

Understanding **Value Decomposition Algorithms** in Deep Cooperative Multi-Agent Reinforcement Learning

Zehao Dou

Department of Statistics and Data Science, Yale University

ZEHAO.DOU@YALE.EDU

Jakub Grudzien Kuba

Department of Statistics, University of Oxford

JAKUB.GRUDZIEN@NEW.OX.AC.UK

Yaodong Yang

Institute for AI, Peking University & BIGAI

YAODONG.YANG@PKU.EDU.CN

Abstract

Value function decomposition is becoming a popular rule of thumb for scaling up multi-agent reinforcement learning (MARL) in cooperative games. For such a decomposition rule to hold, the assumption of the *individual-global max* (IGM) principle must be made; that is, the local maxima on the decomposed value function per every agent must amount to the global maximum on the joint value function. This principle, however, does not have to hold in general. As a result, the applicability of value decomposition algorithms is concealed and their corresponding convergence properties remain unknown. In this paper, we make the first effort to answer these questions. Specifically, we introduce the set of cooperative games in which the value decomposition methods find their validity, which is referred as *decomposable games*. In decomposable games, we theoretically prove that applying the multi-agent fitted Q-Iteration algorithm (MA-FQI) will lead to an optimal Q-function. In non-decomposable games, the estimated Q-function by MA-FQI can still converge to the optimum under the circumstance that the Q-function needs projecting into the decomposable function space at each iteration. In both settings, we consider value function representations by practical deep neural networks and derive their corresponding convergence rates. To summarize, our results, for the first time, offer theoretical insights for MARL practitioners in terms of when value decomposition algorithms converge and why they perform well.

Keywords: Deep Multi-Agent Reinforcement Learning, Value Decomposition Methods, Cooperative Games, Deep Q-Networks, Reinforcement Learning Theory

1. Introduction

Q-learning is one of the most classical approach of solving Markov decision processes in single-agent reinforcement learning (RL) (Sutton and Barto, 2018). At every iteration, a learning agent fits the state-action value critic, and then acts to maximize it. This method, combined with the expressive power of deep neural networks, enabled RL agents to learn to solve complex decision-making problems (Mnih et al., 2015; Silver et al., 2016). Although this hybrid approach serves as the template for designing deep RL methods, the difficulty of analyzing deep neural network models makes the understanding of it still lacking. However, recently, the first steps towards demystifying it were made by Fan et al. (2020), who derive the convergence rate of *Deep Q-Network* (DQN) (Mnih et al., 2015). Their analysis uncovered that one of the keys behind the success of DQN is over-parameterization of the critic network, thus bridging the Q-learning framework and the deep-learning components of the algorithm.

In multi-agent reinforcement learning (MARL) (Yang and Wang, 2020), applying effective Q-learning based method is no longer straightforward. If agents act independently, greedy policies with respect to their local state-action value functions do not necessarily maximize the global value function. This impedes agents from performing the policy improvement step of Q-learning. To tackle this issue, the framework of *Centralized Training with Decentralized Execution (CTDE)* was introduced (Foerster et al., 2018; Wen et al., 2018, 2020). In the CTDE framework, the agents have access to the global critic during training, which enables the improvement of the joint policy. Afterwards, upon execution (once the training has ended), the learned joint critic is no longer accessible. Therefore, naively relying on the joint critic, not having learned adequate decentralized ones, the agents are put back at the starting point, let alone the large variance issue (Kuba et al., 2021b,a).

A possible solution to the above issue is through enforcing the individual-global max (IGM) principle (Suneahg et al., 2017; Rashid et al., 2018; Yang et al., 2020) within the CTDE framework. IGM states that the global maximizer of the joint state-action value function is the concatenation of the maximizers of the agents’ individual value components. The agents learn their local value functions with Q-learning by combining them monotonically to form the joint value function estimate. As we show in this paper, although this approach can work well in practice (Mahajan et al., 2019), value function decompositions derived from the IGM principle do not hold in general. The lack of their generality increases the difficulty of their analysis and may have impeded us from demystifying the keys behind their empirical success, as well as the methods’ limitations.

This work takes the first step towards understanding the state-of-the-art value-based algorithms. Its purpose is to describe the settings in which these algorithms can be employed, and settle their properties in these settings. With this aim, we first derive the set of cooperative games in which the value decomposition methods find their validity, which is referred as *decomposable games*. Within the decomposable games, we then prove that applying the multi-agent fitted Q-Iteration algorithm (MA-FQI) can in fact lead to the optimal Q-function. This result offers theoretical insights for MARL practitioners in terms of when value decomposition algorithms converge and why they perform well. The second part of our contribution lies in the non-decomposable games, wherein we show that the estimated Q-function by MA-FQI can still converge to the optimum, despite the fact that the estimated Q-function needs projecting into the decomposable function space at each iteration. In both decomposable and non-decomposable games, we consider value function representations by over-parameterized deep neural networks and derive the corresponding convergence rates for MA-FQI. Our work fills the research gap by providing theoretical insights, in terms of convergence guarantee, for the popular value decomposition algorithms in cooperative MARL.

2. Preliminaries & Background

In this section, we provide the background for MARL by introducing the fundamental definitions, and surveying the most important solution approaches. In Subsection 2.1, we introduce the basic nomenclature for Markov Games, and in Subsection 2.2, we review value decomposition algorithms, such as VDN and QMIX. In Subsection 2.3, we review the multi-agent fitted Q-iteration (MA-FQI) framework, which is the multi-agent version of the widely known FQI method.

2.1. Multi-Agent Markov Games

We start by defining the *cooperative multi-agent Markov games* (MAMG) (Littman, 1994). Formally, we consider the tabular episodic framework of the form $\mathcal{MG}(N, \mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma, \pi_0)$, where

N is the number of agents (a.k.a players). $\mathcal{S}^{(i)}$ is the state space of agent $i \in [N]$; without loss of generality, $\mathcal{S} = [0, 1]^d$ for $d \in \mathbb{N}$. We write $\mathcal{S} \triangleq \mathcal{S}^{(1)} \times \dots \times \mathcal{S}^{(N)}$ to denote the joint state space. $\mathcal{A}^{(i)}$ is the action space of player i and $\mathcal{A} \triangleq \mathcal{A}^{(1)} \times \dots \times \mathcal{A}^{(N)}$ denotes the joint action space. \mathbb{P} is the transition probability function, so that $\mathbb{P}(\cdot | \mathbf{s}, \mathbf{a})$ gives the distribution over states if the joint action $\mathbf{a} = (a^1, \dots, a^N)$ is taken at the (joint) state $\mathbf{s} = (s^1, \dots, s^N)$. $R : \mathcal{S} \times \mathcal{A} \rightarrow [-R_{\max}, R_{\max}]$ is the reward function. $\gamma \in [0, 1)$ is the discount factor. π_0 is the initial state distribution. In each episode of MAMG, an initial state \mathbf{s}_0 is drawn from the distribution π_0 . Then, at every time step $t \in \mathbb{N}$, each player $i \in [N]$ observes the local state $s_t^i \in \mathcal{S}^{(i)}$, and takes an action $a_t^i \in \mathcal{A}^{(i)}$ according to its policy $\pi^i(\cdot | s_t^i)$, simultaneously with others. Equivalently, the agents take the joint action $\mathbf{a}_t \in \mathcal{A}$ at state $\mathbf{s}_t \in \mathcal{S}$, according to their joint policy $\pi(\cdot | \mathbf{s}_t) \triangleq \prod_{i \in [N]} \pi^i(\cdot | s_t^i)$. After that, the players receive the joint reward $R(\mathbf{s}_t, \mathbf{a}_t)$ and transit to the next state $\mathbf{s}_{t+1} \sim \mathbb{P}(\cdot | \mathbf{s}_t, \mathbf{a}_t)$. We define the maximization objective of the collaborative agents, which is known as the joint return:

$$J(\pi) \triangleq \mathbb{E}_{\mathbf{s}_0 \sim \pi_0, \mathbf{a}_{0:\infty} \sim \pi, \mathbf{s}_{1:\infty} \sim \mathbb{P}} \left[\sum_{t=0}^{\infty} \gamma^t R(\mathbf{s}_t, \mathbf{a}_t) \right]. \quad (1)$$

Crucially, as a proxy to the joint return, the agents guide their behavior with the joint state-action value function $Q^\pi : \mathcal{S} \times \mathcal{A} \rightarrow [-\frac{R_{\max}}{1-\gamma}, \frac{R_{\max}}{1-\gamma}] \triangleq [-Q_{\max}, Q_{\max}]$, defined as

$$Q^\pi(\mathbf{s}, \mathbf{a}) \triangleq \mathbb{E}_{\mathbf{s}_{1:\infty} \sim \mathbb{P}, \mathbf{a}_{1:\infty} \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t R(\mathbf{s}_t, \mathbf{a}_t) \mid \mathbf{s}_0 = \mathbf{s}, \mathbf{a}_0 = \mathbf{a} \right], \quad (2)$$

on top of which one can define the state value function $V^\pi(\mathbf{s}) \triangleq \mathbb{E}_{\mathbf{a} \sim \pi} [Q^\pi(\mathbf{s}, \mathbf{a})]$. In this paper, **we are interested in the Q-learning type of approach to policy training**. Ideally, at every iteration $k \in \mathbb{N}$, the agents would make the joint policy update $\pi_{k+1}(\arg \max_{\mathbf{a} \in \mathcal{A}} Q^{\pi_k}(\mathbf{s}, \mathbf{a}) | \mathbf{s}) = 1$, i.e., act greedily with respect to Q^{π_k} . Then, the sequence of state-action value functions $\{Q^{\pi_k}\}_{k \in \mathbb{N}}$ would converge to the unique optimal joint state-action value function Q^* . The greedy joint policy, $\pi^*(\arg \max_{\mathbf{a} \in \mathcal{A}} Q^*(\mathbf{s}, \mathbf{a}) | \mathbf{s}) = 1$, is the optimal joint policy, and maximizes the joint return $J(\pi^*)$ (Sutton and Barto, 2018). Unfortunately, in MARL, the agents learn distributed (independent) policies. **Therefore, even though the CTDE framework allows for implementation of the greedy joint policy during training, it does not scale to the execution phase**. To circumvent this, a novel family of value decomposition methods has emerged, which we describe in the next subsection.

2.2. Value Decomposition Algorithms

We start by introducing the pivotal IGM condition that value decomposition algorithms rely on; it enables global maximization in a decentralized manner.

Definition 2.1 (IGM (Individual Global Max) Condition) *For a joint action-value function $Q_{\text{tot}} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, if there exist N individual Q -functions $\{Q_i : \mathcal{S}^{(i)} \times \mathcal{A}^{(i)} \rightarrow \mathbb{R}\}$, such that:*

$$\arg \max_{\mathbf{a} \in \mathcal{A}} Q_{\text{tot}}(\mathbf{s}, \mathbf{a}) = \left(\arg \max_{a^1 \in \mathcal{A}^{(1)}} Q_1(s^1, a^1), \dots, \arg \max_{a^N \in \mathcal{A}^{(N)}} Q_N(s^N, a^N) \right) \quad (3)$$

then Q_{tot} satisfies the IGM condition with $\{Q_i\}_{i \in [N]}$ decomposition. If the IGM condition is met for all valid value functions Q_{tot} , then the MAMG is said to be decomposable.

As its name suggests, the condition states that individual optimal actions of the agents will constitute the optimal joint action. Without the IGM condition, we have to list all the $\sum_{i=1}^N |\mathcal{A}^{(i)}|$ possible joint actions in order to obtain the maximal Q_{tot} value. However, if the IGM condition holds, we only need to find the optimal action corresponding to the value function Q_i for each $i \in [N]$, which only requires $\sum_{i=1}^N |\mathcal{A}^{(i)}|$ computational steps. Most crucially, this condition enables the agents to learn decentralized value functions which, once trained, can successfully be used in the execution phase. These potential benefits brought by “decomposable” games invoke three theoretical questions: **1)** How to decide whether a game is decomposable? **2)** How to find jointly optimal decentralized policies for decomposable games? **3)** How efficient the solutions to decomposable games are?

Although MARL researchers have not been indifferent about decomposability, they have only studied the problem via the second of the above questions. The first question would be skipped by an implicit assumption on the game’s decomposability. Then, to tackle the second question, a solution to the game would be proposed, and its performance would be verified empirically (Suneag et al., 2017; Rashid et al., 2018; Son et al., 2019). The last point remained ignored, leaving us without an idea of an explanation of the empirical efficacy of value decomposition algorithms. Nevertheless, the discovery of these methods is becoming a big step towards taming decomposable MARL problems. Below, we briefly introduce the first algorithm of this kind—VDN.

Value-Decomposition Network (Suneag et al., 2017, VDN) is a method which assumes that the global state-action value function satisfies the additive decomposition: for any $s \in \mathcal{S}$, $a \in \mathcal{A}$,

$$Q_{\text{tot}}(s, a) = \sum_{i=1}^N Q_i(s^i, a^i). \quad (4)$$

The above structure implies that as, for any agent i , the value $Q_i(s^i, a^i)$ increases, so does $Q_{\text{tot}}(s, a)$. Hence, the IGM principle holds for any state-(joint)action pair, meaning that the game is decomposable. With this decomposition, VDN trains the decentralized critics by extending the Deep Q-Network (DQN) algorithm (Mnih et al., 2015). The greedy action selection with respect to Q_{tot} step is performed by all agents i acting greedily with respect to their local critics Q_i . Next, the critics are trained with TD-learning (Sutton and Barto, 2018) with target networks, *i.e.*, by minimizing

$$\begin{aligned} & \mathbb{E}_{s, a, s' \sim \mathcal{B}} \left[\left(Q_{\text{tot}}(s, a) - R(s, a) - \gamma \max_{a'} Q_{\text{tot}}^{\text{tar}}(s', a') \right)^2 \right] \\ &= \mathbb{E}_{s, a, s' \sim \mathcal{B}} \left[\left(\sum_{i=1}^N Q_i(s^i, a^i) - R(s, a) - \gamma \sum_{i=1}^N \max_{a^i} Q_i^{\text{tar}}(s'^i, a'^i) \right)^2 \right], \end{aligned} \quad (5)$$

where \mathcal{B} is the replay buffer. Intuitively, given empirical results from Mnih et al. (2015); Suneag et al. (2017) and building upon the analysis of Fan et al. (2020), we should expect convergence guarantees of this algorithm, as long as the decomposition from Equation (4) is valid. In this paper, we affirm this intuition theoretically, and provide the key factors of the algorithm’s efficacy.

One of the most popular extension of VDN is the QMIX algorithm (Rashid et al., 2018). The key novelty of the method is its general, IGM-compliant, value function decomposition,

$$Q_{\text{tot}}(s, a) = \Phi(s)(Q_1(s^1, a^1), \dots, Q_N(s^N, a^N)). \quad (6)$$

Here, for every $s \in \mathcal{S}$, the function $\Phi(s) : \mathbb{R}^N \rightarrow \mathbb{R}$ is the trainable *mixing network*, whose weights are computed for every state by the network $\Phi(\cdot)$. Crucially, it satisfies the monotonicity assumption $\frac{\partial \Phi(s)(Q_1, \dots, Q_N)}{\partial Q_i} \geq 0$, which implies that the value of $Q_{\text{tot}}(s, a)$ increases monotonically with

$Q_i(s^i, a^i)$. To guarantee this condition, the architecture of the network $\Phi(\cdot)$ is constructed so that, for every state s , the weights of $\Phi(s)$ are non-negative. VDN is a special case of QMIX, with the mixing network taking form $\Phi^{\text{VDN}}(s)(Q_1, \dots, Q_N) = \sum_{i=1}^N Q_i$, for every state s . Monotonicity of QMIX, again, implies the IGM principle and decomposability of the game. Hence, the agents can learn their critics by acting greedily with respect to them, and repetitively minimizing the loss from Equation (5), substituting Equation (6) into Q_{tot} . As verified empirically, this method achieves substantially superior performance to that of VDN. However, there exist simple problems where QMIX fails utterly (Mahajan et al., 2019). This warns us that the deployment of value decomposition algorithms, even those as powerful as QMIX, requires care and understanding.

2.3. Multi-Agent Fitted Q-Iteration (MA-FQI) Framework

Before we demystify the properties of the value decomposition algorithms, we specify the framework which generalizes all of them. Concretely, the core of these algorithms is the minimization of the (empirical) loss from Equation (5) within a function class \mathcal{F} . In practice, \mathcal{F} is a family of neural networks with a specific architecture. The data (s, \mathbf{a}) on which the minimization takes place is drawn from a large replay buffer \mathcal{B} . As argued by Fan et al. (2020), in the case of large state spaces and buffer sizes, independent draws of (s, \mathbf{a}) from \mathcal{B} constitute a marginal distribution $\sigma \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ which is fixed throughout training. These two steps of an empirical approximation and minimization of the squared TD-error is summarized by Algorithm 1—MA-FQI.

Algorithm 1 Multi-Agent Fitted Q-Iteration Algorithm (MA-FQI)

Input: MAMG $\mathcal{MG}(N, \mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma, \pi_0)$, number of iterations K , function classes $\{\mathcal{F}_k\}_{k \in [K]}$, state-action sampling distribution σ , sample size n , initial Q-function estimate \tilde{Q}_0 .

