Journal of Machine Learning Research 24 (2023) 1-39 Submitted 7/22; Revised 11/23; Published 12/23

Over-parameterized Deep Nonparametric Regression for Dependent Data with Its Applications to Reinforcement Learning

Xingdong Feng FENG.XINGDONG@MAIL.SHUFE.EDU.CN
School of Statistics and Management
Shanghai University of Finance and Economics

Yuling Jiao YULINGJIAOMATH@WHU.EDU.CN
School of Mathematics and Statistics
Hubei Key Laboratory of Computational Science
Wuhan University

Lican Kang [*] KANGLICAN@WHU.EDU.CN
School of Mathematics and Statistics
Wuhan University

Baqun Zhang ZHANG.BAQUN@MAIL.SHUFE.EDU.CN
School of Statistics and Management
Shanghai University of Finance and Economics

Fan Zhou [†] ZHOUFAN@MAIL.SHUFE.EDU.CN
School of Statistics and Management
Shanghai University of Finance and Economics

Editor: Csaba Szepesvari

Abstract

In this paper, we provide statistical guarantees for over-parameterized deep nonparametric regression in the presence of dependent data. By decomposing the error, we establish non-asymptotic error bounds for deep estimation, which is achieved by effectively balancing the approximation and generalization errors. We have derived an approximation result for Hölder functions with constrained weights. Additionally, the generalization error is bounded by the weight norm, allowing for a neural network parameter number that is much larger than the training sample size. Furthermore, we address the issue of the curse of dimensionality by assuming that the samples originate from distributions with low intrinsic dimensions. Under this assumption, we are able to overcome the challenges posed by high-dimensional spaces. By incorporating an additional error propagation mechanism, we derive oracle inequalities for the over-parameterized deep fitted Q-iteration.

Keywords: Deep reinforcement learning, Low-dimensional Riemannian manifold, Penalized regression, β-mixing

[*] Corresponding author.
[†] Corresponding author.

1 Introduction

Consider the following regression model

$$Y = f_0(X) + \varepsilon, \tag{1}$$

where $Y \in \mathbb{R}$ is the response, $X \in \mathbb{R}^d$ is the covariate, ε is the random error with mean zero, and $f_0 : \mathbb{R}^d \to \mathbb{R}$ denotes the underlying regression function. Our purpose is to estimate f_0 with the over-parameterized ReLU neural networks given observations $\{Z_i\}_{i=1}^n := \{X_i, Y_i\}_{i=1}^n$ which may not be independently and identically distributed (i.i.d.).

In this paper, we consider dependent samples assumed to be strictly stationary β-mixing (Yu, 1994; Antos et al., 2007, 2008; Lazaric et al., 2012). Here, strictly stationarity indicates that the samples admit the same distribution. This assumption generalizes applications of model (1) in i.i.d. cases, which allows for many statistical and machine learning problems including longitudinal data analysis, time series analysis, estimations in stochastic differential equation and reinforcement learning, etc. To estimate f_0 at the population level, we consider the expected risk

$$\mathcal{R}(f) = \mathbb{E}_Z \ \left[\ell(f(X), Y)\right],$$

where $\ell(\cdot, \cdot) : \mathbb{R}^2 \to [0, \infty)$ is the loss function. The global minimizer f^* is defined as

$$f^* = \underset{f}{\operatorname{argmin}}\ \mathcal{R}(f).$$

It can be deduced that $f_0 = f^*$ for certain cases including the mean and quantile regressions where $\mathbb{E}[\varepsilon|X] = 0$ or the conditional quantile of ε given X is zero, respectively. In order to estimate f_0 in (1) from samples, we consider the empirical risk minimization (ERM) problem as given by

$$\hat{f}_n = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \mathcal{R}_n(f) := \frac{1}{n}\sum_{i=1}^n \ell(f(X_i), Y_i), \tag{2}$$

where $\mathcal{F}$ denotes the deep neural network class with certain structures. A goal of our theoretical study is to estimate the excess risk, which is given as

$$\mathcal{R}(\hat{f}_n) - \mathcal{R}\left(f^*\right) := \mathbb{E}_Z \left[\ell(\hat{f}_n(X), Y)\right] - \mathbb{E}_Z \left[\ell(f^*(X), Y)\right].$$

More recently, regression using deep neural network has attracted much attention (Bauer and Kohler, 2019; Kohler and Langer, 2021; Schmidt-Hieber, 2020; Nakada and Imaizumi, 2020; Farrell et al., 2021; Jiao et al., 2021; Suzuki, 2018; Suzuki and Nitanda, 2021; Shen et al., 2021; Fan et al., 2022, among others) in the framework of nonparametric estimation (Stone, 1982; Gyorfi et al., 2002; Tsybakov, 2009). The convergence results given in the aforementioned elegant works provided an under-parameterized statistical guarantee on those deep estimators as the amount of samples is more than that of parameters. However, over-parametrization is one of the key tricks for model training in deep learning, see Jacot et al. (2018); Allen-Zhu et al. (2019); Du et al. (2019); Zou and Gu (2019); Liu et al. (2022); Chizat et al. (2019) and the reference therein. The reason why over-parameterized reinforcement learning works well is not fully understood, and it is still a theoretically fundamental but challenging problem to provide statistical guarantees under over-parameterized regimes in deep learning (Belkin, 2021; Bartlett et al., 2021; Berner et al., 2021).

Many efforts have been made to understand the role of over-parametrization in linear and kernel models (Belkin et al., 2018, 2019a; Hastie et al., 2022; Belkin et al., 2019b; Liang and Rakhlin, 2020; Nakkiran et al., 2020; Bartlett et al., 2020; Tsigler and Bartlett, 2023; Belkin, 2021; Bartlett et al., 2021; Tsigler and Bartlett, 2023). However, Kohler and Krzyzak (2021) gave an negative result by showing that the empirical risk minimization estimator in deep non-parametric regression with over-parametrization can be inconsistent.

Technically, the challenge of theoretical analysis of those over-parameterized deep non-parametric estimators roots in the current bias variance trade-off between the approximation and statistical errors. Modern neural network approximation results use those network parameters such as depth, width and size to bound the approximation error which decreases as the values of these parameters increase (Telgarsky, 2016; Yarotsky, 2017, 2018; Petersen and Voigtlaender, 2018; Zhou, 2020; Shen et al., 2019; Shen, 2020; Lu et al., 2021). The statistical error which measures the supremum of the empirical process indexed by neural network class is bounded by a ratio of a certain complexity measure of the considered neural network, such as the pseudo dimension, to the corresponding sample size using localized methods (Bartlett et al., 2005) (or square root of the sample size using the chaining directly (Van Der Vaart et al., 1996)). Taking deep ReLU neural network as an example, its pseudo dimension is further bounded by its parameter numbers (Bartlett et al., 2019). Hence, only those results for under-parameterization can be established by choosing depth, width and size of neural networks in terms of sample sizes to balance those two errors.

The development of linear regression from low dimensional models to high dimensional ones may provide some insight into generalizing statistical guarantees for deep learning from under-parametrization to over-parametrization. In fact, regularization controlling certain norms of regression coefficients plays an important role in high-dimensional regression. Motivated by this idea, we demystify the reason why over-parameterized deep neural networks works well by considering nonparametric regression model (1) using norm-constrained deep ReLU neural networks.

Furthermore, we analyze the over-parameterized deep fitted Q-iteration (ODFQI) method in reinforcement learning (Kaelbling et al., 1996; Sutton and Barto, 2018, RL) as an application example, and also establish its theoretical results. RL is one of the most important research areas of machine learning that deals with sequential decision-making problems. In online reinforcement learning, an agent learns to maximize the expected future return by interacting with the environment, which can be mathematically modeled as a Markov decision process (MDP). Recently, deep reinforcement learning, which employs deep neural networks to approximate value functions (Li, 2017; Henderson et al., 2018; François-Lavet et al., 2018), has made significant progress in a wide range of areas including natural language processing (Ranzato et al., 2015; Brakel et al., 2017; Bubeck et al., 2023), robotics (Levine et al., 2016, 2018), video games (Mnih et al., 2015), AlphaGo method (Silver et al., 2016), among others. However, the theoretical development of deep reinforcement learning is far behind its empirical success. In this work, we establish the oracle inequalities for ODFQI, a representative value-based RL algorithm, where the learner takes transition data as its input and approximates the target value function with a properly chosen class of deep neural networks (Ernst et al., 2005; Riedmiller, 2005).

The main contributions of this work are summarized as follows.

- We have established the statistical guarantees for both over-parameterized deep nonparametric regression and ODFQI, and provided a prior rule on setting the hyperparameters of depth, width and number of iterations to achieve the desired convergence rate in terms of training sample size.
- We have shown that the deep estimation is adaptive to the smoothness of f_0 and the dimension of covariates. Thus, it circumvents the curse of dimensionality if the distribution of samples is supported on a low-dimensional Riemannian manifold.

1.1 Outlines

The rest of the paper is organized as follows. In Section 2, we introduce the ReLU neural networks with certain weight constraints. In Section 3, we carry out the error analysis by establishing the oracle inequality error bound for the over-parameterized deep nonparametric regression and reduce the curse of dimensionality by exploring the data structure with possible low intrinsic dimension. In Section 4, we consider the application in RL to formulate ODFQI in details and establish the corresponding oracle inequality. Concluding remarks are then given in Section 5. Proofs for all the theorems are deferred to Appendix A.

1.2 Notations

We end this section by introducing some notations used throughout this paper. For any $a, b \in \mathbb{R}$, $\lceil a \rceil$ denotes the smallest integer no less than a, $\lfloor a \rfloor$ denotes the largest integer less than a, $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$, $a \lesssim b$ or $b \gtrsim a$ denotes $a \leq Cb$ for some constant $C > 0$ and $a \asymp b$ when $a \lesssim b \lesssim a$. Let $\mathbb{N}_0, \mathbb{N}$ denote non-negative and strictly positive integers, respectively. For a multi-index $\boldsymbol{s} = (s_1, \ldots, s_d) \in \mathbb{N}_0^d$, the symbol $\partial^{\boldsymbol{s}}$ denotes the partial differential operator $\partial^{\boldsymbol{s}} := (\frac{\partial}{\partial x_1})^{s_1} \ldots (\frac{\partial}{\partial x_d})^{s_d}$ and we use the convention that $\partial^{\boldsymbol{s}}$ is the identity operator when $\boldsymbol{s} = \boldsymbol{0}$. $\|x\|_q = (\sum_{i=1}^d |x_i|^q)^{\frac{1}{q}}$ is the usual q-norm ($q \in [1, \infty]$) of a vector $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$. For probability measure μ and measurable function $Q : \mathbb{R}^d \to \mathbb{R}^1$, we write $\|Q\|_{L^q(\mu)}^q = \mathbb{E}_{x \sim \mu} |Q(x)|^q$.

2 ReLU neural networks with constraint weight

Let $L, N_1, \ldots, N_L \in \mathbb{N}$. We consider the function $\psi : \mathbb{R}^d \to \mathbb{R}^k$ that can be parameterized by a ReLU neural network of the following form

$$\begin{aligned} &\psi_0(\boldsymbol{x}) = \boldsymbol{x}, \\ &\psi_{\ell+1}(\boldsymbol{x}) = \sigma\left(A_\ell \psi_\ell(\boldsymbol{x}) + \boldsymbol{b}_\ell\right), \quad \ell = 0, \ldots, L-1, \\ &\psi(\boldsymbol{x}) = A_L \psi_L(\boldsymbol{x}), \end{aligned} \tag{3}$$

with $A_\ell \in \mathbb{R}^{N_{\ell+1} \times N_\ell}, \boldsymbol{b}_\ell \in \mathbb{R}^{N_{\ell+1}}$, $N_0 = d$ and $N_{L+1} = k$. The activation function $\sigma(x) = x \vee 0$ is the ReLU function which operates element-wisely. The numbers $G = \max\{N_1, \ldots, N_L\}$ and L are the width and depth of the neural network, respectively. We use $\mathcal{F}_{d,k}(G, L)$ to denote the space of functions that can be parameterized by ReLU neural networks with width G and depth L. When the input dimension d and the output dimension k are clear from contexts, we simplify the notation as $\mathcal{F}(G, L)$. Sometimes, we use the notation $\psi_\theta \in \mathcal{F}(G, L)$ to emphasize that the neural network function ψ_θ is parameterized by $\theta = \left(\left(A_0, \boldsymbol{b}_0\right), \ldots, \left(A_{L-1}, \boldsymbol{b}_{L-1}\right), A_L\right)$. Let $\|A\| = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{i,j}|$ for $A \in \mathbb{R}^{m\times n}$. We define the norm-constrained neural network $\mathcal{F}(G, L, M)$ as the set of functions $\psi_\theta \in \mathcal{F}(G, L)$ satisfying the following constraint

$$\xi(\theta) = \|A_L\| \prod_{\ell=0}^{L-1} \max\left\{\left\|(A_\ell, \boldsymbol{b}_\ell)\right\|, 1\right\} \leq M, \tag{4}$$

where M is a positive constant. Note that we can truncate the output of $\psi \in \mathcal{F}_{d,k}(G, L, M)$ by applying $\chi_B(x) = (x \vee -B) \wedge B$ element-wisely. Note that

$$\chi_B(x) = \sigma(x) - \sigma(-x) - (B+1)\sigma(\tfrac{x}{B+1} - \tfrac{B}{B+1}) + (B+1)\sigma(-\tfrac{x}{B+1} - \tfrac{B}{B+1}).$$

By Lemma 25, we can conclude that $\chi_B(\psi) \in \mathcal{F}_{d,k}(\max\{G, 4k\}, L+1, (2B+4)\max\{M, 1\})$. This truncation procedure will not change the rate of approximation bounds in Theorem 4, in which we give the approximation error of $\mathcal{F}(G, L, M)$ within Hölder class as given in Definition 3. Hence, without loss of generality we can assume that $B = 1$ and $\mathcal{F}(G, L, M)$ is bounded by 1 since we can always rescale the truncated version.

3 Error analysis

In order to bound the excess risk of the ERM $\hat{f}_n$ in (2), we first decompose it into two terms referring to statistical and approximation errors as shown in the following Lemma 2. To this end, we introduce the following assumption on the loss ℓ, which holds for the mostly used losses in regression.

Assumption 1 *$\ell(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}^+ \cup \{0\}$ is continuous, and $\ell(a, y) = 0$ if $a = y$ for $(a, y) \in \mathbb{R}^2$. Moreover, $\ell(\cdot,\cdot)$ is λ-Lipschitz continuous in its first argument, where λ is a positive constant. In other words, for any $a_1, a_2 \in \mathbb{R}^2$, we have*

$$|\ell(a_1, \cdot) - \ell(a_2, \cdot)| \leq \lambda |a_1 - a_2|.$$

Lemma 2 *Given random samples $\{Z_i\}_{i=1}^n = \{(X_i, Y_i)\}_{i=1}^n$. Under Assumption 1, the excess risk of ERM $\hat{f}_n$ satisfies*

$$\mathcal{R}(\hat{f}_n) - \mathcal{R}\left(f^*\right) \leq 2 \sup_{f\in\mathcal{F}(G,L,M)} |\mathcal{R}(f) - \mathcal{R}_n(f)| + \lambda \inf_{f\in\mathcal{F}(G,L,M)} \left\|f - f^*\right\|_{L^1(\nu)}, \tag{5}$$

where ν denotes the marginal probability measure of X.

Proof From the definition of ERM $\hat{f}_n$ in (2), for any $f \in \mathcal{F}(G, L, M)$, it follows that $\mathcal{R}_n(\hat{f}_n) \leq \mathcal{R}_n(f)$. Then, by Assumption 1 we have

$$\begin{aligned}\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*) =& \mathcal{R}(\hat{f}_n) - \mathcal{R}_n(\hat{f}_n) + \mathcal{R}_n(\hat{f}_n) - \mathcal{R}_n(f) \\ &+ \mathcal{R}_n(f) - \mathcal{R}(f) + \mathcal{R}(f) - \mathcal{R}(f^*) \\ \leq& \mathcal{R}(\hat{f}_n) - \mathcal{R}_n(\hat{f}_n) + \mathcal{R}_n(f) - \mathcal{R}(f) + \mathcal{R}(f) - \mathcal{R}(f^*) \\ =& \left\{\mathcal{R}(\hat{f}_n) - \mathcal{R}_n(\hat{f}_n)\right\} + \{\mathcal{R}_n(f) - \mathcal{R}(f)\} + \{\mathcal{R}(f) - \mathcal{R}(f^*)\} \\ \leq& 2 \sup_{f \in \mathcal{F}(G,L,M)} |\mathcal{R}(f) - \mathcal{R}_n(f)| + \lambda \|f - f^*\|_{L^1(\nu)},\end{aligned}$$

where ν denotes the marginal probability measure of X. Since the above inequality holds for any $f \in \mathcal{F}(G, L, M)$, then the desirable result can be obtained by taking infimum over $f \in \mathcal{F}(G, L, M)$. ■

By Lemma 2, we can impose the bounds on the approximation error

$$\inf_{f \in \mathcal{F}(G,L,M)} \|f - f^*\|_{L^1(\nu)}$$

and the statistical error

$$\sup_{f \in \mathcal{F}(G,L,M)} |\mathcal{R}(f) - \mathcal{R}_n(f)|,$$

respectively. Next, we establish an upper bound on the approximation error for the ReLU network $\mathcal{F}(G, L, M)$ introduced in Section 3.1. Furthermore, based on this approximation result we give the size-independent statistical error in Section 3.2.

