Q^* - Q^{\pi_\kappa}$, where π_κ is the greedy policy with respect to $\widehat{Q}^\lambda_\kappa$. Following from (B.13) and the linearity of Bellman operators, it holds that

$$Q^* - Q^{\pi_\kappa} = T^{\pi^*} Q^* - T^{\pi_\kappa} Q^{\pi_\kappa} = T^{\pi^*}(Q^* - \widehat{Q}^\lambda_\kappa) + (T^{\pi^*} - T^{\pi_\kappa})\widehat{Q}^\lambda_\kappa + T^{\pi_\kappa}(\widehat{Q}^\lambda_\kappa - Q^{\pi_\kappa})$$
$$\leq T^{\pi^*}(Q^* - \widehat{Q}^\lambda_\kappa) + T^{\pi_\kappa}(\widehat{Q}^\lambda_\kappa - Q^{\pi_\kappa}). \tag{B.17}$$

Meanwhile, it follows from the definition of Q^* that $Q^* \geq Q^{\pi_\kappa}$. Combining with (B.17), we obtain that

$$0 \leq Q^* - Q^{\pi_\kappa} \leq T^{\pi^*}(Q^* - \widehat{Q}^\lambda_\kappa) + T^{\pi_\kappa}(\widehat{Q}^\lambda_\kappa - Q^{\pi_\kappa})$$
$$= \gamma \cdot E^{\pi^*}(Q^* - \widehat{Q}^\lambda_\kappa) + \gamma \cdot E^{\pi_\kappa}(\widehat{Q}^\lambda_\kappa - Q^{\pi_\kappa})$$
$$= \gamma \cdot (E^{\pi^*} - E^{\pi_\kappa})(Q^* - \widehat{Q}^\lambda_\kappa) + \gamma \cdot E^{\pi_\kappa}(Q^* - Q^{\pi_\kappa}). \tag{B.18}$$

where the first equality follows from the definition of Bellman operator in (B.10). Note that the Bellman operator T^π is contractive with respect to any given policy π and bounded reward. It thus holds that the operator $I - \gamma E^{\pi_k}$ is invertible, where I is the identity mapping. Following from (B.18), we obtain that

$$Q^* - Q^{\pi_\kappa} \leq \gamma \cdot (I - \gamma E^{\pi_\kappa})^{-1}(E^{\pi^*} - E^{\pi_\kappa})(Q^* - \widehat{Q}^\lambda_\kappa), \tag{B.19}$$

which holds for any $\kappa > 0$. Following from the definition in (B.10), it holds for any π and $Q_1 \geq Q_2$ that $E^\pi Q_1 \geq E^\pi Q_2$. Then following from the infinite expansion of $(I - \gamma E^{\pi_k})^{-1}$, it holds for any π and $Q_1 \geq Q_2$ that

$$(I - \gamma E^{\pi_\kappa})^{-1}Q_1 \geq (I - \gamma E^{\pi_\kappa})^{-1}Q_2.$$

Therefore, combining (B.16) and (B.19), we obtain for $\kappa > 0$ that

$$0 \leq Q^* - Q^{\pi_\kappa} \leq (I - \gamma E^{\pi_\kappa})^{-1}\left\{\sum_{i=0}^{\kappa-1} \gamma^{\kappa-i} \cdot \left((E^{\pi^*})^{\kappa-i} - \prod_{j=i+1}^{\kappa} E^{\pi_j}\right)\epsilon_{i+1}\right.$$
$$\left. + \gamma^{\kappa+1} \cdot \left((E^{\pi^*})^{\kappa+1} - \prod_{j=0}^{\kappa} E^{\pi_j}\right)(Q^* - \widehat{Q}^\lambda_0)\right\}. \tag{B.20}$$

For notational simplicity, we introduce the following shorthands,

$$J_\kappa = (I - \gamma E^{\pi_\kappa})^{-1}\left((E^{\pi^*})^{\kappa+1} - \prod_{j=0}^{\kappa} E^{\pi_j}\right),$$

$$F_{i,\kappa} = (I - \gamma E^{\pi_k})^{-1}\left((E^{\pi^*})^{\kappa-i} - \prod_{j=i+1}^{\kappa} E^{\pi_j}\right), \quad i = 0, \ldots, \kappa - 1. \tag{B.21}$$

Based on the shorthands defined in (B.21), we rewrite (B.20) as follows,

$$0 \leq Q^* - Q^{\pi_\kappa} \leq \gamma^{\kappa+1} \cdot J_\kappa(Q^* - \widehat{Q}_0^\lambda) + \sum_{i=0}^{\kappa-1} \gamma^{\kappa-i} F_{i,\kappa}(\epsilon_{i+1}). \tag{B.22}$$

In what follows, we bound the error $\|Q^* - Q^{\pi_\kappa}\|_{1,\mu}$ based on (B.20), where μ is a given distribution on $\widetilde{\mathcal{M}}(\Omega)$. Recall that for a function f defined on the support of a distribution μ, we define the shorthand $\mu(f) = \mathbb{E}_{\omega \sim \mu}[f(\omega)]$. Following from (B.22) and the linearity of operators $F_{i,\kappa}$, J_κ, and $\mu(\cdot)$, we obtain that

$$\|Q^* - Q^{\pi_\kappa}\|_{1,\mu} = \mu(|Q^* - Q^{\pi_\kappa}|) \leq \gamma^{\kappa+1} \cdot \mu(J_\kappa|Q^* - \widehat{Q}_0^\lambda|) + \sum_{i=0}^{\kappa-1} \gamma^{\kappa-i} \mu(F_{i,\kappa}|\epsilon_{i+1}|). \tag{B.23}$$

Note that the operator E^π defined in (B.10) is a Markov transition operator defined on $\widetilde{\mathcal{M}}(\Omega)$ with transition dynamics given by

$$P_{s,t+1} \sim P(\cdot \mid A_t, P_{s,t}), \quad A_{t+1} \sim \pi(\cdot \mid P_{s,t+1}),$$

Thus, the operator E^π is a Markov transition kernel defined on $\widetilde{\mathcal{M}}(\Omega)$. The composition $\mu(E^\pi) = E^\pi \cdot \mu$ defines a probability measure, which is a transition over the initial probability distribution μ on $\widetilde{\mathcal{M}}(\Omega)$. Indeed, for any $f(\cdot)$ defined on $\widetilde{\mathcal{M}}(\Omega)$, we denote by $X_0 \sim \mu$ the initial distribution of state-action configurations. It then holds that

$$\mu(E^\pi f) = \mathbb{E}\big[f(X_1)\big] = \mu_1(f), \tag{B.24}$$

where X_1 is the transition of X_0 following the Markov transition operator E^π, and μ_1 is the marginal distribution of X_1. The composite of operators $\mu(J_K)$ can be expanded in the following infinite sum,

$$\mu(J_K) = \sum_{\ell=0}^\infty \gamma^\ell \cdot \mu\big((\mathbb{E}^{\pi_\kappa})^\ell \circ (E^{\pi^*})^{\kappa+1}\big) - \gamma^\ell \cdot \mu\bigg((E^{\pi_\kappa})^\ell \circ \prod_{j=0}^\kappa E^{\pi_j}\bigg), \tag{B.25}$$

where each term of the summation is a difference of probability measures multiplied by γ^ℓ. Note that $\|Q\|_\infty \leq Q_{\max}$ and $\|\widehat{Q}_0^\lambda\|_\infty \leq Q_{\max}$. Therefore, following from (B.25), we obtain that

$$\mu(J_\kappa|Q^* - \widehat{Q}_0^\lambda|) \leq \sum_{\ell=0}^\infty \gamma^\ell \cdot \bigg\{ \mu\big((E^{\pi_\kappa})^\ell \circ (E^{\pi^*})^{\kappa+1}|Q^* - \widehat{Q}_0^\lambda|\big) - \mu\bigg((\mathbb{E}^{\pi_\kappa})^\ell \circ \prod_{j=0}^\kappa E^{\pi_j}|Q^* - \widehat{Q}_0^\lambda|\bigg)\bigg\}$$

$$\leq \sum_{\ell=0}^\infty \gamma^\ell \cdot 4Q_{\max} \leq 4Q_{\max}/(1-\gamma). \tag{B.26}$$