- 1: **for** episode $k = 0, 1, 2, \dots, K - 1$ **do**
- 2: Sample i.i.d observations $\{(s_j, \mathbf{a}_j, R_j, s'_j)\}_{j \in [n]}$ with (s_j, \mathbf{a}_j) drawn from distribution σ .
- 3: Compute targets $Y_j = R_j + \gamma \tilde{Q}_k(s'_j, a_*^1, \dots, a_*^N)$, where $\forall i \in [N]$,

$$a_*^i = \arg \max_{a^i \in \mathcal{A}^{(i)}} \tilde{Q}_k^i(s_j^i, a^i) \quad \backslash \backslash \text{IGM condition}$$
- 4: Update the joint action-value function: $\tilde{Q}_{k+1} \leftarrow \arg \min_{f \in \mathcal{F}_{k+1}} \frac{1}{n} \sum_{j=1}^n [Y_j - f(s_j, \mathbf{a}_j)]^2$.
- 5: **end for**
- 6: Define the policy π_K as the product of the greedy policies $\{\pi_K^i\}_{i \in [N]}$ with respect to $\{\tilde{Q}_K^i\}_{i \in [N]}$.

Output: An estimator \tilde{Q}_K of Q^* and its greedy policy π_K .

Compared with the Factorized Multi-Agent Fitted Q-Iteration (FMA-FQI) proposed by Wang et al. (2021), the state spaces are continuous thus infinite in our game setting, which is far beyond tabular case. Under the IGM condition, MA-FQI share certain similarities to its single-agent variant of FQI (Munos and Szepesvári, 2008): the step of computing targets through decentralized maximization gives the actual max-target, and the resulting distributed greedy policies result in a greedy joint policy. Hence, we can expect that the theoretical guarantees of FQI find their extension in MARL. Indeed, in the following sections, we show that the presence of multiple agents, does not prevent, yet slows down, the framework from convergence under the VDN model.

3. Decomposable Games

A preliminary step that we must take before we analyze the value decomposition algorithms is the analysis of frameworks that they are applicable to. Specifically, we characterize a class of MAMGs in which the additive value decomposition (i.e., Equation (4)) holds.

Definition 3.1 (Decomposable Game) *A multi-agent Markov Game $\mathcal{MG}(N, \mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma, d_0)$ is a decomposable game if its reward function $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ can be decomposed as:*

$$R(\mathbf{s}, \mathbf{a}) = R_1(s^1, a^1) + R_2(s^2, a^2) + \dots + R_N(s^N, a^N)$$

(here, $R_i : \mathcal{S}^{(i)} \times \mathcal{A}^{(i)} \rightarrow \mathbb{R}$ can be regarded as independent reward for the i -th agent) and the transition kernel \mathbb{P} can be decomposed as:

$$\mathbb{P}(\mathbf{s}' | \mathbf{s}, \mathbf{a}) = F_1(\mathbf{s}' | s^1, a^1) + F_2(\mathbf{s}' | s^2, a^2) + \dots + F_N(\mathbf{s}' | s^N, a^N).$$

As we can see, a game is decomposable when both its reward function and its transition kernel can be distributed across individual agents and their local interactions with the game. The key property of a decomposable game is that the state-action value function Q^π can also be decomposed, regardless of the policy π . This fact can be easily proved by expanding $Q^\pi(\mathbf{s}, \mathbf{a})$ with the Bellman equation (Sutton and Barto, 2018):

$$\begin{aligned} Q^\pi(\mathbf{s}, \mathbf{a}) &= R(\mathbf{s}, \mathbf{a}) + \gamma \cdot \mathbb{E}_{\mathbf{s}' \sim \mathbb{P}}[V^\pi(\mathbf{s}')] = R(\mathbf{s}, \mathbf{a}) + \gamma \int_{\mathcal{S}} V^\pi(\mathbf{s}') \mathbb{P}(\mathbf{s}' | \mathbf{s}, \mathbf{a}) d\mathbf{s}' \\ &= \sum_{i=1}^N R_i(s^i, a^i) + \gamma \int_{\mathcal{S}} V^\pi(\mathbf{s}') \cdot \left(\sum_{i=1}^N F_i(\mathbf{s}' | s^i, a^i) \right) d\mathbf{s}' \\ &= \sum_{i=1}^N \left[R_i(s^i, a^i) + \gamma \int_{\mathcal{S}} V^\pi(\mathbf{s}') \cdot F_i(\mathbf{s}' | s^i, a^i) d\mathbf{s}' \right] \triangleq \sum_{i=1}^N Q_i^\pi(s^i, a^i). \end{aligned}$$

Therefore, the decomposability of a game is a sufficient condition for the decomposability of the Q-value functions, which establishes the IGM principle in the game. In our studies, however, we pay most of our attention to the image of Q under the Bellman operator T (Sutton and Barto, 2018) defined as $[TQ](\mathbf{s}, \mathbf{a}) = R(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathbf{s}' \sim \mathbb{P}}[\max_{\mathbf{a}'} Q(\mathbf{s}', \mathbf{a}')]]$, because in Algorithm 1 the critic \tilde{Q}_{k+1} is trained to match $T\tilde{Q}_k$. Fortunately, in a decomposable game, TQ is also decomposable. In fact, decomposable games are the **only** type of games in which this property holds, as given by the following proposition.

Proposition 3.1 *For a MAMG $\mathcal{MG}(N, \mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma, \pi_0)$, these two statements are equivalent:*

- (1) \mathcal{MG} is a decomposable game.
- (2) For any state-action value critic Q , and any discount factor $\gamma \in [0, 1)$, TQ is a decomposable function, i.e., there exist $G_1^{(\gamma)}, G_2^{(\gamma)}, \dots, G_N^{(\gamma)}$ such that:

$$[TQ](\mathbf{s}, \mathbf{a}) = G_1^{(\gamma)}(s^1, a^1) + G_2^{(\gamma)}(s^2, a^2) + \dots + G_N^{(\gamma)}(s^N, a^N) \quad (7)$$

See Appendix A for proof. Hence, the algorithms which follow the framework of MA-FQI (Algorithm 1) implicitly make an assumption not simply about the decomposability of the Q-function, but also on the decomposability of the reward and transition functions. Although this setting might be rare in reality, it may be considered as its approximation through Taylor expansion up to the first order. The empirical success of VDN supports this point of view. Nevertheless, under this exact decomposable setting, we study the properties of VDN in the next section.

4. Convergence Analysis in Decomposable Games

In this section, we study the convergence of VDN in the decomposable game. Precisely, we show that, in decomposable games, by considering the agents' joint action, VDN can be interpreted as DQN with a different (decomposed) function class. This similarity enables us to extend the analysis from single-agent deep RL (Munos and Szepesvári, 2008) to MARL.

We start by setting up the framework for function approximators. Firstly, for the purpose of convenience and clarity, we introduce the following assumptions.

Assumption 4.1 *All agents have identical state and action spaces, i.e., $\mathcal{S}^{(i)} = \mathcal{S}^{(j)} \triangleq \mathcal{S}$ and $\mathcal{A}^{(i)} = \mathcal{A}^{(j)} \triangleq \mathcal{A}$ hold for $\forall i, j \in [N]$.*

This assumption, although presented in practical applications, does not influence our analysis. It only enables us to simplify writing and notation, and allows us to replace quantities including summations with simple multiplication by the number of agents N . We proceed by defining the set of functions that are sums over maps of the decomposed input.

Definition 4.1 *Let \mathcal{M} be a set of maps $m : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. Then, the N -composition set of \mathcal{M} is defined as*

$$\mathcal{M}^{\oplus N} \triangleq \left\{ m^{(N)} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \mid m^{(N)}(\mathbf{s}, \mathbf{a}) = \sum_{i=1}^N m^i(s^i, a^i), \text{ and } m^i \in \mathcal{M}, \forall i \in [N] \right\}.$$

The role the above definition plays is that it captures the output of the joint critic of VDN into one function. It may be tempting to think that VDN simply adds N decentralized and uncorrelated functions together, while the procedure of it is subtler. The algorithm, first, splits the state-action input (\mathbf{s}, \mathbf{a}) into N parts, $\{(s^i, a^i)\}_{i \in [N]}$, then lets the parts pass through corresponding critics $\{Q^i\}_{i \in [N]}$, and computes their sum $Q_{\text{tot}} = \sum_{i=1}^N Q_i$ at the end. Thus, we can think of Q_{tot} as of one joint critic, whose computation can be partially decentralized.

With this definition, and the intuition behind it, we continue our analysis. Crucially, as the joint critic Q_{tot} is an element of an N -composition set, we must be able to study such sets. In particular, covering numbers of function classes play the key role in our considerations. One way to settle them is to take advantage of studies of neural networks, by relating the covering number of the N -composition set to that of its components, which we do in the following lemma.

Lemma 4.1 *Let $\mathcal{N}(\mathcal{M}, \delta)$ denote the cardinality of the minimal δ covering of set \mathcal{M} of maps $m : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. Then we have*

$$\mathcal{N}(\mathcal{M}^{\oplus N}, \delta) \leq \mathcal{N}(\mathcal{M}, \delta/N)^N.$$

For proof see Appendix B. Furthermore, we need a notion to describe the discrepancy between two function classes \mathcal{M}_1 and \mathcal{M}_2 . In our analysis, most of the time we will need to study the worst-case scenario, of the mismatch between an approximator $m_1 \in \mathcal{M}_1$ and the ground-truth $m_2 \in \mathcal{M}_2$ be maximal possible. Therefore, we use the following notion of distance.

Definition 4.2 *Let \mathcal{M}_1 and \mathcal{M}_2 be two classes of bounded functions with the domain $\mathcal{S} \times \mathcal{A}$ and image \mathbb{R} . Then, the distance between \mathcal{M}_1 and \mathcal{M}_2 is defined as*

$$\text{dist}(\mathcal{M}_1, \mathcal{M}_2) \triangleq \sup_{m_1 \in \mathcal{M}_1} \inf_{m_2 \in \mathcal{M}_2} \|m_1 - m_2\|_{\infty}.$$

As before, having more control over the particular components $\{Q_i\}_{i \in [N]}$ of Q_{tot} , we are interested in relating the distance of two N -composition sets to the distance between their components. In the following lemma we obtain an elegant linear relation, derived in Appendix B.

Lemma 4.2 *Let \mathcal{M}_1 and \mathcal{M}_2 be instances of function classes from Definition 4.2. Then*

$$\text{dist}(\mathcal{M}_1^{\oplus N}, \mathcal{M}_2^{\oplus N}) \leq N \cdot \text{dist}(\mathcal{M}_1, \mathcal{M}_2).$$

Knowing the relations between distributed functions and their N -composition, we possess tools that can unroll the properties of DQN to the multi-agent VDN algorithm. To give exact bounds, however, we must specify precisely the function approximators that the agents use. As it is often implemented by deep MARL practitioners, we let every agent train a deep neural network, and the resulting joint critic is a summation over them.

Definition 4.3 (Deep Sparse ReLU Network) *For any depth $L \in \mathbb{N}$, sparsity parameter $s \in \mathbb{N}$, and sequence of widths $\{d_j\}_{j=0}^{L+1} \subseteq \mathbb{N}$, and $U > 0$, the function class $\mathcal{F}(L, \{d_j\}_{j=0}^{L+1}, s, U)$ is the set of maps $f : \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_{L+1}}$, defined as*

$$f(x) = W_{L+1}\sigma(W_L\sigma(\dots(W_2\sigma(W_1x + v_1) + v_2) \dots v_{L-1}) + v_L),$$

where for $j = 0, \dots, L$, $W_{j+1} \in \mathbb{R}^{d_{j+1} \times d_j}$ are weight matrices, v_j are bias vectors, and $\sigma(x) = \max(0, x)$ is the ReLU activation function. Furthermore, for this class we require that the weights of the network are not too large, i.e., $\|(W_l, v_l)\|_{\max} \leq 1, \forall l \in [L+1]$, not too many of the weights are non-zero, i.e., $\sum_{l=1}^{L+1} \|(W_l, v_l)\|_{\max} \leq s$, and that $\max_{j \in d_{L+1}} \|f_j\|_{\infty} \leq U$.

In our analysis, the efficacy of a network is related to its smoothness properties. To study them, we introduce the notion of Hölder smoothness—a tool considered in deep learning and reinforcement learning literature (Chen and Jiang, 2019; Fan et al., 2020).

Definition 4.4 (Hölder Smooth Function) *Let $d \in \mathbb{N}$, and $\mathcal{D} \subset \mathbb{R}^d$ be a compact set, and let $\beta, B > 0$. The Hölder smooth functions on \mathcal{D} are elements of the set*

$$\mathcal{C}_d(\mathcal{D}, \beta, B) = \left\{ f : \mathcal{D} \rightarrow \mathbb{R} : \sum_{|\alpha| < \beta} \|\partial^\alpha f\|_{\infty} + \sum_{\|\alpha\|_1 \leq \lfloor \beta \rfloor} \sum_{x \neq y \in \mathcal{D}} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x - y\|_{\infty}^{\beta - \lfloor \beta \rfloor}} \leq B \right\},$$

where $\lfloor \beta \rfloor$ is the floor of β , $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$, and $\partial^\alpha = \partial^{\alpha_1} \dots \partial^{\alpha_d}$.

Furthermore, to study fine compositions of real mappings, we must define a class of compositions of Hölder smooth functions.

Definition 4.5 (Composition of Hölder smooth functions) *Let $q \in \mathbb{N}$ and $\{p_j\}_{j \in [q]}$ be integers, and let $\{[a_j, b_j]\}_{j \in [q]}$ non-empty real intervals. For any $j \in [q]$, consider a vector-valued function $g_j : [a_j, b_j]^{p_j} \rightarrow [a_{j+1}, b_{j+1}]^{p_{j+1}}$, such that each of its components $g_{j,k}$ ($k \in [p_{j+1}]$) is Hölder smooth, and depends on $t_j \leq p_j$ components of its input. We set $p_{q+1} = 1$, and define the class of compositions of Hölder smooth functions $\mathcal{G}(\{p_j, t_j, a_j, b_j\}_{j \in [q]})$ as functions f , that can be written in a form $f = g_q \circ g_{q-1} \circ \dots \circ g_1$, where g_1, \dots, g_q follow the rules listed above.*

In our study, it is important the type of neural network stays close to the above class, wherein their training targets happen to find themselves. Next, we lay out the characterization of the networks used by the agents, and that of compositions of Hölder smooth functions that the networks track.

Definition 4.6 (Function Classes) Let $d_0 = d$, and $d_{L+1} = 1$. Then, any agent has access to the function class of neural network (Definition 4.3) critics

$$\mathcal{F}_{\text{net}} \triangleq \left\{ f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} : f(\cdot, a) \in \mathcal{F} \left(L, \{d_j\}_{j=0}^{L+1}, s, Q_{\max}/N \right), \forall a \in \mathcal{A} \right\}.$$

Correspondingly, employing compositions of Hölder smooth functions (Definition 4.5), we define the class

$$\mathcal{G}_H \triangleq \left\{ g : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} : g(\cdot, a) \in \mathcal{G} \left(\{p_j, t_j, \beta_j, B_j\}_{j \in [q]} \right), \forall a \in \mathcal{A} \right\},$$

and refer to it as the Hölder class for brevity. It follows that the joint critic Q_{tot} belongs to the class $\mathcal{F}_{\text{net}}^{\oplus N}$, which tracks the corresponding Hölder class $\mathcal{G}_H^{\oplus N}$.

In the following, we make a standard assumption on approximate closure of $\mathcal{F}_{\text{net}}^{\oplus N}$ under the Bellman operator T , where the vicinity of $\mathcal{F}_{\text{net}}^{\oplus N}$ is considered to be $\mathcal{G}_H^{\oplus N}$ (Chen and Jiang, 2019; Fan et al., 2020). Note that, if the joint critic was able to learn the optimal value Q^* , then by the Bellman optimality equation $TQ^* = Q^*$. Hence, we would have $Q^* \in \mathcal{F}_{\text{net}}^{\oplus N}$ and $TQ^* \in \mathcal{F}_{\text{net}}^{\oplus N}$, which suggests the approximate closure.

Assumption 4.2 For any $f \in \mathcal{F}_{\text{net}}^{\oplus N}$, we have $Tf \in \mathcal{G}_H^{\oplus N}$, where T is the Bellman operator.

Lastly, we make an assumption about the concentration coefficients (Munos and Szepesvári, 2008), which provide some notion of distance between two probability distributions on $\mathcal{S} \times \mathcal{A}$ in a MAMG.

Assumption 4.3 (Concentration Coefficients) Let $\mathcal{P}(\mathcal{S} \times \mathcal{A})$ be the set of probability measures that are absolutely continuous with respect to the Lebesgue measure on $\mathcal{S} \times \mathcal{A}$. Let $\nu_1, \nu_2 \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$, and the initial state-(joint)action pair has distribution $(s_0, a_0) \sim \nu_1$. Let $\{\pi_t\}_{t=1}^{\infty}$ be a sequence of joint policies so that, for $t \geq 1$, $\mathbf{a}_t \sim \pi_t(\cdot | s_t)$, and $\mathbb{P}^{\pi_t} \dots \mathbb{P}^{\pi_1} \nu_1$ is the marginal distribution of (s_t, \mathbf{a}_t) . We define the concentration coefficient at time t as

$$\kappa_t(\nu_1, \nu_2) \triangleq \sup_{\pi_1, \dots, \pi_t} \left[\mathbb{E}_{\mathbf{s}, \mathbf{a} \sim \nu_2} \left(\left| \frac{d(\mathbb{P}^{\pi_t} \dots \mathbb{P}^{\pi_1} \nu_1)}{d\nu_2}(\mathbf{s}, \mathbf{a}) \right|^2 \right) \right]$$

We assume that for $\nu_2 = \sigma$, the sampling distribution of Algorithm 1, for any $\nu_1 = \mu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$, there exists a finite constant $\phi_{\mu, \sigma}$ such that $\phi_{\mu, \sigma} = (1 - \gamma)^2 \sum_{t=1}^{\infty} t\gamma^{t-1} \kappa_t(\mu, \sigma)$.