3.1 Approximation error

The term $\inf_{f \in \mathcal{F}(G,L,M)} \|f - f^*\|_{L^1(\nu)}$ can be bounded by the approximation error of the function class $\mathcal{F}(G, L, M)$ to Hölder continuous class, see Definition 3. To that end, we assume that the distribution of the predictor X is supported on the bounded set $[0, 1]^d$ without loss of generality. Therefore, we consider the target function f^* defined on $\mathcal{X} = [0, 1]^d$.

Definition 3 ***(Hölder classes)*** *For $\zeta > 0$ with $\zeta = r + \omega$, where $r \in \mathbb{N}_0$ and $\omega \in (0, 1]$ and $d \in \mathbb{N}$, we denote the Hölder class $\mathcal{H}^\zeta\left(\mathbb{R}^d\right)$ as*

$$\mathcal{H}^\zeta\left(\mathbb{R}^d\right) := \left\{f : \mathbb{R}^d \to \mathbb{R}, \max_{\|\boldsymbol{s}\|_1 \leq r} \|\partial^{\boldsymbol{s}} f\|_\infty \leq 1, \max_{\|\boldsymbol{s}\|_1 = r} \sup_{x \neq y} \frac{|\partial^{\boldsymbol{s}} f(x) - \partial^{\boldsymbol{s}} f(x)|}{\|x - y\|_\infty^\omega} \leq 1\right\}.$$

Given the hypercube $[0, 1]^d \subseteq \mathbb{R}^d$, we denote $\mathcal{H}^\zeta := \left\{f : [0, 1]^d \to \mathbb{R}, f \in \mathcal{H}^\zeta\left(\mathbb{R}^d\right)\right\}$.

We use

$$\mathcal{E}\left(\mathcal{H}^\zeta, \mathcal{F}(G, L, M)\right) = \sup_{f \in \mathcal{H}^\zeta} \inf_{\psi \in \mathcal{F}(G,L,M)} \|f - \psi\|_\infty$$

as the measure for the approximation error. Now we are ready for imposing a bound on $\mathcal{E}\left(\mathcal{H}^\zeta, \mathcal{F}(G, L, M)\right)$ in the following theorem. This proof technique follows from Jiao et al. (2023).

Theorem 4 *Let $d \in \mathbb{N}$ and $\zeta = r + \omega > 0$, where $r \in \mathbb{N}_0$ and $\omega \in (0, 1]$. For any width $G \gtrsim M^{d/(d+1)} \log M$ and depth $L \gtrsim \log M$, we have*

$$\mathcal{E}(\mathcal{H}^\zeta, \mathcal{F}(G, L, M)) \lesssim M^{-\zeta/(d+1)}.$$

In Theorem 4, the weight constraint as given in (4) is used to bound the approximation error, which is important to establish the statistical error and oracle inequalities for over-parameterized ReLU neural networks in the following sections. Note that this is the technical novelty of this work. In Yarotsky (2017), the approximation capacity of neural networks has been constructed. However, the result given in Theorem 4 is established by using norm-constrained neural networks explicitly constructed to approximate the local Taylor polynomials by adopting the idea of Yarotsky (2017), where the first step is to approximate the quadratic monominal x^2.

Lemma 5 *For any $k \in \mathbb{N}$, there exists a function $\psi_k \in \mathcal{F}\left(2k+1, 2k, \frac{4}{3}(\frac{7}{4})^{k+1} - \frac{4}{3}\right)$ such that $\psi_k(0) = 0$ and*

$$\left|x^2 - \psi_k(x)\right| \leq 2^{-2(k+1)}, \quad x \in [0, 1].$$

Proof Following Lemma 2.4 of Telgarsky (2015) and Proposition 2 of Yarotsky (2017), the teeth function $T_i = T_1 \circ T_{i-1} = T_1 \circ \cdots \circ T_1$ is used to construct the approximator

$$\psi_k(x) = x - \sum_{i=1}^{k} 4^{-i} T_i(x),$$

where $T_1(x) = 2x$ for $x \in [0, 1/2]$ and $T_1(x) = 2(1-x)$ for $x \in [1/2, 1]$. As shown in Proposition 2 of Yarotsky (2017) that ψ_k achieves the approximation error $|x^2 - \psi_k(x)| \leq 2^{-2(k+1)}$. Obviously, $T_1 \in \mathcal{F}(2, 2, 7)$, by (b) in Lemma 25, $T_i \in \mathcal{F}(2, 2i, 7^i)$ and consequently $\psi_k \in \mathcal{F}\left(2k+1, 2k, \frac{4}{3}(\frac{7}{4})^{k+1} - \frac{4}{3}\right)$. ■

Based on Lemma 5, we can approximate other monomials and local Taylor expansion with norm constraint ReLU network, see Section A.1 for more details.

3.2 Statistical error

The term $\sup_{f \in \mathcal{F}(G,L,M)} |\mathcal{R}(f) - \mathcal{R}_n(f)|$ is the statistical error of the ReLU neural networks $\mathcal{F}(G, L, M)$ with dependent data $\{Z_i\}_{i=1}^n$. We first introduce the definition of β-mixing for describing the dependence of a general stochastic process $\{W_t\}_{t \geq 1}$.

Definition 6 ***(β-mixing)*** *Let $\{W_t\}_{t \geq 1}$ be a stochastic process and denote the collection $(W_1, \ldots, W_n)$ as $W^{1:n}$, where $n = \infty$ is allowed. Moreover, denote the σ-algebra generated by $W^{i:j} (i \leq j)$ as $\sigma\left(W^{i:j}\right)$. The s-th β-mixing coefficient of $\{W_t\}_{t \geq 1}$, denoted as β_s, is given by*

$$\beta_s = \sup_{t \geq 1} \mathbb{E}\left[\sup_{B \in \sigma(W^{t+s:\infty})} \left|P\left(B \mid W^{1:t}\right) - P(B)\right|\right]. \tag{6}$$

$\{W_t\}_{t \geq 1}$ is said to be β-mixing if $\beta_s \to 0$ as $s \to \infty$. In particular, we say that a β-mixing process mixes at an exponential rate with parameters $\bar{\beta}, b, \eta > 0$ if $\beta_s \leq \bar{\beta} \exp\left(-b s^\eta\right)$ holds for all $s \geq 0$.

By adopting the independent block (IB) technique of Yu (1994) for the strictly β-mixing n-sequence $\{Z_i\}_{i=1}^n$, we divide $\{Z_i\}_{i=1}^n$ into $2\mu_n$ blocks of length a_n ($n = 2a_n\mu_n$) and use the independent copy to substitute half of the blocks. Then, we can transform the original problem to the analysis of the IB sequence to which the standard tools for the independent case can be used to obtain the Rademacher complexity of a function class $\mathcal{F}(G, L, M)$, see Section A.2 in Appendix for more details. Thus we obtain the upper bound of the statistical error $\sup_{f\in\mathcal{F}(G,L,M)} |\mathcal{R}(f) - \mathcal{R}_n(f)|$ when the samples are β-mixng, shown in the following theorem.

Theorem 7 *If $\{Z_i\}_{i=1}^n$ is strictly stationary β-mixing and Assumption 1 holds, then*

$$\mathbb{E} \sup_{f\in\mathcal{F}(G,L,M)} |\mathcal{R}(f) - \mathcal{R}_n(f)| \lesssim \frac{\lambda M\sqrt{L + 2 + \log(d + 1)}}{\sqrt{\mu_n}} + \lambda\mu_n\beta_{a_n},$$

where β_{a_n} is defined in (6).

Remark 8 *β-mixing condition has been introduced to characterize the temporal dependency in time series and Markov decision process sequences in RL, see Lazaric et al. (2012); Antos et al. (2007, 2008); Wong et al. (2020); Chen and Fan (2006) for more details. Furthermore, the exponential β-mixing condition holds if a sequence is geometrically ergodic (Davydov, 1974; Douc et al., 2018).*

Remark 9 *In Theorem 7, the weight constraint given in* (4) *plays a pivotal role in bounding the statistical error by using Theorem 1 in Golowich et al. (2018). See Section A.2 for more details. The underlying principle of norm-based capacity control can be traced back to Bartlett (1996). Furthermore, it is noteworthy that this weight constraint exhibits a close connection with that of Bartlett et al. (2017). Additionally, Golowich et al. (2018) presented an in-depth discussion of the relationship between these constraints.*

Theorem 7 implies that the statistical error bound is determined by μ_n, β_{a_n}, depth L, the norm constraint parameter M, and log *transform of the dimension d. For the fixed parameters M, L, d, if we set $\mu_n = \frac{n}{(\log n)^\tau}$ for some constant $\tau > 0$ (Liang et al., 2009; Hang and Steinwart, 2017) and assume that $\{Z_i\}_{i=1}^n$ is β-mixing with an exponential rate, the error bound becomes $\frac{(\log n)^{\frac{\tau}{2}}}{\sqrt{n}} + \frac{\bar{\beta}n}{(\log n)^\tau} e^{-b(\log n)^{\eta\tau}/2^\eta}$ with $\bar{\beta}, b, \eta$ defined in Definition 6.*

3.3 Oracle inequalities with over-parameterization

Combing the approximation and statistical error bounds as given in Theorems 4 and 7, respectively, we can establish the non-asymptotic error bound for the excess risk $\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*)$ by properly choosing the constraint parameter M and depth L for the function class, and arbitrary large width G as shown in the following Theorem 10.

Theorem 10 *(Oracle inequality) Suppose that $f^* \in \mathcal{H}^\zeta$ with $\zeta = r + \omega, r \in \mathbb{N}_0$ and $\omega \in (0, 1]$, $\{Z_i\}_{i=1}^n$ is strictly stationary β-mixing and Assumption 1 holds. If we set the width and depth as $G \gtrsim M^{d/(d+1)} \log M$ and $L \asymp \log M$, respectively, where the norm constraint parameter is given by $M \asymp \mu_n^{(d+1)/(2\zeta+2d+2)}$, then the excess risk satisfies*

$$\mathbb{E}\left[\mathcal{R}(\hat{f}_n)\right] - \mathcal{R}(f^*) \lesssim \lambda\mu_n^{-\zeta/(2\zeta+2d+2)}\sqrt{\log(d\mu_n)} + \lambda\mu_n\beta_{a_n}.$$

Moreover, if we set $\mu_n = \frac{n}{(\log n)^\tau}$ for some constant $\tau > 0$ and assume that $\{Z_i\}_{i=1}^n$ is exponentially β-mixing with parameters $\bar{\beta}, b, \eta$ defined in Definition 6, then we obtain that

$$\mathbb{E}\left[\mathcal{R}(\hat{f}_n)\right] - \mathcal{R}\left(f^*\right) \lesssim \lambda \left(\frac{n}{(\log n)^\tau}\right)^{-\zeta/(2\zeta+2d+2)} \sqrt{\log(dn)} + \frac{\lambda\bar{\beta}n}{(\log n)^\tau} e^{-b(\log n)^{\eta\tau}/2^\eta}.$$

Theorem 10 primarily relies on the Lipschitz continuity of the loss function ℓ, without imposing any boundedness assumption for the response variable Y. This result can also be extended to the squared loss by employing the truncation technique introduced by Bauer and Kohler (2019) and Kohler and Langer (2021) to avoid the necessity of assuming boundedness of the response Y. In Theorem 10, the smoothness parameter ζ is an absolute constant. To ensure the convergence, we require that the mixing parameters $\bar{\beta}, b, \eta$ satisfy $\log n - b(\log n)^{\eta\tau}/2^\eta \leq 0$ such that $\lim_{n\to\infty} \frac{\bar{\beta}n}{(\log n)^\tau} e^{-b(\log n)^{\eta\tau}/2^\eta} = 0$. Moreover, the non-asymptotic error bound is $\mathcal{O}\left(n^{\frac{-\zeta}{2\zeta+2d+2}}\right)$ by ignoring those logarithmic terms if the mixing parameters $\bar{\beta}, b, \eta$ also satisfy $ne^{-b(\log n)^{\eta\tau}/2^\eta} \leq n^{-\zeta/(2\zeta+2d+2)}$. We observe that this convergence rate $\mathcal{O}\left(n^{\frac{-\zeta}{2\zeta+2d+2}}\right)$ can not achieve the optimal rate in nonparametric regression for i.i.d. data with the squared loss. To shed light on this discrepancy, it is essential to undertake a comprehensive analysis that encompasses both approximation and statistical errors. In the approximation error analysis in Theorem 4, we impose a norm constraint on the weights of neural networks, and we obtain an approximation error determined by the weight norm M, specifically $M^{-\zeta/(d+1)}$, as given in Theorem 4. This is suboptimal when compared to Yarotsky (2017). In the statistical error analysis in Theorem 7, the rate is also suboptimal since the tools of local Rademacher complexity (Bartlett et al., 2005) can not be directly used, resulting in a suboptimal result. Combining the analysis of approximation and statistical errors indeed generates a suboptimal rate. However, the oracle inequalities given in Theorem 10 hold if the width G is taken faster than the order $\mathcal{O}(n^{d/(2\zeta+2d+2)})$ by omitting the logarithm factor, indicating that the convergence results hold even when the number of parameters in the neural network is much larger than the sample size n. Our results in Theorem 10 significantly improve recent results for understanding deep learning (Bauer and Kohler, 2019; Kohler and Langer, 2021; Imaizumi and Fukumizu, 2019; Schmidt-Hieber, 2020; Nakada and Imaizumi, 2020; Farrell et al., 2021; Jiao et al., 2021; Suzuki, 2018; Suzuki and Nitanda, 2021; Shen et al., 2021; Fan et al., 2022), where the number of parameters is strictly smaller than n. This over-parameterized result constitutes a significant and notable contribution to the field of deep learning. Over-parametrization emerges as a pivotal technique in the training of models within the deep learning paradigm, as attested by several prominent studies (Jacot et al., 2018; Allen-Zhu et al., 2019; Du et al., 2019; Zou and Gu, 2019; Liu et al., 2022; Chizat et al., 2019). In this work, we expound upon an over-parameterized framework employed in deep estimation problems. This framework allows for a holistic exploration of the interplay between generalization, approximation, and optimization errors, providing valuable insights into these critical facets of deep learning. To better understand the optimization, we provide a brief description of the optimization procedure here. For further investigation, we have included this as a topic for future research, which is discussed in Section 5. As shown in (2), it is a constrained optimization problem. However, as an alternative, we can consider its regularized version. Subsequently, we can proceed to obtain the theoretical guarantees for the regularized optimization procedure by leveraging the neural tangent kernel analysis from Zou et al. (2020).

However, the convergence rates in Theorem 10 suffer from the curse of dimensionality when d is large. To address this issue, we leverage the low-dimensional data structure in the following section.

3.4 Circumvent the curse of dimensionality

To circumvent the curse of dimensionality, we further assume that the distribution of the predicator X is supported on an low-dimensional Riemannian manifold. High-dimensional data such as images and natural languages are empirically verified to be supported on approximately lower-dimensional manifolds in computer vision and natural language processing. We first briefly review manifolds, partition of unity, and function spaces defined on smooth manifolds; see Federer (1959), Lee (2006), Tu (2011), Chen et al. (2022), Aamari et al. (2019) for more details.

Definition 11 *(Chart) Let $\mathcal{M}$ be a d^*-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^d$. A chart for $\mathcal{M}$ is a pair (U, ϕ) such that $U \subset \mathcal{M}$ is open and $\phi : U \mapsto \mathbb{R}^{d^*}$, where ϕ is a homeomorphism, i.e., bijective, ϕ and ϕ^{-1} are both continuous.*

We say two charts (U, ϕ) and (V, ψ) on $\mathcal{M}$ are C^k compatible if and only if the transition functions,

$$\phi \circ \psi^{-1} : \psi(U \cap V) \mapsto \phi(U \cap V) \quad \text{and} \quad \psi \circ \phi^{-1} : \phi(U \cap V) \mapsto \psi(U \cap V)$$

are both C^k.