Similarly, we bound the term $\mu(F_{i,\kappa}|\epsilon_{i+1}|)$. Following from the definition of $F_{i,\kappa}$ in (B.21), it holds that

$$\mu(F_{j,\kappa}|\epsilon_{i+1}|)$$
$$= \sum_{\ell=0}^\infty \gamma^\ell \cdot \bigg\{ \mu\big((\mathbb{E}^{\pi_\kappa})^\ell \circ (E^{\pi^*})^{\kappa-i}|\epsilon_{i+1}|\big) - \mu\bigg((E^{\pi_\kappa})^\ell \circ \prod_{j=i+1}^\kappa E^{\pi_j}|\epsilon_{i+1}|\bigg)\bigg\}. \tag{B.27}$$

We first bound the term $\mu((E^{\pi_\kappa})^\ell \circ (E^{\pi^*})^{\kappa-i}|\epsilon_{i+1}|)$. We define a Markov process X_t with initial state $X_0 \sim \mu_0$. We define the transition operator to be E^{π^*} for $0 < t \leq \kappa - i$ and E^{π_κ} for $\kappa - i < t \leq \kappa - i + \ell$. We then denote by $\widetilde{\mu}$ the marginal distribution of $X_{\kappa-i+\ell}$. Following from the Cauchy-Schwartz inequality, we obtain that

$$\mu\big((E^{\pi_\kappa})^\ell \circ (E^{\pi^*})^{\kappa-i}|\epsilon_{i+1}|\big) = \widetilde{\mu}(|\epsilon_{i+1}|) = \int_{\widetilde{\mathcal{M}}(\Omega)} |\epsilon_{i+1}(\omega)| \, d\widetilde{\mu}(\omega)$$

$$\leq \bigg(\int_{\widetilde{\mathcal{M}}(\Omega)} |\epsilon_{i+1}(\omega)|^2 \, d\nu(\omega)\bigg)^{1/2} \cdot \bigg(\int_{\widetilde{\mathcal{M}}(\Omega)} \Big|\frac{d\widetilde{\mu}}{d\nu}(\omega)\Big|^2 \, d\nu(\omega)\bigg)^{1/2}, \tag{B.28}$$

where $\mathrm{d}\widetilde{\mu}/\mathrm{d}\nu$ is the Radon-Nikodym derivative. Following from Assumption 3.1, it then holds that

$$\mu\big((E^{\pi_\kappa})^\ell \circ (E^{\pi^*})^{\kappa-i}|\epsilon_{i+1}|\big) \leq \phi(\kappa - i + \ell; \mu, \nu) \cdot \|\epsilon_{i+1}\|_\nu. \tag{B.29}$$

Note that the same bound holds if we change the transition operators. Therefore, combining (B.27) and (B.29), it then holds that

$$\mu(F_{j,\kappa}|\epsilon_{i+1}|) \leq \sum_{\ell=0}^\infty 2\gamma^\ell \cdot \phi(\kappa - i + \ell; \mu, \nu) \cdot \|\epsilon_{i+1}\|_\nu. \tag{B.30}$$

We denote by $\epsilon_{\max,\kappa} = \max_{i \in [\kappa]} \|\epsilon_i\|_\nu$. Following from (B.30), it then holds that

$$\sum_{i=0}^{\kappa-1} \gamma^{\kappa-i} \mu(F_{i,\kappa}|\epsilon_{i+1}|) \leq 2 \sum_{i=0}^{\kappa-1} \sum_{\ell=0}^\infty \gamma^{\kappa-i+\ell} \cdot \phi(\kappa - i + \ell; \mu, \nu) \cdot \|\epsilon_{i+1}\|_\nu$$

$$\leq 2\epsilon_{\max,\kappa} \cdot \sum_{\ell=0}^\infty \sum_{m=\ell+1}^{\ell+\kappa} \gamma^m \cdot \phi(m; \mu, \nu)$$

$$\leq 2\epsilon_{\max,\kappa} \cdot \sum_{m=1}^\infty m \cdot \gamma^m \cdot \phi(m; \mu, \nu), \tag{B.31}$$

Combining (B.31) and Assumption 3.1, we obtain that

$$\sum_{i=0}^{\kappa-1} \gamma^{\kappa-i} \mu(F_{i,\kappa}|\epsilon_{i+1}|) \leq 2\epsilon_{\max,\kappa} \cdot \sum_{m=1}^\infty m \cdot \gamma^m \cdot \phi(m; \mu, \nu) \leq \frac{2\gamma \cdot \Phi(\mu, \nu)}{(1-\gamma)^2} \cdot \epsilon_{\max,\kappa}. \tag{B.32}$$

Finally, combining (B.23), (B.26), and (B.32), we conclude that

$$\|Q^* - Q^{\pi_\kappa}\|_{1,\mu} \leq \frac{2\gamma \cdot \Phi(\mu, \nu)}{(1-\gamma)^2} \cdot \epsilon_{\max,\kappa} + \frac{4\gamma^{\kappa+1} \cdot Q_{\max}}{1-\gamma}, \tag{B.33}$$

which completes the proof of Theorem 3.2.

B.4. Proof of Theorem 3.4

Proof. We define $\widehat{Q} = \widehat{Q}_k^\lambda$ and $Q = Q_k^\lambda$ for notational simplicity, where $\widehat{Q}_k^\lambda$ and Q_k^λ are define in (2.25). Recall that we denote by $\{\omega_i\}_{i\in[n]}$ the sample that is drawn independently from the sampling distribution ν, where $\omega_i = \delta_{a_i} \times p_{i,s}$, $\widehat{p}_{i,s}$ the empirical approximation of $p_{i,s}$ with N observations, and $\widehat{\omega}_i = \delta_{a_i} \times \widehat{p}_{i,s}$. We further denote by $p_{i,s'}$ the mean-field state after transition, which follows the distribution $P(\cdot \,|\, \omega_{a_i,s_i})$, and $\widehat{p}_{i,s'}$ the empirical approximation of $p_{i,s'}$ with N observations. In what follows, we define $\widehat{Q}^\lambda = \widehat{Q}_{k+1}^\lambda$ and $Q^\lambda = Q_{k+1}^\lambda$ for notational simplicity. More specifically, we define

$$Q^\lambda = \underset{f \in \mathcal{H}(K)}{\mathrm{argmin}} \frac{1}{n} \sum_{i=1}^n \big(f(\mu_{\widehat{\omega}_i}) - \widehat{y}_i\big)^2 + \lambda\|f\|_{\mathcal{H}(K)}^2, \quad \widehat{Q}^\lambda = \max\{Q^\lambda, Q_{\max}\}, \tag{B.34}$$

where

$$\widehat{y}_i = r_i + \gamma \cdot \sup_{a \in \mathcal{A}} \widehat{Q}(\mu_{\delta_a \times \widehat{p}_{i,s'}}). \tag{B.35}$$

