REPRESENTATION AND STATISTICAL PROPERTIES OF DEEP NEURAL NETWORKS ON STRUCTURED DATA

A Dissertation
Presented to
The Academic Faculty

By

Minshuo Chen

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Engineering
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology

August 2022

REPRESENTATION AND STATISTICAL PROPERTIES OF DEEP NEURAL NETWORKS ON STRUCTURED DATA

Thesis committee:

Dr. Tuo Zhao, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Wenjing Liao, Co-advisor
Department of Mathematics
Georgia Institute of Technology

Dr. Alexander Shapiro
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yajun Mei
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Hongyuan Zha
School of Data Science
Chinese University of Hong Kong, Shen Zhen

Date approved: June 22nd, 2022

To my family.

ACKNOWLEDGMENTS

First and foremost I would like to express deep gratitude to my advisors, Dr. Tuo Zhao and Dr. Wenjing Liao. They not only gave me guidelines for acquiring advanced knowledge in interdisciplinary areas, but encouraged me to continuously learn and practice. Their research vision, patience, and enthusiasm are indispensable enzyme to my growth during the five-year study. I am extraordinarily grateful for the opportunity they have given to me as a part of their group over the past exciting years.

I would like to extend my gratitude to each and every member in my thesis committee. I am sincerely appreciating Dr. Alexander Shapiro for his willingness to serve on my thesis committee. I still remembered the encourage from Dr. Yajun Mei, when I first came to Georgia Tech for a campus visit five years ago. I have benefited a lot from collaborating with Dr. Hongyuan Zha, absorbing his insightful advice.

Special thanks are due to faculty outside my dissertation committee over the years – Dr. Jason Lee, Dr. Molei Tao, Dr. Hua Wang, Dr. Mengdi Wang, Dr. Zhaoran Wang, Dr. Yao Xie, Dr. Lin F. Yang, Dr. Zhuoran Yang, Dr. Haomin Zhou – I have learned much from all of you, and I look forward with anticipation to our future interactions. I am also indebted to my supervisors, Yu Bai and Pengcheng He, during my internships at Salesforces research and Microsoft. They exposed me to the broader area of machine learning and deep learning.

In my graduate study, I am fortunate to interact with a number of talented graduate/undergraduate students, within or outside my research group. I would like to thank Zhehui Chen, Siawpeng Er, Xiang Ji, Haoming Jiang, Yan Li, Hao Liu, Tianyi Liu, Yuqing Wang, and Yujia Xie for interesting and fruitful collaborations. I am more than eager to see our paths crossing in the near future. In addition, I extend my thanks to Zaiwei Chen, Chen Liang, Guanyi Wang, Liyan Xie, and Shixiang Zhu for many joint activities in Atlanta, e.g., hiking and badminton. These are enjoyable diversions from intensive days.

The author gratefully acknowledges the support from H. Milton Steward School of

Industrial and Systems Engineering and Machine Learning program at Georgia Tech. Last but not the least, my deep thanks go to my family, without whom I could not complete any endeavor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Significant success of deep learning has brought unprecedented challenges to conventional wisdom in statistics, optimization, and applied mathematics. In many high-dimensional applications, e.g., image data of hundreds of thousands of pixels, deep learning is remarkably scalable and mysteriously generalizes well. Although such appealing behavior stimulates wide applications, a fundamental theoretical challenge – curse of data dimensionality – naturally arises. Roughly put, the sample complexity in practical applications is significantly smaller than that predicted by theory. It is a common belief that deep neural networks are good at learning various geometric structures hidden in data sets. However, little theory has been established to explain such a power. This thesis aims to bridge the gap between theory and practice by studying function approximation and statistical theories of deep neural networks in exploitation of geometric structures in data.

$\star$ **Function Approximation Theories on Low-dimensional Manifolds using Deep Neural Networks**. Function approximation by neural networks in Euclidean spaces has been extensively studied in literature; the obtained rate of approximation, however, is extremely slow in high dimensions.

In Chapter 3, we first develop an efficient universal approximation theory for α-Hölder functions on a d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$ ($d \ll D$). A feedforward network architecture is constructed for function approximation, where the size of the network grows like $\epsilon^{-d/\alpha}$ with ϵ being the approximation error.