Definition 12 *(C^k Atlas) A C^k atlas for $\mathcal{M}$ is a collection of pairwise C^k compatible charts $\{(U_i, \phi_i)\}_{i \in \mathcal{A}}$ such that $\bigcup_{i \in \mathcal{A}} U_i = \mathcal{M}$.*

Definition 13 *(Smooth manifold) A smooth manifold is a manifold together with a C^∞ atlas.*

Definition 14 *(Hölder functions on $\mathcal{M}$) Let $\mathcal{M}$ be a d^*-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^d$. Let $\{(U_i, P_i)\}_{i \in \mathcal{A}}$ be an atlas of $\mathcal{M}$ where the P_i's are orthogonal projections onto tangent space. For a positive number $\zeta > 0$, a function $f : \mathcal{M} \mapsto \mathbb{R}$ belonging to Hölder class $\mathcal{H}^\zeta(\mathcal{M})$ is ζ-Hölder smooth if for each chart (U_i, P_i) in the atlas, we have $f \circ P_i^{-1} \in C^r$ with $\max_{\|\boldsymbol{s}\|_1 \le r} |\partial^{\boldsymbol{s}} f(x)| \le 1$; And for any $\|\boldsymbol{s}\|_1 = r$ and $x, y \in U_i$, $\sup_{x \neq y} \frac{|\partial^{\boldsymbol{s}} f(x) - \partial^{\boldsymbol{s}} f(y)|}{\|x-y\|_\infty^s} \le 1$, where r is the largest integer strictly smaller than ζ and $s = \zeta - r$.*

Definition 15 *(Partition of Unity, Definition 13.4 in Tu (2011)) A C^∞ partition of unity on a manifold $\mathcal{M}$ is a collection of nonnegative C^∞ functions $\rho_i : \mathcal{M} \mapsto \mathbb{R}^+$ for $i \in \mathcal{A}$ such that the collection of the supports, $\{\mathrm{supp}(\rho_i)\}_{i \in \mathcal{A}}$ is locally finite, i.e., every point on $\mathcal{M}$ has a neighborhood that meets only finitely many of $\mathrm{supp}(\rho_i)$'s; And $\sum_{i \in \mathcal{A}} \rho_i = 1$.*

It follows from Theorem 13.7 in Tu (2011) that there exists a C^∞ partition of unity for a smooth manifold, which leads to a decomposition $f = \sum_{i \in \mathcal{A}} f_i$ with $f_i = f\rho_i$ where the same regularity holds for f_i and f due to the equality $f_i \circ \phi_i^{-1} = \left(f \circ \phi_i^{-1}\right) \times \left(\rho_i \circ \phi_i^{-1}\right)$ for a chart (U_i, ϕ_i). Thus, the function f can be written as the sum of the functions f_i, $i \in \mathcal{A}$ and f_i is only supported in a single chart.

Assumption 16 *The distribution of the predicator X is supported on $\mathcal{M} \subset [0,1]^d$, where $\mathcal{M}$ is a compact d^*-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^d$ with condition number $(1/\widetilde{\tau})$ and area of surface $S_{\mathcal{M}}$ (Lee, 2006).*

For a compact Riemannian manifold $\mathcal{M}$, the condition number $(1/\widetilde{\tau})$ controls both local (such as curvature) and global properties (such as self-avoidance) of a compact Riemannian manifold $\mathcal{M}$ (Baraniuk and Wakin, 2009). Moreover, $\widetilde{\tau}$ also represents the geometric concept "reach" (Federer, 1959; Aamari et al., 2019), which is the largest number having the following property: the open normal bundle about $\mathcal{M}$ of radius r is embedded in $\mathbb{R}^d$ for all $r < \widetilde{\tau}$ (Niyogi et al., 2008; Baraniuk and Wakin, 2009). Condition number $(1/\widetilde{\tau})$ or the reach $\widetilde{\tau}$ influences the complexity of function approximation on $\mathcal{M}$ using neural networks.

The surface area $S_{\mathcal{M}}$ of a manifold $\mathcal{M}$ is defined as the integral of 1 over the manifold with respect to the Riemannian volume element (Chapter 10, Lee (2003); Chapter 8, Lee (2006); and Chapter 5, Hubbard and Hubbard (2015)). For example, for the surface area of a d-dimensional unit ball, this definition gives the well-known result $2\pi^{d/2}/\Gamma(d/2)$, where Γ is Gamma function. For function approximation on $\mathcal{M}$ by neural networks, we approximate the target function on a finite number of charts which cover $\mathcal{M}$. Larger surface area $S_{\mathcal{M}}$ only leads to a larger number of charts, which further leads to a wider (linearly in $S_{\mathcal{M}}$) neural network width and larger prefactor of the approximation error.

We are ready for introducing the second main result about the oracle inequality that circumvents the curse of dimensionality as shown in Theorem 17.

Theorem 17 *(Oracle inequality circumvents the curse of dimensionality) Suppose that Assumptions 1 and 16 holds, $\{Z_i\}_{i=1}^n$ is strictly stationary β-mixing, $f^* \in \mathcal{H}^\zeta$ with $\zeta = r + \omega,\ r \in \mathbb{N}_0$ and $\omega \in (0,1]$. For any width $G \gtrsim S_{\mathcal{M}}(2/\widetilde{\tau})^{d^*} d^* \log(d^*)\, M^{d^*/(d^*+1)} \log M$ and depth $L \asymp \log M$, if we set $M \asymp \mu_n^{(d^*+1)/(2\zeta+2d^*+2)}$, then we have*

$$\mathbb{E}\left[\mathcal{R}(\hat{f}_n)\right] - \mathcal{R}\left(f^*\right) \lesssim \lambda \mu_n^{-\zeta/(2\zeta+2d^*+2)} \sqrt{\log(d^*\mu_n)} + \lambda \mu_n \beta_{a_n}.$$

Moreover, if we set $\mu_n = \frac{n}{(\log n)^\tau}$ for some constant $\tau > 0$, and assume that $\{Z_i\}_{i=1}^n$ is exponentially β-mixing with parameters $\bar{\beta}, b, \eta$ satisfying $ne^{-b(\log n)^{\eta\tau}/2^\eta} \lesssim n^{-\zeta/(2\zeta+2d^*+2)}$, we have

$$\mathbb{E}\left[\mathcal{R}(\hat{f}_n)\right] - \mathcal{R}\left(f^*\right) \lesssim \lambda \left(\frac{n}{(\log n)^\tau}\right)^{-\zeta/(2\zeta+2d^*+2)} \sqrt{\log(d^*n)}.$$

The low dimensional data structures are considered by Nakada and Imaizumi (2020), Schmidt-Hieber (2019), Chen et al. (2022) and Jiao et al. (2021) to reduce the influence of the curse of dimensionality. By Theorem 17, the non-asymptotic error bound is at the rate $\mathcal{O}\left(n^{\frac{-\zeta}{2\zeta+2d^*+2}}\right)$ by ignoring those logarithmic factors. Thus, this error bound is adaptive to the low-dimensional data structure if we properly choose the depths and the weight constraint for those considered networks. Hence it circumvents the curse of dimensionality if the intrinsic dimension is small compared with the ambient dimension.

4 Applications in Reinforcement Learning

In this section, we make an extension to explore the offline RL by specifically formulating ODFQI in RL and constructing the oracle inequality for ODFQI. Although algorithmic and statistical properties of traditional fitted Q-iteration (FQI) are well studied by existing works (Murphy, 2005; Munos and Szepesvári, 2008; Antos et al., 2007; Farahmand et al., 2009, 2016; Tosatto et al., 2017; Geist et al., 2019), very few of them can be directly applied to deep FQI when deep neural networks are used to estimate the value function. Fan et al. (2019) provided some theoretical results for deep FQI in the scenario where the number of parameters is smaller than the sample size, and further assume that the batch data are i.i.d., which ignores the temporal dependence existing in MDPs.

In comparison with the result given in Theorem 10 of over-parameterized deep regression, ODFQI needs further error propagation. Specifically, the main idea of obtaining the finite sample bound for FQI is that we first bound the non-parametric fitting error at each iteration and then control the error prorogation across iterations (Antos et al., 2008; Farahmand et al., 2010; Scherrer et al., 2015; Lazaric et al., 2016; Farahmand et al., 2016). The main assumptions used in the theoretical development are the mild distribution shift condition and the realizability-type condition. The necessity of these two conditions are discussed recently (Xie and Jiang, 2020, 2021; Chen and Jiang, 2019). To bound the fitting error at each iteration, we turn to obtain the approximation error of over-parameterized deep ReLU neural networks on Hölder class in Theorem 4 and derive the related generalization error on the dependent data in Theorem 7.

4.1 Markov Decision Process

A discounted MDP is defined by a quintuple $(\mathcal{X}, \mathcal{A}, P, \mathcal{R}, \gamma)$, where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{X} \times \mathcal{A} \subseteq \mathbb{R}^d \to \mathcal{P}(\mathcal{X})$ is the transition probability kernel, $\mathcal{R}(\cdot \mid x, a)$ refers to the distribution of immediate reward $R(x, a)$, and $\gamma \in [0, 1)$ is the discount factor. $\mathcal{P}(\mathcal{X})$ here denotes the sets of probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, such that $P(\cdot|x, a)$ is a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ for each pair $(x, a) \in \mathcal{X} \times \mathcal{A}$, which defines the next-state distribution upon taking action a in state x, and $P(D|\cdot, \cdot)$ is some measurable function on $\mathcal{X} \times \mathcal{A}$ for every $D \in \mathcal{B}(\mathcal{X})$. Moreover, let $\pi(\cdot|x)$ denote the stochastic policy which is an associated distribution of the action at state x. Given an initial distribution $\nu \in \mathcal{P}(\mathcal{X})$, i.e., $X_1 \sim \nu$, the batch data $\{Z_i\}_{i=1}^n = \{X_i, A_i, R_i, X'_i\}_{i=1}^n$ with $X'_i = X_{i+1}$ are generated by

$$X_1 \sim \nu,\ A_i \sim \pi(\cdot \mid X_i),\ R_i \sim \mathcal{R}(\cdot \mid X_i, A_i),\ X'_i \sim P(\cdot \mid X_i, A_i),\ i = 1, \ldots, n.$$

Furthermore, the joint distribution of $\{X_i, A_i\}_{i=1}^n$ is given by

$$\nu(x_1) \prod_{i=2}^{n} \pi(a_{i-1}|x_{i-1}) P(x_i|x_{i-1}, a_{i-1}).$$

We assume that the samples $\{Z_i\}_{i=1}^n$ are strictly stationary β-mixing. Let μ be the distribution of (X_i, A_i) for each $i \in \{1, \ldots, n\}$, then $\mu = \nu \circ \pi$ is the the stationary distribution of this Markov chain $\{X_i, A_i\}_{i=1}^n$, where $\mu = \nu \circ \pi$ is defined by $\mu(E) = \int_E \pi(da|x) d\nu(x)$ for any $E \in \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{A})$.

Denote the action-value function as

$$Q^{\pi}(x,a) := \mathbb{E}\left[\sum_{i=1}^{\infty}\gamma^{i-1}R_i \mid X_1 = x, A_1 = a, \pi\right].$$

For a given policy π, Q^{π} is the unique fixed point of the Bellman operator

$$\mathcal{T}^{\pi}Q(x,a) := \mathbb{E}R(x,a) + \gamma P^{\pi}Q(x,a),$$

with

$$P^{\pi}Q(x,a) := \int P(dx'|x,a)\pi(da'|x')Q(x',a').$$

Without loss of generality, suppose that $R(x,a) \in [0, R_{\max}]$ for each pair $(x,a) \in \mathcal{X} \times \mathcal{A}$, thus Q^{π} takes values in $\left[0, \frac{R_{\max}}{1-\gamma}\right]$. There exists a policy π^* (Agarwal et al., 2019) that maximizes Q^{π}, such that $Q^* := Q^{\pi^*}$, which implies that Q^* satisfies the optimal Bellman equation $Q^* = \mathcal{T}^*Q^*$, where the optimal Bellman operator $\mathcal{T}^*$ is given by

$$\mathcal{T}^*Q(x,a) = \mathbb{E}[R(x,a)] + \gamma\mathbb{E}_{X'\sim P(\cdot|x,a)}\max_{a'\in\mathcal{A}}[Q\left(X',a'\right)].$$

It is straightforward to check that $\mathcal{T}^*$ is a γ-contraction in the sup-norm. We define the greedy policy of an action-value function Q as

$$\pi(x;Q) \in \underset{a\in\mathcal{A}}{\operatorname{argmax}}\ Q(x,a),\ x \in \mathcal{X}.$$

4.2 Deep Fitted Q-iteration

Since $\mathcal{T}^*$ is a γ-contraction, at the population level, we can apply fixed point iteration to approximate the optimal action-value function Q^*. To be precise, suppose that $\mathcal{R}(\cdot|x,a)$ and $P(\cdot|x,a)$ are known, the following iteration

$$Q_0 \to Q_1 = \mathcal{T}^*Q_0 \to Q_2 = \mathcal{T}^*Q_1 \to \ldots \to Q_{J-1} = \mathcal{T}^*Q_{J-2} \to Q_J = \mathcal{T}^*Q_{J-1}, \tag{7}$$

approximate Q^* well when J is large enough. In practice, we only have the batch data $\{Z_i\}_{i=1}^n = \{X_i, A_i, R_i, X_i'\}_{i=1}^n$, and then ODFQI (Ernst et al., 2005; Riedmiller, 2005) mimics the iteration (7) via replacing $Q_j, j = 1, \ldots, J$ with $\widehat{Q}_j$, an estimator in $\mathcal{F}(G,L,M)$ given by the following regression problem

$$\widehat{Q}_j \in \underset{Q\in\mathcal{F}(G,L,M)}{\operatorname{argmin}}\ \widehat{\mathcal{L}}(Q) = \frac{1}{n}\sum_{i=1}^{n}\left(Q(X_i,A_i) - Y_i\right)^2, \tag{8}$$

where $Y_i := R_i + \gamma\max_{a'\in\mathcal{A}}\widehat{Q}_{j-1}(X_{i+1},a')$, and $\widehat{Q}_0 \in \mathcal{F}(G,L,M)$ is an initial guess. Let $\mathcal{L}(Q)$ be the expectation of $\widehat{\mathcal{L}}(Q)$ given $\widehat{Q}_{j-1}$. It is easy to check that for any measurable Q defined on $\mathcal{X}\times\mathcal{A}$, we have

$$\mathcal{L}(Q) = \|Q - \mathcal{T}^*\widehat{Q}_{j-1}\|_{L^2(\mu)}^2 + \mathbb{E}[(\mathcal{T}^*\widehat{Q}_{j-1}(X,A) - Y)^2],$$

where μ denotes the distribution of the state-action (X,A) and $Y := R+\gamma\max_{a'\in\mathcal{A}}\widehat{Q}_{j-1}(X',a')$. The detailed architecture of ODFQI is summarized in Algorithm 1.

Algorithm 1 Over-parameterized Deep Fitted Q-Iteration Algorithm (ODFQI)

```
1: Input: Initial value $\widehat{Q}_0 \in \mathcal{F}(G, L, M)$.
2: for $j = 1, \ldots, J$ do
3:     Input $(X_i, A_i, R_i, X_i')\,, i = 1, \ldots, n$.
4:     Compute $Y_i = R_i + \gamma \max_{a' \in \mathcal{A}} \widehat{Q}_{j-1}(X_i', a')$.
5:     Obtain the $j$-step action-value function $\widehat{Q}_j$ via solving (8), that is,
           $$\widehat{Q}_j \in \underset{Q \in \mathcal{F}(G,L,M)}{\operatorname{argmin}} \widehat{\mathcal{L}}(Q).$$
6: end for
7: Output: The estimate $\widehat{Q}_J$ of $Q^*$ and the greed policy $\pi_J = \pi(\cdot; \widehat{Q}_J)$.
```

In Algorithm 1, each iteration necessitates a data sample of size n, as stipulated in the third step of Algorithm 1. Consequently, we have J datasets, and the cumulative data sample size utilized is nJ. Additionally, we maintain the critical assumption that the J datasets employed in Algorithm 1 are mutually independent. This assumption plays a pivotal role in our subsequent theoretical analysis.

Next, we present the error analysis for ODFQI by bounding $\|Q^* - Q^{\pi_J}\|_{L_1(\nu)}$ for any admissible distribution ν. We first introduce the definition of concentration coefficients that controls the distribution shift because certain concentratability is necessary for the theoretical development of the batch mode RL (Munos, 2003; Chen and Jiang, 2019; Fan et al., 2019; Xie and Jiang, 2020, 2021).

Definition 18 (Concentration Coefficients, Assumption 4.3 of Fan et al. (2019)) *Let $\nu_1, \nu_2 \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$ be two probability measures that are absolutely continuous with respect to the Lebesgue measure on $\mathcal{X} \times \mathcal{A}$. Let $\{\pi_t\}_{t \geq 1}$ be a sequence of policies. Suppose the initial state-action pair (X_0, A_0) of the MDP has distribution ν_1, and we take action A_t according to the policy π_t. For any integer m, we denote the distribution of (X_m, A_m) by $\nu_1 P^{\pi_1} \cdots P^{\pi_m}$. The m-th concentration coefficient is defined as*

$$c_{\nu_1,\nu_2}(m) = \sup_{\pi_1,\ldots,\pi_m} \left[\mathbb{E}_{\nu_2} \left| \frac{d\left(\nu_1 P^{\pi_1} \ldots P^{\pi_m}\right)}{d\nu_2} \right|^2 \right]^{1/2},$$

where the supremum is taken over all possible policies. Furthermore, let μ be the distribution of (X_i, A_i) in Algorithm 1 and let ν be a fixed distribution on $\mathcal{X} \times \mathcal{A}$. Denote

$$C_{\nu,\mu} := (1-\gamma)^2 \cdot \sum_{m \geq 1} m \gamma^{m-1} c_{\nu,\mu}(m), \tag{9}$$

and assume $C_{\nu,\mu} < \infty$, where $(1-\gamma)^2$ in (9) is a normalization term due to the equation $\sum_{m \geq 1} \gamma^{m-1} \cdot m = (1-\gamma)^{-2}$.