We now analyze the one-step approximation error $\|\widehat{Q}^\lambda - T\widehat{Q}\|_\nu$, where T is the Bellman optimality operator. Note that by the truncation, it holds that $\widehat{Q} \leq Q_{\max}$ and therefore $T\widehat{Q} \leq Q_{\max}$. Hence, we have

$$\|\widehat{Q}^\lambda - T\widehat{Q}\|_\nu \leq \|Q^\lambda - T\widehat{Q}\|_\nu, \tag{B.36}$$

which holds since $\widehat{Q}^\lambda = \max\{Q^\lambda, Q_{\max}\}$. In the sequel, we upper bound the right-hand side of (B.36). To this end, we define

$$\overline{Q}^\lambda = \underset{f \in \mathcal{H}(K)}{\mathrm{argmin}} \frac{1}{n} \sum_{i=1}^n \big(f(\mu_{\omega_i}) - y_i\big)^2 + \lambda\|f\|_{\mathcal{H}(K)}^2, \tag{B.37}$$

where

$$y_i = r_i + \gamma \cdot \sup_{a \in \mathcal{A}} \widehat{Q}(\mu_{\delta_a \times p_{i,s'}}). \tag{B.38}$$

Here $\overline{Q}^\lambda$ corresponds to the regression in (B.34) if we use the mean embedding of exact mean-field states $p_{i,s}$ and $p_{i,s'}$ in place of their finite sample approximations $\widehat{p}_{i,s}$ and $\widehat{p}_{i,s'}$, respectively. Meanwhile, for random variables $A_i \in \mathcal{A}$ and $P_{i,s} \in \mathcal{M}(\mathcal{S})$, we define the random variable $Y_i = r(\delta_{A_i} \times P_{i,s}) + \max_{a \in \mathcal{A}} Q(\mu_{\delta_a \times P_{i,s'}})$, where $P_{i,s'} \sim P(\cdot \,|\, \delta_A \times P_{i,s})$. It then holds that y_i is a realization of the random variable $Y_i \,|\, A_i = a, P_{i,s} = p_s$. We denote by $\omega_{a,s} = \delta_a \times p_s$ and $\rho(\cdot \,|\, \omega_{a,s})$ the distribution of $Y_i \,|\, A_i = a, P_{i,s} = p_s$. The function TQ is thus defined as

$$TQ(\omega_{a,s}) = \mathbb{E}[Y_i \,|\, A_i = a, P_{i,s} = p_s] = \mathbb{E}\Big[r(\omega_{a,s}) + \max_{a' \in \mathcal{A}} Q(\delta_{a'} \times P_s) \,\Big|\, P_s \sim P(\cdot \,|\, \omega_{a,s})\Big].$$

Note that under the universality Assumption 2.3, each state-action configuration $\omega \in \widetilde{\mathcal{M}}(\Omega)$ uniquely characterizes a mean embedding μ_ω (Gretton et al., 2007; 2012). For each $\omega \in \widetilde{\mathcal{M}}(\Omega)$, we define $\rho(\cdot \,|\, \mu_\omega) = \rho(\cdot \,|\, \omega)$. Meanwhile, following from (Szabó et al., 2015), the probability measure ν on $\widetilde{\mathcal{M}}(\Omega)$ equivalently defines a probability measure on the space of mean embeddings $\mathcal{X}$. With a slight abuse of notations, we denote by $\nu(\mu_\omega)$ such a measure defined by $\nu(\omega)$, and do not distinguish between them in the sequel. The function TQ then equivalently defines a function on $\mathcal{X}$ that takes the form,

$$TQ(\mu_{\omega_i}) = \mathbb{E}[Y_i \,|\, \mu_{\omega_i}] = \mathbb{E}_{Y \sim \rho(\cdot \,|\, \mu_{\omega_i})}[Y].$$

Recall that we define for $f \in \mathcal{H}(K)$ the integral operator $\mathcal{C} : \mathcal{H}(K) \mapsto \mathcal{H}(K)$ as follows,

$$\mathcal{C}f(x) = \int_{\widetilde{\mathcal{M}}(\Omega)} K(x, \mu_\omega) f(\mu_\omega) \, \mathrm{d}\nu(\mu_\omega).$$

We introduce empirical integral operators $\mathcal{C}_\omega$ and $\mathcal{C}_{\widehat{\omega}}$, which is defined as follows,

$$\mathcal{C}_\omega f(x) = \frac{1}{n} \sum_{i=1}^n K(x, \mu_{\omega_i}) f(\mu_{\omega_i}), \quad \mathcal{C}_{\widehat{\omega}} f(x) = \frac{1}{n} \sum_{i=1}^n K(x, \mu_{\widehat{\omega}_i}) f(\mu_{\widehat{\omega}_i}). \tag{B.39}$$

The operators $\mathcal{C}_\omega$ and $\mathcal{C}_{\widehat{\omega}}$ are estimates of $\mathcal{C}$ with the samples $\{\mu_{\omega_i}\}_{i \in [n]}$ and $\{\mu_{\widehat{\omega}_i}\}_{i \in [n]}$, respectively. The following proposition characterizes the exact form of Q^λ and $\overline{Q}^\lambda$ defined in (B.34) and (B.37), respectively.

Proposition B.1 (Exact Solutions (Caponnetto & De Vito, 2007)). *For Q^λ and $\overline{Q}^\lambda$ defined in (B.34) and (B.37), respectively, it holds that*

$$Q^\lambda = (\mathcal{C}_{\widehat{\omega}} + \lambda)^{-1} g_{\widehat{\omega}}, \quad \overline{Q}^\lambda = (\mathcal{C}_\omega + \lambda)^{-1} g_\omega,$$

where

$$g_{\widehat{\omega}}(\cdot) = \frac{1}{n} \sum_{i=1}^n K(\cdot, \mu_{\omega_i}) \widehat{y}_i, \quad g_\omega(\cdot) = \frac{1}{n} \sum_{i=1}^n K(\cdot, \mu_{\widehat{\omega}_i}) y_i. \tag{B.40}$$

Proof. See (Caponnetto & De Vito, 2007) for a detailed proof. □

In the sequel, we denote by $Q_{\mathcal{H},T}$ the projection of $T\widehat{Q}$ onto the RKHS $\mathcal{H}(K)$ with respect to the norm $\|\cdot\|_\nu$, which is defined as follows,

$$Q_{\mathcal{H},T} \in \underset{f \in \mathcal{H}(K)}{\operatorname{argmin}} \|f - T\widehat{Q}\|_\nu.$$

The following proposition characterizes the exact risk of the one-step approximation.

Proposition B.2 (Exact Risk (Caponnetto & De Vito, 2007)). *It holds that*

$$\|Q^\lambda - T\widehat{Q}\|_\nu^2 - \|Q_{\mathcal{H},T} - T\widehat{Q}\|_\nu^2 = \|\sqrt{\mathcal{C}}(Q^\lambda - Q_{\mathcal{H},T})\|_{\mathcal{H}(K)}^2. \tag{B.41}$$

Proof. See (Caponnetto & De Vito, 2007) for a detailed proof. □

Propositions B.1 and B.2 are standard results of kernel ridge regression. Under Assumptions 2.3 and 3.3, the following lemma adapted from (Szabó et al., 2015) upper bounds the right-hand side of (B.41).