Furthermore, we prove efficient approximation theory for convolutional residual networks in approximating Besov functions. We highlight two important contributions: 1) Besov functions generalize Hölder functions by allowing inhomogeneous spatially varying regularity; 2) Convolutional residual networks share the universal approximation ability, while are free of cardinality constraints on weight parameters. As a consequence, convolutional residual networks are shown to enjoy adaptability to both data intrinsic structures

and function non-uniform regularity.

The aforementioned theories focus on function approximation in terms of function value (function L^∞-norm). Nonetheless, the constructed neural network approximator can be highly nonsmooth. In this regard, we demonstrate the benefit of overparameterized neural networks in function approximation. Specifically, we show that large neural networks are capable of accurately approximating a target function, and the network itself enjoys benign Lipschitz continuity. This theory partially justifies the appealing performance and robustness of deep and wide networks in practice.

$\star$ **Statistical Theories on Low-dimensional Data Structures using Deep Neural Networks**. Efficient approximation theories of neural networks provide valuable guidelines to properly choose network architectures, when data exhibit geometric structures. In combination with statistical tools, we prove that neural networks can circumvent the curse of data dimensionality and enjoy fast statistical convergence in various learning problems.

In Chapter 4, we consider nonparametric regression/classification problems. Suppose for example, the ground truth regression function is α-Hölder continuous and the response is contaminated by sub-Gaussian noise. By minimizing the empirical squared loss over a proper neural network class, the obtained empirical risk minimizer converges to the ground truth at a rate $\widetilde{O}\left(n^{-\frac{\alpha}{2\alpha+d}}\right)$, where n is the sample size. Similar results are extended to binary classification problems.

In Chapter 5, we study distribution estimation using Generative Adversarial Networks (GANs). The target data distribution has either a nonparametric density function or intrinsic low-dimensional structures. In both cases, we show that the data distribution can be represented as a pushforward distribution by the generator network. We in addition establish sample complexity bounds of GANs in estimating these distributions under the strong Wasserstein distance. In particular, we show that GANs are adaptive to intrinsic structures in data.

In Chapter 6, we consider doubly-robust policy learning using neural networks, when

the covariate has low-dimensional structures. We show nonasymptotic regret bounds of the learned policy competing with optimal oracle policies in both discrete action and continuous action scenarios. This result amplifies the adaptability and flexibility of neural networks.

CHAPTER 1
INTRODUCTION

Deep learning has made astonishing breakthroughs in various real-world applications, such as computer vision [1, 2, 3], natural language processing [4, 5, 6], healthcare [7, 8], robotics [9], etc. For example, in image classification, the winner of the 2017 ImageNet challenge retained a top-5 error rate of 2.25% [10], while the data set consists of about 1.2 million labeled high resolution images in 1000 categories. In speech recognition, [11] reported that deep neural networks outperformed humans with a 5.15% word error rate on the LibriSpeech corpus constructed from audio books [12]. Such a data set consists of approximately 1000 hours of 16kHz read English speech from 8000 audio books.

The empirical success of deep learning brings new challenges to the conventional wisdom of machine learning. Data sets in these applications are in high-dimensional spaces. In existing literature, a minimax lower bound has been established for the optimal algorithm of learning C^s functions in $\mathbb{R}^D$ [13, 14]. Denote the underlying function by f_0. The minimax lower bound suggests a pessimistic sample complexity: To obtain an estimator $\widehat{f}$ for each C^s function f_0 with an ϵ-error, uniformly for all C^s functions (i.e., $\sup_{f_0 \in C^s} \|\widehat{f} - f_0\|_{L_2} \leq \epsilon$ with $\|\cdot\|_{L_2}$ denoting the function L_2 norm), the optimal algorithm requires the sample size $n \gtrsim \epsilon^{-\frac{2s+D}{s}}$ in the worst scenario (i.e., when f_0 is the most difficult for the algorithm to estimate). We instantiate such a sample complexity bound to the ImageNet data set, which consists of RGB images with a resolution of 224×224. The theory above suggests that, to achieve an ϵ-error, the number of samples has to scale as $\epsilon^{-224 \times 224 \times 3/s}$, where the smoothness parameter s is significantly smaller than $224 \times 224 \times 3$. Setting $\epsilon = 0.1$ already gives rise to a huge number of samples far beyond practical applications, which well exceeds 1.2 million labeled images in ImageNet. This is known as the *curse of data dimensionality*.