The following proposition on the error propagation (Antos et al., 2008; Farahmand et al., 2010; Scherrer et al., 2015; Lazaric et al., 2016; Farahmand et al., 2016; Fan et al., 2019) connects the error bound of $\|Q^* - Q^{\pi_J}\|_{L_1(\nu)}$ with that of $\|\widehat{Q}_j - \mathcal{T}^* \widehat{Q}_{j-1}\|_{L_2(\mu)}$ which is the estimation error of the deep regression (8) in each iteration.

Proposition 19 (Error propagation) *Let π_J be the greedy policy of $\widehat{Q}_J$ in Algorithm 1 and Q^{π_J} be the action-value function corresponding to π_J, then*

$$\mathbb{E}\left\|Q^* - Q^{\pi_J}\right\|_{L_1(\nu)} \leq \frac{2\gamma}{(1-\gamma)^2}\left(C_{\nu,\mu}\max_{1\leq j\leq J}\mathbb{E}\left\|\varepsilon_j\right\|_{L_2(\mu)} + 2\gamma^J R_{\max}\right),$$

where $\varepsilon_j = \widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}$, $j = 1, \ldots, J$.

Proof Proposition 19 is a known result given in Theorem 6.1 of Fan et al. (2019). ■

By Proposition 19, it suffices to bound $\|\widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}\|_{L_2(\mu)}$. To this end, we first decompose the excess risk $\mathcal{L}(\widehat{Q}_j) - \mathcal{L}(\mathcal{T}^*\widehat{Q}_{j-1})$ into the approximation and statistical errors as given in Lemma 20, and then impose the bound on each error by Theorems 4 and 7, respectively.

Lemma 20 *Provided with a random sample $\{Z_i\}_{i=1}^n$, the excess risk satisfies*

$$\mathcal{L}(\widehat{Q}_j) - \mathcal{L}(\mathcal{T}^*\widehat{Q}_{j-1}) \leq 2\sup_{Q\in\mathcal{F}(G,L,M)}\left|\mathcal{L}(Q) - \widehat{\mathcal{L}}(Q)\right| + \inf_{Q\in\mathcal{F}(G,L,M)}\|Q - \mathcal{T}^*\widehat{Q}_{j-1}\|^2_{L^2(\mu)}.$$

Proof From the definition of $\widehat{Q}_j$ in (8), for any $Q \in \mathcal{F}(G, L, M)$, it follows that $\widehat{\mathcal{L}}(\widehat{Q}_j) \leq \widehat{\mathcal{L}}(Q)$. Then, we have

$$\begin{aligned}\mathcal{L}(\widehat{Q}_j) - \mathcal{L}(\mathcal{T}^*\widehat{Q}_{j-1}) =& \mathcal{L}(\widehat{Q}_j) - \widehat{\mathcal{L}}(\widehat{Q}_j) + \widehat{\mathcal{L}}(\widehat{Q}_j) - \widehat{\mathcal{L}}(Q) + \widehat{\mathcal{L}}(Q) - \mathcal{L}(Q) + \mathcal{L}(Q) - \mathcal{L}(\mathcal{T}^*\widehat{Q}_{j-1})\\ \leq& \mathcal{L}(\widehat{Q}_j) - \widehat{\mathcal{L}}(\widehat{Q}_j) + \widehat{\mathcal{L}}(Q) - \mathcal{L}(Q) + \mathcal{L}(Q) - \mathcal{L}(\mathcal{T}^*\widehat{Q}_{j-1})\\ \leq& 2\sup_{Q\in\mathcal{F}(G,L,M)}\left|\mathcal{L}(Q) - \widehat{\mathcal{L}}(Q)\right| + \left\{\mathcal{L}(Q) - \mathcal{L}(\mathcal{T}^*\widehat{Q}_{j-1})\right\}\\ \leq& 2\sup_{Q\in\mathcal{F}(G,L,M)}\left|\mathcal{L}(Q) - \widehat{\mathcal{L}}(Q)\right| + \|Q - \mathcal{T}^*\widehat{Q}_{j-1}\|^2_{L_2(\mu)}.\end{aligned}$$

Since the above inequality holds for any $Q \in \mathcal{F}(G, L, M)$, then the desired result can be obtained by taking infimum over $Q \in \mathcal{F}(G, L, M)$. ■

The term $\inf_{Q\in\mathcal{F}(G,L,M)}\|Q - \mathcal{T}^*\widehat{Q}_{j-1}\|^2_{L^2(\mu)}$ can be bounded by the approximation error of the function class $\mathcal{F}(G, L, M)$ to Hölder class in Theorem 4 under the assumption that the target function $\mathcal{T}^*\widehat{Q}_{j-1}$ lies in Hölder class. To that end, we assume that the distribution of the state-action pair (X, A) is supported on the bounded set $[0, 1]^d$ without loss of generality. The term $\sup_{Q\in\mathcal{F}(G,L,M)}\left|\mathcal{L}(Q) - \widehat{\mathcal{L}}(Q)\right|$ is the statistical error of the ReLU neural networks $\mathcal{F}(G, L, M)$ with dependent data $\{Z_i\}_{i=1}^n$. Similar to Theorem 7, we can obtain the upper bound of the statistical error $\sup_{Q\in\mathcal{F}(G,L,M)}\left|\mathcal{L}(Q) - \widehat{\mathcal{L}}(Q)\right|$ assuming that $\{Z_i\}_{i=1}^n$ are strictly stationary β-mixng. Especially, the target function Q^* is bounded by $\frac{R_{\max}}{1-\gamma}$ instead of 1 in Section 3, without loss of generality we can assume that both $\mathcal{F}(G, L, M)$ and $\mathcal{H}^\zeta$ are bounded by $\frac{R_{\max}}{1-\gamma}$ in the following theorem. Then, T^*Q, $Q \in \mathcal{F}(G, L, M)$, is also bounded by $\frac{R_{\max}}{1-\gamma}$. This leads to a multiple factor $\frac{1}{1-\gamma}$ in the error analysis, see Theorems 21 and 23 for details. It is worthy of pointing out that our method is technically different from those of Antos et al. (2008), Farahmand (2011), and Fan et al. (2019) to bound the statistical error. Actually, in addition to using IB techniques for analysing β-mixing sequences, we turn to controlling the Rademacher complexity of ReLU neural networks with weight constraints to obtain the statistical error.

Therefore, we can establish the non-asymptotic error bound for the excess risk $\mathcal{L}(\widehat{Q}_j) - \mathcal{L}(\mathcal{T}^*\widehat{Q}_{j-1})$ $\left(\|\widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}\|^2_{L_2(\mu)}\right)$ by properly choosing the constraint parameter M and depth L for the function class and arbitrary large width G as shown in the following Theorem 21.

Theorem 21 *Suppose that $\{\mathcal{T}^*\widehat{Q}_{j-1}\}_{j=1}^J \in \mathcal{H}^\zeta$ with $\zeta = r + \omega, r \in \mathbb{N}_0$ and $\omega \in (0,1]$, $\{Z_i\}_{i=1}^n$ is strictly stationary β-mixing. If we set the norm constraint parameter, the width and depth as $M \asymp \mu_n^{(d+1)/(4\zeta+2d+2)}$, $G \gtrsim M^{d/(d+1)} \log M$, and $L \asymp \log M$, respectively, then the excess risk satisfies*

$$\mathbb{E}[\|\widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}\|^2_{L_2(\mu)}] \lesssim \frac{R^2_{\max}}{(1-\gamma)^2}\left(\mu_n^{-\zeta/(2\zeta+d+1)}\sqrt{\log(d\mu_n)} + \mu_n\beta_{a_n}\right).$$

Moreover, if we set $\mu_n = \frac{n}{(\log n)^\tau}$ for some constant $\tau > 0$ and assume that $\{Z_i\}_{i=1}^n$ is exponentially β-mixing with parameters $\bar{\beta}, b, \eta$ defined in Definition 6, then we obtain that

$$\mathbb{E}[\|\widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}\|^2_{L_2(\mu)}] \lesssim \frac{R^2_{\max}}{(1-\gamma)^2}\left[\left(\frac{n}{(\log n)^\tau}\right)^{-\zeta/(2\zeta+d+1)}\sqrt{\log(dn)} + \frac{\bar{\beta}n}{(\log n)^\tau}e^{-b(\log n)^{\eta\tau}/2^\eta}\right].$$

Remark 22 *The completeness assumption $\{\mathcal{T}^*\widehat{Q}_{j-1}\}_{j=1}^J \in \mathcal{H}^\zeta$ is widely used, see Chen and Jiang (2019) and Fan et al. (2019) and references therein. Recall that*

$$\mathcal{T}^*Q(x,a) = \mathbb{E}[R(x,a)] + \gamma\mathbb{E}_{X'\sim P(\cdot|x,a)}\max_{a'\in\mathcal{A}}[Q\left(X',a'\right)],\ Q \in \mathcal{F}(G,L,M).$$

Let $r(x,a) := \mathbb{E}R(x,a)$ be the expected reward function and assume $P(dx'|x,a) = f(x'|x,a)dx'$ for each pair $(x,a) \in \mathcal{X}\times\mathcal{A}$, where $f(x'|x,a)$ denotes the density function of $P(dx'|x,a)$ with respect to Lebesgue measure. Then, we have

$$\mathcal{T}^*Q(x,a) = r(x,a) + \gamma\int f(x'|x,a)\max_{a'\in\mathcal{A}}[Q\left(x',a'\right)]dx',\ Q \in \mathcal{F}(G,L,M).$$

As pointed out by Fan et al. (2019), $\mathcal{T}^*Q(x,a)$ is Hölder continuous if both $r(x,a)$ and $f(x'|x,a)$ are Hölder continuous. Therefore, the completeness assumption holds if both $r(x,a)$ and $f(x'|x,a)$ are sufficiently smooth.

This completeness assumption holds particular significance in our theoretical development, especially in the context of bounding the approximation error $\inf_{Q\in\mathcal{F}(G,L,M)}\|Q - \mathcal{T}^*\widehat{Q}_{j-1}\|^2_{L^2(\mu)}$. In a similar vein, Munos and Szepesvári (2008) utilized the inherent Bellman error to represent the approximation error. Recall that the inherent Bellman error of a function class $\mathcal{F}$ as defined by Munos and Szepesvári (2008) is expressed as

$$d(\mathcal{T}\mathcal{F},\mathcal{F}) = \sup_{g\in\mathcal{F}}\inf_{f\in\mathcal{F}}\|f - \mathcal{T}^*g\|_{L^2(\mu)}.$$

It is evident that the approximation error can be controlled by the inherent Bellman error of function class $\mathcal{F}(G,L,M)$, which can be represented as

$$d(\mathcal{T}\mathcal{F}(G,L,M),\mathcal{F}(G,L,M)) = \sup_{g\in\mathcal{F}(G,L,M)} \inf_{f\in\mathcal{F}(G,L,M)} \|f-\mathcal{T}^* g\|_{L^2(\mu)}.$$

Hence, it is plausible to follow the approach of Munos and Szepesvári (2008) to bound the inherent Bellman error of $\mathcal{F}(G,L,M)$ and subsequently bound the approximation error. To establish the upper bound for the inherent Bellman error, Munos and Szepesvári (2008) introduced specific smoothness assumptions on the transition probability kernel and reward function. These assumptions align with the smoothness conditions discussed earlier, ensuring the validity of our completeness assumption. In contrast to Munos and Szepesvári (2008), we introduce a direct smoothness assumption on $\mathcal{T}^*$, specifically that $\{\mathcal{T}^*\widehat{Q}_{j-1}\}_{j=1}^J \in \mathcal{H}^\zeta$. In this context, we proceed to bound

$$\mathcal{E}\left(\mathcal{H}^\zeta,\mathcal{F}(G,L,M)\right) = \sup_{f\in\mathcal{H}^\zeta} \inf_{\psi\in\mathcal{F}(G,L,M)} \|f-\psi\|_\infty.$$

This approach allows for the control of the approximation error, as demonstrated in Theorem 4.

Now, we give one of the main results in this paper, an oracle inequality of ODFQI.

Theorem 23 *(Oracle inequality) Assume that the conditions of Theorem 21 hold. Then,*

$$\mathbb{E}\left[\|Q^*-Q^{\pi_J}\|_{L_1(\nu)}\right] \lesssim \frac{C_{\nu,\mu}\gamma R_{\max}}{(1-\gamma)^3}\left(\mu_n^{-\zeta/(4\zeta+2d+2)}(\log(d\mu_n))^{1/4} + \sqrt{\mu_n\beta_{a_n}}\right) + \frac{\gamma^{J+1}R_{\max}}{(1-\gamma)^2}.$$

Moreover, if we set $\mu_n = \frac{n}{(\log n)^\tau}$ for some constant $\tau>0$, and assume that $\{Z_i\}_{i=1}^n$ is exponentially β-mixing with parameters $\bar{\beta}, b, \eta$ satisfying $ne^{-b(\log n)^{\eta\tau}/2^\eta} \lesssim n^{\frac{-\zeta}{2\zeta+d+1}}$, we have

$$\mathbb{E}\left[\|Q^*-Q^{\pi_J}\|_{L_1(\nu)}\right] \lesssim \frac{C_{\nu,\mu}\gamma R_{\max}}{(1-\gamma)^3}\left(\frac{n}{(\log n)^\tau}\right)^{-\zeta/(4\zeta+2d+2)}(\log(dn))^{1/4} + \frac{\gamma^{J+1}R_{\max}}{(1-\gamma)^2}.$$

By Theorem 23, when J is large enough, the non-asymptotic error bound is $\mathcal{O}\left(n^{\frac{-\zeta}{4\zeta+2d+2}}\right)$ by ignoring logarithmic terms. Thus, we get the consistency result of ODFQI when n and J go to infinity. The oracle inequalities given in Theorem 23 are still over-parameterized results since the width G can be larger than $\mathcal{O}(n^{d/(4\zeta+2d+2)})$. These oracle inequalities are not direct results of Fan et al. (2019) where an under-parameterized framework is considered. Meanwhile, the convergence rates in Theorem 23 still suffer from the curse of dimensionality as in Theorem 10. Similar to Theorem 17, the convergence rates lessening the curse of dimensionality can also be established when the state-action pair (X,A) always remains on a low-dimensional manifold for some certain dynamical systems.

5 Conclusion

We establish the error bound of over-parameterized deep nonparametric regression with dependent data and make an extension to over-parameterized deep fitted Q-iteration. In over-parameterized deep nonparametric regression, with the error decomposition, we transform the desired error bound into controlling the statistical and approximation errors through Hölder functions using ReLU neural networks with norm-constrained weights. The bound explicitly depends on the sample size, the ambient dimension, and the width and depth of the neural network, which provides an insight into how to choose these hyper-parameters in model training to achieve a desired convergence rate. Furthermore, we show that the curse of dimensionality can be circumvented under the assumption that the distribution of observations is supported on a low-dimensional Riemannian manifold. Moreover, a non-asymptotic error bound of ODFQI is similarly obtained through error propagation. Extending the current results to other scenarios such as time series analysis and exploring the optimization challenges within deep estimation problems will be considered as our future work.

Acknowledgment

We would like to express our sincere gratitude to the editor, the associate editor, and three anonymous reviewers for their constructive comments, which have greatly contributed to the significant improvement of the manuscript. This research was partly supported by the National Natural Science Foundation of China grant (12371270, 12371441, 12001356), Shanghai Science and Technology Development Funds (23JC1402100), Zhejiang Lab (Open Research Project NO. 2022RC0AB06), Fundamental Research Funds for the Central Universities: High-Quality Development of Digital Economy: An Investigation of Characteristics and Driving Strategies (Grant Numbers 2023110139), Shanghai Research Center for Data Science and Decision Technology, Shanghai Education Development Foundation and Shanghai Municipal Education Commission (Chenguang Program), the Fundamental Research Funds for the Central Universities, and the research fund of KLATASDSMOE of China.

Appendix A. Appendix

In this appendix, we give the detailed proofs of Theorems 4, 7, 10, 17, 21, and 23.

A.1 Proof of Theorem 4

Let us first introduce some basic operations on neural networks. These operations will be useful for the construction of neural networks when we study the approximation capacity.