With the definitions and assumptions set up, we finally reaching to discovering the theoretical properties of Algorithm 1.

4.1. Theoretical Properties of Algorithm 1

The following theorem describes how the error in MA-FQI (Algorithm 1) propagates, and holds regardless of the function class used for critics. Extending the error propagation theorem for single-agent FQI to cooperative decomposable MAMGs, we have the following error bound.

Theorem 4.1 (Error Propagation) *Let $K \in \mathbb{N}$, and $\{\tilde{Q}_k\}_{k \in [K]}$ be iterates of Algorithm 1 in a decomposable MAMG. Let π_K be the greedy joint policy with respect to \tilde{Q}_K , and Q^{π_K} be the actual state-action value function of π_K . Recall that Q^* denotes the optimal state-action value function. Under Assumption 4.3,*

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} \leq \frac{2\phi_{\mu,\sigma}\gamma}{(1-\gamma)^2} \cdot \epsilon_{\max} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} R_{\max}, \text{ where } \epsilon_{\max} = \max_{k \in [K]} \|T\tilde{Q}_{k-1} - \tilde{Q}_k\|_{\sigma}.$$

This theorem decomposes the final error into the *approximation error*, $2\phi_{\mu,\sigma}\gamma\epsilon_{\max}/(1-\gamma)^2$, and the *algorithmic error*, $4\gamma^{K+1}R_{\max}/(1-\gamma)^2$. The latter term does not depend on the function approximators used, and vanishes fast as the number of iterations K increases. Therefore, the problem is the former term—the error arising from the function approximator, and in particular, its approximate closure under the Bellman operator T . Hereafter, we focus our analysis on it, which we begin with the following theorem, proved by Fan et al. (2020).

Theorem 4.2 (One-step Approximation Error) *Let σ be a probability distribution on $\mathcal{S} \times \mathcal{A}$, and let $\{(s_i, a_i)\}_{i \in [n]}$ be a sample drawn from σ . Let R_i and s'_i be the reward and the next state corresponding to (s_i, a_i) . Let $Q^{\text{tar}} \in \mathcal{F}_{\text{net}}^{\oplus N}$. For every $i \in [n]$, we define the training target $Y_i = R_i + \gamma \max_{a \in \mathcal{A}} Q^{\text{tar}}(s'_i, a)$. Let*

$$\hat{Q} = \arg \min_{f \in \mathcal{F}_{\text{net}}^{\oplus N}} \frac{1}{n} \sum_{i=1}^n [f(s_i, a_i) - Y_i]^2,$$

and for any $\delta > 0$ and function class \mathcal{F} , let $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_{\infty})$ denote the cardinality of the minimal δ -covering of \mathcal{F} , with respect to l_{∞} -norm. Then, for some absolute constant $C > 0$,

$$\|\hat{Q} - TQ^{\text{tar}}\|_{\sigma}^2 \leq 4\text{dist}(\mathcal{F}_{\text{net}}^{\oplus N}, \mathcal{G}_H^{\oplus N})^2 + C \cdot (Q_{\max}^2/n) \cdot \log \mathcal{N}(\mathcal{F}_{\text{net}}^{\oplus N}, \delta, \|\cdot\|_{\infty}).$$

The theorem decomposes the approximation error into quantities that are properties of the function approximator class and the (target) Hölder class. The first term, involving $\text{dist}(\mathcal{F}_{\text{net}}^{\oplus N}, \mathcal{G}_H^{\oplus N})$, can be thought of as a metric of mismatch between the class $\mathcal{F}_{\text{net}}^{\oplus N}$ and the class of targets $\mathcal{G}_H^{\oplus N}$. The better neural networks approximate the Hölder functions, the smaller this metric is. The second term, involving $\mathcal{N}(\mathcal{F}_{\text{net}}^{\oplus N}, \delta, \|\cdot\|_{\infty})$, can be thought of as a measure of sparsity of the class. The less expressible the networks are, the bigger its δ -covering. By considering sparse ReLU networks, and their N -composition that the agents use during learning, we provide the main theorem of this section, which reveals the convergence property of VDN in decomposable games.

Theorem 4.3 (Main Theorem 1: Decomposable Setting) *Let \mathcal{F}_{net} and \mathcal{G}_H be defined as in Definition 4, based on the class of neural networks $\mathcal{F}_1 = \dots = \mathcal{F}_K = \mathcal{F}(L^*, \{d_j\}_{j=0}^{L^*+1}, s^*, Q_{\max}/N)$, and the class of Hölder smooth functions $\mathcal{G}_H(\{p_j, t_j, \beta_j, B_j\}_{j \in [q]})$. For any $j \in [q-1]$, we define $\beta_j^* = \beta_j \prod_{l=j+1}^q \min(\beta_l, 1)$, and $\beta_q^* = 1$. In addition, let $\alpha^* = \max_{j \in [q]} \frac{t_j}{2\beta_j^* + t_j} < 1$. We assume that the sample size is large, relative to the parameters of \mathcal{G}_H , so that there exists a constant $\xi > 0$, such that*

$$\max \left\{ \sum_{j=1}^q (t_j + \beta_j + 1)^{3+t_j}, \sum_{j \in [q]} \log(t_j + \beta_j), \max_{j \in [q]} p_j \right\} = \mathcal{O}((\log n)^{\xi}). \quad (8)$$

Moreover, we assume that the hyper-parameters of the neural networks satisfy

$$L^* = \mathcal{O}\left((\log n)^{\xi^*}\right), \quad d \leq \min_{j \in [L^*]} d_j^* \leq \max_{j \in [L^*]} d_j^* = \mathcal{O}(n^{\xi^*}), \quad \text{and } s^* = \Theta\left(n^{\alpha^*} (\log n)^{\xi^*}\right), \quad (9)$$

for some absolute constant $\xi^* > 1 + 2\xi$. Let π_K be the output joint policy of Algorithm 1, and Q^{π_K} be its (true) joint state-action value function. Then, for some absolute constant $C > 0$,

$$\begin{aligned} & \|Q^* - Q^{\pi_K}\|_{1,\mu} \\ & \leq \frac{C\phi_{\mu,\sigma}\gamma}{(1-\gamma)^2} \left(N \cdot n^{-(1-\alpha^*)/2} + \sqrt{|\mathcal{A}| \cdot N \cdot \log N} \cdot n^{-(1-\alpha^*)/2} (\log n)^{(1+2\alpha^*)/2} \right) + \frac{4\gamma^{K+1}}{(1-\gamma)^2} \cdot R_{\max}. \end{aligned}$$

For proof see Appendix C.

5. Convergence Analysis in Non-decomposable Games

In this section, we extend the decomposable games to the general non-decomposable games, which is a more challenging setting. For simplicity, here instead of using multi-layer networks for training, we consider the 2-layer ReLU networks. Our proof can be easily applied to more complicated function classes (such as multi-layer networks). Under this setting, we make the function class used in the k -th iteration to be $\mathcal{F}_k = \mathcal{F}(B_k, M)^{\oplus N}$, the set of decomposable 2-layer ReLU networks with weight M and their path norm bounded by B_k , to be rigorous:

$$\mathcal{F}(B, M) = \left\{ f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \mid f(s, a) = \sum_{i=1}^M \alpha_i^a \cdot (\langle \beta_i^a, s \rangle + \gamma_i^a), \max_{a \in \mathcal{A}} \sum_{i=1}^M |\alpha_i^a| \cdot (\|\beta_i^a\|_1 + |\gamma_i^a|) \leq B \right\},$$

and $\mathcal{F}(B, M)^{\oplus N} \in \{F : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}\}$ is its N -composition. In the following parts, we are going to show that even for non-decomposable game where $T\tilde{Q}_k$ may not be close to any decomposable functions for a decomposable \tilde{Q}_k , the MA-FQI Algorithm will still be able to converge to the optimal value function Q^* as long as Q^* itself is a decomposable function, which is in fact a counterfactual result since we need to project our estimator onto the decomposable function class in each iteration, which may cause divergence by our intuition. In the following paragraphs, we are going to show that \tilde{Q}_k will provably converge to Q^* when following Algorithm 1. First, we are going to bridge the gap between the value function of the greedy policy π_k , denoted by Q^{π_k} and the estimated Q-value \tilde{Q}_k by the following lemma:

Lemma 5.1

$$\|Q^* - Q^{\pi_k}\|_{\infty} \leq \frac{2\gamma}{1-\gamma} \|Q^* - \tilde{Q}_k\|_{\infty}.$$

Therefore, in order to control the error $\|Q^* - Q^{\pi_k}\|_{\infty}$, we only need to upper bound the estimation error of Q-function, which is $\|Q^* - \tilde{Q}_k\|_{\infty}$. Since \tilde{Q}_k is generated in an iterative manner and $\tilde{Q}_{k+1} = \text{Proj}\left(T\tilde{Q}_k, \mathcal{F}(B_k)^{\oplus N}, \|\cdot\|_{\sigma}\right)$, we can upper bound the last iteration error $\|Q^* - \tilde{Q}_K\|_{\infty}$ in a cumulative way.

Lemma 5.2

$$\|Q^* - \tilde{Q}_K\|_{\infty} \leq \frac{\varepsilon_{\max}}{1-\eta} + \frac{4\gamma^K}{(1-\gamma)^2} R_{\max}.$$

Here, $\eta = (N+1)\gamma$ and $\varepsilon_{\max} = \max_{k \in [K]} \left\| \tilde{Q}_{k+1} - \text{Proj}\left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_{\sigma}\right) \right\|_{\infty}$.

From the two lemmas above, we know that in order to upper bound the gap $\|Q^* - Q^{\pi_K}\|_{\infty}$, we only need to upper bound the ε_{\max} . Next, we are going to prove that, with high probability over the sampling of $(s, a) \sim \sigma$ in each iteration, the discrepancy ε_{\max} can be well upper bounded by using the approximation properties as well as the generalization properties of 2-layer ReLU networks (which are introduced in detail in Appendix E).

Lemma 5.3 *With probability at least $1 - \delta$ over the sampling of $(s, a) \sim \sigma$ in all the K iterations, when the discount ratio $\gamma \ll \frac{1}{N^2}$, we can let $B_k = \frac{8Nc_2R_{\max}}{1-4N^2\gamma} := B$ for $\forall k \in [K]$, such that:*

$$\begin{aligned} \varepsilon_{\max} := \max_{k \in [K]} \left\| \tilde{Q}_{k+1} - \text{Proj} \left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_{\sigma} \right) \right\|_{\infty} &\leq c_1 B d \cdot \left[\frac{B^2}{M} + 8|\mathbf{A}| \cdot \frac{\log(8|\mathbf{A}|/\delta)}{n} \cdot Q_{\max}^2 \right. \\ &\quad \left. + \sqrt{|\mathbf{A}|} \cdot \left(16Q_{\max}(2B+2) \sqrt{\frac{2\log(2d)}{n}} + 8Q_{\max}^2 \sqrt{\frac{8|\mathbf{A}| \log(2c(B+1)^2/\delta)}{n}} \right) \right]^{\frac{1}{d+2}}, \end{aligned}$$

where c, c_1, c_2 are constants.

Finally, after combining all the three lemmas above, we conclude that:

Theorem 5.1 (Main Theorem 2: Non-decomposable Setting) *Assume the optimal Q -function $Q^* \in \mathcal{C}^{\oplus N}$ is a decomposable function. Let the function classes $\mathcal{F}_1 = \dots = \mathcal{F}_K = \mathcal{F}(B, M)^{\oplus N}$. When the discount ratio $\gamma \ll \frac{1}{N^2}$, we can choose the path norm bound $B = \frac{8Nc_2R_{\max}}{1-4N^2\gamma}$. Then, by running MA-FQI (Algorithm 1), for some constant $c, c_1, c_2 > 0$, we have:*

$$\begin{aligned} \|Q^* - \tilde{Q}^{\pi_K}\|_{\infty} &\leq \frac{8\gamma^{K+1}}{(1-\gamma)^3} R_{\max} + \frac{c_1 B d \gamma}{(1-(N+1)\gamma)(1-\gamma)} \cdot \left[\frac{B^2}{M} + 8|\mathbf{A}| \cdot \frac{\log(8|\mathbf{A}|/\delta)}{n} \cdot Q_{\max}^2 \right. \\ &\quad \left. + \sqrt{|\mathbf{A}|} \cdot \left(16Q_{\max}(2B+2) \sqrt{\frac{2\log(2d)}{n}} + 8Q_{\max}^2 \sqrt{\frac{8|\mathbf{A}| \log(2c(B+1)^2/\delta)}{n}} \right) \right]^{\frac{1}{d+2}}. \end{aligned}$$

For proofs see Appendix D. As we can see, the first term $\frac{8\gamma^{K+1}}{(1-\gamma)^3} R_{\max}$ exponentially shrinks to 0 since $\gamma < 1$. For the second term, after treating all the instance-based parameters (such as $B, d, \gamma, N, Q_{\max}$) as constants, has order $\mathcal{O}(\frac{1}{M}) + \mathcal{O}(\frac{1}{\sqrt{n}})$. Here, $\mathcal{O}(\frac{1}{M})$ and $\mathcal{O}(\frac{1}{\sqrt{n}})$ come from the approximation error and generalization error of 2-layer ReLU networks respectively. For sufficiently large width M (which stands for the over-parameterization) and large sample size n (which stands for the small gap between the empirical mean and population mean in the sampling process of each iteration), the l_{∞} error between Q^* and Q^{π_K} converges to 0. Although the sample complexity $\mathcal{O}(1/\varepsilon^{2d+4})$ suffers the curse of dimension, the convergence itself is a huge step for understanding the MA-FQI algorithm in cooperative multi-agent reinforcement learning.

6. Conclusion

Although value decomposition methods for cooperative MARL has great promise for addressing coordination problems in a variety of applications (Yang et al., 2017; Zhou et al., 2020, 2021), theoretical understandings for these approaches are still limited. This paper makes the initial effort to bridge this gap by considering a general framework for theoretical studies. Central to our findings is the decomposable games where value decomposition methods can be applied safely. Specifically, we show that the multi-agent fitted Q-Iteration algorithm (MA-FQI), parameterized by multi-layer deep ReLU networks, can lead to the optimal Q-function. Moreover, for non-decomposable games, the estimated Q-function parameterized by wide 2-layer ReLU networks, can still converge to the optimum by using MA-FQI, despite the fact that the Q-function needs projecting into the decomposable function space at each iteration. In our future works, we are going to extend the 2-layer ReLU networks to a much broader function class, and see whether we can reduce the sample complexity and avoid the curse of dimension. Also, mean-field game setting will be taken into consideration and we will see whether the convergence guarantee can still be provided in the sense of distribution.

References

- Martin Anthony, Peter L Bartlett, and Peter L Bartlett. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.
- Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.
- Leo Breiman. Hinging hyperplanes for regression, classification, and function approximation. *IEEE Transactions on Information Theory*, 39(3):999–1013, 1993.
- Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.
- Zehao Don, E Weinan, and Chao Ma. A priori estimates of the generalization error for autoencoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3327–3331. IEEE, 2020.
- Weinan E, Chao Ma, and Lei Wu. A priori estimates of the generalization error for two-layer neural networks. *arXiv preprint arXiv:1810.06397*, 2018.
- Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR, 2020.
- Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- Jason M Klusowski and Andrew R Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016.
- Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251*, 2021a.
- Jakub Grudzien Kuba, Muning Wen, Yaodong Yang, Linghui Meng, Shangding Gu, Haifeng Zhang, David Henry Mguni, and Jun Wang. Settling the variance of multi-agent policy gradients. *arXiv preprint arXiv:2108.08612*, 2021b.
- Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. pages 157–163, 1994.
- Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. Maven: Multi-agent variational exploration. *arXiv preprint arXiv:1910.07483*, 2019.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.
- Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

- Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401, 2015.
- Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 4295–4304. PMLR, 2018.
- David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 5887–5896. PMLR, 2019.
- Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.
- Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- Jianhao Wang, Zhizhou Ren, Beining Han, Jianing Ye, and Chongjie Zhang. Towards understanding cooperative multi-agent q-learning with value factorization. *Advances in Neural Information Processing Systems*, 34, 2021.
- Ying Wen, Yaodong Yang, Rui Luo, Jun Wang, and Wei Pan. Probabilistic recursive reasoning for multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2018.
- Ying Wen, Yaodong Yang, and Jun Wang. Modelling bounded rationality in multi-agent interactions by generalized recursive reasoning. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 414–421. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.
- Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*, 2020.
- Yaodong Yang, Lantao Yu, Yiwei Bai, Jun Wang, Weinan Zhang, Ying Wen, and Yong Yu. A study of ai population dynamics with million-agent reinforcement learning. *arXiv preprint arXiv:1709.04511*, 2017.
- Yaodong Yang, Ying Wen, Jun Wang, Liheng Chen, Kun Shao, David Mguni, and Weinan Zhang. Multi-agent determinantal q-learning. In *International Conference on Machine Learning*, pages 10757–10766. PMLR, 2020.
- Ming Zhou, Jun Luo, Julian Vilella, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadakar, Zheng Chen, et al. Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving. *arXiv e-prints*, pages arXiv–2010, 2020.