To bridge the aforementioned huge gap between theory and practice, we take the *low* *dimensional geometric structures* in data sets into consideration. This is motivated by the fact that real-world data sets often exhibit low dimensional structures. Many images consist of projections of a three-dimensional object followed by some transformations, such as rotation, translation, and skeleton. This generating mechanism induces a small number of intrinsic parameters [15, 16]. Speech data are composed of words and sentences following the grammar, and therefore have small degrees of freedom [17]. More broadly, visual, acoustic, textual, and many other types of data often have low dimensional geometric structures due to rich local regularities, global symmetries, repetitive patterns, or redundant sampling [18, 19, 20, 21]. It is therefore reasonable to assume that data lie on a manifold $\mathcal{M}$ of dimension $d \ll D$.

This thesis shows that deep neural networks are adaptive to data intrinsic structures in function approximation and statistical learning problems, and partially explains why deep learning is free of the curse of data ambient dimensionality. We summarize the main contributions as follows.

Function Approximation Theories on Low-dimensional Manifolds using Deep Neural Networks. In Chapter 3, we first develop an efficient universal approximation theory for α-Hölder functions on a d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$ ($d \ll D$). A feedforward network architecture is constructed for function approximation, where the size of the network grows like $\epsilon^{-d/\alpha}$ with ϵ being the approximation error.

Furthermore, we prove efficient approximation theory for convolutional residual networks in approximating Besov functions. We highlight two important contributions: 1) Besov functions generalize Hölder functions by allowing inhomogeneous spatially varying regularity; 2) Convolutional residual networks share the universal approximation ability, while are free of cardinality constraints on weight parameters. As a consequence, convolutional residual networks are shown to enjoy adaptability to both data intrinsic structures and function non-uniform regularity.

The last result goes beyond function value approximation, and ensures the obtained approximator having good continuity. Specifically, we demonstrate the benefit of overparameterized neural networks in function approximation. We show that large neural networks are capable of accurately approximating a target function in L_∞ norm, and the network itself enjoys desired Lipschitz continuity. This theory partially justifies the appealing performance and robustness of deep and wide networks in practice.

Statistical Theories on Low-dimensional Data Structures using Deep Neural Networks. In Chapter 4, we consider nonparametric regression/classification problems. Suppose for example, the ground truth regression function is α-Hölder continuous and the response is contaminated by sub-Gaussian noise. By minimizing the empirical squared loss over a proper neural network class, the obtained empirical risk minimizer converges to the ground truth at a rate $\widetilde{O}\left(n^{-\frac{\alpha}{2\alpha+d}}\right)$, where n is the sample size. Similar results are extended to binary classification problems.

In Chapter 5, we study distribution estimation using Generative Adversarial Networks (GANs). The target data distribution has either a nonparametric density function or intrinsic low dimensional structures. In both cases, we show that the data distribution can be represented as a pushforward distribution by the generator network. We in addition establish sample complexity bounds of GANs in estimating these distributions under the strong Wasserstein distance. In particular, we show GANs are adaptive to intrinsic structures in data.

In Chapter 6, we consider doubly-robust policy learning using neural networks, when the covariate has low-dimensional structures. We show nonasymptotic regret bounds of the learned policy competing with optimal oracle policies in both discrete action and continuous action scenarios. This result amplifies the adaptability and flexibility of neural networks.

1.1 List of Notations

Frequently used notations are listed in the following table.

Table 1.1: List of notations.

$\mathbb{R}^D$	D-dimensional Euclidean space
$\mathcal{M}$	Low-dimensional Riemannian manifold
$\mathbf{x}, \mathbf{y}, \mathbf{z}$	Multi-dimensional vectors
$\mathbf{s}$	Multi-index of positive integers
$\\|\cdot\\|_{L^2}$	Function L^2 norm
$\\|\cdot\\|_2$	Vector Euclidean norm or matrix operator norm
$\\|\cdot\\|_\infty$	Function L^∞ norm or vector/matrix/tensor ℓ_∞ norm
$\lfloor a \rfloor$	The largest integer smaller than a
$\lceil a \rceil$	The smallest integer no smaller than a
C^k	k-th order continuously differentiable functions
$L^p(\mathcal{X})$	p-th order integrable functions on domain $\mathcal{X}$
$\mathcal{H}^\alpha(\mathcal{X})$	Hölder functions on domain $\mathcal{X}$
$W^{\alpha,p}(\mathcal{X})$	Sobolev functions on domain $\mathcal{X}$
$B^\alpha_{p,q}(\mathcal{X})$	Besov functions on domain $\mathcal{X}$
$a \vee b$	The smaller one of a and b
$a \wedge b$	The larger one of a and b
$W_1(\mu, \nu)$	Wasserstein-1 distance between μ and ν
$d_{\mathcal{H}^\beta}(\mu, \nu)$	β-Hölder IPM between μ and ν
$T_\sharp \rho$	Pushforward distribution of ρ under T