Lemma B.3 (Exact Risk Bound (Szabó et al., 2015)). *Let* $0 < \eta + \tau < 1$ *and* $C(\eta) = 32\log^2(6/\eta)$. *Under Assumptions 2.3 and 3.3, for*

$$N \geq 2\varrho \cdot (1 + \sqrt{\log n + \delta})^2 \cdot (64L^2\varsigma^2/\lambda^2)^{1/h}, \quad n \geq \frac{2C(\eta)\varsigma\beta b}{(b-1)\lambda^{1+1/b}}, \quad \lambda \leq \|\mathcal{C}\|_{\mathcal{H}(K)}, \tag{B.42}$$

it holds with probability at least $1 - \eta - \tau$ *that*

$$\|\sqrt{\mathcal{C}}(Q^\lambda - Q_{\mathcal{H},T})\|^2_{\mathcal{H}(K)} \leq \frac{8L^2 Q^2_{\max}\left(1 + \sqrt{\log(|\mathcal{A}| \cdot n/2\tau)}\right)^{2h} \cdot (2\varrho)^h}{\lambda \cdot N^h} \cdot \left(1 + \frac{5\varsigma^2}{\lambda^2}\right) \tag{B.43}$$
$$+ C(\eta) \cdot \left(R\lambda^c + \frac{\varsigma^2 R}{\lambda^{2-c}n^2} + \frac{\varsigma R\lambda^{c-1}}{4n} + \frac{\varsigma M^2}{\lambda n^2} + \frac{\Sigma^2\beta b}{(b-1)n\lambda^{1/b}}\right).$$

Proof. See §C for a detailed proof. □

We define ψ_T as follows,

$$\psi_T = \sup_{k \in [\kappa]} \|T\widehat{Q}^\lambda_k - \Pi_{\mathcal{H}(K)}(TQ^\lambda_k)\|_\nu. \tag{B.44}$$

Therefore, combining (B.44), Proposition B.2, and Lemma B.3, for

$$N \geq 2\varrho \cdot (1 + \sqrt{\log n + \delta})^2 \cdot (64L^2\varsigma^2/\lambda^2)^{1/h}, \quad n \geq \frac{2C(\eta)\varsigma\beta b}{(b-1)\lambda^{1+1/b}}, \quad \lambda \leq \|\mathcal{C}\|_{\mathcal{H}(K)}, \tag{B.45}$$

it holds with probability at least $1 - \tau - \eta$ that

$$\|Q^\lambda - T\widehat{Q}\|^2_\nu \leq \frac{8L^2 Q^2_{\max}\left(1 + \sqrt{\log(|\mathcal{A}| \cdot n/2\tau)}\right)^{2h} \cdot (2\varrho)^h}{\lambda \cdot N^h} \cdot \left(1 + \frac{5\varsigma^2}{\lambda^2}\right) + \psi^2_T \tag{B.46}$$
$$+ C(\eta) \cdot \left(R\lambda^c + \frac{\varsigma^2 R}{\lambda^{2-c}n^2} + \frac{\varsigma R\lambda^{c-1}}{4n} + \frac{\varsigma M^2}{\lambda n^2} + \frac{\Sigma^2\beta b}{(b-1)n\lambda^{1/b}}\right),$$

which concludes the proof of Theorem 3.4. □

B.5. Proof of Theorem 3.5

Proof. Throughout the proof, we denote by C and C' positive absolute constants, whose value may vary from lines to lines. It suffices to study the dominating error of the one-step approximation error in (3.8). Recall that that we set $n = N^a$. We split the proof into two cases in terms of a.

Case 1: We first consider the case when $a > h \cdot (1 + 1/b)/ \cdot (c + 3)$. Note that the dominating term of $\mathcal{G}_1$ has the order of $(\log n)^h/(N^h \cdot \lambda^3)$. Thus if $\mathcal{G}_1$ in (3.9) converges to zero as N goes to infinity, it holds that $(\log n)^h/(N^h \cdot \lambda^3)$ converges to zero. Therefore, for any positive absolute constant C, it holds for a sufficiently large N that $\log n/(N \cdot \lambda^{2/h}) \leq C$. The requirement on N in (3.7) thus holds as long as $\mathcal{G}_1$ converges. We set $\lambda = (\log n/N)^{h/(c+3)} = (a \cdot \log N/N)^{h/(c+3)}$. It then holds that

$$n \cdot \lambda^{1+1/b} = N^a \cdot (a \cdot \log N/N)^{h \cdot (1+1/b)/(c+3)} = \frac{(\log N)^{h \cdot (1+1/b)/(c+3)}}{N^{h \cdot (1+1/b)/(c+3)-a}}. \tag{B.47}$$

Therefore, for $a > h \cdot (1 + 1/b)/(c + 3)$, it holds for arbitrary constant C that $n \cdot \lambda^{1+1/b} > C$ with a sufficently large n. Thus the requirement on n in (3.7) holds. Upon computation, it holds that the dominating parts of the right-hand side of (3.8) are $\mathcal{G}_1$ and $C(\eta) \cdot R\lambda^c$. We then obtain that

$$\|\widehat{Q}^\lambda_{\kappa+1} - T\widehat{Q}^\lambda_\kappa\|_\nu \leq (C + C(\eta) \cdot R) \cdot \left(\frac{\log(|\mathcal{A}| \cdot N/\tau)}{N}\right)^{\frac{h \cdot c}{2 \cdot (c+3)}} + \psi_T, \tag{B.48}$$

which holds with probability at least $1 - \eta - \tau$. Following from (B.48) and Theorem 3.2, it holds with probability at least $1 - \eta - \tau$ that

$$\|Q^* - Q^{\pi_\kappa}\|_{1,\mu} \le \frac{2\gamma \cdot \Phi(\mu, \nu)}{(1-\gamma)^2} \cdot \left(C' \cdot \left(\log(|\mathcal{A}| \cdot N/\tau)/N\right)^{\frac{h \cdot c}{2 \cdot (c+3)}} + \mathrm{dist}_T \right) + \frac{4\gamma^{\kappa+1} \cdot Q_{\max}}{1-\gamma}, \tag{B.49}$$

which concludes half of the proof of Theorem 3.5.

Case 2: We consider the case when $a < h \cdot (1 + 1/b)/ \cdot (c+3)$. We fix $\lambda = 1/N^{ab/(bc+1)}$. It then holds that

$$n \cdot \lambda^{1+1/b} = n \cdot (1/n^{\frac{b}{bc+1}})^{\frac{b+1}{b}} = N^{1 - \frac{b+1}{bc+1}}. \tag{B.50}$$

Following from Assumption 3.3, it holds that $b > 1$ and $c \ge 1$, which implies $(b+1)/(bc+1) < 1$. Therefore, following from (B.50), it holds for any C that $n \cdot \lambda^{1+1/b} \ge C$ for a sufficiently large n, and the requirement on n in (3.7) holds. Upon computation, for $a < h \cdot (1 + 1/b)/ \cdot (c+3)$, the dominating parts of the right-hand side of (3.8) are λ^c and $1/(n \cdot \lambda^{1+1/b})$. Therefore, following from Theorem 3.4, it holds with probability at least $1 - \eta - \tau$ that

$$\|\widehat{Q}^\lambda_{\kappa+1} - T\widehat{Q}^\lambda_\kappa\|^2_\nu \le C'/N^{\frac{abc}{bc+1}}. \tag{B.51}$$

Then following from Theorems 3.2 and (B.51), it holds with probability at least $1 - \eta - \tau$ that

$$\|Q^* - Q^{\pi_\kappa}\|_{1,\mu} \le \frac{2\gamma \cdot \Phi(\mu, \nu)}{(1-\gamma)^2} \cdot \left(C'/N^{\frac{abc}{2(bc+1)}} + \mathrm{dist}_T \right) + \frac{4\gamma^{\kappa+1} \cdot Q_{\max}}{1-\gamma}, \tag{B.52}$$

which together with (B.49) concludes the proof of Theorem 3.5. $\qquad\square$

C. Proof of Lemma B.3

Proof. The proof strategy is similar to that of Theorem 1 by (Caponnetto & De Vito, 2007) and Main Theorem by (Szabó et al., 2015). We split the error into the following terms,

$$\|\sqrt{\mathcal{C}}(Q^\lambda - Q_{\mathcal{H},T})\|^2_{\mathcal{H}(K)} \le 2 \underbrace{\|\sqrt{\mathcal{C}}(Q^\lambda - \overline{Q}^\lambda)\|^2_{\mathcal{H}(K)}}_{S_1} + 2 \underbrace{\|\sqrt{\mathcal{C}}(\overline{Q}^\lambda - Q_{\mathcal{H},T})\|^2_{\mathcal{H}(K)}}_{S_2}, \tag{C.1}$$

where Q^λ and $\overline{Q}^\lambda$ are defined in (B.34) and (B.37), respectively. In what follows, we establish upper bounds for S_1 and S_2.