Ming Zhou, Ziyu Wan, Hanjing Wang, Muning Wen, Runzhe Wu, Ying Wen, Yaodong Yang, Weinan Zhang, and Jun Wang. Malib: A parallel framework for population-based multi-agent reinforcement learning. *arXiv preprint arXiv:2106.07551*, 2021.

Appendix A. Proof of Proposition 3.1

Proof Let us first prove the implication (1) \implies (2). For a decomposable MAMG $\mathcal{MG}(N, \mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma, \pi_0)$, we have

$$\begin{aligned} [TQ](s, a) &= R(s, a) + \gamma \cdot \mathbb{E}_{s' \sim \mathbb{P}}[\max_{a'} Q(s', a')] \\ &= \sum_{i=1}^N R_i(s^i, a^i) + \gamma \int_{\mathcal{S}} \max_{a'} Q(s', a') \left(\sum_{i=1}^N F_i(s' | s^i, a^i) \right) ds' \\ &= \sum_{i=1}^N \left[R_i(s^i, a^i) + \gamma \int_{\mathcal{S}} \max_{a'} Q(s', a') \cdot F_i(s' | s^i, a^i) ds' \right] \triangleq \sum_{i=1}^N G_i^{(\gamma)}(s^i, a^i). \end{aligned}$$

On the other hand, if statement (2) holds for any γ and π , by setting $\gamma = 0$ and expanding the Bellman operator $[TQ](s, a) = R(s, a) + \gamma \cdot \mathbb{E}_{s' \sim \mathbb{P}}[\max_{a'} Q(s', a')]$, we obtain $\sum_{i=1}^N G_i^{(0)}(s^i, a^i) = R(s, a)$. Hence, $R(s, a)$ is a decomposable function, meaning that there exist functions R_1, R_2, \dots, R_N such that:

$$R(s, a) = R_1(s^1, a^1) + R_2(s^2, a^2) + \dots + R_N(s^N, a^N).$$

With this decomposition, for an arbitrary $\gamma > 0$, we can rewrite the Bellman operator as

$$\begin{aligned} [TQ](s, a) &= \sum_{i=1}^N G_i^{(\gamma)}(s^i, a^i) \\ &= \sum_{i=1}^N R_i(s^i, a^i) + \gamma \mathbb{E}_{s' \sim \mathbb{P}}[\max_{a'} Q(s', a')] = \sum_{i=1}^N R_i(s^i, a^i) + \gamma \mathbb{E}_{s' \sim \mathbb{P}}[\max_{a'} V^{\pi_Q}(s')], \end{aligned}$$

where π_Q is a greedy policy with respect to Q . Let us set $g_i^{(\gamma)}(s^i, a^i) = G_i^{(\gamma)}(s^i, a^i) - R_i(s^i, a^i)$. The above equality implies that

$$\sum_{i=1}^N g_i^{(\gamma)}(s^i, a^i) = \gamma \mathbb{E}_{s' \sim \mathbb{P}}[V^{\pi_Q}(s')] = \gamma \langle \mathbb{P}(\cdot | s, a), V^{\pi_Q}(\cdot) \rangle_{\mathcal{S}}, \quad (10)$$

where $\langle \mathbb{P}(\cdot | s, a), v(\cdot) \rangle_{\mathcal{S}} = \int_{\mathcal{S}} \mathbb{P}(s' | s, a) v(s') ds'$ is a linear functional of $v : \mathcal{S} \rightarrow \mathbb{R}$. Hence, the decomposability of $\mathbb{P}(\cdot | s, a)$ follows from taking a functional derivative of Equation (10) with respect to $v(\cdot)$, which finishes the proof. \blacksquare

Appendix B. Proofs of results relating functions and their N -compositions

B.1. Proof of Lemma 4.1

Proof Let \mathcal{M}^* be a minimal δ/N -covering of \mathcal{M} . Let $m^{(N)} \in \mathcal{M}^{\oplus N}$. Then, there exist functions $m_1, \dots, m_N \in \mathcal{M}$, such that for any $s = s^{1:N} \in \mathcal{S}$ and $a = a^{1:N} \in \mathcal{A}$, we have

$$m^{(N)}(s, a) = \sum_{i=1}^N m_i(s^i, a^i).$$

From the definition of a δ/N -covering, we know that there exist m_1^*, \dots, m_N^* , such that for any $i \in [N]$, we have $\|m_i - m_i^*\|_\infty \leq \delta/N$. Hence, for any $\mathbf{s} = s^{1:N}$ and $\mathbf{a} = a^{1:N}$,

$$\delta \geq \sum_{i=1}^N |m_i(s^i, a^i) - m_i^*(s^i, a^i)| \geq \left| \sum_{i=1}^N [m_i(s^i, a^i) - m_i^*(s^i, a^i)] \right| = \left| m^{(N)}(\mathbf{s}, \mathbf{a}) - \sum_{i=1}^N m_i^*(s^i, a^i) \right|.$$

As $\sum_{i=1}^N m_i^*(\cdot, \cdot) \in \mathcal{M}^{\oplus N}$, it follows that $(\mathcal{M}^*)^{\oplus N}$ is a δ -covering of $\mathcal{M}^{\oplus N}$. We also have

$$|(\mathcal{M}^*)^{\oplus N}| \leq |\mathcal{M}^*|^N,$$

which finishes the proof. ■

B.2. Proof of Lemma 4.2

Proof Let $m_1^{(N)} \in \mathcal{M}_1^{\oplus N}$ and $m_2^{(N)} \in \mathcal{M}_2^{\oplus N}$. For any $\mathbf{s} = s^{1:N} \in \mathcal{S}$, $\mathbf{a} = a^{1:N} \in \mathcal{A}$, we have

$$\begin{aligned} |m_1^{(N)}(\mathbf{s}, \mathbf{a}) - m_2^{(N)}(\mathbf{s}, \mathbf{a})| &= \left| \sum_{i=1}^N m_{1,i}(s^i, a^i) - \sum_{i=1}^N m_{2,i}(s^i, a^i) \right| \\ &\leq \sum_{i=1}^N |m_{1,i}(s^i, a^i) - m_{2,i}(s^i, a^i)| \leq \sum_{i=1}^N \|m_{1,i} - m_{2,i}\|_\infty. \end{aligned}$$

Therefore, taking supremum over (\mathbf{s}, \mathbf{a}) , we have

$$\|m_1^{(N)} - m_2^{(N)}\|_\infty \leq \sum_{i=1}^N \|m_{1,i} - m_{2,i}\|_\infty. \quad (11)$$

Let us now fix $m_1^{(N)} = \tilde{m}_1^{(N)}$. For every $i \in [N]$, let $(m_{2,i,k})_{k \in \mathbb{N}}$ be a sequence in \mathcal{M}_2 such that

$$\lim_{k \rightarrow \infty} \|\tilde{m}_{1,i} - m_{2,i,k}\|_\infty = \inf_{m_{2,i} \in \mathcal{M}_2} \|\tilde{m}_{1,i} - m_{2,i}\|_\infty. \quad (12)$$

The Inequality (11) implies that

$$\|\tilde{m}_1^{(N)} - m_{2,k}^{(N)}\|_\infty \leq \sum_{i=1}^N \|\tilde{m}_{1,i} - m_{2,i,k}\|_\infty. \quad (13)$$

As the right-hand side of the above inequality has a finite limit, given in Equation (12), the sequence on the left-hand side above is bounded. Therefore, by Bolzano-Weierstrass Theorem, it has a convergent subsequence $(\|\tilde{m}_1^{(N)} - m_{2,k_j}^{(N)}\|_\infty)_{j \in \mathbb{N}}$. This and Inequality (13) imply that

$$\lim_{j \rightarrow \infty} \|\tilde{m}_1^{(N)} - m_{2,k_j}^{(N)}\|_\infty \leq \lim_{j \rightarrow \infty} \sum_{i=1}^N \|\tilde{m}_{1,i} - m_{2,i,k_j}\|_\infty$$

$$= \sum_{i=1}^N \lim_{j \rightarrow \infty} \|\tilde{m}_{1,i} - m_{2,i,k_j}\|_\infty = \sum_{i=1}^N \inf_{m_{2,i} \in \mathcal{M}_2} \|\tilde{m}_{1,i} - m_{2,i}\|_\infty.$$

We can therefore conclude that

$$\inf_{m_2^{(N)}} \|\tilde{m}_1^{(N)} - m_2^{(N)}\|_\infty \leq \sum_{i=1}^N \inf_{m_{2,i}} \|\tilde{m}_{1,i} - m_{2,i}\|_\infty \quad (14)$$

(Here we dropped the sets from *inf* for brevity.)

Now, unfreezing $\tilde{m}_1^{(N)}$ and taking the supremum over $m_1^{(N)} \in \mathcal{M}_1^{\oplus N}$,

$$\sup_{m_1^{(N)}} \inf_{m_2^{(N)}} \|m_1^{(N)} - m_2^{(N)}\|_\infty \leq \sup_{m_1^{(N)}} \sum_{i=1}^N \inf_{m_{2,i}} \|m_{1,i} - m_{2,i}\|_\infty \leq \sum_{i=1}^N \sup_{m_{1,i}} \inf_{m_{2,i}} \|m_{1,i} - m_{2,i}\|_\infty.$$

Recalling that suprema and infima over $m_{1,i}$ and $m_{2,i}$, for all $i \in [N]$, are taken over sets \mathcal{M}_1 and \mathcal{M}_2 , respectively, allows us to rewrite the above as

$$\text{dist}(\mathcal{M}_1^{\oplus N}, \mathcal{M}_2^{\oplus N}) \leq N \cdot \text{dist}(\mathcal{M}_1, \mathcal{M}_2),$$

which finishes the proof. ■

Remark B.1 We would like to highlight that this result (Lemma 4.2) is quite surprising. The presence of infimum in Definition 4.2 had thrown doubt on the possibility of decomposing the distance to $\mathcal{M}_2^{\oplus N}$ over the summing N copies of \mathcal{M}_2 , as it happened in Inequality (14). Indeed, for any collection of sets $\{\mathcal{X}_1, \dots, \mathcal{X}_N\}$, and any subset \mathcal{Y} of $\mathcal{X}_1 \times \dots \times \mathcal{X}_N$, we have

$$\inf_{(x_1, \dots, x_N) \in \mathcal{Y}} \sum_{i=1}^N x_i \geq \sum_{i=1}^N \inf_{x_i \in \mathcal{X}_i} x_i. \quad (15)$$

What enabled us to arrive there was the trick with a sequence of independent maps in \mathcal{M}_2 from Equation (12), which always have a representant (composition map) in $\mathcal{M}_2^{\oplus N}$, for which they provide the upper bound from Inequality (13).

Appendix C. Proof of Theorem 4.3

Proof Let us recall that by Theorems 4.1 & 4.2, we have

$$\begin{aligned} \|Q^* - Q^{\pi_K}\|_{1,\mu} &\leq \frac{2\phi_{\mu,\sigma}\gamma}{(1-\gamma)^2} \cdot \epsilon_{\max} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} R_{\max} \\ &\leq \frac{2\phi_{\mu,\sigma}\gamma}{(1-\gamma)^2} \cdot \left[4\text{dist}(\mathcal{F}_{\text{net}}^{\oplus N}, \mathcal{G}_{\text{H}}^{\oplus N})^2 + C \cdot (Q_{\max}^2/n) \cdot \log \mathcal{N}(\mathcal{F}_{\text{net}}^{\oplus N}, \delta, \|\cdot\|_\infty) \right]^{\frac{1}{2}} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} R_{\max}. \end{aligned} \quad (16)$$

Hence, to prove the theorem it remains to provide bounds for

$$\text{dist}(\mathcal{F}_{\text{net}}^{\oplus N}, \mathcal{G}_{\text{H}}^{\oplus N}) \text{ and } \log \mathcal{N}(\mathcal{F}_{\text{net}}^{\oplus N}, \delta, \|\cdot\|_\infty).$$

Step 1 (Covering Numbers). We begin with the latter. By Lemma 4.1, we have

$$\log \mathcal{N}(\mathcal{F}_{\text{net}}^{\oplus N}, \delta, \|\cdot\|_{\infty}) \leq N \log \mathcal{N}(\mathcal{F}_{\text{net}}, \delta/N). \quad (17)$$

Furthermore, by Theorem 14.5 in (Anthony et al., 1999), setting $D = \prod_{l=1}^{L^*+1} (d_l^* + 1)$, we have

$$\log \left[\mathcal{N} \left(\frac{\delta}{N}, \mathcal{F} \left(L^*, \{d_j^*\}_{j=0}^{L^*+1}, s^*, \frac{Q_{\max}}{N} \right), \|\cdot\|_{\infty} \right) \right] \leq (s^* + 1) \cdot \log \left[2 \frac{N}{\delta} \cdot (L^* + 1) \cdot D^2 \right]. \quad (18)$$

Let the covering number of the above be denoted as $\mathcal{N}_{\delta/N}$, so that the left-hand side equals $\log \mathcal{N}_{\delta/N}$. The class \mathcal{F}_{net} consists of $|\mathcal{A}|$ components, each being a copy of the class $\mathcal{F}(L^*, \{d_j^*\}_{j=0}^{L^*+1}, s^*, \frac{Q_{\max}}{N})$. Hence, by copying the δ/N -covering of cardinality $\mathcal{N}_{\delta/N}$ to each of the class copies, we obtain a δ/N -covering of \mathcal{F}_{net} (by composing elements from all component classes). The resulting δ/N -covering of \mathcal{F}_{net} has $\mathcal{N}_{\delta/N}^{|\mathcal{A}|}$ elements. Hence

$$\mathcal{N}(\mathcal{F}_{\text{net}}, \delta/N) \leq \mathcal{N}_{\delta/N}^{|\mathcal{A}|},$$

which combined with Inequality (18), and with $\delta = \frac{1}{n}$ (Theorem 4.2 holds for any δ) gives

$$\log \mathcal{N}(\mathcal{F}_{\text{net}}, \delta/N) \leq |\mathcal{A}| \cdot (s^* + 1) \cdot \log [2N \cdot n \cdot (L^* + 1) \cdot D^2]. \quad (19)$$

Furthermore, we have

$$\begin{aligned} \log [2N \cdot n \cdot (L^* + 1) \cdot D^2] &= \log [2n \cdot (L^* + 1) \cdot D^2] + \log(N) \\ &\leq \log [2n \cdot (L^* + 1) \cdot D^2] (1 + \log(N)) \leq C_0 \cdot \log [2n \cdot (L^* + 1) \cdot D^2] \cdot \log(N), \end{aligned}$$

where $C_0 > 0$ is an absolute constant. Recall the choice of hyper-parameters (Equation 9). We have

$$\begin{aligned} \log \mathcal{N}(\mathcal{F}_{\text{net}}, \delta/N) &\leq C_0 \cdot |\mathcal{A}| \cdot (s^* + 1) \cdot \log [2n \cdot (L^* + 1) \cdot D^2] \cdot \log(N) \\ &= \mathcal{O} \left(|\mathcal{A}| \cdot s^* \cdot L^* \cdot \left(\log n + \max_{j \in [L^*]} \log(d_j^*) \right) \cdot \log(N) \right) \\ &= \mathcal{O} \left(|\mathcal{A}| \cdot n^{\alpha^*} (\log n)^{\xi^*} \cdot (\log n)^{\xi^*} (\log n + \xi^* \log n) \cdot \log(N) \right) \\ &= \mathcal{O} \left(|\mathcal{A}| \cdot n^{\alpha^*} \cdot (\log n)^{2\xi^*+1} \cdot \log(N) \right). \end{aligned}$$

Combining this with Inequality (17), we get that for some absolute constant $C_1 > 0$,

$$\log \mathcal{N} \left(\mathcal{F}_{\text{net}}^{\oplus N}, \frac{1}{n}, \|\cdot\|_{\infty} \right) \leq C_1 \cdot N \cdot |\mathcal{A}| \cdot n^{\alpha^*} \cdot (\log n)^{2\xi^*+1} \cdot \log(N). \quad (20)$$

Step 2 (Distance). We now bound the distance

$$\text{dist}(\mathcal{F}_{\text{net}}^{\oplus N}, \mathcal{G}_{\text{H}}^{\oplus N}).$$

By Lemma 4.2, we have

$$\text{dist}(\mathcal{F}_{\text{net}}, \mathcal{G}_{\text{H}}^{\oplus N}) \leq N \cdot \text{dist}(\mathcal{F}_{\text{net}}, \mathcal{G}_{\text{H}}), \quad (21)$$

which implies that it suffices to study the distance between the agents' local function classes. We invoke the following lemma.