CHAPTER 2
PRELIMINARY

This chapter divides into three parts: 1) an introduction to manifolds; 2) definitions of function spaces in Euclidean spaces and manifolds; 3) neural network architectures. These preliminaries appear constantly in later chapters.

2.1 Riemannian Manifold

We briefly review manifolds and partition of unity defined on smooth manifolds. Details can be found in [22] and [23]. Let $\mathcal{M}$ be a d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$.

Definition 2.1 (Chart)**.** *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).*

The open set U is called a coordinate neighborhood, and ϕ is called a coordinate system on U. A chart essentially defines a local coordinate system on $\mathcal{M}$. Given a suitable coordinate neighborhood U around a point $\mathbf{c}$ on the manifold $\mathcal{M}$, we denote $\mathsf{P}_{\mathbf{c}}$ as the orthogonal projection onto the tangent space at $\mathbf{c}$, which gives a particular coordinate system on U.

Example 2.1 (Projection to Tangent Space)**.** *Let $T_{\mathbf{c}}(\mathcal{M})$ be the tangent space of $\mathcal{M}$ at the point $\mathbf{c} \in \mathcal{M}$ (see a formal definition in [22, Section 8.1]). We denote $\mathbf{v}_1, \ldots, \mathbf{v}_d$ as an orthonormal basis of $T_{\mathbf{c}}(\mathcal{M})$. Then the orthogonal projection onto the tangent space $T_{\mathbf{c}}(\mathcal{M})$ is defined as $\mathsf{P}_{\mathbf{c}}(\mathbf{x}) = V^\top(\mathbf{x} - \mathbf{c})$ for $\mathbf{x} \in U$ with $V = [\mathbf{v}_1, \ldots, \mathbf{v}_d] \in \mathbb{R}^{D \times d}$.*

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 2.2 (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 2.3 (Smooth Manifold)**.** *A smooth manifold is a manifold together with a C^∞ atlas.*

Classical examples of smooth manifolds are the Euclidean space $\mathbb{R}^D$, the torus, and the unit sphere. We further define a Riemannian manifold as a pair $(\mathcal{M}, g)$, where $\mathcal{M}$ is a smooth manifold and g is a Riemannian metric [24, Chapter 2]. To better interpret Definition 2.2 and 2.3, we give an example of a C^∞ atlas on the unit sphere in $\mathbb{R}^3$.

Example 2.2. *We denote $\mathbb{S}^2$ as the unit sphere in $\mathbb{R}^3$, i.e., $x^2 + y^2 + z^2 = 1$. The following atlas of $\mathbb{S}^2$ consists of 6 overlapping charts $(U_1, \mathsf{P}_1), \ldots, (U_6, \mathsf{P}_6)$ corresponding to hemispheres:*

$$\begin{aligned}
U_1 &= \{(x,y,z) \mid x > 0\},\ \mathsf{P}_1(x,y,z) = (y,z),\\
U_2 &= \{(x,y,z) \mid x < 0\},\ \mathsf{P}_2(x,y,z) = (y,z),\\
U_3 &= \{(x,y,z) \mid y > 0\},\ \mathsf{P}_3(x,y,z) = (x,z),\\
U_4 &= \{(x,y,z) \mid y < 0\},\ \mathsf{P}_4(x,y,z) = (x,z),\\
U_5 &= \{(x,y,z) \mid z > 0\},\ \mathsf{P}_5(x,y,z) = (x,y),\\
U_6 &= \{(x,y,z) \mid z < 0\},\ \mathsf{P}_6(x,y,z) = (x,y).
\end{aligned}$$

Here P_i is the orthogonal projection onto the tangent space at the pole of each hemisphere. Moreover, all the six charts are C^∞ compatible, and therefore, $(U_1, \mathsf{P}_1), \ldots, (U_6, \mathsf{P}_6)$ form an atlas of $\mathbb{S}^2$.