Lemma 24 *Let $\psi \in \mathcal{F}(G, L, M)$, then it can be written in the form* (3) *such that $\|A_L\| \leq M$ and $\|(A_\ell, \boldsymbol{b}_\ell)\| \leq 1$ for $0 \leq \ell \leq L-1$.*

Proof First, we formulate ψ in the form (3) and let $k_\ell := \max\{\|(A_\ell, \boldsymbol{b}_\ell)\|, 1\}$ for all $0 \leq \ell \leq L-1$. Let $\tilde{A}_\ell = A_\ell / k_\ell, \tilde{\boldsymbol{b}}_\ell = \boldsymbol{b}_\ell / \left(\prod_{i=0}^{\ell} k_i\right), \tilde{A}_L = A_L \prod_{i=0}^{L-1} k_i$ and consider the new parameterization of ψ :

$$\tilde{\psi}_{\ell+1}(\boldsymbol{x}) = \sigma\left(\tilde{A}_\ell \tilde{\psi}_\ell(\boldsymbol{x}) + \tilde{\boldsymbol{b}}_\ell\right), \quad \tilde{\psi}_0(\boldsymbol{x}) = \boldsymbol{x}.$$

Then we have $\left\|\tilde{A}_L\right\| \leq M$ and

$$\left\|\left(\tilde{A}_\ell, \tilde{\boldsymbol{b}}_\ell\right)\right\| = \frac{1}{k_\ell}\left\|\left(A_\ell, \frac{\boldsymbol{b}_\ell}{\prod_{i=0}^{\ell-1} k_i}\right)\right\| \leq \frac{1}{k_\ell}\left\|(A_\ell, \boldsymbol{b}_\ell)\right\| \leq 1,$$

where the second inequality holds by $k_i \geq 1$.

Second, we conclude that $\psi_\ell(\boldsymbol{x}) = \left(\prod_{i=0}^{\ell-1} k_i\right)\tilde{\psi}_\ell(\boldsymbol{x})$ by induction. For $\ell = 1$, since the ReLU function is absolutely homogeneous, then it yields that

$$\psi_1(\boldsymbol{x}) = \sigma\left(A_0\boldsymbol{x} + \boldsymbol{b}_0\right) = k_0\sigma\left(\tilde{A}_0\boldsymbol{x} + \tilde{\boldsymbol{b}}_0\right) = k_0\tilde{\psi}_1(\boldsymbol{x}).$$

By induction, we have

$$\begin{aligned}\psi_{\ell+1}(\boldsymbol{x}) = \sigma\left(A_\ell\psi_\ell(\boldsymbol{x}) + \boldsymbol{b}_\ell\right) &= \left(\prod_{i=0}^{\ell} k_i\right)\sigma\left(\tilde{A}_\ell\frac{\psi_\ell(\boldsymbol{x})}{\prod_{i=0}^{\ell-1} k_i} + \tilde{\boldsymbol{b}}_\ell\right) \\ &= \left(\prod_{i=0}^{\ell} k_i\right)\sigma\left(\tilde{A}_\ell\tilde{\psi}_\ell(\boldsymbol{x}) + \tilde{\boldsymbol{b}}_\ell\right) = \left(\prod_{i=0}^{\ell} k_i\right)\tilde{\psi}_{\ell+1}(\boldsymbol{x}).\end{aligned}$$

Thus it follows that

$$\psi(\boldsymbol{x}) = A_L\psi_L(\boldsymbol{x}) = A_L\left(\prod_{i=0}^{L-1} k_i\right)\tilde{\psi}_L(\boldsymbol{x}) = \tilde{A}_L\tilde{\psi}_L(\boldsymbol{x}),$$

which yields that ψ can be parameterized by $\left(\left(\tilde{A}_0, \tilde{\boldsymbol{b}}_0\right), \ldots, \left(\tilde{A}_{L-1}, \tilde{\boldsymbol{b}}_{L-1}\right), \tilde{A}_L\right)$. This completes the proof.

■

Lemma 25 *Let $\psi_1 \in \mathcal{F}_{d_1,k_1}(G_1, L_1, M_1)$ and $\psi_2 \in \mathcal{F}_{d_2,k_2}(G_2, L_2, M_2)$, then the following properties hold.*

(a) If $d_1 = d_2$, $k_1 = k_2$, $G_1 \leq G_2$, $L_1 \leq L_2$ and $M_1 \leq M_2$, then $\mathcal{F}_{d_1,k_1}(G_1, L_1, M_1) \subseteq \mathcal{F}_{d_2,k_2}(G_2, L_2, M_2)$.

(b) If $k_1 = d_2$, then $\psi_2 \circ \psi_1 \in \mathcal{F}_{d_1,k_2}(\max\{G_1, G_2\}, L_1 + L_2, M_2\max\{M_1, 1\})$. Let $A \in \mathbb{R}^{d_2\times d_1}$ and $\boldsymbol{b} \in \mathbb{R}^{d_2}$. Define the function $\psi(\boldsymbol{x}) := \psi_2(A\boldsymbol{x} + \boldsymbol{b})$ for $\boldsymbol{x} \in \mathbb{R}^{d_1}$, then $\psi \in \mathcal{F}_{d_1,k_2}(G_2, L_2, M_2\max\{\|(A, \boldsymbol{b})\|, 1\})$.

(c) If $d_1 = d_2$, define $\psi(\boldsymbol{x}) := (\psi_1(\boldsymbol{x}), \psi_2(\boldsymbol{x}))$, then $\psi \in \mathcal{F}_{d_1,k_1+k_2}(G_1+G_2, \max\{L_1, L_2\}, \max\{M_1, M_2\})$.

(d) If $d_1 = d_2$ and $k_1 = k_2$, then, for any $c_1, c_2 \in \mathbb{R}$, $c_1\psi_1 + c_2\psi_2 \in \mathcal{F}_{d_1,k_1}(G_1 + G_2, \max\{L_1, L_2\}, |c_1|M_1 + |c_2|M_2)$.

Proof By Lemma 24, $\psi_i, i=1,2$ can be parameterized in the form (3) with parameters $\left(\left(A_0^{(i)}, \boldsymbol{b}_0^{(i)}\right), \ldots, \left(A_{L_i-1}^{(i)}, \boldsymbol{b}_{L_i-1}^{(i)}\right), A_{L_i}^{(i)}\right)$ such that $\left\|A_{L_i}^{(i)}\right\| \leq M_i$ and $\left\|\left(A_\ell^{(i)}, \boldsymbol{b}_\ell^{(i)}\right)\right\| \leq 1$ for $\ell \neq L_i$.

(a) We assume that $A_\ell^{(1)} \in \mathbb{R}^{G_2 \times G_2}$ and $\boldsymbol{b}_\ell^{(1)} \in \mathbb{R}^{G_2}, 0 \leq \ell \leq L_1 - 1$, by adding suitable zero rows and columns to $A_\ell^{(1)}$ and $\boldsymbol{b}_\ell^{(1)}$ if necessary (this operation does not change the norm). Then, ψ_1 can be parameterized by the parameters

$$\Big((A_0^{(1)}, \boldsymbol{b}_0^{(1)}), \ldots, (A_{L_1-1}^{(1)}, \boldsymbol{b}_{L_1-1}^{(1)}), \underbrace{(\,\mathrm{Id}\,, \mathbf{0}), \ldots, (\,\mathrm{Id}\,, \mathbf{0})}_{L_2-L_1 \text{ times}}, A_{L_1}^{(1)}\Big),$$

where Id is the identity matrix. Thus we have $\psi_1 \in \mathcal{F}_{d_2,k_2}(G_2, L_2, M_2)$.

(b) We assume $G_1 = G_2$ without loss of generality. Then, $\psi_2 \circ \psi_1$ can be parameterized by

$$\left(\left(A_0^{(1)}, \boldsymbol{b}_0^{(1)}\right), \ldots, \left(A_{L_1-1}^{(1)}, \boldsymbol{b}_{L_1-1}^{(1)}\right), \left(A_0^{(2)} A_{L_1}^{(1)}, \boldsymbol{b}_0^{(2)}\right), \left(A_1^{(2)}, \boldsymbol{b}_1^{(2)}\right), \ldots, \left(A_{L_2-1}^{(2)}, \boldsymbol{b}_{L_2-1}^{(2)}\right), A_{L_2}^{(2)}\right).$$

Then it follows that

$$\begin{aligned}
\left\|\left(A_0^{(2)} A_{L_1}^{(1)}, \boldsymbol{b}_0^{(2)}\right)\right\| &= \left\|\left(A_0^{(2)}, \boldsymbol{b}_0^{(2)}\right)\begin{pmatrix} A_{L_1}^{(1)} & 0 \\ 0 & 1 \end{pmatrix}\right\| \\
&\leq \left\|\left(A_0^{(2)}, \boldsymbol{b}_0^{(2)}\right)\right\| \left\|\begin{pmatrix} A_{L_1}^{(1)} & 0 \\ \mathbf{0} & 1 \end{pmatrix}\right\| \\
&\leq \max\{M_1, 1\}.
\end{aligned}$$

Therefore, it can be concluded that $\psi_2 \circ \psi_1 \in \mathcal{F}_{d_1,k_2}(G_1, L_1 + L_2, M_2 \max\{M_1, 1\})$. As for the function $\psi(\boldsymbol{x}) := \psi_2(A\boldsymbol{x} + \boldsymbol{b})$, it can also be parameterized by

$$\left(\left(A_0^{(2)} A, A_0^{(2)} \boldsymbol{b} + \boldsymbol{b}_0^{(2)}\right), \left(A_1^{(2)}, \boldsymbol{b}_1^{(2)}\right), \ldots, \left(A_{L_2-1}^{(2)}, \boldsymbol{b}_{L_2-1}^{(2)}\right), A_{L_2}^{(2)}\right).$$

Due to $\left\|\left(A_0^{(2)} A, A_0^{(2)} \boldsymbol{b} + \boldsymbol{b}_0^{(2)}\right)\right\| = \left\|\left(A_0^{(2)}, \boldsymbol{b}_0^{(2)}\right)\begin{pmatrix} A & \boldsymbol{b} \\ \mathbf{0} & 1 \end{pmatrix}\right\| \leq \max\{\|(A, \boldsymbol{b})\|, 1\}$, it yields that $\psi \in \mathcal{F}(G_2, L_2, M_2 \max\{\|(A, \boldsymbol{b})\|, 1\})$.

(c) We can assume that $L_1 = L_2$. Then, ψ can be parameterized by the parameters $((A_0, \boldsymbol{b}_0), \ldots, (A_{L_1-1}, \boldsymbol{b}_{L_1-1}), A_{L_1})$, where

$$A_\ell := \begin{pmatrix} A_\ell^{(1)} & \mathbf{0} \\ \mathbf{0} & A_\ell^{(2)} \end{pmatrix}, \quad \boldsymbol{b}_\ell := \begin{pmatrix} \boldsymbol{b}_\ell^{(1)} \\ \boldsymbol{b}_\ell^{(2)} \end{pmatrix}.$$

Notice that

$$\|(A_\ell, \boldsymbol{b}_\ell)\| = \left\|\begin{pmatrix} A_\ell^{(1)} & \mathbf{0} & \boldsymbol{b}_\ell^{(1)} \\ \mathbf{0} & A_\ell^{(2)} & \boldsymbol{b}_\ell^{(2)} \end{pmatrix}\right\| \leq 1$$

and $\|A_{L_1}\| = \max\left\{\left\|A_{L_1}^{(1)}\right\|, \left\|A_{L_1}^{(2)}\right\|\right\} \leq \max\{M_1, M_2\}$.

(d) Replacing the matrix A_{L_1} in (c) by $\left(c_1 A_{L_1}^{(1)}, c_2 A_{L_1}^{(2)}\right)$ yields the conclusion following from

$$\left\|\left(c_1 A_{L_1}^{(1)}, c_2 A_{L_1}^{(2)}\right)\right\| \leq |c_1| \left\|A_{L_1}^{(1)}\right\| + |c_2| \left\|A_{L_1}^{(2)}\right\| \leq |c_1| M_1 + |c_2| M_2.$$

■

Lemma 26 *For any $k \in \mathbb{N}$, there exists $\psi_k \in \mathcal{F}\left(6k+3, 2k+2, 96(\frac{7}{4})^{k+1} - 96\right)$ such that $\psi_k : [-1,1]^2 \to [-1,1]$ and $\psi_k(x,y) = 0$ if $xy = 0$ and*

$$|xy - \psi_k(x,y)| \leq 3 \cdot 2^{-2k-1}, \quad x, y \in [-1,1].$$

Proof By Lemma 5, there exists $\phi_k \in \mathcal{F}(2k+1, 2k, \frac{4}{3}(\frac{7}{4})^{k+1} - \frac{4}{3})$

such that $\phi_k(0) = 0$ and $|x^2 - \phi_k(x)| \leq 2^{-2(k+1)}$ for $x \in [0,1]$. By $xy = 2\left((\frac{|x+y|}{2})^2 - (\frac{|x|}{2})^2 - (\frac{|y|}{2})^2\right)$, we can consider the function

$$\begin{aligned}\widetilde{\psi}_k(x,y) &= 2\phi_k\left(\tfrac{1}{2}|x+y|\right) - 2\phi_k\left(\tfrac{1}{2}|x|\right) - 2\phi_k\left(\tfrac{1}{2}|y|\right) \\ &= 2\phi_k\left(\tfrac{1}{2}\sigma(x+y) + \tfrac{1}{2}\sigma(-x-y)\right) - 2\phi_k\left(\tfrac{1}{2}\sigma(x) + \tfrac{1}{2}\sigma(-x)\right) - 2\phi_k\left(\tfrac{1}{2}\sigma(y) + \tfrac{1}{2}\sigma(-y)\right).\end{aligned}$$

Then, we have $\widetilde{\psi}_k(x,y) = 0$ if $xy = 0$. For any $x, y \in [-1,1]$, we also obtain that

$$\begin{aligned}\left|xy - \widetilde{\psi}_k(x,y)\right| &\leq 2\left|\left(\tfrac{|x+y|}{2}\right)^2 - \phi_k\left(\tfrac{|x+y|}{2}\right)\right| + 2\left|\left(\tfrac{|x|}{2}\right)^2 - \phi_k\left(\tfrac{|x|}{2}\right)\right| + 2\left|\left(\tfrac{|y|}{2}\right)^2 - \phi_k\left(\tfrac{|y|}{2}\right)\right| \\ &\leq 3 \cdot 2^{-2k-1}.\end{aligned}$$

Thence we have $\widetilde{\psi}_k \in \mathcal{F}\left(6k+3, 2k+1, \frac{32}{3}(\frac{7}{4})^{k+1} - \frac{32}{3}\right)$ by (b) and (d) in Lemma 25. Finally, let $\chi(x) = \sigma(x) - \sigma(-x) - 2\sigma(\frac{1}{2}x - \frac{1}{2}) + 2\sigma(-\frac{1}{2}x - \frac{1}{2}) = (x \vee -1) \wedge 1$, then $\chi \in \mathcal{F}(4,1,6)$. Denote the target function as

$$\psi_k(x,y) = \chi(\widetilde{\psi}_k(x,y)) = (\widetilde{\psi}_k(x,y) \vee -1) \wedge 1.$$

For any $x, y \in [-1,1]$, we then have

$$|xy - \psi_k(x,y)| \leq |xy - \widetilde{\psi}_k(x,y)| \leq 3 \cdot 2^{-2k-1}.$$

Therefore, it can be deduced that $\psi_k \in \mathcal{F}\left(6k+3, 2k+2, 96(\frac{7}{4})^{k+1} - 96\right)$ by (b) in Lemma 25. ■

Lemma 27 *For any $d \geq 2$ and $k \in \mathbb{N}$, there exists $\psi \in \mathcal{F}\left((6k+3)d, (k+1)d, d^7(2d)^{k+1}\right)$ such that $\psi : [-1,1]^d \to [-1,1]$ and*

$$|x_1 \cdots x_d - \psi(\boldsymbol{x})| \leq 3d2^{-2k}, \quad \boldsymbol{x} = (x_1, \ldots, x_d)^\top \in [-1,1]^d.$$

Moreover, $\psi(\boldsymbol{x}) = 0$ if $x_1 \cdots x_d = 0$.

Proof Firstly, we consider the case $d = 2^m$ for some $m \in \mathbb{N}$. For $m = 1$, by Lemma 26, there exists $\psi_1 \in \mathcal{F}\left(6k+3, 2k+2, 96(\frac{7}{4})^{k+1}\right)$ such that $\psi_1 : [-1,1]^2 \to [-1,1]$ and $|x_1x_2 - \psi_1(x_1,x_2)| \leq 3 \cdot 2^{-2k-1}$ for any $x_1, x_2 \in [-1,1]$. We define $\psi_m : [-1,1]^{2^m} \to [-1,1]$ inductively by

$$\psi_{m+1}(x_1, \dots, x_{2^{m+1}}) = \psi_1(\psi_m(x_1, \dots, x_{2^m}), \psi_m(x_{2^m+1}, \dots, x_{2^{m+1}})).$$

Then, we have $\psi_m(x_1, \dots, x_{2^m}) = 0$ if $x_1 \cdots x_{2^m} = 0$ since it holds for $m = 1$.