Lemma C.1 (Inequality 4.18, (Fan et al., 2020)) *For function classes \mathcal{F}_{net} and \mathcal{G}_H defined as in Definition 4, with hyper-parameters specified in Equations (8) & (9),*

$$\text{dist}(\mathcal{F}_{net}, \mathcal{G}_H)^2 = \mathcal{O}(n^{\alpha^*-1}).$$

Combining the lemma with Inequality (21), we obtain that for some absolute constant $C_2 > 0$, we have

$$\text{dist}(\mathcal{F}_{net}^{\oplus N}, \mathcal{G}_H^{\oplus N})^2 \leq C_2 \cdot N^2 \cdot n^{\alpha^*-1}. \quad (22)$$

Combining Inequalities (20) & (22) with Inequality (16), we have

$$\begin{aligned} \|Q^* - Q^{\pi_K}\|_{1,\mu} &\leq \frac{2\phi_{\mu,\sigma}\gamma}{(1-\gamma)^2} \cdot \epsilon_{\max} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} R_{\max} \\ &\leq \frac{2\phi_{\mu,\sigma}\gamma}{(1-\gamma)^2} \cdot \left[4C_2 \cdot N^2 \cdot n^{\alpha^*-1} \right. \\ &\quad \left. + C \cdot (Q_{\max}^2/n) \cdot C_1 \cdot N \cdot |\mathcal{A}| \cdot n^{\alpha^*} \cdot (\log n)^{2\xi^*+1} \cdot \log(N) \right]^{\frac{1}{2}} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} R_{\max} \\ &\leq \frac{2\phi_{\mu,\sigma}\gamma}{(1-\gamma)^2} \cdot \left[2\sqrt{C_2} \cdot N \cdot n^{(\alpha^*-1)/2} \right. \\ &\quad \left. + \sqrt{C \cdot C_1 \cdot N \cdot \log(N) \cdot |\mathcal{A}|} \cdot Q_{\max} \cdot n^{(\alpha^*-1)/2} \cdot (\log n)^{\xi^*+1/2} \right] + \frac{4\gamma^{K+1}}{(1-\gamma)^2} R_{\max}, \end{aligned}$$

simplifying, taking into account that $1 - \alpha^* > 0$, and bounding,

$$\begin{aligned} &\leq \frac{\tilde{C}\phi_{\mu,\sigma}\gamma}{(1-\gamma)^2} \cdot \left[N \cdot n^{-(1-\alpha^*)/2} \right. \\ &\quad \left. + \sqrt{N \cdot \log(N) \cdot |\mathcal{A}|} \cdot Q_{\max} \cdot n^{-(1-\alpha^*)/2} \cdot (\log n)^{\xi^*+1/2} \right] + \frac{4\gamma^{K+1}}{(1-\gamma)^2} R_{\max}, \end{aligned}$$

where $\tilde{C} > 0$ is an absolute constant. This completes the proof. ■

Appendix D. Proofs of Lemmas and Theorems in Section 5

In this section, we propose the complete proofs for the lemmas and theorems in Section 5.

D.1. Proof of Lemma 5.1

According to the Bellman Equation, we can obtain that:

$$\begin{aligned} Q^*(s, a) &= R(s, a) + \gamma \mathbb{E}_{s'|s,a} V^*(s') := R(s, a) + \gamma P^{\pi^*} Q^*(s, a) \\ Q^{\pi^k}(s, a) &= R(s, a) + \gamma \mathbb{E}_{s'|s,a} V^{\pi^k}(s') := R(s, a) + \gamma P^{\pi^k} Q^{\pi^k}(s, a) \end{aligned}$$

Therefore, we subtract the first equation with the second, and know that for $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\begin{aligned}
 (Q^* - Q^{\pi_k})(s, a) &= \gamma \cdot \left(P^{\pi^*} Q^*(s, a) - P^{\pi_k} Q^{\pi_k}(s, a) \right) \\
 &\stackrel{(a)}{\leq} \gamma \cdot \left(P^{\pi^*} Q^* - P^{\pi_k} Q^* + P^{\pi_k} Q^* - P^{\pi_k} Q^{\pi_k} - P^{\pi^*} \tilde{Q}_k + P^{\pi_k} \tilde{Q}_k \right)(s, a) \\
 &= \gamma \cdot \left(P^{\pi^*} - P^{\pi_k} \right) (Q^* - \tilde{Q}_k)(s, a) + \gamma \cdot P^{\pi_k} (Q^* - Q^{\pi_k})(s, a) \\
 &\stackrel{(b)}{\leq} 2\gamma \cdot \|Q^* - \tilde{Q}_k\|_\infty + \gamma \cdot \|Q^* - Q^{\pi_k}\|_\infty.
 \end{aligned} \tag{23}$$

Since the $(s, a) \in \mathcal{S} \times \mathcal{A}$ can be chosen arbitrarily, so we conclude that:

$$\|Q^* - Q^{\pi_k}\|_\infty \leq 2\gamma \cdot \|Q^* - \tilde{Q}_k\|_\infty + \gamma \cdot \|Q^* - Q^{\pi_k}\|_\infty \Rightarrow \|Q^* - Q^{\pi_k}\|_\infty \leq \frac{2\gamma}{1-\gamma} \|Q^* - \tilde{Q}_k\|_\infty,$$

which comes to our conclusion. In Equation (23), (a) and (b) hold because of two properties of the operator T^π , and we list them below as two lemmas. The first lemma shows us that for a given action-value function $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, the policy π that maximizes $T^\pi Q$ is exactly the greedy policy for Q , i.e., π_Q .

Lemma 1 *For any action value function $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ and any policy π , denote π_Q as the greedy policy for Q , then we have:*

$$P^{\pi_Q} Q = PQ \geq P^\pi Q.$$

Proof [Proof of Lemma 1] By the definition of the P^π operator, we know that:

$$P^\pi Q(s, a) = \mathbb{E}_{s'|s, a} \mathbb{E}_{a' \sim \pi(s')} Q(s', a') \leq \mathbb{E}_{s'|s, a} \max_{a'} Q(s', a') = PQ(s, a)$$

holds for $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$. Therefore, we can conclude that $PQ \geq P^\pi Q$ holds for any policy π and action value function Q . On the other hand, since π_Q is the greedy policy with regard to Q , we have:

$$\mathbb{P} \left[a \in \arg \max_{a'} Q(s, a') \mid a \sim \pi_Q(s) \right] = 1,$$

which leads to

$$P^{\pi_Q} Q(s, a) = \mathbb{E}_{s'|s, a} \mathbb{E}_{a' \sim \pi_Q(s')} Q(s', a') = \mathbb{E}_{s'|s, a} \max_{a'} Q(s', a') = PQ(s, a).$$

Combine the two equations above, and it comes to our conclusion. ■

The second lemma shows us that for any policy π , the operator P^π has Lipschitz constant 1 under the l_∞ norm.

Lemma 2 *For any policy π and any two action value functions $Q_1, Q_2 : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, we have:*

$$\|P^\pi Q_1 - P^\pi Q_2\|_\infty \leq \|Q_1 - Q_2\|_\infty.$$

Proof [Proof of Lemma 2] Since

$$P^\pi Q_1(s, a) = \mathbb{E}_{s'|s, a} \mathbb{E}_{a' \sim \pi(s')} Q_1(s', a'), P^\pi Q_2(s, a) = \mathbb{E}_{s'|s, a} \mathbb{E}_{a' \sim \pi(s')} Q_2(s', a'),$$

we have:

$$\begin{aligned} |P^\pi Q_1(s, a) - P^\pi Q_2(s, a)| &\leq \mathbb{E}_{s'|s, a} \mathbb{E}_{a' \sim \pi(s')} |Q_1(s', a') - Q_2(s', a')| \\ &\leq \mathbb{E}_{s'|s, a} \mathbb{E}_{a' \sim \pi(s')} \|Q_1 - Q_2\|_\infty = \|Q_1 - Q_2\|_\infty. \end{aligned}$$

Since the state-action pair (s, a) can be arbitrarily chosen, we obtain that:

$$\|P^\pi Q_1 - P^\pi Q_2\|_\infty \leq \|Q_1 - Q_2\|_\infty,$$

which comes to our conclusion. ■

Since π_k is the greedy policy with regard to \tilde{Q}_k , we know that $P^{\pi_k} \tilde{Q}_k \geq P^{\pi^*} \tilde{Q}_k$ by Lemma 1, which explains why (a) holds. Also, (b) is a direct extension of Lemma 2.

D.2. Proof of Lemma 5.2

As we know that, the estimations of optimal action value function Q^* are iteratively updated by:

$$\tilde{Q}_{k+1} = \arg \min_{f \in \mathcal{F}_{k+1}} \frac{1}{n} \sum_{i=1}^n \left[T\tilde{Q}_k(s_i, a_i) - f(s_i, a_i) \right]^2,$$

where $TQ(s, a) = R(s, a) + \gamma \cdot PQ(s, a)$. Also, we denote When n is sufficiently large and \mathcal{F} is closed in the set of decomposable continuous functions $\mathcal{C}^{\oplus N}$, we know that

$$\tilde{Q}_{k+1} \approx \arg \min_{f \in \mathcal{C}^{\oplus N}} \mathbb{E}_{(s, a) \sim \sigma} \left[T\tilde{Q}_k(s, a) - f(s, a) \right]^2 := \text{Proj} \left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma \right).$$

After we denote

$$\varepsilon_{\max} = \max_{k \in [K]} \left\| \tilde{Q}_{k+1} - \text{Proj} \left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma \right) \right\|_\infty,$$

we have:

$$\begin{aligned} \|Q^* - \tilde{Q}_{k+1}\|_\infty &\leq \left\| \tilde{Q}_{k+1} - \text{Proj} \left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma \right) \right\|_\infty + \left\| Q^* - \text{Proj} \left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma \right) \right\|_\infty \\ &\leq \varepsilon_{\max} + \left\| \text{Proj} \left(Q^*, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma \right) - \text{Proj} \left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma \right) \right\|_\infty \\ &\stackrel{(a)}{\leq} \varepsilon_{\max} + (2N - 1) \cdot \|Q^* - T\tilde{Q}_k\|_\infty \stackrel{(b)}{\leq} \varepsilon_{\max} + (2N - 1)\gamma \cdot \|Q^* - \tilde{Q}_k\|_\infty \end{aligned} \quad (24)$$

Here, (b) holds because by using Lemma 2:

$$\|Q^* - T\tilde{Q}_k\|_\infty = \|TQ^* - T\tilde{Q}_k\|_\infty = \gamma \cdot \|PQ^* - P\tilde{Q}_k\|_\infty \leq \gamma \cdot \|Q^* - \tilde{Q}_k\|_\infty.$$

(a) holds because of the Lipschitz property for the projection operator, and we are going to explain this in the following lemma, and meanwhile we will give an explicit form for the projection operator.

Lemma 3 (Explicit form of Projection Operator) *For the projection operator above, we have the explicit expression when distribution σ is separable, which means $\sigma \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ can be written as $\sigma_1 \times \sigma_2 \times \dots \times \sigma_N$ where $\sigma_i \in \mathcal{P}(\Upsilon \times \mathcal{A}^{(i)})$ is a distribution over the subspace. Actually, for any C^1 continuous function $f : [a, b]^N \rightarrow \mathbb{R}$, the closet decomposable C^1 continuous function is:*

$$\text{Proj}(f, \mathcal{C}^{\oplus N}, \|\cdot\|_{\sigma})(x_1, x_2, \dots, x_N) = \sum_{i=1}^N f_i(x_i) - (N-1)C.$$

where $f_i(x_i) = \mathbb{E}_{x_{-i} \sim \sigma_{-i}}[f(x_i, x_{-i})]$, $\forall i \in [N]$ and $C = \mathbb{E}_{x \sim \sigma} f(x)$.

Proof [Proof of Lemma 3] For brevity, we denote σ_i as the marginal distribution of x_i , and σ_{-i} as the marginal distribution of $x_{-i} \in \mathbb{R}^{N-1}$. We have

$$\begin{aligned} \mathbb{E}_{x \sim \sigma} \left[\left(\sum_{i=1}^N f_i(x_i) - f(x) \right)^2 \right] &= \mathbb{E}_{x_i \sim \sigma_i} \left[\mathbb{E}_{x_{-i} \sim \sigma_{-i}} \left[\left(\sum_{i=1}^N f_i(x_i) - f(x) \right)^2 \right] \right] \\ &= \mathbb{E}_{x_i \sim \sigma_i} \left[\mathbb{E}_{x_{-i} \sim \sigma_{-i}} \left[\left(f_i(x_i) + \left(\sum_{j \neq i} f_j(x_j) - f(x) \right) \right)^2 \right] \right] \\ &= \mathbb{E}_{x_i \sim \sigma_i} \left[f_i(x_i)^2 + 2f_i(x_i) \mathbb{E}_{x_{-i} \sim \sigma_{-i}} \left[\sum_{j \neq i} f_j(x_j) - f(x) \right] + \mathbb{E}_{x_{-i} \sim \sigma_{-i}} \left[\left(\sum_{j \neq i} f_j(x_j) - f(x) \right)^2 \right] \right]. \end{aligned}$$

The minimum is thus attained if, for every i ,

$$f_i(x_i) = \mathbb{E}_{x_{-i} \sim \sigma_{-i}} \left[f(x_i, x_{-i}) - \sum_{j \neq i} f_j(x_j) \right].$$

Denoting $c_i := \mathbb{E}_{x_i \sim \sigma_i}[f_i(x_i)]$, then we have:

$$f_i(x_i) = \mathbb{E}_{x_{-i} \sim \sigma_{-i}} [f(x_i, x_{-i})] - \sum_{j \neq i} c_j. \quad (25)$$

Taking expectation under $x_i \sim \sigma_i$ on both sides,

$$c_i = \mathbb{E}_{x \sim \sigma} [f(x)] - \sum_{j \neq i} c_j,$$

which leads to

$$C := \sum_{j=1}^N c_j = \mathbb{E}_{x \sim \sigma} [f(x)]. \quad (26)$$

Combining this with equation 25 and aggregating constants, we conclude that the closest decomposable function under distribution σ is

$$\sum_{i=1}^N f_i(x_i) - (N-1)C,$$

where $f_i(x^i) = E_{x_{-i} \sim \sigma_{-i}}[f(x_i, x_{-i})]$ and $C = \mathbb{E}_{x \sim \sigma}[f(x)]$, which comes to our conclusion. \blacksquare

From this lemma, we can obtain two properties of the projection operator. First, for two functions $f, g : \mathbb{R}^N \rightarrow \mathbb{R}$, we know that

$$\|\text{Proj}(f, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma) - \text{Proj}(g, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma)\|_\infty \leq (2N - 1) \cdot \|f - g\|_\infty.$$

Second, if function f has Lipschitz constant L , then its projection $\text{Proj}(f, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma)$ is also Lipschitz continuous, and its Lipschitz constant is $\sqrt{N}L$ since for $\forall i \in [N]$:

$$|f_i(x_i) - f_i(x'_i)| = |\mathbb{E}_{x_{-i} \sim \sigma_{-i}}[f(x_i, x_{-i}) - f(x'_i, x_{-i})]| \leq L \cdot |x_i - x'_i|.$$

After taking $i = 1, 2, \dots, N$ and summing them up:

$$|\text{Proj}(f, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma)(x) - \text{Proj}(f, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma)(x')| \leq L\|x - x'\|_1 \leq L\sqrt{N}\|x - x'\|_2.$$

D.3. Proof of Lemma 5.3

For each iteration $k \in [K]$, we are going to upper bound the discrepancy of the k -th iteration

$$\varepsilon_k := \left\| \tilde{Q}_{k+1} - \text{Proj}(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma) \right\|_\infty.$$

We know that:

$$\tilde{Q}_{k+1} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \left[f(s_i, a_i) - T\tilde{Q}_k(s_i, a_i) \right]^2.$$

If we make our sample size n large enough, the \tilde{Q}_{k+1} be closer to:

$$\arg \min_{f \in \mathcal{F}} \mathbb{E}_{(s,a) \sim \sigma} \left[f(s, a) - T\tilde{Q}_k(s, a) \right]^2.$$

By using the posterior generalization bound proposed by Theorem 13, we know that: for $\forall \mathbf{a} \in \mathcal{A}$ and any distribution σ_s over state space \mathcal{S} , with probability at least $1 - \delta$ over the choice of training data, it holds that:

$$\left| \left\| \tilde{Q}_{k+1}(\cdot, \mathbf{a}) - T\tilde{Q}_k(\cdot, \mathbf{a}) \right\|_n^2 - \left\| \tilde{Q}_{k+1}(\cdot, \mathbf{a}) - T\tilde{Q}_k(\cdot, \mathbf{a}) \right\|_{\sigma_s}^2 \right| \leq 16Q_{\max}(\|\tilde{Q}_{k+1}(\cdot, \mathbf{a})\|_P + 1) \sqrt{\frac{2 \log(2d)}{n}} + 4Q_{\max}^2 \sqrt{\frac{2 \log(2c(\|\tilde{Q}_{k+1}(\cdot, \mathbf{a})\|_P + 1)^2/\delta)}{n}}$$