Bounding S_1: Following from Proposition B.1, it holds that

$$
\begin{aligned}
Q^\lambda - \overline{Q}^\lambda &= (\mathcal{C}_{\widehat{\omega}} + \lambda)^{-1} g_{\widehat{\omega}} - (\mathcal{C}_\omega + \lambda)^{-1} g_\omega \\
&= (\mathcal{C}_{\widehat{\omega}} + \lambda)^{-1} (g_{\widehat{\omega}} - g_\omega) + \left((\mathcal{C}_{\widehat{\omega}} + \lambda)^{-1} - (\mathcal{C}_\omega + \lambda)^{-1}\right) g_\omega \\
&= (\mathcal{C}_{\widehat{\omega}} + \lambda)^{-1} (g_{\widehat{\omega}} - g_\omega) + (\mathcal{C}_{\widehat{\omega}} + \lambda)^{-1}(\mathcal{C}_{\widehat{\omega}} - \mathcal{C}_\omega)(\mathcal{C}_\omega + \lambda)^{-1} g_\omega \\
&= (\mathcal{C}_{\widehat{\omega}} + \lambda)^{-1} \left(g_{\widehat{\omega}} - g_\omega + (\mathcal{C}_{\widehat{\omega}} - \mathcal{C}_\omega)Q^\lambda\right).
\end{aligned}
$$

The error term S_1 is then upper bounded as follows,

$$
\begin{aligned}
S_1 &= \|\sqrt{\mathcal{C}}(Q^\lambda - \overline{Q}^\lambda)\|^2_{\mathcal{H}(K)} \\
&\le \|\sqrt{\mathcal{C}}(\mathcal{C}_{\widehat{\omega}} + \lambda)^{-1}\|^2_{\mathcal{H}(K)} \cdot \left(2\|g_{\widehat{\omega}} - g_\omega\|^2_{\mathcal{H}(K)} + 2\|\mathcal{C}_{\widehat{\omega}} - \mathcal{C}_\omega\|^2_{\mathcal{H}(K)} \cdot \|\overline{Q}^\lambda\|^2_{\mathcal{H}(K)}\right),
\end{aligned} \tag{C.2}
$$

where, with a slight abuse of notation, we denote by $\|\cdot\|_{\mathcal{H}(K)}$ the operator norm when applying to operators defined on $\mathcal{H}(K)$. We first establish an upper bound for the term $\|\overline{Q}^\lambda\|^2_{\mathcal{H}(K)}$. Following from Proposition B.1, it holds that

$$
\begin{aligned}
\|\overline{Q}^\lambda\|^2_{\mathcal{H}(K)} &= \|(\mathcal{C}_\omega + \lambda)^{-1} g_\omega\|^2_{\mathcal{H}(K)} = \|\frac{1}{n}\sum_{i=1}^n (\mathcal{C}_\omega + \lambda)^{-1} K(\cdot, \mu_{\widehat{\omega}_i}) y_i\|^2_{\mathcal{H}(K)} \\
&\le \|(\mathcal{C}_\omega + \lambda)^{-1}\|^2_{\mathcal{H}(K)} \cdot \frac{1}{n}\sum_{i=1}^n \|K(\cdot, \mu_{\widehat{\omega}_i})\|^2_{\mathcal{H}(K)} \cdot |y_i|^2 \\
&= \|(\mathcal{C}_\omega + \lambda)^{-1}\|^2_{\mathcal{H}(K)} \cdot \frac{1}{n}\sum_{i=1}^n K(\mu_{\widehat{\omega}_i}, \mu_{\widehat{\omega}_i}) \cdot |y_i|^2,
\end{aligned}
$$

where the last inequality follows from the definition of operator norm and the Cauchy-Schwartz inequality, and the last equality follows from the reproducing property of kernel $K(\cdot, \cdot)$. Note that $\|(\mathcal{C}_\omega + \lambda)^{-1}\|^2_{\mathcal{H}(K)} \leq 1/\lambda$. Therefore, following from Assumption 2.3, it holds that

$$\|\overline{Q}^\lambda\|^2_{\mathcal{H}(K)} \leq (1/\lambda)^2 \cdot \frac{1}{n} \sum_{i=1}^{n} \varsigma Q_{\max}^2 \leq \varsigma Q_{\max}^2 / \lambda^2. \tag{C.3}$$

We now bound the remaining terms in $\mathcal{S}_1$. The following lemmas hold, which establish the upper bounds for terms in the right-hand side of (C.2).

Lemma C.1 (Concentration of Empirical Mean (Altun & Smola, 2006))**.** Let ϵ be a positive absolute constant. Let $\widehat{\omega}$ be the empirical approximation of $\omega \in \widetilde{\mathcal{M}}(\Omega)$, which takes the form

$$\widehat{\omega} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\omega_i},$$

where $\{\omega_i\}_{i=1}^N$ are independent samples of ω. Under Assumption 2.3, it holds with probability at least $1 - \exp(-2\epsilon)$ that

$$\|\mu_{\widehat{\omega}} - \mu_\omega\|_{\mathcal{H}(k)} \leq (1 + \sqrt{\epsilon}) \cdot \sqrt{2\varrho/N}.$$

Proof. See §D.1 for a detailed proof. □

Lemma C.2. Let ϵ be a positive absolute constant. It holds with probability at least $1 - |\mathcal{A}| \cdot n \exp(-2\epsilon)$ that

$$\|g_{\widehat{\omega}} - g_\omega\|^2_{\mathcal{H}(K)} \leq (1 + \varsigma/\lambda^2) \cdot L^2 Q_{\max}^2 \cdot (1 + \sqrt{\epsilon})^{2h} \cdot (2\varrho/N)^h. \tag{C.4}$$

Proof. See §D.2 for a detailed proof. □

Lemma C.3. Let ϵ be a positive absolute constant. Under Assumption 2.3, it holds with probability at least $1 - n \exp(-2\epsilon)$ that

$$\|\mathcal{C}_\omega - \mathcal{C}_{\widehat{\omega}}\|^2_{\mathcal{H}(K)} \leq 4L^2 \varsigma \cdot (1 + \sqrt{\epsilon})^{2h} \cdot (2\varrho/N)^h. \tag{C.5}$$

Proof. See §D.3 for a detailed proof. □

Lemma C.4. Let ϵ and η be positive absolute constants. Under Assumption 2.3, for

$$\lambda \leq \|\mathcal{C}\|_{\mathcal{H}(K)}, \quad N \geq 2\varrho \cdot (1 + \sqrt{\epsilon})^2 \cdot (64L^2\varsigma/\lambda^2)^{1/h}, \tag{C.6}$$

it holds with probability at least $1 - \eta/3 - n \exp(-2\epsilon)$ that

$$\|\sqrt{\mathcal{C}}(\mathcal{C}_{\widehat{\omega}} + \lambda)^{-1}\|_{\mathcal{H}(K)} \leq 2/\sqrt{\lambda}. \tag{C.7}$$

Proof. See §D.4 for a detailed proof. □

Finally, combining Lemma C.2, C.3, C.4, and (C.3), it follows from (C.2) that

$$\begin{aligned}
S_1 &\leq \|\sqrt{\mathcal{C}}(\mathcal{C}_{\widehat{\omega}} + \lambda)^{-1}\|^2_{\mathcal{H}(K)} \cdot \left(2\|g_{\widehat{\omega}} - g_\omega\|^2_{\mathcal{H}(K)} + 2\|\mathcal{C}_{\widehat{\omega}} - \mathcal{C}_\omega\|^2_{\mathcal{H}(K)} \cdot \|\overline{Q}^\lambda\|^2_{\mathcal{H}(K)}\right) \\
&\leq \frac{8L^2 Q_{\max}^2 (1 + \sqrt{\epsilon})^{2h} \cdot (2\varrho)^h}{\lambda \cdot N^h} \cdot \left(1 + \frac{5\varsigma^2}{\lambda^2}\right),
\end{aligned} \tag{C.8}$$

which holds with probability at least $1 - \eta/3 - n \cdot |\mathcal{A}| \cdot \exp(-2\epsilon)$.