Secondly, by induction, we can show that $\psi_m \in \mathcal{F}\left((6k+3)2^{m-1}, (2k+2)m, (96)^m(\frac{7}{4})^{(k+1)m}\right)$ and

$$|x_1 \cdots x_{2^m} - \psi_m(x_1, \dots, x_{2^m})| \leq (2^m - 1)\epsilon,$$

where $\epsilon = 3{\cdot}2^{-2k-1}$. Indeed, the assertion holds for $m = 1$ by construction. Assume that it is true for $m \in \mathbb{N}$, we need to prove it also holds for $m+1$. By (b)–(c) in Lemma 25 and the definition of ψ_{m+1}, we have $\psi_{m+1} \in \mathcal{F}\left((6k+3)2^m, (2k+2)(m+1), (96)^{(m+1)}(\frac{7}{4})^{(k+1)(m+1)}\right)$. For any $x_1, \dots, x_{2^{m+1}} \in [-1,1]$, we denote $s_1 = x_1 \cdots x_{2^m}$, $t_1 = x_{2^m+1} \cdots x_{2^{m+1}}$, $s_2 = \psi_m(x_1, \dots, x_{2^m})$ and $t_2 = \psi_m(x_{2^m+1}, \dots, x_{2^{m+1}})$, then $s_1, t_1, s_2, t_2 \in [-1,1]$. By induction, one obtains that

$$|s_1 - s_2|, |t_1 - t_2| \leq (2^m - 1)\epsilon.$$

Then,

$$\begin{aligned}
&|x_1 \cdots x_{2^{m+1}} - \psi_{m+1}(x_1, \dots, x_{2^{m+1}})| \\
=&|s_1t_1 - \psi_1(s_2, t_2)| \\
\leq&|s_1t_1 - s_1t_2| + |s_1t_2 - s_2t_2| + |s_2t_2 - \psi_1(s_2, t_2)| \\
\leq&|t_1 - t_2| + |s_1 - s_2| + \epsilon \\
\leq&(2^{m+1} - 1)\epsilon,
\end{aligned}$$

i.e., the assertion holds for $m + 1$. For general $d \geq 2$, we choose $m = \lceil \log_2 d \rceil$, then $2^{m-1} < d \leq 2^m$. We define the target function $\psi : [-1,1]^d \to [-1,1]$ by

$$\psi(\boldsymbol{x}) := \psi_m\left(\begin{pmatrix} \mathrm{Id}_d \\ \mathbf{0}_{(2^m-d)\times d} \end{pmatrix} \boldsymbol{x} + \begin{pmatrix} \mathbf{0}_{d\times 1} \\ \mathbf{1}_{(2^m-d)\times 1} \end{pmatrix}\right),$$

where Id_d is $d \times d$ identity matrix, $\mathbf{0}_{p\times q}$ is $p \times q$ zero matrix and $\mathbf{1}_{(2^m-d)\times 1}$ is all ones vector. By (a)–(b) in Lemma 25,

$$\psi \in \mathcal{F}\left((6k+3)2^{m-1}, (2k+2)m, (96)^m(\frac{7}{4})^{(k+1)m}\right) \subseteq \mathcal{F}\left((6k+3)d, (k+1)d, d^7(2d)^{k+1}\right)$$

and the approximation error is

$$|x_1 \cdots x_d - \psi(\boldsymbol{x})| \leq (2^m - 1)\epsilon \leq 2d\epsilon = 3d2^{-2k}.$$

Obviously, $\psi(\boldsymbol{x}) = 0$ if $x_1 \cdots x_d = 0$ since ψ_m has the same property. ■

By Lemma 27, we can then approximate any $f \in \mathcal{H}^\zeta$ by approximating the local Taylor expansion

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{n} \in \{0,1,\dots,N\}^d} \psi_{\boldsymbol{n}}(\boldsymbol{x}) \sum_{\|\boldsymbol{s}\|_1 \le r} \frac{\partial^{\boldsymbol{s}} f(\frac{\boldsymbol{n}}{N})}{\boldsymbol{s}!} \left(\boldsymbol{x} - \frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{s}}, \tag{A.1}$$

where we use the usual conventions $\boldsymbol{s}! = \prod_{i=1}^d s_i!$ and $(\boldsymbol{x} - \frac{\boldsymbol{n}}{N})^{\boldsymbol{s}} = \prod_{i=1}^d (x_i - \frac{n_i}{N})^{s_i}$. The functions $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}}$ form a partition of unity of $[0,1]^d$ and each $\psi_{\boldsymbol{n}}$ is supported on a sufficiently small neighborhood of $\boldsymbol{n}/N$.

Lemma 28 *For any $N, k \in \mathbb{N}$ and $f \in \mathcal{H}^\zeta$ with $\zeta = r + \omega$, where $r \in \mathbb{N}_0$ and $\omega \in (0,1]$, there exists $\psi \in \mathcal{F}(G, L, M)$ where*

$$\begin{aligned} G &= (r+1)d^r(N+1)^d(6k+3)(d+r), \\ L &= (k+1)(d+r), \\ M &= 6(r+1)d^r(N+1)^d N(d+r)^7(2(d+r))^{k+1}, \end{aligned}$$

such that

$$\|f - \psi\|_{L^\infty([0,1]^d)} \le 2^d d^r (N^{-\zeta} + 3(r+1)(d+r)2^{-2k}).$$

Proof Let

$$\psi(t) = \sigma(1 - |t|) = \sigma(1 - \sigma(t) - \sigma(-t)) \in [0,1], \quad t \in \mathbb{R},$$

then $\psi \in \mathcal{F}(2,2,3)$ and the support of ψ is $[-1,1]$. For any $\boldsymbol{n} = (n_1, \dots, n_d)^\top \in \{0,1,\dots,N\}^d$, define

$$\psi_{\boldsymbol{n}}(\boldsymbol{x}) := \prod_{i=1}^d \psi(Nx_i - n_i), \quad \boldsymbol{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d,$$

then $\psi_{\boldsymbol{n}}$ is supported on $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \frac{\boldsymbol{n}}{N}\|_\infty \le \frac{1}{N}\}$. The functions $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}}$ form a partition of unity of the domain $[0,1]^d$:

$$\sum_{\boldsymbol{n} \in \{0,1,\dots,N\}^d} \psi_{\boldsymbol{n}}(\boldsymbol{x}) = \prod_{i=1}^d \sum_{n_i=0}^N \psi(Nx_i - n_i) \equiv 1, \quad \boldsymbol{x} \in [0,1]^d.$$

Let $p(\boldsymbol{x})$ be the local Taylor expansion (A.1). For convenience, we denote $p_{\boldsymbol{n},\boldsymbol{s}}(\boldsymbol{x}) := \psi_{\boldsymbol{n}}(\boldsymbol{x})(\boldsymbol{x} - \frac{\boldsymbol{n}}{N})^{\boldsymbol{s}}$ and $c_{\boldsymbol{n},\boldsymbol{s}} := \partial^{\boldsymbol{s}} f(\frac{\boldsymbol{n}}{N})/\boldsymbol{s}!$. Then, $p_{\boldsymbol{n},\boldsymbol{s}}$ is supported on $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \frac{\boldsymbol{n}}{N}\|_\infty \le \frac{1}{N}\}$ and

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{n} \in \{0,1,\dots,N\}^d} \sum_{\|\boldsymbol{s}\|_1 \le r} c_{\boldsymbol{n},\boldsymbol{s}} p_{\boldsymbol{n},\boldsymbol{s}}(\boldsymbol{x}).$$

By Lemma A.8 of Petersen and Voigtlaender (2018), the approximation error is

$$\begin{aligned}
|f(\boldsymbol{x}) - p(\boldsymbol{x})| &= \left| \sum_{\boldsymbol{n}} \psi_{\boldsymbol{n}}(\boldsymbol{x}) f(\boldsymbol{x}) - \sum_{\boldsymbol{n}} \psi_{\boldsymbol{n}}(x) \sum_{\|\boldsymbol{s}\|_1 \le r} c_{\boldsymbol{n},\boldsymbol{s}} \left(\boldsymbol{x} - \frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{s}} \right| \\
&\le \sum_{\boldsymbol{n}} \psi_{\boldsymbol{n}}(\boldsymbol{x}) \left| f(\boldsymbol{x}) - \sum_{\|\boldsymbol{s}\|_1 \le r} c_{\boldsymbol{n},\boldsymbol{s}} \left(\boldsymbol{x} - \frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{s}} \right| \\
&= \sum_{\boldsymbol{n}: \|\boldsymbol{x} - \frac{\boldsymbol{n}}{N}\|_\infty < \frac{1}{N}} \left| f(\boldsymbol{x}) - \sum_{\|\boldsymbol{s}\|_1 \le r} c_{\boldsymbol{n},\boldsymbol{s}} \left(\boldsymbol{x} - \frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{s}} \right| \\
&\le \sum_{\boldsymbol{n}: \|\boldsymbol{x} - \frac{\boldsymbol{n}}{N}\|_\infty < \frac{1}{N}} d^r \left\| \boldsymbol{x} - \frac{\boldsymbol{n}}{N} \right\|_\infty^\zeta \\
&\le 2^d d^r N^{-\zeta}.
\end{aligned}$$

Let $\Phi_D \in \mathcal{F}((6k+3)D, (k+1)D, D^7(2D)^{k+1})$ be the D-product function constructed in Lemma 27. Then, we can approximate $p_{\boldsymbol{n},\boldsymbol{s}}$ by

$$\psi_{\boldsymbol{n},\boldsymbol{s}}(\boldsymbol{x}) := \Phi_{d+\|\boldsymbol{s}\|_1}(\psi(Nx_1 - n_1), \dots, \psi(Nx_d - n_d), \dots, x_i - \tfrac{n_i}{N}, \dots),$$

where the term $x_i - n_i/N$ appears in the input only when $s_i \neq 0$ and it repeats s_i times. (When $d = 1$ and $\boldsymbol{s} = \boldsymbol{0}$, we simply let $\psi_{n,\boldsymbol{0}}(x) = \psi(Nx - n)$.). Since $x_i - n_i/N = \sigma(x_i - n_i/N) - \sigma(-x_i + n_i/N)$ and $\|\boldsymbol{s}\|_1 \le r$, by (b)–(c) in Lemma 25, we have $\psi_{\boldsymbol{n},\boldsymbol{s}} \in \mathcal{F}((6k+3)(d+r), (k+1)(d+r), 6N(d+r)^7(2(d+r))^{k+1})$. By Lemma 27, the approximation error is

$$|p_{\boldsymbol{n},\boldsymbol{s}}(\boldsymbol{x}) - \psi_{\boldsymbol{n},\boldsymbol{s}}(\boldsymbol{x})| \le 3(d+r)2^{-2k}.$$

Since $\Phi_D(t_1, \dots, t_D) = 0$ when $t_1 t_2 \cdots t_D = 0$, $\psi_{\boldsymbol{n},\boldsymbol{s}}$ is supported on $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \frac{\boldsymbol{n}}{N}\|_\infty \le \frac{1}{N}\}$.

Now, we can approximate $p(\boldsymbol{x})$ by

$$\psi(\boldsymbol{x}) = \sum_{\boldsymbol{n} \in \{0,1,\dots,N\}^d} \sum_{\|\boldsymbol{s}\|_1 \le r} c_{\boldsymbol{n},\boldsymbol{s}} \psi_{\boldsymbol{n},\boldsymbol{s}}(\boldsymbol{x}).$$

Observe that $|c_{\boldsymbol{n},\boldsymbol{s}}| = |\partial^{\boldsymbol{s}} f(\frac{\boldsymbol{n}}{N})/\boldsymbol{s}!| \le 1$ and the number of terms in the inner summation is

$$\sum_{\|\boldsymbol{s}\|_1 \le r} 1 = \sum_{j=0}^{r} \sum_{\|\boldsymbol{s}\|_1 = j} 1 \le \sum_{j=0}^{r} d^j \le (r+1)d^r.$$

The approximation error is, for any $\boldsymbol{x} \in [0,1]^d$,

$$\begin{aligned}
|p(\boldsymbol{x}) - \psi(\boldsymbol{x})| =& \left| \sum_{\boldsymbol{n}} \sum_{\|\boldsymbol{s}\|_1 \le r} c_{\boldsymbol{n},\boldsymbol{s}} p_{\boldsymbol{n},\boldsymbol{s}}(\boldsymbol{x}) - \sum_{\boldsymbol{n}} \sum_{\|\boldsymbol{s}\|_1 \le r} c_{\boldsymbol{n},\boldsymbol{s}} \psi_{\boldsymbol{n},\boldsymbol{s}}(\boldsymbol{x}) \right| \\
\le& \sum_{\boldsymbol{n}} \sum_{\|\boldsymbol{s}\|_1 \le r} |c_{\boldsymbol{n},\boldsymbol{s}}| |p_{\boldsymbol{n},\boldsymbol{s}}(\boldsymbol{x}) - \psi_{\boldsymbol{n},\boldsymbol{s}}(\boldsymbol{x})| \\
\le& \sum_{\boldsymbol{n}:\|\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\|_\infty < \frac{1}{N}} \sum_{\|\boldsymbol{s}\|_1 \le r} |p_{\boldsymbol{n},\boldsymbol{s}}(\boldsymbol{x}) - \psi_{\boldsymbol{n},\boldsymbol{s}}(\boldsymbol{x})| \\
\le& 3 \cdot 2^d (r+1)(d+r) d^r 2^{-2k}.
\end{aligned}$$

Hence, the total approximation error is

$$|f(\boldsymbol{x}) - \psi(\boldsymbol{x})| \le |f(\boldsymbol{x}) - p(\boldsymbol{x})| + |p(\boldsymbol{x}) - \psi(\boldsymbol{x})| \le 2^d d^r (N^{-\zeta} + 3(r+1)(d+r)2^{-2k}).$$

By (d) in Lemma 25, we obtain the desirable result. ■

Using the construction in Lemma 28, we can give a proof of the approximation bound in Theorem 4 as follows.

Proof We choose $N = \lceil 2^{2k/\zeta} \rceil$ in the Lemma 28, then there exist $\psi \in \mathcal{F}(G, L, M)$ with

$$\begin{aligned}
G &= (r+1)d^r(N+1)^d(6k+3)(d+r) \asymp 2^{2dk/\zeta} k, \\
L &= (k+1)(d+r), \\
M &= 6(r+1)d^r(N+1)^d N (d+r)^7 (2(d+r))^{k+1} \asymp 2^{2(d+1)k/\zeta},
\end{aligned}$$

such that $\|f - \psi\|_{C([0,1]^d)} \le 2^d d^r (N^{-\zeta} + 3(r+1)(d+r)2^{-2k}) \lesssim 2^{-2k}$. Then, $k \asymp \log M$, $G \asymp 2^{2dk/\zeta} k \asymp M^{d/(d+1)} \log M$, $L = (k+1)(d+r) \asymp \log M$ and we have the approximation bound

$$\|f - \psi\|_{C([0,1]^d)} \lesssim 2^{-2k} \lesssim M^{-\zeta/(d+1)}.$$

Since increasing G and L can only decrease the approximation error, the bound holds for any $G \gtrsim M^{d/(d+1)} \log M$ and $L \gtrsim \log M$. ■

A.2 Proof of Theorem 7

In order to prove Theorem 7, we first introduce some lemmas in Yu (1994). Recall the main idea of the construction of independent block for the strictly stationary β-mixing n-sequence $\{Z_i\}_{i=1}^n$. This is the key technique to obtain the Rademacher complexity of a function class $\mathcal{F}(G, L, M)$ with strictly stationary β-mixing data $\{Z_i\}_{i=1}^n$. Without loss of generality, for any integer pair (a_n, μ_n) with $n = 2a_n\mu_n$, we divide the strictly stationary n-sequence $\{Z_i\}_{i=1}^n$ into $2\mu_n$ blocks of length a_n. Denote the indices in the blocks alternately by H's and T's. Note that these indices depend on n, but for simplicity we suppress n. That is,

$$\begin{aligned}
H_1 &:= \{i : 1 \le i \le a_n\}, \\
T_1 &:= \{i : a_n + 1 \le i \le 2a_n\}.
\end{aligned}$$

In general, for $1 \leq j \leq \mu_n$, let

$$\begin{aligned} H_j &:= \left\{ i : 2(j-1)a_n + 1 \leq i \leq (2j-1)a_n \right\}, \\ T_j &:= \left\{ i : (2j-1)a_n + 1 \leq i \leq (2j)a_n \right\}, \end{aligned}$$

and define $H = \bigcup_{j=1}^{\mu_n} \{H_j\}$. Denote the random variables that correspond to the H_j and T_j indices as

$$\begin{aligned} Z\left(H_j\right) &= \{Z_i, i \in H_j\} = \{Z_{2(j-1)a_n+1}, \cdots, Z_{(2j-1)a_n}\}, \\ Z\left(T_j\right) &= \{Z_i, i \in T_j\} = \{Z_{(2j-1)a_n+1}, \cdots, Z_{(2j)a_n}\}. \end{aligned}$$

Furthermore, let the whole sequence of H-blocks be denoted by $Z_{a_n} := \left\{ Z\left(H_j\right) : \ j = 1, 2, \ldots, \mu_n \right\}$, and the whole sequence of T-blocks is defined as $Z_{1,a_n} := \{Z(T_j) : j = 1, 2, \ldots, \mu_n\}$. Take a sequence of i.i.d blocks $\left\{\Xi\left(H_j\right) : j = 1, \ldots, \mu_n\right\}$, where $\Xi\left(H_j\right) = \left\{Z_i' : i \in H_j\right\}$, such that the sequence is independent of $\{Z_i\}_{i=1}^n$ and each block has the same distribution as a block from the original sequence, that is, $\mathscr{L}\left(\Xi\left(H_j\right)\right) = \mathscr{L}\left(Z\left(H_j\right)\right), \ j = 1, \ldots, \mu_n$. We call this constructed sequence the independent block a_n-sequence (IB sequence), and denote the IB sequence as Ξ_{a_n}. We then transform the original problem to the analysis of the IB sequence to which the standard tools for the independent case can be used.