From the inequality above, we try to establish an upper bound of $\left| \|\tilde{Q}_{k+1} - T\tilde{Q}_k\|_n^2 - \|\tilde{Q}_{k+1} - T\tilde{Q}_k\|_\sigma^2 \right|$ where σ is a distribution over $\mathcal{S} \times \mathcal{A}$. According to our assumption, action space \mathcal{A} is a discrete space. Denote $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{|\mathcal{A}|}\}$, then with probability at least $1 - |\mathcal{A}|\delta$:

$$\begin{aligned} & \left| \|\tilde{Q}_{k+1} - T\tilde{Q}_k\|_n^2 - \|\tilde{Q}_{k+1} - T\tilde{Q}_k\|_\sigma^2 \right| \\ &= \left| \sum_{\mathbf{a} \in \mathcal{A}} \hat{p}_{\mathbf{a}} \cdot \|\tilde{Q}_{k+1}(\cdot, \mathbf{a}) - T\tilde{Q}_k(\cdot, \mathbf{a})\|_{n\hat{p}_{\mathbf{a}}}^2 - \sum_{\mathbf{a} \in \mathcal{A}} p_{\mathbf{a}} \cdot \|\tilde{Q}_{k+1}(\cdot, \mathbf{a}) - T\tilde{Q}_k(\cdot, \mathbf{a})\|_{\sigma_{\mathbf{a}}}^2 \right| \end{aligned}$$

$$\begin{aligned}
 &\leq \sum_{\mathbf{a} \in \mathbf{A}} \hat{p}_{\mathbf{a}} \cdot \left| \|\tilde{Q}_{k+1}(\cdot, \mathbf{a}) - T\tilde{Q}_k(\cdot, \mathbf{a})\|_{n\hat{p}_{\mathbf{a}}}^2 - \|\tilde{Q}_{k+1}(\cdot, \mathbf{a}) - T\tilde{Q}_k(\cdot, \mathbf{a})\|_{\sigma_{\mathbf{a}}}^2 \right| \\
 &\quad + \sum_{\mathbf{a} \in \mathbf{A}} |\hat{p}_{\mathbf{a}} - p_{\mathbf{a}}| \cdot \|\tilde{Q}_{k+1}(\cdot, \mathbf{a}) - T\tilde{Q}_k(\cdot, \mathbf{a})\|_{\sigma_{\mathbf{a}}}^2 \\
 &\leq \sum_{\mathbf{a} \in \mathbf{A}} \hat{p}_{\mathbf{a}} \cdot \left(16Q_{\max}(\|\tilde{Q}_{k+1}(\cdot, \mathbf{a})\|_P + 1) \sqrt{\frac{2\log(2d)}{n\hat{p}_{\mathbf{a}}}} + 4Q_{\max}^2 \sqrt{\frac{2\log(2c(\|\tilde{Q}_{k+1}(\cdot, \mathbf{a})\|_P + 1)^2/\delta)}{n\hat{p}_{\mathbf{a}}}} \right) \\
 &\quad + \sum_{\mathbf{a} \in \mathbf{A}} |\hat{p}_{\mathbf{a}} - p_{\mathbf{a}}| \cdot \|\tilde{Q}_{k+1}(\cdot, \mathbf{a}) - T\tilde{Q}_k(\cdot, \mathbf{a})\|_{\sigma_{\mathbf{a}}}^2 \tag{27}
 \end{aligned}$$

Here, for $\forall \mathbf{a}' \in \mathbf{A}$, $p_{\mathbf{a}'} := \mathbb{P}_{(s, \mathbf{a}) \sim \sigma}[\mathbf{a} = \mathbf{a}']$ and $\hat{p}_{\mathbf{a}'} := \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{\mathbf{a}^i = \mathbf{a}'\}$ stand for the population probability and the empirical probability of joint action \mathbf{a}' under distribution $\sigma \in \mathcal{P}(\mathcal{S} \times \mathbf{A})$. By using Hoeffding inequality, we know that: for $\forall \mathbf{a} \in \mathbf{A}$,

$$\mathbb{P}[|\hat{p}_{\mathbf{a}} - p_{\mathbf{a}}| > t] \leq 2\exp(-2nt^2),$$

which means with probability at least $1 - \delta$, it holds that $|\hat{p}_{\mathbf{a}} - p_{\mathbf{a}}| \leq \sqrt{\frac{\log(2/\delta)}{n}}$. To sum up, with probability at least $1 - 2|\mathbf{A}|\delta$, we have:

$$\begin{aligned}
 &\left| \|\tilde{Q}_{k+1} - T\tilde{Q}_k\|_n^2 - \|\tilde{Q}_{k+1} - T\tilde{Q}_k\|_{\sigma}^2 \right| \\
 &\leq \sum_{\mathbf{a} \in \mathbf{A}} \sqrt{\hat{p}_{\mathbf{a}}} \cdot \left(16Q_{\max}(\|\tilde{Q}_{k+1}(\cdot, \mathbf{a})\|_P + 1) \sqrt{\frac{2\log(2d)}{n}} + 4Q_{\max}^2 \sqrt{\frac{2\log(2c(\|\tilde{Q}_{k+1}(\cdot, \mathbf{a})\|_P + 1)^2/\delta)}{n}} \right) \\
 &\quad + \sum_{\mathbf{a} \in \mathbf{A}} \sqrt{\frac{\log(2/\delta)}{n}} \cdot 4Q_{\max}^2. \tag{28}
 \end{aligned}$$

According to the way to construct \tilde{Q}_{k+1} , we know that: $\|\tilde{Q}_{k+1}(\cdot, \mathbf{a})\|_P \leq B_{k+1} \forall \mathbf{a} \in \mathbf{A}$. Therefore, from Equation (28), we obtain that:

$$\begin{aligned}
 &\left| \|\tilde{Q}_{k+1} - T\tilde{Q}_k\|_n^2 - \|\tilde{Q}_{k+1} - T\tilde{Q}_k\|_{\sigma}^2 \right| \\
 &\leq \sqrt{|\mathbf{A}|} \cdot \left(16Q_{\max}(B_{k+1} + 1) \sqrt{\frac{2\log(2d)}{n}} + 4Q_{\max}^2 \sqrt{\frac{2\log(2c(B_{k+1} + 1)^2/\delta)}{n}} \right) \\
 &\quad + |\mathbf{A}| \cdot \sqrt{\frac{\log(2/\delta)}{n}} \cdot 4Q_{\max}^2 := \Delta_1 \tag{29}
 \end{aligned}$$

holds with probability at least $1 - 2|\mathbf{A}|\delta$ over the sampling. Next, we are going to conduct the same upper bound for function \hat{Q}_{k+1} . Denote the decomposable function:

$$\text{Proj} \left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_{\sigma} \right) = f_k^1(\gamma_1, a_1) + f_k^2(\gamma_2, a_2) + \dots + f_k^N(\gamma_N, a_N).$$

According to Theorem 5, we know that for $\forall i \in [N]$, $a_i \in \mathcal{A}^{(i)}$, there exists a two-layer network $\hat{f}^i : \Upsilon^{(i)} \times \mathcal{A}^{(i)} \rightarrow \mathbb{R}$ of width M such that $\|\hat{f}^i(\cdot, a_i) - f_k^i(\cdot, a_i)\|_P \leq 4\gamma(f_k^i(\cdot, a_i))$ and for any distribution σ_{γ} :

$$\mathbb{E}_{\gamma_i \sim \sigma_{\gamma}} \left(\hat{f}^i(\gamma_i, a_i) - f_k^i(\gamma_i, a_i) \right)^2 \leq \frac{16\gamma^2(f_k^i(\cdot, a_i))}{M}.$$

Then, for any joint action $\mathbf{a} = (a_1, a_2, \dots, a_N)$, there exists a function $\widehat{Q}_{k+1}(s, \mathbf{a}) := \widehat{f}^1(\gamma_1, a_1) + \widehat{f}^2(\gamma_2, a_2) + \dots + \widehat{f}^N(\gamma_N, a_N)$, which satisfies the following two properties:

- Function $\widehat{Q}_{k+1}(\cdot, \mathbf{a})$ is a two-layer ReLU network with width $M|\mathbf{A}|$ and its path norm

$$\|\widehat{Q}_{k+1}(\cdot, \mathbf{a})\|_P \leq 4 \sum_{i=1}^N \gamma(f_k^i(\cdot, a_i)).$$

In order to make \widehat{Q}_{k+1} contained in the function class $\mathcal{F}(B_{k+1})^{\oplus N}$, we have to make:

$$B_{k+1} \geq 4N \cdot \max_{i, a_i} \gamma(f_k^i(\cdot, a_i)),$$

so that we can guarantee that $\widehat{Q}_{k+1}(\cdot, \mathbf{a}) \in \mathcal{F}(B_{k+1})^{\oplus N}$.

- For any distribution σ_s over the state space \mathcal{S} , the mean squared error can be upper bounded as:

$$\begin{aligned} \mathbb{E}_{s \sim \sigma_s} \left(\widehat{Q}_{k+1}(s, \mathbf{a}) - \text{Proj} \left(T\widetilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma \right) \right)^2 &\leq N \cdot \sum_{i=1}^N \mathbb{E}_{s \sim \sigma_s} \left(\widehat{f}^i(\gamma_i, a_i) - f_k^i(\gamma_i, a_i) \right)^2 \\ &\leq \frac{16N^2}{M} \cdot \max_{i, a_i} \gamma^2(f_k^i(\cdot, a_i)) \leq \frac{B_{k+1}^2}{M}. \end{aligned} \quad (30)$$

Again, by using the same technique as Equation (27), we know that

$$\begin{aligned} &\left| \|\widehat{Q}_{k+1} - T\widetilde{Q}_k\|_n^2 - \|\widehat{Q}_{k+1} - T\widetilde{Q}_k\|_\sigma^2 \right| \\ &\leq \sum_{\mathbf{a} \in \mathbf{A}} \sqrt{\widehat{p}_{\mathbf{a}}} \cdot \left(16Q_{\max}(\|\widehat{Q}_{k+1}(\cdot, \mathbf{a})\|_P + 1) \sqrt{\frac{2 \log(2d)}{n}} + 4Q_{\max}^2 \sqrt{\frac{2 \log(2c\|\widehat{Q}_{k+1}\|)}{n}} \right) \\ &\quad + \sum_{\mathbf{a} \in \mathbf{A}} \sqrt{\frac{\log(2/\delta)}{n}} \cdot 4Q_{\max}^2 \end{aligned} \quad (31)$$

holds with probability at least $1 - 2|\mathbf{A}|\delta$. Therefore, according to the two properties above, we can conclude that: with probability at least $1 - 2|\mathbf{A}|\delta$, it holds that

$$\begin{aligned} &\left| \|\widehat{Q}_{k+1} - T\widetilde{Q}_k\|_n^2 - \|\widehat{Q}_{k+1} - T\widetilde{Q}_k\|_\sigma^2 \right| \\ &\leq \sqrt{|\mathbf{A}|} \cdot \left(16Q_{\max}(C_k + 1) \sqrt{\frac{2 \log(2d)}{n}} + 4Q_{\max}^2 \sqrt{\frac{2 \log(2c(C_k + 1)^2/\delta)}{n}} \right) \\ &\quad + |\mathbf{A}| \cdot \sqrt{\frac{\log(2/\delta)}{n}} \cdot 4Q_{\max}^2 \\ &\leq \sqrt{|\mathbf{A}|} \cdot \left(16Q_{\max}(B_{k+1} + 1) \sqrt{\frac{2 \log(2d)}{n}} + 4Q_{\max}^2 \sqrt{\frac{2 \log(2c(B_{k+1} + 1)^2/\delta)}{n}} \right) \end{aligned}$$

$$+ |\mathbf{A}| \cdot \sqrt{\frac{\log(2/\delta)}{n}} \cdot 4Q_{\max}^2 := \Delta_2 \quad (32)$$

where $C_k = 4 \sum_{i=1}^N \gamma(f_k^i(\cdot, a_i)) \leq B_{k+1}$. After summing up Equation (29) and Equation (32), we know that with probability at least $1 - 4|\mathbf{A}|\delta$ over sampling, the following two inequalities hold simultaneously:

$$\begin{aligned} & \left| \|\tilde{Q}_{k+1} - T\tilde{Q}_k\|_n^2 - \|\tilde{Q}_{k+1} - T\tilde{Q}_k\|_\sigma^2 \right| \leq |\mathbf{A}| \cdot \sqrt{\frac{\log(2/\delta)}{n}} \cdot 4Q_{\max}^2 \\ & + \sqrt{|\mathbf{A}|} \cdot \left(16Q_{\max}(B_{k+1} + 1) \sqrt{\frac{2\log(2d)}{n}} + 4Q_{\max}^2 \sqrt{\frac{2\log(2c(B_{k+1} + 1)^2/\delta)}{n}} \right) := \Delta_1, \\ & \left| \|\hat{Q}_{k+1} - T\tilde{Q}_k\|_n^2 - \|\hat{Q}_{k+1} - T\tilde{Q}_k\|_\sigma^2 \right| \leq |\mathbf{A}| \cdot \sqrt{\frac{\log(2/\delta)}{n}} \cdot 4Q_{\max}^2 \\ & + \sqrt{|\mathbf{A}|} \cdot \left(16Q_{\max}(B_{k+1} + 1) \sqrt{\frac{2\log(2d)}{n}} + 4Q_{\max}^2 \sqrt{\frac{2\log(2c(B_{k+1} + 1)^2/\delta)}{n}} \right) := \Delta_2 \end{aligned}$$

Then: under the events above, by the definition of \tilde{Q}_{k+1} , we have:

$$\|\tilde{Q}_{k+1} - T\tilde{Q}_k\|_\sigma^2 \leq \|\tilde{Q}_{k+1} - T\tilde{Q}_k\|_n^2 + \Delta_1 \leq \|\hat{Q}_{k+1} - T\tilde{Q}_k\|_n^2 + \Delta_1 \leq \|\hat{Q}_{k+1} - T\tilde{Q}_k\|_\sigma^2 + \Delta_1 + \Delta_2.$$

Note that for any decomposable continuous function $f \in \mathcal{C}^{\oplus N}$:

$$\|f - T\tilde{Q}_k\|_\sigma^2 = \left\| f - \text{Proj} \left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma \right) \right\|_\sigma^2 + \left\| \text{Proj} \left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma \right) - T\tilde{Q}_k \right\|_\sigma^2.$$

Therefore, we conclude that:

$$\left\| \tilde{Q}_{k+1} - \text{Proj} \left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma \right) \right\|_\sigma^2 \leq \left\| \hat{Q}_{k+1} - \text{Proj} \left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma \right) \right\|_\sigma^2 + \Delta_1 + \Delta_2. \quad (33)$$

We have already obtained upper bound for all the three terms. After adding them up, we obtain the following bound:

$$\begin{aligned} & \left\| \tilde{Q}_{k+1} - \text{Proj} \left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma \right) \right\|_\sigma^2 \leq \frac{B_{k+1}^2}{M} + 8|\mathbf{A}| \cdot \frac{\log(2/\delta)}{n} \cdot Q_{\max}^2 \\ & + \sqrt{|\mathbf{A}|} \cdot \left(16Q_{\max}(2B_{k+1} + 2) \sqrt{\frac{2\log(2d)}{n}} + 8Q_{\max}^2 \sqrt{\frac{2\log(2c(B_{k+1} + 1)^2/\delta)}{n}} \right), \end{aligned}$$

holds with probability at least $1 - 4|\mathbf{A}|\delta$ over the sampling. In the next step, we are going to upper bound the l_∞ norm of the function differences above. Notice that for a two-layer ReLU network \tilde{Q}_{k+1} , its Lipschitz constant can be upper bounded by its path norm, which is because the ReLU activation function $\sigma(\cdot)$ itself is 1-Lipschitz continuous. Denote L_{k+1} as the Lipschitz constant of $\tilde{Q}_{k+1} - \text{Proj} \left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma \right)$ on which we will analyze later. Notice that if f is a continuous function with Lipschitz constant L , there is a relation on its l_2 norm and its l_∞ norm.

Lemma 4 Suppose $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a continuous function with Lipschitz constant L , then we have:

$$\|f\|_2^2 \geq \frac{\|f\|_\infty^{d+2} \cdot \pi^{d/2}}{3L^d \cdot d^2 \Gamma\left(\frac{d}{2} + 1\right)}.$$

Proof [Proof of Lemma 4] By the definition of $\|f\|_\infty$, we know that for $\forall \varepsilon > 0$, there exists $x_0 \in \mathbb{R}^d$ such that $|f(x_0)| > \|f\|_\infty - \varepsilon$. Then for other $x \in \mathbb{R}^d$, it holds that

$$|f(x)| \geq \min(0, |f(x_0)| - L\|x - x_0\|),$$

since $|f(x)| \geq |f(x_0)| - |f(x_0) - f(x)| \geq |f(x_0)| - L\|x - x_0\|$. Therefore, the l_2 norm of f can be lower bounded by:

$$\begin{aligned} \|f\|_2^2 &\geq \int_{\mathbb{R}^d} \min(0, |f(x_0)| - L\|x - x_0\|)^2 dx = \int_{B(x_0, |f(x_0)|/L)} (|f(x_0)| - L\|x - x_0\|)^2 dx \\ &= \int_0^{|f(x_0)|/L} r^{d-1} \cdot (|f(x_0)| - Lr)^2 dr \cdot \frac{d\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)} \\ &= \frac{|f(x_0)|^{d+2}}{L^d} \cdot \frac{2\pi^{d/2}}{(d+1)(d+2)\Gamma(\frac{d}{2} + 1)} \geq \frac{|f(x_0)|^{d+2} \cdot \pi^{d/2}}{3L^d \cdot d^2\Gamma(\frac{d}{2} + 1)}. \end{aligned}$$

After making $\varepsilon \rightarrow 0$, we can conclude that:

$$\|f\|_2^2 \geq \frac{\|f\|_\infty^{d+2} \cdot \pi^{d/2}}{3L^d \cdot d^2\Gamma(\frac{d}{2} + 1)}.$$

■

With the lemma above, we can finally get the upper bound for $\tilde{Q}_{k+1} - \text{Proj}(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma)$. Assume that the density function of distribution σ over $\mathbf{S} \times \mathbf{A}$ has a universal lower bound c_σ , then:

$$\left\| \tilde{Q}_{k+1} - \text{Proj}(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma) \right\|_\sigma^2 \geq c_\sigma^2 \cdot \left\| \tilde{Q}_{k+1} - \text{Proj}(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma) \right\|_2^2, \quad (34)$$

which leads to:

$$\begin{aligned} \left\| \tilde{Q}_{k+1} - \text{Proj}(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma) \right\|_\infty^{d+2} &\leq \frac{3L_{k+1}^d \cdot d^2\Gamma(\frac{d}{2} + 1)}{c_\sigma^2 \cdot \pi^{d/2}} \cdot \left[\frac{B_{k+1}^2}{M} + 8|\mathbf{A}| \cdot \frac{\log(2/\delta)}{n} \cdot Q_{\max}^2 \right. \\ &\quad \left. + \sqrt{|\mathbf{A}|} \cdot \left(16Q_{\max}(2B_{k+1} + 2) \sqrt{\frac{2\log(2d)}{n}} + 8Q_{\max}^2 \sqrt{\frac{2\log(2c(B_{k+1} + 1)^2/\delta)}{n}} \right) \right]. \quad (35) \end{aligned}$$