For a general compact smooth manifold $\mathcal{M}$, we can construct an atlas using orthogonal projections to tangent spaces as local coordinate systems. Let $\mathsf{P}_{\mathbf{c}}$ be the orthogonal *projection to the tangent space $T_{\mathbf{c}}(\mathcal{M})$ for $\mathbf{c} \in \mathcal{M}$. Let $U_{\mathbf{c}}$ be an open coordinate neighborhood containing $\mathbf{c}$ such that $\mathsf{P}_{\mathbf{c}}$ is a homeomorphism. Since $\mathcal{M}$ is compact, there exist a finite number of points $\{\mathbf{c}_i\}$ such that the charts $\{(U_{\mathbf{c}_i}, \mathsf{P}_{\mathbf{c}_i})\}$ form an atlas of $\mathcal{M}$.*

We next introduce the partition of unity, which plays a crucial role in our construction of neural network function approximators.

Definition 2.4 (Partition of Unity, Definition 13.4 in [22])**.** *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*

1. *the collection of supports, $\{supp(\rho_i)\}_{i\in\mathcal{A}}$ is locally finite, i.e., every point on $\mathcal{M}$ has a neighborhood that meets only finitely many of $\operatorname{supp}(\rho_i)$'s;*
2. *$\sum \rho_i = 1$.*

For a smooth manifold, a C^∞ partition of unity always exists.

Proposition 2.1 (Existence of a C^∞ partition of unity, Theorem 13.7 in [22])**.** *Let $\{U_i\}_{i\in\mathcal{A}}$ be an open cover of a compact smooth manifold $\mathcal{M}$. Then there is a C^∞ partition of unity $\{\rho_i\}_{i\in\mathcal{A}}$ where every ρ_i has a compact support such that $supp(\rho_i) \subset U_i$.*

Proposition 2.1 gives rise to the decomposition $f = \sum_{i=1}^{\infty} f_i$ with $f_i = f\rho_i$. Note that the f_i's have the same regularity as f, since

$$f_i \circ \phi_i^{-1} = (f \circ \phi_i^{-1}) \times (\rho_i \circ \phi_i^{-1})$$

for a chart (U_i, ϕ_i). This decomposition implies that we can express f as a sum of the f_i's, where every f_i is only supported in a single chart.

To characterize the curvature of a manifold, we adopt the following geometric concept.

Definition 2.5 (Reach [25], Definition 2.1 in [26])**.** *Denote*

$$\mathcal{C}(\mathcal{M}) = \left\{ \mathbf{x} \in \mathbb{R}^D : \exists\, \mathbf{p} \neq \mathbf{q} \in \mathcal{M}, \|\mathbf{p} - \mathbf{x}\|_2 = \|\mathbf{q} - \mathbf{x}\|_2 = \inf_{\mathbf{y}\in\mathcal{M}} \|\mathbf{y} - \mathbf{x}\|_2 \right\}$$

as the set of points that have at least two nearest neighbors on $\mathcal{M}$. The reach $\tau > 0$ is defined as

$$\tau = \inf_{\mathbf{x} \in \mathcal{M}, \mathbf{y} \in \mathcal{C}(\mathcal{M})} \|\mathbf{x} - \mathbf{y}\|_2 .$$

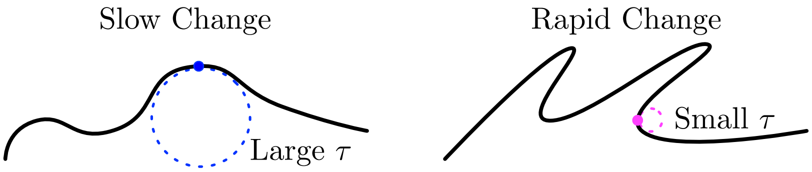


Figure 2.1: Manifolds with large and small reaches.

Reach has a straightforward geometrical interpretation: At each point $\mathbf{x} \in \mathcal{M}$, the radius of the osculating circle is greater or equal to τ. Intuitively, a large reach for $\mathcal{M}$ requires the manifold $\mathcal{M}$ not to change "rapidly" as shown in Figure 2.1.