Next, we give some definitions. Let σ_i's be i.i.d Rademacher random variable, and assume σ_i's is independent of Z_i's and Z_i''s. For some measurable function g, denote

$$P_n g := \frac{1}{n} \sum_{i=1}^{n} \sigma_i g\left(Z_i\right).$$

For the sequence $\{Z_i\}_{i=1}^n$, we write

$$\widetilde{Y}_{j,g}\left(Z_{a_n}\right) := \sum_{i \in H_j} \sigma_i g\left(Z_i\right).$$

For the constructed IB sequence Ξ, define

$$W_{j,g}\left(\Xi_{a_n}\right) := \sum_{i \in H_j} \sigma_i g\left(Z_i'\right).$$

Lemma 29 *(Lemma 4.1 in Yu (1994)) Let the distributions of Z_{a_n} and Ξ_{a_n} be $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$, respectively. For any measurable function h on $\mathbb{R}^{\mu_n a_n}$ with bound $\widetilde{M}$,*

$$\left| \mathcal{Q} h\left(Z_{a_n}\right) - \tilde{\mathcal{Q}} h\left(\Xi_{a_n}\right) \right| \leq \widetilde{M}\left(\mu_n - 1\right)\beta_{a_n}.$$

Lemma 30 *Suppose that $\mathbf{F}_{\widetilde{M}}$ is a function class bounded by $\widetilde{M}$, then*

$$\mathbb{E}_{\{Z_i,\sigma_i\}_{i=1}^n}\left(\sup_{g \in \mathbf{F}_{\widetilde{M}}} \left|P_n g\right| \right) \leq \mathbb{E}_{\{Z_i',\sigma_i\}_{i=1}^n}\left(\sup_{g \in \mathbf{F}_{\widetilde{M}}} \left| \frac{1}{\mu_n} \sum_{j=1}^{\mu_n} \frac{W_{j,g}(\Xi_{a_n})}{a_n} \right| \right) + 2\widetilde{M}\mu_n \beta_{a_n}.$$

Proof Note that the strictly β-mixing process $Z_{a_n} = \{Z(H_j); j = 1, \dots, \mu_n\}$ has the same distribution as $Z_{1,a_n} = \{Z(T_j); j = 1, \dots, \mu_n\}$, where

$$\begin{aligned} Z(H_j) &= \{Z_{2(j-1)a_n+1}, \dots, Z_{(2j-1)a_n}\}, \\ Z(T_j) &= \{Z_{(2j-1)a_n+1}, \dots, Z_{(2j)a_n}\}. \end{aligned}$$

Then we have

$$\begin{aligned} &\mathbb{E}_{\{Z_i,\sigma_i\}_{i=1}^n}\left(\sup_{g\in\mathbf{F}_{\widetilde{M}}} |P_n g|\right) \\ &= \mathbb{E}_{\{Z_i,\sigma_i\}_{i=1}^n}\left(\sup_{g\in\mathbf{F}_{\widetilde{M}}} \left|\frac{1}{n}\sum_{j=1}^{\mu_n} \widetilde{Y}_{j,g}(Z_{a_n}) + \frac{1}{n}\sum_{j=1}^{\mu_n} \widetilde{Y}_{j,g}(Z_{1,a_n})\right|\right) \\ &\le \mathbb{E}_{\{Z_i,\sigma_i\}_{i=1}^n}\left(\sup_{g\in\mathbf{F}_{\widetilde{M}}} \left|\frac{1}{n}\sum_{j=1}^{\mu_n} \widetilde{Y}_{j,g}(Z_{a_n})\right|\right) + \mathbb{E}_{\{Z_i,\sigma_i\}_{i=1}^n}\left(\sup_{g\in\mathbf{F}_{\widetilde{M}}} \left|\frac{1}{n}\sum_{j=1}^{\mu_n} \widetilde{Y}_{j,g}(Z_{1,a_n})\right|\right) \\ &= 2\mathbb{E}_{\{Z_i,\sigma_i\}_{i=1}^n}\left(\sup_{g\in\mathbf{F}_{\widetilde{M}}} \left|\frac{1}{n}\sum_{j=1}^{\mu_n} \widetilde{Y}_{j,g}(Z_{a_n})\right|\right) \\ &\le 2\mathbb{E}_{\{Z_i',\sigma_i\}_{i=1}^n}\left(\sup_{g\in\mathbf{F}_{\widetilde{M}}} \left|\frac{1}{n}\sum_{j=1}^{\mu_n} W_{j,g}(\Xi_{a_n})\right|\right) + 2\widetilde{M}\mu_n\beta_{a_n} \\ &= \mathbb{E}_{\{Z_i',\sigma_i\}_{i=1}^n}\left(\sup_{g\in\mathbf{F}_{\widetilde{M}}} \left|\frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \frac{W_{j,g}(\Xi_{a_n})}{a_n}\right|\right) + 2\widetilde{M}\mu_n\beta_{a_n}, \end{aligned}$$

where the last inequality follows from Lemma 29 and σ_i's being independent of Z_i's and Z_i''s. ■

Based on the above lemmas, we give the proof of Theorem 7 as follows.

Proof Let $\widetilde{Z}_i$ be a independent copy of Z_i, and σ_i's be the i.i.d. Rademacher random variables that are independent with $\widetilde{Z}_i$ and Z_i, $i = 1, \dots, n$. Denote the Rademacher complexity of $\mathcal{F}(G, L, M)$ as

$$\mathcal{G}(\mathcal{F}(G,L,M)) = \mathbb{E}_{\{X_i,\sigma_i\}_{i=1}^n}\left[\sup_{f\in\mathcal{F}(G,L,M)} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i))\right|\right].$$

Then we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}(G,L,M)}|\mathcal{R}(f)-\mathcal{R}_n(f)|\right] &= \mathbb{E}\left[\sup_{\ell_f\in\ell\circ\mathcal{F}(G,L,M)}\left|\frac{1}{n}\sum_{i=1}^{n}\ell_f(Z_i)-\mathbb{E}[\ell_f(Z_1)]\right|\right]\\
&= \mathbb{E}\left[\sup_{\ell_f\in\ell\circ\mathcal{F}(G,L,M)}\left|\frac{1}{n}\sum_{i=1}^{n}\ell_f(Z_i)-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\ell_f(\widetilde{Z}_i)]\right|\right]\\
&\leq \mathbb{E}\left[\sup_{\ell_f\in\ell\circ\mathcal{F}(G,L,M)}\left|\frac{1}{n}\sum_{i=1}^{n}\ell_f(Z_i)-\frac{1}{n}\sum_{i=1}^{n}\ell_f(\widetilde{Z}_i)\right|\right]\\
&= \mathbb{E}\left[\sup_{\ell_f\in\ell\circ\mathcal{F}(G,L,M)}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i(\ell_f(Z_i)-\ell_f(\widetilde{Z}_i))\right|\right]\\
&\leq 2\mathbb{E}\left[\sup_{\ell_f\in\ell\circ\mathcal{F}(G,L,M)}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell_f(Z_i)\right|\right]\\
&\leq 4\lambda\mathcal{G}(\mathcal{F}(G,L,M))\\
&\leq 8\lambda\mathbb{E}\left(\sup_{f\in\mathcal{F}(G,L,M)}\left|\frac{1}{\mu_n}\sum_{j=1}^{\mu_n}\frac{W_{j,f}(\Xi_{a_n})}{a_n}\right|\right)+8\lambda\mu_n\beta_{a_n}\\
&\lesssim \frac{\lambda M\sqrt{L+2+\log(d+1)}}{\sqrt{\mu_n}}+\lambda\mu_n\beta_{a_n},
\end{aligned}$$

where the first inequality follows from the Jensen's inequality, and the second equality holds since both $\sigma_i\ell_f(Z_i)$ and $\sigma_i\ell_f(\widetilde{Z}_i)$ are governed by the same law, the third inequality holds by Lemma 5 of Meir and Zhang (2003), the fourth inequality holds by Lemma 30 and $\mathcal{F}(G,L,M)$ is bounded by 1, and the last inequality directly follows from Theorem 2 of Golowich et al. (2018) since $(W_{j,f}(\Xi_{a_n})/a_n)$'s are i.i.d. and bounded by 1. ■

A.3 Proof of Theorem 10

Proof By Theorem 4, for any $f\in\mathcal{H}^{\zeta}$, there exists a function $\psi\in\mathcal{F}(G,L,M)$ with width $G\gtrsim M^{d/(d+1)}\log M$ and depth $L\asymp\log M$ such that

$$|f(x)-\psi(x)|\lesssim M^{-\zeta/(d+1)},$$

for all $x\in[0,1]^d$. By Lemma 20 and Theorem 7, it yields that

$$\mathbb{E}\left[\mathcal{R}(\hat{f}_n)\right]-\mathcal{R}\left(f^*\right)\lesssim\lambda M^{-\zeta/(d+1)}+\frac{\lambda M\sqrt{L+2+\log(d+1)}}{\sqrt{\mu_n}}+\lambda\mu_n\beta_{a_n}.$$

Setting $M\asymp\mu_n^{(d+1)/(2\zeta+2d+2)}$, then it follows that

$$\mathbb{E}\left[\mathcal{R}(\hat{f}_n)\right]-\mathcal{R}\left(f^*\right)\lesssim\lambda\mu_n^{-\zeta/(2\zeta+2d+2)}\sqrt{\log(d\mu_n)}+\lambda\mu_n\beta_{a_n}.$$

Moreover, if we set $\mu_n = \frac{n}{(\log n)^\tau}$ for some constant $\tau > 0$ and assume that $\{Z_i\}_{i=1}^n$ is exponentially β-mixing with parameters $\bar{\beta}, b, \eta$ defined in Definition 6, then we obtain that

$$\mathbb{E}\left[\mathcal{R}(\hat{f}_n)\right] - \mathcal{R}\left(f^*\right) \lesssim \lambda \left(\frac{n}{(\log n)^\tau}\right)^{-\zeta/(2\zeta+2d+2)} \sqrt{\log(dn)} + \frac{\lambda\bar{\beta}n}{(\log n)^\tau} e^{-b(\log n)^{\eta\tau}/2^\eta}.$$

■

A.4 Proof of Theorem 17

Proof We prove Theorem 17 with the following three steps. We first construct an finite atlas that covers the manifold $\mathcal{M}$. Then similar to Chen et al. (2022) we project each chart linearly to a d^*-dimensional hypercube on which we approximate the low-dimensional Hölder smooth functions respectively. Lastly, we combine the approximation results on all charts to get an error bound of the approximation on the whole manifold. This procedure is similar to those of Schmidt-Hieber (2019); Chen et al. (2019, 2022), but we apply our new approximation result when approximating the low-dimensional Hölder smooth functions on each projected chart, which leads to a better prefactor of error compared to most existing results.

Step 1: Atlas Construction and Projection. Let $B(x, \tilde{r})$ denote the open Euclidean ball with radius $\tilde{r} > 0$ and center $x \in \mathbb{R}^d$. Given any $\tilde{r} > 0$, we have an open cover $\{B(x, \tilde{r})\}_{x\in\mathcal{M}}$ of $\mathcal{M}$. By the compactness of $\mathcal{M}$, there exists a finite cover $\{B\left(x_i, \tilde{r}\right)\}_{i=1,\ldots,C_\mathcal{M}}$ for some finite integer $C_\mathcal{M}$ such that $\mathcal{M} \subset \bigcup_i B\left(x_i, \tilde{r}\right)$. Let $(1/\widetilde{\tau})$ denote the condition number of $\mathcal{M}$, then we can choose proper radius $\tilde{r} < \widetilde{\tau}/2$ such that $U_i = \mathcal{M} \cap B\left(x_i, \tilde{r}\right)$ is diffeomorphic to a ball in $\mathbb{R}^{d^*}$ (Niyogi et al., 2008). Besides, the number of charts $C_\mathcal{M}$ satisfies

$$C_\mathcal{M} \leq \left\lceil S_\mathcal{M} T_{d^*}/\tilde{r}^{d^*}\right\rceil,$$

where $S_\mathcal{M}$ is the area of the surface of $\mathcal{M}$ and T_{d^*} is the thickness of U_i's, which is defined as the average number of U_i's that contains a point on $\mathcal{M}$. By (19) in Chapter 2 of Conway and Sloane (2013), the thickness T_{d^*} scales approximately linear in d^* and there exist coverings such that $T_{d^*} \leq d^* \log\left(d^*\right) + d^* \log\log\left(d^*\right) + 5d^* \leq 7d^* \log\left(d^*\right)$. Let the tangent space of $\mathcal{M}$ at x_i be denoted by $T_{x_i}(\mathcal{M})$ and let $V_i \in \mathbb{R}^{d\times d^*}$ be the matrix concatenating the orthonormal basis of the tangent space as column vectors. Then for any $x \in U_i$ we can define the projection

$$\phi_i(x) = a_i\left(V_i^\top\left(x - x_i\right) + b_i\right),$$

where $a_i \in (0, 1]$ and b_i are proper scalar and vector such that $\phi_i(x) \in [0, 1]^{d^*}$ for any $x \in U_i$. Note that each projection ϕ_i is a linear function, which can be computed by a one-hidden layer ReLU network.

Step 2: Approximate low-dimensional functions. For charts $\{(U_i, \phi_i)\}_{i=1}^{C_\mathcal{M}}$, we can approximate the function on each chart by approximating the projected function in the low-dimensional space. By Theorem 13.7 in Tu (2011), the target function f can be written as

$$f = \sum_{i=1}^{C_\mathcal{M}} f\rho_i := \sum_{i=1}^{C_\mathcal{M}} f_i,$$

where ρ_i's are elements in C^∞ partition of unity on $\mathcal{M}$ being supported in U_i's. Note that the manifold $\mathcal{M}$ is compact and smooth and ρ_i's are C^∞, so f_i's have the same smoothness as f itself for $i = 1, \ldots, C_{\mathcal{M}}$. Note that the collection of the supports, $\{\operatorname{supp}(\rho_i)\}_{i \in \mathcal{A}}$ is locally finite, and let C_ρ denote the maximum number of $\operatorname{supp}(\rho_i)$'s that a point on $\mathcal{M}$ can belong to. Besides, since ϕ_i is a linear projection operator, it is not hard to show that $f_i \circ \phi_i^{-1}$ is a Hölder smooth function with order $\zeta > 0$ on $\phi_i(U_i) \subset [0,1]^{d^*}$, i.e., $f_i \circ \phi_i^{-1} \in \mathcal{H}^\zeta(\phi_i(U_i))$ is bounded by some universal constant $C_0 > 0$ over $i = 1, \ldots, C_{\mathcal{M}}$, where a detailed proof can be found in Lemma 2 of Chen et al. (2022). By the extended version of Whitney's extension theorem in Fefferman (2006), we can approximate the smooth extension of $f_i \circ \phi_i^{-1}$ on $[0,1]^{d^*}$. By Theorem 4, for $G \gtrsim M^{d^*/(d^*+1)} \log M$ and $L \gtrsim \log M$, we have

$$\left| f_i \circ \phi_i^{-1}(x) - g_i(x) \right| \lesssim C_0 M^{-\zeta/(d^*+1)},$$

for any $x \in \phi_i(U_i) \subset [0,1]^{d^*}$.