Notice that \tilde{Q}_{k+1} is a two-layer ReLU network with its path norm: $\max_{\mathbf{a}} \|\tilde{Q}_{k+1}(\cdot, \mathbf{a})\|_P \leq B_{k+1}$. For a two-layer ReLU network, its Lipschitz constant can be upper bounded by its path norm since its activation function $\sigma(\cdot)$ is 1-Lipschitz continuous, which leads to the fact that:

$$L_{k+1} = \text{Lip}(\tilde{Q}_{k+1}) \leq \|\tilde{Q}_{k+1}(\cdot, \mathbf{a})\|_P \leq B_{k+1}.$$

Next, we are going to determine the choice of the sequence of path norm upper bound $\{B_k\}$. Since $\tilde{Q}_0 \equiv 0$, we only need $B_0 \geq 0$. From the proof above, we have already known that the sequence $\{B_k\}$ needs to satisfy:

$$B_{k+1} \geq 4N \cdot \max_{i, a_i} \gamma(f_k^i(\cdot, a_i)),$$

where: $\text{Proj}\left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma\right) = f_k^1(\gamma_1, a_1) + f_k^2(\gamma_2, a_2) + \dots + f_k^N(\gamma_N, a_N)$. Notice that when $\tilde{Q}_k \in \mathcal{F}(B_k)^{\oplus N}$, we have: for $\forall i \in [N], a_i \in \mathcal{A}$,

$$\|f_k^i(\cdot, a_i)\|_\infty \leq 2\|T\tilde{Q}_k\|_\infty \leq 2(R_{\max} + \gamma\|\tilde{Q}_k\|_\infty) \leq 2(R_{\max} + \gamma\|\tilde{Q}_k\|_P) \leq 2(R_{\max} + N\gamma B_k).$$

Therefore, we only need to make $B_{k+1} = 4N \cdot 2c(R_{\max} + N\gamma B_k)$ where c is the upper bound of ratio between the spectral norm and the l_∞ norm of a continuous function defined on $[0, 1]^d$, which is a pure constant. So that we can guarantee that once $\tilde{Q}_k \in \mathcal{F}(B_k)^{\oplus N}$, we have $B_{k+1} \geq 4N \cdot \max_{i, a_i} \gamma(f_k^i(\cdot, a_i))$ holds. When the discount ratio γ is small enough, we can make

$$B_k = \frac{8NcR_{\max}}{1 - 4N^2\gamma} := B > 0, \quad \forall k \in [K].$$

Now, the Equation (35) becomes: with probability $p \geq 1 - 4|\mathbf{A}|\delta$,

$$\begin{aligned} \left\|\tilde{Q}_{k+1} - \text{Proj}\left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma\right)\right\|_\infty^{d+2} &\leq \frac{3B^d \cdot d^2 \Gamma\left(\frac{d}{2} + 1\right)}{c_\sigma^2 \cdot \pi^{d/2}} \cdot \left[\frac{B^2}{M} + 8|\mathbf{A}| \cdot \frac{\log(2/\delta)}{n} \cdot Q_{\max}^2\right. \\ &\quad \left.+ \sqrt{|\mathbf{A}|} \cdot \left(16Q_{\max}(2B + 2)\sqrt{\frac{2\log(2d)}{n}} + 8Q_{\max}^2\sqrt{\frac{2\log(2c(B + 1)^2/\delta)}{n}}\right)\right] \end{aligned} \quad (36)$$

holds for $\forall k \in [K]$, which leads to the conclusion that:

$$\begin{aligned} \varepsilon_{\max} := \max_{k \in [K]} \left\|\tilde{Q}_{k+1} - \text{Proj}\left(T\tilde{Q}_k, \mathcal{C}^{\oplus N}, \|\cdot\|_\sigma\right)\right\|_\infty &\leq c_1 B d \cdot \left[\frac{B^2}{M} + 8|\mathbf{A}| \cdot \frac{\log(2/\delta)}{n} \cdot Q_{\max}^2\right. \\ &\quad \left.+ \sqrt{|\mathbf{A}|} \cdot \left(16Q_{\max}(2B + 2)\sqrt{\frac{2\log(2d)}{n}} + 8Q_{\max}^2\sqrt{\frac{2\log(2c(B + 1)^2/\delta)}{n}}\right)\right]^{\frac{1}{d+2}}. \end{aligned} \quad (37)$$

It comes to our conclusion after replacing δ with $\frac{\delta}{4|\mathbf{A}|}$.

Appendix E. Existing Understanding on 2-layer ReLU Networks

In this section, we introduce several important properties on neural networks, mainly on their approximation properties and generalization bounds. First, we study the two-layer ReLU networks.

E.1. Approximation Properties

In this section, I will focus on the approximation of 2-layer ReLU networks to a target function. Assume $f^* : \Omega \rightarrow \mathbb{R}$ be the target function, where $\Omega = [-1, 1]^d$, and $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be the training set. Here the data points $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d samples drawn from an underlying distribution π with $\text{supp}(\pi) \subset \Omega$, and $y_i = f^*(\mathbf{x}_i)$. We aim to recover f^* by fitting S using a two-layer fully connected neural network with ReLU (rectified linear units) activation:

$$f(x; \theta) = \sum_{k=1}^m a_k \sigma(\mathbf{b}_k \cdot \mathbf{x} + c_k)$$

Here, function $\sigma(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ denotes the ReLU activation: $\sigma(t) = \max(0, t)$, $\mathbf{b}_k \in \mathbb{R}^d$ and the whole parameter set $\theta = \{(a_k, \mathbf{b}_k, c_k)\}_{k=1}^m$ is to be learned, and m is the width of the network. In order to control the magnitude of learned network. We use the following scale-invariant norm.

Definition E.1 (Path norm ([Neyshabur et al., 2015](#))) For a two-layer ReLU network, the path norm is defined as:

$$\|\theta\|_P = \sum_{k=1}^m |a_k|(\|\mathbf{b}_k\|_1 + |c_k|)$$

Definition E.2 (Spectral norm) Given $f \in L^2(\Omega)$, denote by $F \in L^2(\mathbb{R}^d)$ an extension of f to \mathbb{R}^d . Let \hat{F} be the Fourier transform of F , then:

$$f(\mathbf{x}) = \int_{\mathbb{R}^d} e^{i\langle \mathbf{x}, \mathbf{w} \rangle} \hat{F}(\omega) d\omega \quad \forall \mathbf{x} \in \Omega$$

We define the spectral norm of f by:

$$\gamma(f) = \inf_{F \in L^2(\mathbb{R}^d), F|_{\Omega} = f|_{\Omega}} \int_{\mathbb{R}^d} \|\omega\|_1^2 \cdot |\hat{F}(\omega)| d\omega$$

We also define $\hat{\gamma}(f) = \max\{\gamma(f), 1\}$.

Assumption E.1 We consider target functions that are bounded and have finite spectral norm.

$$F_s = L^2(\Omega) \cap \{f(\mathbf{x}) : \Omega \rightarrow \mathbb{R} | \gamma(f) < \infty, \|f\|_{\infty} \leq 1\}$$

We assume that $f^* \in F_s$.

Since $\|f^*\|_{\infty} \leq 1$, we can truncate the network by $\tilde{f}(x) = \min\{|f(x)|, 1\} \text{sign}(f)$. By an abuse of notation, in the following we still use $f(x)$ to denote $\tilde{f}(x)$. Our goal is to minimize the generalization error (also known as population risk).

$$L(\theta) = \mathbb{E}_{\mathbf{x}, y} [l(f(\mathbf{x}; \theta), y)]$$

However, practically, we only have to minimize the empirical risk

$$\hat{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^n l(f(\mathbf{x}_i, \theta), y_i)$$

Here, the generalization gap is defined as the difference between expected and empirical risk. The loss function is $l(y_1, y_2) = (y_1 - y_2)^2$ and that's why we analyze only regressive problems.

According to ([Barron, 1993](#)), ([Breiman, 1993](#)) and ([Klusowski and Barron, 2016](#)), we can obtain the following approximation properties.

Lemma E.1 For any $F \in F_s$, one has the integral representation:

$$f(\mathbf{x}) - f(0) - \mathbf{x} \cdot \nabla f(0) = v \int_{\{-1, 1\} \times [0, 1] \times \mathbb{R}^d} h(\mathbf{x}; z, t, \mathbf{w}) dp(z, t, \mathbf{w})$$

where:

$$p(z, t, \mathbf{w}) = |\hat{f}(\mathbf{w})| \|\mathbf{w}\|_1^2 |\cos(\|\mathbf{w}\|_1 t - zb(\mathbf{w}))| / v$$

$$s(z, t, \mathbf{w}) = -\text{sign}(\cos(\|\mathbf{w}\|_1 t - zb(\mathbf{w})))$$

$$h(\mathbf{x}, z, t, \mathbf{w}) = s(z, t, \mathbf{w})(z\mathbf{x} \cdot \mathbf{w} / \|\mathbf{w}\|_1 - t)_+$$

v is the normalization constant such that $\int p(z, t, \mathbf{w}) dz dt d\mathbf{w} = 1$, which satisfies $v \leq 2\gamma(f)$.

Proof Since $f \in L^2(\mathbb{R}^d)$, we have:

$$f(\mathbf{x}) - f(0) - \mathbf{x} \cdot \nabla f(0) = \int_{\mathbb{R}^d} (e^{i\mathbf{w} \cdot \mathbf{x}} - i\mathbf{w} \cdot \mathbf{x} - 1) \hat{f}(\mathbf{w}) d\mathbf{w}$$

Note that the identity

$$-\int_0^c [(z-s)_+ e^{is} + (-z-s)_+ e^{-is}] ds = e^{iz} - iz - 1$$

holds when $|z| \leq c$. Choosing $c = \|\mathbf{w}\|_1$, $z = \mathbf{w} \cdot \mathbf{x}$, we have;

$$|z| \leq \|\mathbf{w}\|_1 \|\mathbf{x}\|_\infty \leq c$$

Let $s = \|\mathbf{w}\|_1 t$, $0 \leq t \leq 1$, and $\hat{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|_1$, we have:

$$-\|\mathbf{w}\|_1^2 \int_0^1 [(\hat{\mathbf{w}} \cdot \mathbf{x} - t)_+ e^{i\|\mathbf{w}\|_1 t} + (-\hat{\mathbf{w}} \cdot \mathbf{x} - t)_+ e^{-i\|\mathbf{w}\|_1 t}] dt = e^{i\mathbf{w} \cdot \mathbf{x}} - i\mathbf{w} \cdot \mathbf{x} - 1.$$

Let $\hat{f}(\mathbf{w}) = e^{ib(\mathbf{w})} |f(\mathbf{w})|$, according to the two equations above:

$$f(\mathbf{x}) - f(0) - \mathbf{x} \cdot \nabla f(0) = \int_{\mathbb{R}^d} \int_0^1 g(t, \mathbf{w}) dt d\mathbf{w},$$

where:

$$g(t, \mathbf{w}) = -\|\mathbf{w}\|_1^2 |\hat{f}(\mathbf{w})| \cdot [(\hat{\mathbf{w}} \cdot \mathbf{x} - t)_+ \cos(\|\mathbf{w}\|_1 t + b(\mathbf{w})) + (-\hat{\mathbf{w}} \cdot \mathbf{x} - t)_+ \cos(\|\mathbf{w}\|_1 t - b(\mathbf{w}))].$$

Consider a density on $\{0, 1\} \times [0, 1] \times \mathbb{R}^d$ defined by:

$$p(z, t, \mathbf{w}) = |\hat{f}(\mathbf{w})| \|\mathbf{w}\|_1^2 |\cos(\|\mathbf{w}\|_1 t - zb(\mathbf{w}))| / v$$

where the normalized constant v is given by

$$v = \int_{\mathbb{R}^d} \int_0^1 (|\cos(\|\mathbf{w}\|_1 t + b(\mathbf{w}))| + |\cos(\|\mathbf{w}\|_1 t - b(\mathbf{w}))|) dt d\mathbf{w}$$

Since $f \in F_s$, therefore: $v \leq 2\gamma(f) < +\infty$. So, this density is well-defined. To simplify the notations, denote:

$$s(z, t, \mathbf{w}) = -\text{sign}(\cos(\|\mathbf{w}\|_1 t - zb(\mathbf{w}))), \quad h(\mathbf{x}; z, t, \mathbf{w}) = s(z, t, \mathbf{w}) (z\hat{\mathbf{w}} \cdot \mathbf{x} - t)_+$$

Then we have

$$f(\mathbf{x}) - f(0) - \mathbf{x} \cdot \nabla f(0) = v \int_{\{-1, 1\} \times [0, 1] \times \mathbb{R}^d} h(\mathbf{x}; z, t, \mathbf{w}) dp(z, t, \mathbf{w}).$$

■

For simplicity, in the following part, we assume $f(0) = 0, \nabla f(0) = 0$ because according to the equation above, we can use $f(\mathbf{x}) - f(0) - (\mathbf{x} \cdot \nabla f(0))_+ + (-\mathbf{x} \cdot \nabla f(0))_+$ to replace $f(\mathbf{x})$. This

is a Monte-Carlo scheme. Therefore, we take m samples $T_m = \{(z_1, t_1, \mathbf{w}_1), \dots, (z_m, t_m, \mathbf{w}_m)\}$ with (z_i, t_i, \mathbf{w}_i) randomly drawn from the probability density function $p(z, t, \mathbf{w})$, and consider the empirical average $\hat{f}_m(\mathbf{x}) = \frac{v}{m} \sum_{k=1}^m h(\mathbf{x}; z_k, t_k, \mathbf{w}_k)$, which is exactly a two-layer ReLU network of width m . The central limit theorem tells us that the approximation error:

$$\mathbb{E}_{(z,t,\mathbf{w})} [h(\mathbf{x}; z, t, \mathbf{w})] - \frac{1}{m} \sum_{k=1}^m h(\mathbf{x}; z_k, t_k, \mathbf{w}_k) \approx \sqrt{\frac{\text{Var}_{(z,t,\mathbf{w})}[h(\mathbf{x}; z, t, \mathbf{w})]}{m}}$$

So what we have to do is bounding the variance on the right-hand side of the equation above.

Theorem 5 *For any distribution π with $\text{supp}(\pi) \subset \Omega$ and any $f \in F_s$, there exists a two-layer network $f(\mathbf{x}; \tilde{\theta})$ of width m such that:*

$$\mathbb{E}_{\mathbf{x} \sim \pi} |f(\mathbf{x}) - f(\mathbf{x}; \tilde{\theta})|^2 \leq \frac{16\gamma^2(f)}{m}$$

Furthermore, the path norm of the parameter $\tilde{\theta}$ can be bounded by the spectral norm of the target function: $\|\tilde{\theta}\|_P \leq 4\gamma(f)$.

Proof Let $\hat{f}_m(\mathbf{x}) = \frac{v}{m} \sum_{k=1}^m h(\mathbf{x}; z_k, t_k, \mathbf{w}_k)$ be the Monte-Carlo estimator, then:

$$\begin{aligned} \mathbb{E}_{T_m} \mathbb{E}_{\mathbf{x}} |f(\mathbf{x}) - \hat{f}_m(\mathbf{x})|^2 &= \mathbb{E}_{\mathbf{x}} \mathbb{E}_{T_m} |f(\mathbf{x}) - \hat{f}_m(\mathbf{x})|^2 \\ &= \frac{v^2}{m} \mathbb{E}_{\mathbf{x}} (\mathbb{E}_{(z,t,\mathbf{w})} [h^2(\mathbf{x}; z, t, \mathbf{w})] - f^2(\mathbf{x})) \\ &\leq \frac{v^2}{m} \mathbb{E}_{\mathbf{x}} \mathbb{E}_{(x,t,\mathbf{w})} [h^2(\mathbf{x}; z, t, \mathbf{w})] \end{aligned} \quad (38)$$

For any fixed \mathbf{x} , the variance above can be bounded as:

$$\mathbb{E}_{(x,t,\mathbf{w})} [h^2(\mathbf{x}; z, t, \mathbf{w})] \leq \mathbb{E}_{(x,t,\mathbf{w})} [(z\hat{\mathbf{w}} \cdot \mathbf{x} - t)_+^2] \leq \mathbb{E}_{(x,t,\mathbf{w})} [(|\hat{\mathbf{w}} \cdot \mathbf{x}| + t)^2] \leq 4$$

Hence we have:

$$\mathbb{E}_{T_m} \mathbb{E}_{\mathbf{x}} |f(\mathbf{x}) - \hat{f}_m(\mathbf{x})|^2 \leq \frac{4v^2}{m} \leq \frac{16\gamma^2(f)}{m}$$

So we get the following conclusion: there exists a set of T_m , such that: $\mathbb{E}_{\mathbf{x}} |f - \hat{f}_m|^2 \leq \frac{16\gamma^2(f)}{m}$. Notice the special structure of the Monte-Carlo estimator, we have: $|a_k| = \frac{v}{m}$, $\|\mathbf{b}_k\|_1 = 1$, $|c_k| \leq q$. Therefore, $\|\tilde{\theta}\|_P \leq 2v \leq 4\gamma(f)$. \blacksquare

E.2. Generalization Properties

Definition E.3 (Rademacher Complexity) *Let H be a hypothesis space. The Rademacher Complexity of H with respect to samples $S = (z_1, \dots, z_n)$ is defined as:*

$$\hat{R}(H) = \frac{1}{n} \mathbb{E}_{\xi} \left[\sup_{h \in H} \sum_{i=1}^n h(z_i) \xi_i \right]$$

where $\{\xi_i\}_{i=1}^n$ are independent random variables with probability $P(\xi_i = 1) = P(\xi_i = -1) = \frac{1}{2}$

Before coming to the estimation of Rademacher Complexity, we need to introduce some fundamental properties.