In our proof for the universal approximation theory, reach determines a proper choice of an atlas for $\mathcal{M}$. Specifically, we choose each chart U_i to be contained in a ball of radius no larger than $\tau/4$. For smooth manifolds with a small τ, we need a large number of charts. Therefore, reach of a smooth manifold reflects the complexity of the neural network for function approximation on $\mathcal{M}$.

2.2 Function Space

We provide definitions of Hölder, Sobolev, and Besov functions in both Euclidean spaces and low dimensional Riemannian manifolds.

2.2.1 Regularity in Euclidean Space

We begin with Hölder functions.

Definition 2.6 (Hölder Function)**.** *Given a Hölder index $\alpha > 0$, a function $f : \mathcal{X} \mapsto \mathbb{R}$ belongs to the Hölder class $\mathcal{H}^\alpha(\mathcal{X})$, if and only if, for any multi-index $\mathrm{s} \in \mathbb{N}^d$ with $|\mathrm{s}| = \sum_{i=1}^d s_i \leq \lfloor \alpha \rfloor$, the derivative $\partial^{\mathrm{s}} f = \frac{\partial^{|s|} f}{\partial x_1^{s_1} \dots \partial x_d^{s_d}}$ exists, and for any s satisfying $|\mathrm{s}| = \lfloor \alpha \rfloor$,* *we have*

$$\sup_{x \neq y} \frac{\left|\partial^{\mathbf{s}} f(\mathbf{x}) - \partial^{\mathbf{s}} f(\mathbf{y})\right|}{\|\mathbf{x}-\mathbf{y}\|_2^{\alpha - \lfloor \alpha \rfloor}} < \infty \quad \textit{for any } \mathbf{x}, \mathbf{y} \textit{ in the interior of } \mathcal{X}.$$

When $f \in \mathcal{H}^{\alpha}(\mathcal{X})$, we define its Hölder norm as

$$\|f\|_{\mathcal{H}^{\alpha}(\mathcal{X})} = \sum_{0 \leq s \leq \lfloor \alpha \rfloor} \|\partial^{s} f\|_{\infty} + \sum_{|s| = \lfloor \alpha \rfloor} \sup_{x \neq y} \frac{\left|\partial^{\mathbf{s}} f(x) - \partial^{\mathbf{s}} f(y)\right|}{\|x-y\|_2^{\alpha - \lfloor \alpha \rfloor}}.$$

The Hölder continuity above can be generalized to multi-dimensional mappings. Specifically, for $g = [g_1, \ldots, g_d]^{\top} : \mathcal{X} \mapsto \mathbb{R}^d$, we say it is α-Hölder if and only if each coordinate mapping g_i is α-Hölder. In addition, the Hölder norm of g is defined as $\|g\|_{\mathcal{H}^{\alpha}(\mathcal{X})} = \sum_{i=1}^{d} \|g_i\|_{\mathcal{H}^{\alpha}(\mathcal{X})}$.

Next, we turn to Sobolev functions.

Definition 2.7 (Sobolev Function)**.** *Let $\alpha \geq 0, 1 \leq p \leq \infty$ be integers, and domain $\mathcal{X} \subset \mathbb{R}^D$. We define Sobolev space $W^{\alpha,p}(\mathcal{X})$ as*

$$W^{\alpha,p}(\mathcal{X}) = \left\{ f \in L^p(\mathcal{X}) : D^{\mathbf{s}} f \in L^p(\mathcal{X}) \textit{ for all } |\mathbf{s}| \leq \alpha \right\},$$

where s *is a multi-index.*

For $f \in W^{\alpha,p}(\mathcal{X})$, we define its Sobolev norm as

$$\|f\|_{W^{\alpha,p}(\mathcal{X})} = \Big(\sum_{|\mathbf{s}| \leq \alpha} \|D^{\mathbf{s}} f\|_{L^p(\mathcal{X})}^p \Big)^{1/p}.$$

In the special case of $p = \infty$, the Sobolev norm can be rewritten as $\|f\|_{W^{\alpha,\infty}(\mathcal{X})} = \max_{|\mathbf{s}| \leq \alpha} \|D^{\mathbf{s}} f\|_{L^{\infty}(\mathcal{X})}$. In this case, $\|f\|_{W^{0,\infty}} < \infty$ implies that the function value is bounded, and $\|f\|_{W^{1,\infty}} < \infty$ implies both the function value and its derivatives are bounded.