Bounding S_2: Note that the error S_2 is the excess risk of a standard kernel ridge regression with n samples. Following from Assumptions 2.3 and 3.3, it holds that

$$|y - Q_{\mathcal{H},T}(\mu_\omega)| \leq |y| + |Q_{\mathcal{H},T}(\mu_\omega)| \leq Q_{\max} + \sqrt{\varsigma} \cdot \max\{1, \|\mathcal{C}\|_{\mathcal{H}(K)}\} \cdot R.$$

Therefore, it holds for some positive absolute constants Σ and M that

$$\int_{\mathbb{R}} \exp\big(|y - Q_{\mathcal{H},T}(\mu_\omega)|/M\big) - \frac{|y - Q_{\mathcal{H},T}(\mu_\omega)|}{M} - 1 \, \mathrm{d}\rho(y|\mu_\omega) \leq \frac{\Sigma^2}{2M^2}, \tag{C.9}$$

which holds for ν-almost surely all $\omega \in \widetilde{\mathcal{M}}(\Omega)$. The following lemma follows from Theorem 1 by (Caponnetto & De Vito, 2007), which characterizes an upper bound of $\mathcal{S}_2$.

Lemma C.5 (Excess Risk of Kernel Ridge Regression (Caponnetto & De Vito, 2007)). Let $0 < \eta < 1$, $\lambda > 0$, and $n \in \mathbb{N}$. Let $C(\eta) = 32 \log^2(6/\eta)$ be a positive absolute constant. For

$$n \geq \frac{2C(\eta)\varsigma\beta b}{(b-1)\lambda^{1+1/b}}, \quad \lambda \leq \|\mathcal{C}\|_{\mathcal{H}(K)},$$

under Assumptions 2.3 and 3.3, it holds with probability at least $1 - \eta$ that

$$S_2 = \|\sqrt{\mathcal{C}}(\overline{Q}^\lambda - Q_{\mathcal{H},T})\|_{\mathcal{H}(K)}^2 \leq C(\eta) \cdot \left(R\lambda^c + \frac{\varsigma^2 R}{\lambda^{2-c}n^2} + \frac{\varsigma R\lambda^{c-1}}{4n} + \frac{\varsigma M^2}{\lambda n^2} + \frac{\Sigma^2 \beta b}{(b-1)n\lambda^{1/b}} \right). \tag{C.10}$$

Proof. See (Caponnetto & De Vito, 2007) for a detailed proof. $\square$

We fix the constant $2\epsilon = \log(|\mathcal{A}| \cdot n/\tau)$ in (C.8). The probability that the error bound in (C.8) holds then becomes at least $1 - \eta/3 - \tau$. By combining (C.8) and Lemma C.5, it follows from (C.1) that

$$\|\sqrt{\mathcal{C}}(Q^\lambda - Q_{\mathcal{H},T})\|_{\mathcal{H}(K)}^2 \leq \frac{8L^2 Q_{\max}^2 \big(1 + \sqrt{\log(|\mathcal{A}| \cdot n/\tau)}\big)^{2h} \cdot (2\varrho)^h}{\lambda \cdot N^h} \cdot \left(1 + \frac{5\varsigma^2}{\lambda^2}\right) \tag{C.11}$$
$$+ C(\eta) \cdot \left(R\lambda^c + \frac{\varsigma^2 R}{\lambda^{2-c}n^2} + \frac{\varsigma R\lambda^{c-1}}{4n} + \frac{\varsigma M^2}{\lambda n^2} + \frac{\Sigma^2 \beta b}{(b-1)n\lambda^{1/b}} \right),$$

which holds with probability at least $1 - \eta - \tau$ and thus concludes the proof of Lemma B.3. $\square$

D. Proof of Supporting Lemmas

D.1. Proof of Lemma C.1

Proof. Lemma C.1 is a corollary of Theorem 15 by (Altun & Smola, 2006). We first show that μ_ω is bounded for any $\omega \in \widetilde{\mathcal{M}}(\Omega)$. It holds from the reproducing property of the kernel $k(\cdot, \cdot)$ that

$$|k(x, x')| = |\langle k(\cdot, x), k(\cdot, x')\rangle_{\mathcal{H}(k)}|$$
$$\leq \big(\langle k(\cdot, x), k(\cdot, x)\rangle_{\mathcal{H}(k)} \cdot \langle k(\cdot, x'), k(\cdot, x')\rangle_{\mathcal{H}(k)}\big)^{1/2} = \sqrt{k(x, x) \cdot k(x', x')}. \tag{D.1}$$

Therefore, upon taking the expectation of (D.1), it follows from Assumption 2.3 that

$$\|\mu_\omega\|_{\mathcal{H}(k)}^2 = \mathbb{E}_{x \sim \omega, x' \sim \omega}\big[k(x, x')\big] \leq \mathbb{E}_{x \sim \omega, x' \sim \omega}\big[\sqrt{k(x, x) \cdot k(x', x')}\big] \leq \varrho, \tag{D.2}$$

where the last inequality follows from Assumption 2.3. It then suffices to compute for $\omega \in \widetilde{\mathcal{M}}(\Omega)$ the following risk,

$$R_N = N^{-1/2}\Big(\mathbb{E}_{x \sim \omega, x' \sim \omega}\big[k(x, x) - k(x, x')\big]\Big)^{1/2}.$$

Combining (D.2) and Assumption 2.3, it holds that

$$R_N \leq N^{-1/2}\Big(\mathbb{E}_{x \sim \omega, x' \sim \omega}\big[|k(x, x)| + |k(x, x')|\big]\Big)^{1/2} \leq N^{-1/2}\Big(\mathbb{E}_{x \sim \omega, x' \sim \omega}\big[|k(x, x)| + |k(x, x')|\big]\Big)^{1/2}$$
$$\leq N^{-1/2}\Big(\mathbb{E}_{x \sim \omega}\big[|k(x, x)|\big] + \mathbb{E}_{x \sim \omega, x' \sim \omega}\big[\sqrt{k(x, x) \cdot k(x', x')}\big]\Big)^{1/2} \leq \sqrt{2\varrho/N}. \tag{D.3}$$

Following from (D.3) and Theorem 15 by (Altun & Smola, 2006), it holds for positive absolute constant η_1 that

$$\|\mu_{\widehat{\omega}} - \mu_\omega\|_{\mathcal{H}(k)} \leq R_N + \eta_1 \leq \sqrt{2\varrho/N} + \eta_1,$$

which holds with probability $1 - \exp(-\eta_1^2 N/\varrho)$. Thus, we complete the proof of Lemma C.1 by setting $\eta_1 = \sqrt{2\epsilon\varrho/N}$. $\square$