E.2.1. BASIC PROPERTIES ABOUT RADEMACHER COMPLEXITY

Lemma 6 For any $A \in \mathbb{R}^m$, scalar $c \in \mathbb{R}$, and vector $\mathbf{a}_0 \in \mathbb{R}^m$, we have:

$$R(\{c\mathbf{a} + \mathbf{a}_0 : \mathbf{a} \in A\}) = |c|R(A)$$

Next, we are going to state several more important lemmas about Rademacher Complexity since they explain that the Rademacher Complexity of a finite set grows logarithmically with the size of the set.

Lemma 7 (Massart Lemma) Let $A = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N\}$ be a finite set of vectors in \mathbb{R}^m . Then:

$$R(A) \leq \max_{\mathbf{a} \in A} \|\mathbf{a} - \bar{\mathbf{a}}\| \cdot \frac{\sqrt{2 \log N}}{m}$$

Here: $\bar{\mathbf{a}} = \frac{1}{N} \sum_{i=1}^N \mathbf{a}_i$ is the average of all vectors in A .

Proof According to Lemma 6, we can assume $\bar{\mathbf{a}} = 0$ with loss of generality. Let $\lambda > 0$ and let $A' = \{\lambda \mathbf{a}_1, \lambda \mathbf{a}_2, \dots, \lambda \mathbf{a}_N\}$ where λ is a positive scalar which remains to be determined. Then we calculate the upper bound of Rademacher Complexity of A' .

$$\begin{aligned} mR(A') &= \mathbb{E}_{\sigma} \left[\max_{\mathbf{a} \in A'} \langle \sigma, \mathbf{a} \rangle \right] = \mathbb{E}_{\sigma} \left[\log \left(\max_{\mathbf{a} \in A'} e^{\langle \sigma, \mathbf{a} \rangle} \right) \right] \\ &\leq \mathbb{E}_{\sigma} \left[\log \left(\sum_{\mathbf{a} \in A'} e^{\langle \sigma, \mathbf{a} \rangle} \right) \right] \leq \log \left(\mathbb{E}_{\sigma} \left[\sum_{\mathbf{a} \in A'} e^{\langle \sigma, \mathbf{a} \rangle} \right] \right) \\ &= \log \left(\sum_{\mathbf{a} \in A'} \prod_{i=1}^m \mathbb{E}_{\sigma_i} [e^{\sigma_i a_i}] \right) \end{aligned} \quad (39)$$

Since:

$$\mathbb{E}_{\sigma_i} [e^{\sigma_i a_i}] = \frac{1}{2} (\exp(a_i) + \exp(-a_i)) \leq \exp \left(\frac{a_i^2}{2} \right).$$

Therefore:

$$\begin{aligned} mR(A') &\leq \log \left(\sum_{\mathbf{a} \in A'} \prod_{i=1}^m \exp \left(\frac{a_i^2}{2} \right) \right) = \log \left(\sum_{\mathbf{a} \in A'} \exp(\|\mathbf{a}\|_2^2 / 2) \right) \\ &\leq \log(|A'|) + \max_{\mathbf{a} \in A'} (\|\mathbf{a}\|_2^2 / 2) \end{aligned} \quad (40)$$

According to the definition of A' and Lemma 6, we know that $R(A') = \lambda R(A)$. Then:

$$R(A) \leq \frac{\log(|A|) + \lambda^2 \max_{\mathbf{a} \in A} (\|\mathbf{a}\|_2^2 / 2)}{\lambda m}$$

Finally, set the optimal

$$\lambda = \sqrt{\frac{2 \log |A|}{\max_{\mathbf{a} \in A} (\|\mathbf{a}\|_2^2 / 2)}}$$

and we can come to our conclusion. ■

The following shows that composing A with a Lipschitz function will not blow up the Rademacher Complexity. And this is one of the most important conclusions about Rademacher Complexity.

Lemma 8 (Contraction Lemma) For each $i \in [m]$, let $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$ be a ρ -Lipschitz function, which means for all $x_1, x_2 \in \mathbb{R}$, we have:

$$|\phi_i(x_1) - \phi_i(x_2)| \leq \rho |x_1 - x_2|$$

For $\mathbf{a} \in \mathbb{R}^m$, let $\rho(\mathbf{a})$ denote the vector $(\phi_1(a_1), \phi_2(a_2), \dots, \phi_m(a_m))$ and $\phi \circ A = \{\rho(\mathbf{a}) : \mathbf{a} \in A\}$. Then:

$$R(\phi \circ A) \leq \rho R(A).$$

Proof For simplicity, we can assume $\rho = 1$. Otherwise, we can replace ϕ with $\phi' = \frac{1}{\rho}\phi$ and then use Lemma 6 to prove our conclusion. Let:

$$A_i = \{(a_1, \dots, a_{i-1}, \phi_i(a_i), a_{i+1}, \dots, a_m) : \mathbf{a} \in A\}$$

It is obvious that we only have to prove that for any set A and all i , there holds: $R(A_i) \leq R(A)$. Without loss of generality, we will prove that latter claim for $i = 1$ and to simplify notation, we omit the subscription of ϕ_1 . We have:

$$\begin{aligned} mR(A_1) &= \mathbb{E}_{\sigma} \left[\sup_{\mathbf{a} \in A_1} \sum_{i=1}^m \sigma_i a_i \right] = \mathbb{E}_{\sigma} \left[\sup_{\mathbf{a} \in A} \sigma_1 a_1 + \sum_{i=2}^m \sigma_i a_i \right] \\ &= \frac{1}{2} \mathbb{E}_{\sigma_2, \dots, \sigma_m} \left[\sup_{\mathbf{a} \in A} \left(\phi(a_1) + \sum_{i=2}^m \sigma_i a_i \right) + \sup_{\mathbf{a} \in A} \left(-\phi(a_1) + \sum_{i=2}^m \sigma_i a_i \right) \right] \\ &= \frac{1}{2} \mathbb{E}_{\sigma_2, \dots, \sigma_m} \left[\sup_{\mathbf{a}, \mathbf{a}' \in A} \left(\phi(a_1) - \phi(a'_1) + \sum_{i=2}^m \sigma_i a_i + \sum_{i=2}^m \sigma_i a'_i \right) \right] \\ &\leq \frac{1}{2} \mathbb{E}_{\sigma_2, \dots, \sigma_m} \left[\sup_{\mathbf{a}, \mathbf{a}' \in A} \left(|a_1 - a'_1| + \sum_{i=2}^m \sigma_i a_i + \sum_{i=2}^m \sigma_i a'_i \right) \right] \end{aligned} \quad (41)$$

where in the last inequality, we used the Lipschitz condition of ϕ . Next, we note that the absolute sign can be erased because both \mathbf{a} and \mathbf{a}' are from the same set A . Therefore,

$$mR(A_1) \leq \frac{1}{2} \mathbb{E}_{\sigma_2, \dots, \sigma_m} \left[\sup_{\mathbf{a}, \mathbf{a}' \in A} \left(a_1 - a'_1 + \sum_{i=2}^m \sigma_i a_i + \sum_{i=2}^m \sigma_i a'_i \right) \right]$$

But using the same inequalities in Equation (41), it is easy to see that the right-hand side is equivalent to the occasion where $\phi_1 = \text{Id}$. Therefore, the right size exactly equals $mR(A)$, which comes to our conclusion. \blacksquare

Lemma 9 Let $S = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ be vectors in \mathbb{R}^n . Then, for the hypothesis class $H_1 = \{x \mapsto \langle \mathbf{w}, x \rangle : \|\mathbf{w}\|_2 \leq 1\}$, we have:

$$R(H_1 \circ S) \leq \max_i \|\mathbf{x}_i\|_{\infty} \sqrt{\frac{2 \log(2n)}{m}}$$

Proof Using Holder's Inequality, we know that for any \mathbf{w}, \mathbf{v} , we have: $\langle \mathbf{w}, \mathbf{v} \rangle \leq \|\mathbf{w}\|_1 \|\mathbf{v}\|_\infty$. Therefore:

$$\begin{aligned} mR(H_1 \circ S) &= \mathbb{E}_\sigma \left[\sup_{\mathbf{a} \in H_1 \circ S} \sum_{i=1}^m \sigma_i a_i \right] = \mathbb{E}_\sigma \left[\sup_{\mathbf{w}: \|\mathbf{w}\|_1 \leq 1} \sum_{i=1}^m \sigma_i \langle \mathbf{w}, \mathbf{x}_i \rangle \right] \\ &= \mathbb{E}_\sigma \left[\sup_{\mathbf{w}: \|\mathbf{w}\|_1 \leq 1} \langle \mathbf{w}, \sum_{i=1}^m \sigma_i \mathbf{x}_i \rangle \right] \leq \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \right\|_\infty \right] \end{aligned} \quad (42)$$

For $j \in [n]$, let $\mathbf{v}_j = (x_{1,j}, \dots, x_{m,j}) \in \mathbb{R}^m$. Note that: $\|\mathbf{v}_j\|_2 \leq \sqrt{m} \max_i \|\mathbf{x}_i\|_\infty$. Let $V = \{\mathbf{v}_1, \dots, \mathbf{v}_n, -\mathbf{v}_1, \dots, -\mathbf{v}_n\}$. The right-hand side of Equation 9 is $mR(V)$. Using Mas-sart Lemma (Lemma 7) we have that:

$$R(V) \leq \max_i \|\mathbf{x}_i\|_\infty \sqrt{2 \log(2n)/m}$$

■

Lemma 10 Assume that for all data points $s \in S$ and $h \in H$ where H is a hypothesis set, we all have $l(h, z) \leq c$. Then: with probability of at least $1 - \delta$, for $\forall h \in H$,

$$L_D(h) - L_S(h) \leq 2 \mathbb{E}_{S' \sim D^m} R(l \circ H \circ S') + c \sqrt{\frac{2 \ln(2/\delta)}{m}}$$

According to (E et al., 2018; Don et al., 2020), we can finally get an upper bound of Rademacher Complexity of 2-layer ReLU networks:

Lemma 11 Denote $F_Q = \{f_m(x; \theta) : \mathbb{R}^D \rightarrow \mathbb{R} \mid \|\theta\|_P \leq Q\}$ be the set of two-layer ReLU networks with path norm bounded by Q , then we can bound its Rademacher Complexity.

$$R(F_Q) \leq 2Q \sqrt{\frac{2 \log(2D)}{n}}$$

Proof To simplify the proof, we can assume $c_k = 0$ without loss of generality. Otherwise, we can define $\mathbf{b}_k = (\mathbf{b}_k^T, c_k)^T$, $\mathbf{x} = (\mathbf{x}, 1)^T$.

$$\begin{aligned} n\hat{R}(F_Q) &= \mathbb{E}_\xi \left[\sup_{\|\theta\|_P \leq Q} \sum_{i=1}^n \xi_i \sum_{k=1}^m a_k \|\mathbf{b}_k\|_1 \sigma(\hat{\mathbf{b}}_k^T \mathbf{x}_i) \right] \\ &\leq \mathbb{E}_\xi \left[\sup_{\|\theta\|_P \leq Q, \|\mathbf{u}_k\|_1=1} \sum_{i=1}^n \xi_i \sum_{k=1}^m a_k \|\mathbf{b}_k\|_1 \sigma(\mathbf{u}_k^T \mathbf{x}_i) \right] \\ &\leq \mathbb{E}_\xi \left[\sup_{\|\theta\|_P \leq Q, \|\mathbf{u}_k\|_1=1} \sum_{k=1}^m a_k \|\mathbf{b}_k\|_1 \sum_{i=1}^n \xi_i \sigma(\mathbf{u}_k^T \mathbf{x}_i) \right] \\ &\leq \mathbb{E}_\xi \left[\sup_{\|\theta\|_P \leq Q} \sum_{k=1}^m |a_k \|\mathbf{b}_k\|_1| \sup_{\|\mathbf{u}\|_1=1} \left| \sum_{i=1}^n \xi_i \sigma(\mathbf{u}^T \mathbf{x}_i) \right| \right] \end{aligned}$$

$$\leq Q \mathbb{E}_\xi \left[\sup_{\|\mathbf{u}\|_1=1} \left| \sum_{i=1}^n \xi_i \sigma(\mathbf{u}^T \mathbf{x}_i) \right| \right] \leq Q \mathbb{E}_\xi \left[\sup_{\|\mathbf{u}\|_1 \leq 1} \left| \sum_{i=1}^n \xi_i \sigma(\mathbf{u}^T \mathbf{x}_i) \right| \right]. \quad (43)$$

Due to the symmetry, we have that:

$$\begin{aligned} \mathbb{E}_\xi \left[\sup_{\|\mathbf{u}\|_1 \leq 1} \left| \sum_{i=1}^n \xi_i \sigma(\mathbf{u}^T \mathbf{x}_i) \right| \right] &\leq \mathbb{E}_\xi \left[\sup_{\|\mathbf{u}\|_1 \leq 1} \sum_{i=1}^n \xi_i \sigma(\mathbf{u}^T \mathbf{x}_i) + \sup_{\|\mathbf{u}\|_1 \leq 1} \sum_{i=1}^n -\xi_i \sigma(\mathbf{u}^T \mathbf{x}_i) \right] \\ &= 2 \mathbb{E}_\xi \left[\sup_{\|\mathbf{u}\|_1 \leq 1} \sum_{i=1}^n \xi_i \sigma(\mathbf{u}^T \mathbf{x}_i) \right] \end{aligned} \quad (44)$$

Since the activation function $\sigma(\cdot)$ has Lipschitz constant 1. According to Lemma 8 and Lemma 9, we have:

$$R(F_Q) \leq 2Q \sqrt{\frac{2 \log(2D)}{n}}$$

which comes to our conclusion. \blacksquare

Finally, we can combine Lemma 11 with Lemma 7 and Lemma 8, and obtain the following conclusion, which shows the generalization bound over the 2-layer ReLU networks.

Theorem 12 *Suppose the loss function $l(\cdot, y) = (\cdot - y)^2$ is ρ -Lipschitz continuous and bounded by B . Then with probability at least $1 - \delta$ over the choice of samples, we have:*

$$\sup_{\|f\|_P \leq Q} |L(f) - \hat{L}_n(f)| \leq 4\rho Q \sqrt{\frac{2 \log(2d)}{n}} + B \sqrt{\frac{2 \log(2d/\delta)}{n}}$$

Here:

$$L(f) = \mathbb{E}_{(\mathbf{x}, y) \sim \pi} (f(\mathbf{x}) - y)^2, \quad \hat{L}_n(f) = \mathbb{E}_{\mathbf{x} \in S} (f(\mathbf{x}) - y)^2.$$

Then, by using the union bound, we conclude the following more general result.

Theorem 13 (A posterior generalization bound) *Assume the loss function $l(\cdot, y)$ is ρ -Lipschitz continuous and bounded by B . Then for any $\delta > 0$, with probability at least $1 - \delta$ over the choice the training set S , we have: for any two-layer ReLU network f , it holds that*

$$|L(f) - \hat{L}_n(f)| \leq 4\rho(\|f\|_P + 1) \sqrt{\frac{2 \log(2d)}{n}} + B \sqrt{\frac{2 \log(2c(\|f\|_P + 1)^2/\delta)}{n}}$$

Here: $c = \sum_{k=1}^{+\infty} 1/k^2 = \pi^2/6$.

Proof Consider the decomposition of the full space $\mathcal{F} = \cup_{i=1}^{\infty} \mathcal{F}_i$, where $\mathcal{F}_i = \{f \mid \|f\|_P \leq i\}$. Let $\delta_i = \frac{\delta}{ci^2}$. According to Theorem 12, if we fixed i in advance, then with probability at least $1 - \delta_i$ over the choice of S ,

$$\sup_{\|f\|_P \leq i} |L(f) - \hat{L}_n(f)| \leq 4\rho i \sqrt{\frac{2 \log(2d)}{n}} + B \sqrt{\frac{2 \log(2d/\delta_i)}{n}}$$

So the probability that there exists at least one i to fail the inequality above is at most $\sum_{i=1}^{\infty} \delta_i = \delta$. In other words, with probability at least $1 - \delta$, the inequality above holds for all i . Given any two-layer ReLU network f of width M , let $i_0 = \lceil \|f\|_P \rceil$. Then:

$$\begin{aligned} |L(f) - \hat{L}_n(f)| &\leq 4\rho i_0 \sqrt{\frac{2\log(2d)}{n}} + B \sqrt{\frac{2\log(2ci_0^2/\delta)}{n}} \\ &\leq 4\rho(\|f\|_P + 1) \sqrt{\frac{2\log(2d)}{n}} + B \sqrt{\frac{2\log(2c(\|f\|_P + 1)^2/\delta)}{n}} \end{aligned} \quad (45)$$

which comes to our conclusion. ■