Our later approximation theories will provide error estimate in terms of Sobolev norms. To allow more flexibility, we define fractional Sobolev norms, which can be viewed as a generalization of Sobolev norms to non-integer α. The fractional Sobolev functions are defined as follows.

Definition 2.8 (Sobolev-Slobodeckij space [27]). *For $0 < \alpha < 1$ and $1 \leq p \leq \infty$, we define $W^{\alpha,p}(\mathcal{X})$ as*

$$W^{\alpha,p}(\mathcal{X}) = \big\{f \in L^p(\mathcal{X}) : \|f\|_{W^{\alpha,p}(\mathcal{X})} < \infty\big\}$$

with

$$\|f\|_{W^{\alpha,p}(\mathcal{X})} = \Big(\|f\|^p_{L^p(\mathcal{X})} + \int_{\mathcal{X}}\int_{\mathcal{X}}\Big(\frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\|\mathbf{x}-\mathbf{y}\|_2^{\alpha+D/p}}\Big)^p d\mathbf{x}d\mathbf{y}\Big)^{1/p}$$

for $1 \leq p < \infty$ and

$$\|f\|_{W^{\alpha,\infty}(\mathcal{X})} = \max\left\{\|f\|_{L^\infty(\mathcal{X})}, \text{ess sup}_{\mathbf{x},\mathbf{y}\in\mathcal{X}}\frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\|\mathbf{x}-\mathbf{y}\|_2^{\alpha}}\right\}.$$

We restrict our attention to $\alpha < 1$ for simplicity, as we later focus on approximation guarantees up to first-order continuity.

Lastly, we define Besov functions. The Besov space $B^\alpha_{p,q}$ is a complete quasi-normed space which is a Banach space when $1 \leq p, q \leq \infty$. It generalizes more elementary function spaces such as the Sobolev and Hölder spaces. To define Besov space, we introduce modulus of smoothness.

Definition 2.9 (Modulus of Smoothness [28]). *Let $\mathcal{X} \subset \mathbb{R}^D$. Let a function $f : \mathcal{X} \to \mathbb{R}$ be in $L^p(\mathcal{X})$ for $p > 0$, the r-th modulus of smoothness of f is defined by*

$$w_{r,p}(f, t) = \sup_{\|\mathbf{h}\|_2 \leq t} \|\Delta^r_{\mathbf{h}}(f)\|_{L^p},$$

where

$$\Delta_{\mathbf{h}}^r(f)(\mathbf{x}) = \begin{cases} \sum_{j=0}^r \binom{r}{j}(-1)^{r-j} f(\mathbf{x}+j\mathbf{h}) & \text{if } \mathbf{x} \in \mathcal{X}, \mathbf{x}+r\mathbf{h} \in \mathcal{X}, \\ 0 & \text{otherwise}. \end{cases}$$

Definition 2.10 (Besov Space $B_{p,q}^{\alpha}(\mathcal{X})$)**.** *For $0 < p, q \leq \infty, \alpha > 0, r = \lfloor \alpha \rfloor + 1$, define the seminorm $|\cdot|_{B_{p,q}^{\alpha}}$ as*

$$|f|_{B_{p,q}^{\alpha}} = \begin{cases} \left(\int_0^{\infty} (t^{-\alpha} w_{r,p}(f,t))^q \frac{dt}{t} \right)^{\frac{1}{q}} & \text{if } q < \infty, \\ \sup_{t>0} t^{-\alpha} w_{r,p}(f,t) & \text{if } q = \infty. \end{cases}$$

The norm of the Besov space $B_{p,q}^{\alpha}(\mathcal{X})$ is defined as $\|f\|_{B_{p,q}^{\alpha}} := \|f\|_{L^p} + |f|_{B_{p,q}^{\alpha}}$. The Besov space is $B_{p,q}^{\alpha}(\mathcal{X}) = \{f \in L^p(\mathcal{X}) : \|f\|_{B_{p,q}^{\alpha}} < \infty\}$.

2.2.2 Regularity on Riemannian Manifold

The existence of an atlas on a manifold allows us to generalize function regularity in Euclidean spaces to manifolds. Roughly speaking, regularity on manifold is characterized by local regularity on each neighborhood of the manifold. We first define differentiability of a function supported on a Riemannian manifold.