Step 3: Approximate the target function on the manifold. By the construction of subnetworks, the projected target function $f_i \circ \phi_i^{-1}$ on each region $\phi_i(U_i)$ can be approximated by over-parameterized deep ReLU neural networks g_i. Note that each projection ϕ_i is a linear function can be computed by a one-hidden layer ReLU network. Then we stack two more layer to g_i and get $\tilde{g}_i = g_i \circ \phi_i$ such that for any $x \in U_i$,

$$|f_i(x) - \tilde{g}_i(x)| = |f_i(x) - g_i \circ \phi_i(x)| \lesssim C_0 M^{-\zeta/(d^*+1)},$$

where $\tilde{g}_i$ is a over-parameterized deep ReLU neural network with width $G \gtrsim M^{d^*/(d^*+1)} \log M$ and depth $L \gtrsim \log M$. Since there are $C_{\mathcal{M}}$ charts, we parallelize these subnetworks $\tilde{g}_i$ to get $\tilde{g} = \sum_{i=1}^{C_{\mathcal{M}}} \tilde{g}_i$ such that

$$\begin{aligned} |f(x) - \tilde{g}(x)| &= \left| \sum_{i}^{C_{\mathcal{M}}} f_i(x) - \sum_{i=1}^{C_{\mathcal{M}}} \tilde{g}_i(x) \right| \\ &\leq C_\rho \left| f_i(x) - \tilde{g}_i(x) \right| \\ &\lesssim C_\rho C_0 M^{-\zeta/(d^*+1)}, \end{aligned}$$

for any $x \in \mathcal{M}$. Such a neural network $\tilde{g}$ has width $G \gtrsim C_{\mathcal{M}} M^{d^*/(d^*+1)} \log M$ and depth $L \gtrsim \log M$. Recall that

$$C_{\mathcal{M}} \leq \left\lceil S_{\mathcal{M}} T_{d^*} / r^{d^*} \right\rceil \leq \left\lceil 7 S_{\mathcal{M}} d^* \log(d^*) / r^{d^*} \right\rceil \leq C_1 S_{\mathcal{M}} (2/\widetilde{\tau})^{d^*} d^* \log(d^*)$$

for some universal constant $C_1 > 0$, so the width can be set as

$$G \gtrsim S_{\mathcal{M}} (2/\widetilde{\tau})^{d^*} d^* \log(d^*) M^{d^*/(d^*+1)} \log M.$$

Then we have

$$|f(x) - \tilde{g}(x)| \lesssim M^{-\zeta/(d^*+1)}.$$

By Theorem 7 and Lemma 20, if we set $M \asymp \mu_n^{(d^*+1)/(2\zeta+2d^*+2)}$, then the excess risk satisfies

$$\mathbb{E}\left[\mathcal{R}(\hat{f}_n)\right] - \mathcal{R}(f^*) \lesssim \lambda \mu_n^{-\zeta/(2\zeta+2d^*+2)} \sqrt{\log(d^* \mu_n)} + \lambda \mu_n \beta_{a_n}.$$

Moreover, if we set $\mu_n = \frac{n}{(\log n)^\tau}$ for some constant $\tau > 0$ and assume that $\{Z_i\}_{i=1}^n$ is exponentially β-mixing with parameters $\bar{\beta}, b, \eta$ satisfying $ne^{-b(\log n)^{\eta\tau}/2^\eta} \lesssim n^{\frac{-\zeta}{2\zeta+2d^*+2}}$, then we obtain that

$$\mathbb{E}\left[\mathcal{R}(\hat{f}_n)\right] - \mathcal{R}\left(f^*\right) \lesssim \lambda \left(\frac{n}{(\log n)^\tau}\right)^{-\zeta/(2\zeta+2d^*+2)} \sqrt{\log(d^* n)}.$$

■

A.5 Proof of Theorem 21

Proof Firstly, we bound the statistical error $\sup_{Q\in\mathcal{F}(G,L,M)} \left|\mathcal{L}(Q) - \widehat{\mathcal{L}}(Q)\right|$ with the similar argument of the proof of Theorem 7. Let $\widetilde{Z}_i$ be a independent copy of Z_i, and σ_i's be the i.i.d. Rademacher random variables that are independent with $\widetilde{Z}_i$ and Z_i, $i = 1, \ldots, n$. Denote the composite function class

$$\begin{aligned}\ell \circ \mathcal{F}(G, L, M) :=\Big\{\ell_Q : \ell_Q(x, a, r, x') = \Big(Q(x, a) - r - \gamma \max_{a'\in A} \widehat{Q}_{j-1}(x', a')\Big)^2,\\ Q \in \mathcal{F}(G, L, M)\Big\}.\end{aligned}$$

Thus, it follows that

$$\begin{aligned}\sup_{Q\in\mathcal{F}(G,L,M)} \left|\widehat{\mathcal{L}}(Q) - \mathcal{L}(Q)\right| &= \sup_{Q\in\mathcal{F}(G,L,M)} \left|\frac{1}{n}\sum_{i=1}^n \left(Q(X_i, A_i) - Y_i\right)^2 - \mathbb{E}\left(Q(X_i, A_i) - Y_i\right)^2\right| \\ &=: \sup_{Q\in\mathcal{F}(G,L,M)} \left|\frac{1}{n}\sum_{i=1}^n \ell_Q(X_i, A_i, R_i, X_i') - \mathbb{E}\ell_Q(X_i, A_i, R_i, X_i')\right|.\end{aligned}$$

Therefore we have

$$\begin{aligned}
&\mathbb{E} \sup_{Q\in\mathcal{F}(G,L,M)} \left| \frac{1}{n}\sum_{i=1}^{n} \ell_Q(X_i,A_i,R_i,X_i') - \mathbb{E}\ell_Q(X_i,A_i,R_i,X_i') \right| \\
&= \mathbb{E} \sup_{Q\in\mathcal{F}(G,L,M)} \left| \frac{1}{n}\sum_{i=1}^{n} \ell_Q(X_i,A_i,R_i,X_i') - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\ell_Q(\widetilde{X}_i,\widetilde{A}_i,\widetilde{R}_i,\widetilde{X}_i') \right| \\
&\le \mathbb{E} \sup_{Q\in\mathcal{F}(G,L,M)} \left| \frac{1}{n}\sum_{i=1}^{n} \ell_Q(X_i,A_i,R_i,X_i') - \frac{1}{n}\sum_{i=1}^{n} \ell_Q(\widetilde{X}_i,\widetilde{A}_i,\widetilde{R}_i,\widetilde{X}_i') \right| \\
&= \mathbb{E} \sup_{Q\in\mathcal{F}(G,L,M)} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i(\ell_Q(X_i,A_i,R_i,X_i') - \ell_Q(\widetilde{X}_i,\widetilde{A}_i,\widetilde{R}_i,\widetilde{X}_i')) \right| \\
&\le 2\mathbb{E} \sup_{Q\in\mathcal{F}(G,L,M)} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i\ell_Q(X_i,A_i,R_i,X_i') \right| \\
&\le 2\mathbb{E} \left(\sup_{f\in\ell\circ\mathcal{F}(G,L,M)} \left| \frac{1}{\mu_n}\sum_{j=1}^{\mu_n} \frac{W_{j,f}(\Xi_{a_n})}{a_n} \right| \right) + 4\widetilde{M}\mu_n\beta_{a_n} \\
&\lesssim \frac{\widetilde{M}M\sqrt{L+2+\log(d+1)}}{\sqrt{\mu_n}} + \widetilde{M}\mu_n\beta_{a_n},
\end{aligned}$$

where the first inequality follows from the Jensen's inequality, and the second equality holds since both $\sigma_i\ell_f(Z_i)$ and $\sigma_i\ell_f(\widetilde{Z}_i)$ are governed by the same law, the third inequality holds by Lemma 30 and $\ell\circ\mathcal{F}(G,L,M)$ being bounded by $\widetilde{M} \lesssim \frac{R_{\max}^2}{(1-\gamma)^2}$, and the last inequality directly follows from Theorem 2 of Golowich et al. (2018) since $(W_{j,f}(\Xi_{a_n})/a_n)$'s are i.i.d. and bounded by $\widetilde{M}$.

Secondly, by Theorem 4, for any $f \in \mathcal{H}^\zeta$ bounded by $\frac{R_{\max}}{1-\gamma}$, there exists a function $\psi \in \mathcal{F}(G,L,M)$ bounded by $\frac{R_{\max}}{1-\gamma}$ with width $G \gtrsim M^{d/(d+1)}\log M$ and depth $L \asymp \log M$ such that

$$|f(x)-\psi(x)| \lesssim \frac{R_{\max}M^{-\zeta/(d+1)}}{1-\gamma},$$

for all $x\in[0,1]^d$. By Lemma 20, it yields that

$$\mathbb{E}[\|\widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}\|_{L_2(\mu)}^2] \lesssim \frac{R_{\max}^2 M^{-2\zeta/(d+1)}}{(1-\gamma)^2} + \frac{R_{\max}^2}{(1-\gamma)^2}\left(\frac{M\sqrt{L+2+\log(d+1)}}{\sqrt{\mu_n}} + \mu_n\beta_{a_n} \right).$$

Setting $M \asymp \mu_n^{(d+1)/(4\zeta+2d+2)}$, then it follows that

$$\mathbb{E}[\|\widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}\|_{L_2(\mu)}^2] \lesssim \frac{R_{\max}^2}{(1-\gamma)^2}\left(\mu_n^{-\zeta/(2\zeta+d+1)}\sqrt{\log(d\mu_n)} + \mu_n\beta_{a_n} \right).$$

Moreover, if we set $\mu_n = \frac{n}{(\log n)^\tau}$ for some constant $\tau>0$ and assume that $\{Z_i\}_{i=1}^n$ is exponentially β-mixing with parameters $\bar{\beta}, b, \eta$ defined in Definition 6, then we obtain that

$$\mathbb{E}[\|\widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}\|_{L_2(\mu)}^2] \lesssim \frac{R_{\max}^2}{(1-\gamma)^2}\left[\left(\frac{n}{(\log n)^\tau}\right)^{-\zeta/(2\zeta+d+1)} \sqrt{\log(dn)} + \frac{\bar{\beta}n}{(\log n)^\tau} e^{-b(\log n)^{\eta\tau}/2^\eta} \right].$$

■

A.6 Proof of Theorem 23

Proof By Proposition 19 and Theorem 21, we have

$$\mathbb{E}\left[\|Q^* - Q^{\pi_J}\|_{L_1(\nu)}\right] \lesssim \frac{C_{\nu,\mu}\gamma R_{\max}}{(1-\gamma)^3}\left(\mu_n^{-\zeta/(4\zeta+2d+2)}(\log(d\mu_n))^{1/4} + \sqrt{\mu_n\beta_{a_n}}\right) + \frac{\gamma^{J+1}R_{\max}}{(1-\gamma)^2}.$$

Moreover, if we set $\mu_n = \frac{n}{(\log n)^\tau}$ for some constant $\tau > 0$, and assume that $\{Z_i\}_{i=1}^n$ is exponentially β-mixing with parameters $\bar{\beta}, b, \eta$ satisfying $ne^{-b(\log n)^{\eta\tau}/2^\eta} \lesssim n^{\frac{-\zeta}{2\zeta+d+1}}$, we have

$$\begin{aligned}\mathbb{E}[\|Q^* - Q^{\pi_J}\|_{L_1(\nu)}] \lesssim \frac{C_{\nu,\mu}\gamma R_{\max}}{(1-\gamma)^3}\Big[& (n/(\log n)^\tau)^{-\zeta/(4\zeta+2d+2)}(\log(d\mu_n))^{1/4} \\ & + \frac{\sqrt{\bar{\beta}n}}{(\log n)^{\tau/2}}e^{-b(\log n)^{\eta\tau}/2^{(\eta+1)}}\Big] + \frac{\gamma^{J+1}R_{\max}}{(1-\gamma)^2} \\ \lesssim \frac{C_{\nu,\mu}\gamma R_{\max}}{(1-\gamma)^3}\left(\frac{n}{(\log n)^\tau}\right)^{-\zeta/(4\zeta+2d+2)} & (\log(dn))^{1/4} + \frac{\gamma^{J+1}R_{\max}}{(1-\gamma)^2}.\end{aligned}$$

■

References

Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1): 1359–1399, 2019.

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pages 10–4, 2019.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

András Antos, Csaba Szepesvári, and Rémi Munos. Fitted q-iteration in continuous action-space mdps. *Advances in Neural Information Processing Systems*, 20:9–16, 2007.

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

Richard G Baraniuk and Michael B Wakin. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

Peter Bartlett. For valid generalization the size of the weights is more important than the size of the network. *Advances in neural information processing systems*, 9, 1996.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2018.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019b.

Julius Berner, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. The modern mathematics of deep learning. *arXiv preprint arXiv:2105.04026*, 2021.

Dzmitry Bahdanau Philemon Brakel, Kelvin Xu Anirudh Goyal, Ryan Lowe Joelle Pineau, Aaron Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. In *Conference ICLR*, 2017.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient approximation of deep relu networks for functions on low dimensional manifolds. *Advances in Neural Information Processing Systems*, 32, 2019.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep relu networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4):1203–1253, 2022.

Xiaohong Chen and Yanqin Fan. Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, 130(2):307–335, 2006.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

John Horton Conway and Neil James Alexander Sloane. *Sphere packings, lattices and groups*, volume 290. Springer Science & Business Media, 2013.

Yu A Davydov. Mixing conditions for markov chains. *Theory of Probability & Its Applications*, 18(2):312–328, 1974.

Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer, 2018.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. *arXiv preprint arXiv:1901.00137*, 2019.

Jianqing Fan, Yihong Gu, and Wen-Xin Zhou. How do noise tails impact on deep relu networks? *arXiv preprint arXiv:2203.10418*, 2022.

Amir-massoud Farahmand. Regularization in reinforcement learning. 2011.

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized fitted q-iteration for planning in continuous-space markovian decision problems. In *2009 American Control Conference*, pages 725–730. IEEE, 2009.

Amir Massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, 2010.

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

Charles Fefferman. Whitney's extension problem for. *Annals of Mathematics*, pages 313–359, 2006.

Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G Bellemare, Joelle Pineau, et al. An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, 11(3-4):219–354, 2018.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR, 2019.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

Laszlo Gyorfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

Hanyuan Hang and Ingo Steinwart. A bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics*, 45(2): 708–743, 2017.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2): 949–986, 2022.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

John H Hubbard and Barbara Burke Hubbard. *Vector calculus, linear algebra, and differential forms: a unified approach.* Matrix Editions, 2015.

Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 869–878. PMLR, 2019.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximately low-dimensional manifolds. *arXiv preprint arXiv:2104.06708*, 2021.

Yuling Jiao, Yang Wang, and Yunfei Yang. Approximation bounds for norm constrained neural networks with applications to regression and gans. *Applied and Computational Harmonic Analysis*, 65:249–278, 2023.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

Michael Kohler and Adam Krzyzak. Over-parametrized deep neural networks minimizing the empirical risk do not generalize well. *Bernoulli*, 27(4):2564–2597, 2021.

Michael Kohler and Sophie Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13:3041–3074, 2012.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Analysis of classification-based policy iteration algorithms. *Journal of Machine Learning Research*, 17:1–30, 2016.

J.M. Lee. *Introduction to Smooth Manifolds*. Springer New York, 2003.

John M Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, 2018.

Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

Han-Ying Liang, Deli Li, and Yongcheng Qi. Strong convergence in nonparametric regression with truncated dependent data. *Journal of Multivariate Analysis*, 100(1):162–174, 2009.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel “ridgeless” regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 2022.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.

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. Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567, 2003.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Susan A Murphy. A generalization error for q-learning. *Journal of Machine Learning Research*, 6:1073–1097, 2005.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *The Journal of Machine Learning Research*, 21:174–1, 2020.

Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1):419–441, 2008.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

Martin Riedmiller. Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In *European conference on machine learning*, pages 317–328. Springer, 2005.

Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon, Boris Lesner, and Matthieu Geist. Approximate modified policy iteration and its application to the game of tetris. *The Journal of Machine Learning Research*, 16:1629–1676, 2015.

Johannes Schmidt-Hieber. Deep relu network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*, 2019.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Guohao Shen, Yuling Jiao, Yuanyuan Lin, and Jian Huang. Robust nonparametric regression with deep neural networks. *arXiv preprint arXiv:2107.10343*, 2021.

Zuowei Shen. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Nonlinear approximation via compositions. *Neural Networks*, 119:74–84, 2019.

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.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* MIT press, 2018.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2018.

Taiji Suzuki and Atsushi Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic besov space. *Advances in Neural Information Processing Systems*, 34, 2021.

Matus Telgarsky. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*, 2015.

Matus Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pages 1517–1539. PMLR, 2016.

Samuele Tosatto, Matteo Pirotta, Carlo Eramo, and Marcello Restelli. Boosted fitted q-iteration. In *International Conference on Machine Learning*, pages 3434–3443. PMLR, 2017.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, 24:123–1, 2023.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2009.

Loring W Tu. Manifolds. In *An Introduction to Manifolds*, pages 47–83. Springer, 2011.

Aad W Van Der Vaart, Adrianus Willem van der Vaart, Aad van der Vaart, and Jon Wellner. *Weak convergence and empirical processes: with applications to statistics.* Springer Science & Business Media, 1996.

Kam Chung Wong, Zifan Li, and Ambuj Tewari. Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, 48(2):1124–1142, 2020.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Conference on Learning Theory*, pages 639–649. PMLR, 2018.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Appl. Comput. Harmon. Anal.*, 48(2):787–794, 2020.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109:467–492, 2020.