D.2. Proof of Lemma C.2

Proof. Following from the definition of g_ω and $g_{\widehat\omega}$ in (B.40), it holds that

$$\|g_{\widehat\omega} - g_\omega\|^2_{\mathcal{H}(K)} = \|\frac{1}{n}\sum_{i=1}^n \big(K(\cdot,\mu_{\widehat\omega_i})\cdot \widehat{y}_i - K(\cdot,\mu_{\omega_i})\cdot y_i\big)\|^2_{\mathcal{H}(K)}$$

$$\leq \frac{1}{n}\sum_{i=1}^n \|K(\cdot,\mu_{\widehat\omega_i}) - K(\cdot,\mu_{\omega_i})\|^2_{\mathcal{H}(K)}\cdot |y_i|^2 + \frac{1}{n}\sum_{i=1}^n \|K(\cdot,\mu_{\widehat\omega_i})\|^2_{\mathcal{H}(K)}\cdot |y_i - \widehat{y}_i|^2. \tag{D.4}$$

Combining (D.4) and Assumption 2.3, we obtain that

$$\frac{1}{n}\sum_{i=1}^n \|K(\cdot,\mu_{\widehat\omega_i}) - K(\cdot,\mu_{\omega_i})\|^2_{\mathcal{H}(K)}\cdot |y_i|^2 \leq \frac{L^2 Q^2_{\max}}{n}\sum_{i=1}^n \|\mu_{\widehat\omega_i} - \mu_{\omega_i}\|^{2h}_{\mathcal{H}(k)}. \tag{D.5}$$

Meanwhile, following from the definition of $\widehat{y}_i$ and y_i in (B.35) and (B.38), it holds that

$$|y_i - \widehat{y}_i|^2 \leq |\sup_{a\in\mathcal{A}} \widehat{Q}(\delta_a \times \widehat{p}'_{i,s}) - \sup_{a\in\mathcal{A}} \widehat{Q}(\delta_a \times p'_{i,s})|^2$$

$$\leq \sup_{a\in\mathcal{A}}\big|\min\{Q(\delta_a \times \widehat{p}'_{i,s}), Q_{\max}\} - \min\{Q(\delta_a \times p'_{i,s}), Q_{\max}\}\big|^2$$

$$\leq \sup_{a\in\mathcal{A}} |Q(\delta_a \times \widehat{p}'_{i,s}) - Q(\delta_a \times p'_{i,s})|^2, \tag{D.6}$$

where recall that we denote by $\widehat{Q} = \{Q, Q_{\max}\}$ the previous update of action-value function and $Q \in \mathcal{H}(K)$ the output of the associated regression, and the last inequality follows from the fact that the function $\min\{\cdot, Q_{\max}\}$ is a contraction mapping. Note that for any $a\in\mathcal{A}$ and $\widehat{p}'_{i,s}, p'_{i,s} \in \mathcal{M}(\mathcal{S})$, it holds that

$$|Q(\delta_a \times \widehat{p}'_{i,s}) - Q(\delta_a \times p'_{i,s})|^2 = \langle K(\cdot,\mu_{\omega'_i}) - K(\cdot,\mu_{\widehat\omega'_i}), Q\rangle^2_{\mathcal{H}(K)}$$

$$\leq \|Q\|^2_{\mathcal{H}(K)}\cdot \|K(\cdot,\mu_{\omega'_i}) - K(\cdot,\mu_{\widehat\omega'_i})\|^2_{\mathcal{H}(K)}, \tag{D.7}$$

where we denote by $\omega'_i = \delta_a \times p'_{i,s}$ and $\widehat\omega'_i = \delta_a \times \widehat{p}'_{i,s}$. Following from (C.3) and Assumption 2.3, it then holds that

$$|Q(\delta_a \times \widehat{p}'_{i,s}) - Q(\delta_a \times p'_{i,s})|^2 \leq \varsigma L^2 Q^2_{\max}/\lambda^2 \cdot \|\mu_{\omega'_i} - \mu_{\widehat\omega'_i}\|^{2h}_{\mathcal{H}(k)}. \tag{D.8}$$

Combining (D.6), (D.7), (D.8), and Assumption 2.3, it then holds that

$$\frac{1}{n}\sum_{i=1}^n \|K(\cdot,\mu_{\widehat\omega_i})\|^2_{\mathcal{H}(K)}\cdot |y_i - \widehat{y}_i|^2 \leq \frac{B_K^2 L^2 Q^2_{\max}}{\lambda^2 \cdot n}\sum_{i=1}^n \max_{a\in\mathcal{A}} \|\mu_{\omega'_i} - \mu_{\widehat\omega'_i}\|^{2h}_{\mathcal{H}(k)}. \tag{D.9}$$

By a union bound argument together with Lemma C.1, it follows from (D.5) and (D.9) that

$$\|g_{\widehat\omega} - g_\omega\|^2_{\mathcal{H}(K)} \leq (1+\varsigma/\lambda^2)\cdot L^2 Q^2_{\max}/n \cdot \sum_{i=1}^n (1+\sqrt{\epsilon})^{2h}\cdot (2\varrho/N)^h$$

$$\leq (1+\varsigma/\lambda^2)\cdot L^2 Q^2_{\max}\cdot (1+\sqrt{\epsilon})^{2h}\cdot (2\varrho/N)^h, \tag{D.10}$$

which holds with probability at least $1 - n\cdot|\mathcal{A}|\cdot \exp(-2\epsilon)$. Thus, we complete the proof of Lemma C.2. $\qquad\square$

D.3. Proof of Lemma C.3

Proof. We fix an arbitrary $f \in \mathcal{H}(K)$. Following from the definition of $\mathcal{C}_\omega$ and $\mathcal{C}_{\widehat\omega}$ in (B.39), it holds that

$$\|\mathcal{C}_\omega f - \mathcal{C}_{\widehat\omega} f\|^2_{\mathcal{H}(K)} = \|\frac{1}{n}\sum_{i=1}^n K(\cdot,\mu_{\omega_i})f(\mu_{\omega_i}) - K(\cdot,\mu_{\widehat\omega_i})f(\mu_{\widehat\omega_i})\|^2_{\mathcal{H}(K)}$$

$$\leq \frac{2}{n}\sum_{i=1}^n \|\big(K(\cdot,\mu_{\omega_i}) - K(\cdot,\mu_{\widehat\omega_i})\big)f(\mu_{\omega_i})\|^2_{\mathcal{H}(K)}$$

$$+ \frac{2}{n}\sum_{i=1}^n \|K(\cdot,\mu_{\widehat\omega_i})\big(f(\mu_{\omega_i}) - f(\mu_{\widehat\omega_i})\big)\|^2_{\mathcal{H}(K)}. \tag{D.11}$$

In what follows, we establish upper bounds for each terms in the right-hand side of (D.11). Following from the reproducing property of the kernel $K(\cdot, \cdot)$, it holds that

$$|f(\mu_\omega)|^2 = |\langle f(\cdot), K(\cdot, \mu_\omega)\rangle_{\mathcal{H}(K)}|^2 \leq \|f\|^2_{\mathcal{H}(K)} \cdot \|K(\cdot, \mu_\omega)\|^2_{\mathcal{H}(K)} = \|f\|^2_{\mathcal{H}(K)} \cdot K(\mu_\omega, \mu_\omega). \tag{D.12}$$

Combining (D.12) and Assumption 2.3, we obtain that

$$\|\big(K(\cdot, \mu_{\omega_i}) - K(\cdot, \mu_{\widehat{\omega}_i})\big)f(\mu_{\omega_i})\|^2_{\mathcal{H}(K)} \leq \|K(\cdot, \mu_{\omega_i}) - K(\cdot, \mu_{\widehat{\omega}_i})\|^2_{\mathcal{H}(K)} \cdot \varsigma \cdot \|f\|^2_{\mathcal{H}(K)}$$
$$\leq L^2\varsigma \cdot \|\mu_{\omega_i} - \mu_{\widehat{\omega}_i}\|^{2h}_{\mathcal{H}(k)} \cdot \|f\|^2_{\mathcal{H}(K)}. \tag{D.13}$$

Similarly, following from the reproducing property of the kernel $K(\cdot, \cdot)$, it holds that

$$|f(\mu_\omega) - f(\mu_{\widehat{\omega}})|^2 = |\langle f(\cdot), K(\cdot, \mu_\omega) - K(\cdot, \mu_{\widehat{\omega}})\rangle_{\mathcal{H}(K)}|^2 \leq \|f\|^2_{\mathcal{H}(K)} \cdot \|K(\cdot, \mu_\omega) - K(\cdot, \mu_{\widehat{\omega}})\|^2_{\mathcal{H}(K)}. \tag{D.14}$$

Combining (D.14) and Assumption 2.3, we obtain that

$$\|K(\cdot, \mu_{\widehat{\omega}_i})\big(f(\mu_{\omega_i}) - f(\mu_{\widehat{\omega}_i})\big)\|^2_{\mathcal{H}(K)} \leq \|K(\cdot, \mu_{\widehat{\omega}_i})\|^2_{\mathcal{H}(K)} \cdot \|f\|^2_{\mathcal{H}(K)} \cdot \|K(\cdot, \mu_{\omega_i}) - K(\cdot, \mu_{\widehat{\omega}_i})\|^2_{\mathcal{H}(K)}$$
$$\leq L^2\varsigma \cdot \|\mu_{\omega_i} - \mu_{\widehat{\omega}_i}\|^{2h}_{\mathcal{H}(k)} \cdot \|f\|^2_{\mathcal{H}(K)}. \tag{D.15}$$

Finally, combining (D.11), (D.13), and (D.15), we obtain that

$$\|\mathcal{C}_\omega f - \mathcal{C}_{\widehat{\omega}} f\|^2_{\mathcal{H}(K)} \leq \frac{4L^2\varsigma}{n} \cdot \sum_{i=1}^n \|\mu_{\omega_i} - \mu_{\widehat{\omega}_i}\|^{2h}_{\mathcal{H}(k)} \cdot \|f\|^2_{\mathcal{H}(K)},$$

which is equivalent with the followings,

$$\|\mathcal{C}_\omega - \mathcal{C}_{\widehat{\omega}}\|^2_{\mathcal{H}(K)} \leq \frac{4L^2\varsigma}{n} \cdot \sum_{i=1}^n \|\mu_{\omega_i} - \mu_{\widehat{\omega}_i}\|^{2h}_{\mathcal{H}(k)}. \tag{D.16}$$

Following from (D.16) and Lemma C.1 together with a union bound argument, it then holds for any positive absolute constant ϵ that

$$\|\mathcal{C}_\omega - \mathcal{C}_{\widehat{\omega}}\|^2_{\mathcal{H}(K)} \leq 4L^2\varsigma \cdot (1 + \sqrt{\epsilon})^{2h} \cdot (2\varrho/N)^h, \tag{D.17}$$

which holds with probability at least $1 - n\exp(-2\epsilon)$. Thus, we complete the proof of Lemma C.3. $\qquad \square$

D.4. Proof of Lemma C.4

Proof. It holds that

$$\sqrt{\mathcal{C}}(\mathcal{C}_{\widehat{\omega}} + \lambda)^{-1} = \sqrt{\mathcal{C}}(\mathcal{C} + \lambda)^{-1}(I - (\mathcal{C} - \mathcal{C}_{\widehat{\omega}})(\mathcal{C} + \lambda)^{-1})^{-1}.$$

Therefore, following from the Neumann sequence, it holds that

$$\|\sqrt{\mathcal{C}}(\mathcal{C}_{\widehat{\omega}} + \lambda)^{-1}\|_{\mathcal{H}(K)} \leq \|\sqrt{\mathcal{C}}(\mathcal{C} + \lambda)^{-1}\|_{\mathcal{H}(K)} \sum_{n=0}^\infty \|(\mathcal{C} - \mathcal{C}_{\widehat{\omega}})(\mathcal{C} + \lambda)^{-1}\|^n_{\mathcal{H}(K)}.$$

Following form (C.6), it holds that $\|\sqrt{\mathcal{C}}(\mathcal{C} + \lambda)^{-1}\|_{\mathcal{H}(K)} \leq 1/(2\sqrt{\lambda})$. It then suffices to upper bound the norm $\|(\mathcal{C} - \mathcal{C}_{\widehat{\omega}})(\mathcal{C} + \lambda)^{-1}\|_{\mathcal{H}(K)}$. Note that

$$\|(\mathcal{C} - \mathcal{C}_{\widehat{\omega}})(\mathcal{C} + \lambda)^{-1}\|_{\mathcal{H}(K)} \leq \|(\mathcal{C} - \mathcal{C}_\omega)(\mathcal{C} + \lambda)^{-1}\|_{\mathcal{H}(K)} + \|(\mathcal{C}_\omega - \mathcal{C}_{\widehat{\omega}})(\mathcal{C} + \lambda)^{-1}\|_{\mathcal{H}(K)}. \tag{D.18}$$

Following from (Caponnetto & De Vito, 2007), it holds with probability at least $1 - \eta/3$ that

$$\|(\mathcal{C} - \mathcal{C}_\omega)(\mathcal{C} + \lambda)^{-1}\|_{\mathcal{H}(K)} \leq 1/2. \tag{D.19}$$

Meanwhile, following from Lemma C.3, it holds with probability at least $1 - n\exp(-2\epsilon)$ that

$$\|\mathcal{C}_\omega - \mathcal{C}_{\widehat{\omega}}\|^2_{\mathcal{H}(K)} \leq 4L^2\varsigma \cdot (1 + \sqrt{\epsilon})^{2h} \cdot (2\varrho/N)^h. \tag{D.20}$$

Note that we set N to be such that

$$N \geq 2\varrho \cdot (1 + \sqrt{\epsilon})^2 \cdot (64L^2\varsigma/\lambda^2)^{1/h}. \tag{D.21}$$

Meanwhile, it holds that $\|(\mathcal{C} + \lambda)^{-1}\|_{\mathcal{H}(K)} \leq 1/\lambda$. Therefore, following from (D.20) and (D.21), it holds with probability at least $1 - n\exp(-2\epsilon)$ that

$$\|(\mathcal{C}_\omega - \mathcal{C}_{\widehat{\omega}})(\mathcal{C} + \lambda)^{-1}\|_{\mathcal{H}(K)} \leq \|\mathcal{C}_\omega - \mathcal{C}_{\widehat{\omega}}\|_{\mathcal{H}(K)} \cdot \|(\mathcal{C} + \lambda)^{-1}\|_{\mathcal{H}(K)} \leq \lambda/4 \cdot 1/\lambda = 1/4, \tag{D.22}$$

Finally, combining (D.18), (D.19), and (D.22), it holds with probability at least $1 - \eta/3 - n\exp(-2\epsilon)$ that

$$\|(\mathcal{C} - \mathcal{C}_{\widehat{\omega}})(\mathcal{C} + \lambda)^{-1}\|_{\mathcal{H}(K)} \leq \frac{1}{2\sqrt{\lambda}} \cdot \sum_{n=0}^{\infty}(3/4)^n = 2/\sqrt{\lambda}, \tag{D.23}$$

which concludes the proof of Lemma C.4. $\qquad\square$