Definition 2.11 (C^s Function on Manifold)**.** *Let $\mathcal{M}$ be a d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$. A function $f : \mathcal{M} \mapsto \mathbb{R}$ is C^s if for any chart (U, ϕ), the composition $f \circ \phi^{-1} : \phi(U) \mapsto \mathbb{R}$ is continuously differentiable up to order s.*

Remark 2.1. *The definition of C^s functions is independent of the choice of the chart (U, ϕ). Suppose (V, ψ) is another chart and $V \bigcap U \neq \emptyset$. Then we have*

$$f \circ \psi^{-1} = (f \circ \phi^{-1}) \circ (\phi \circ \psi^{-1}).$$

Since $\mathcal{M}$ is a smooth manifold, (U, ϕ) and (V, ψ) are C^{∞} compatible. Thus, $f \circ \phi^{-1}$ is C^s

and $\phi \circ \psi^{-1}$ is C^∞, and their composition is C^s.

We next define Hölder functions on a smooth manifold $\mathcal{M}$.

Definition 2.12 (Hölder Function on $\mathcal{M}$)**.** *Let $\mathcal{M}$ be a d-dimensional compact Riemannian manifold isometrically embedded in $\mathbb{R}^D$. Let $\{(U_i, \mathsf{P}_i)\}_{i\in\mathcal{A}}$ be an atlas of $\mathcal{M}$ where the P_i's are orthogonal projections onto tangent spaces. For a positive index $\alpha > 0$, a function $f : \mathcal{M} \mapsto \mathbb{R}$ is α-Hölder continuous if for each chart (U_i, P_i) in the atlas, we have*

1. *$f \circ \mathsf{P}_i^{-1} \in C^s$ with $|D^{\mathbf{s}}(f \circ \mathsf{P}_i^{-1})| \leq 1$ for any $|\mathbf{s}| \leq \lfloor \alpha \rfloor, \mathbf{x} \in U_i$;*

2. *for any $|\mathbf{s}| = \lfloor \alpha \rfloor$ and $\mathbf{x}_1, \mathbf{x}_2 \in U_i$,*

$$\left| D^{\mathbf{s}}(f \circ \mathsf{P}_i^{-1})\big|_{\mathsf{P}_i(\mathbf{x}_1)} - D^{\mathbf{s}}(f \circ \mathsf{P}_i^{-1})\big|_{\mathsf{P}_i(\mathbf{x}_2)} \right| \leq \|\mathsf{P}_i(\mathbf{x}_1) - \mathsf{P}_i(\mathbf{x}_2)\|_2^{\alpha - \lfloor \alpha \rfloor}. \tag{2.1}$$

Moreover, we denote the collection of α-Hölder functions on $\mathcal{M}$ as $\mathcal{H}^\alpha(\mathcal{M})$.

Definition 2.12 requires that all s-th order derivatives of $f \circ \mathsf{P}_i^{-1}$ are Hölder continuous. We recover the standard Hölder class on a Euclidean space if P_i is the identity mapping. Following the same spirit, we define Sobolev functions on a manifold.

Definition 2.13 (Sobolev Function on Manifold)**.** *Let $\mathcal{M}$ be a compact Riemannian manifold of dimension d. Let $\{(U_i, \phi_i)\}_{i=1}^{C_\mathcal{M}}$ be a finite atlas on $\mathcal{M}$ and $\{\rho_i\}_{i=1}^{C_\mathcal{M}}$ be a partition of unity on $\mathcal{M}$ such that $\operatorname{supp}(\rho_i) \subset U_i$. For integers $\alpha \geq 0$ and $1 \leq p \leq \infty$, a function $f : \mathcal{M} \to \mathbb{R}$ is in the Sobolev space $W^{\alpha,p}(\mathcal{M})$ if*

$$\|f\|_{W^{\alpha,p}(\mathcal{M})} = \sum_{i=1}^{C_\mathcal{M}} \|(f\rho_i) \circ \phi_i^{-1}\|_{W^{\alpha,p}(\phi_i(U_i))} < \infty.$$

Here we no longer restrict our attention to projections used in Definition 2.12. In fact, we can replace P_i's in Definition 2.12 by any given diffeomorphism on a chart, since they are compatible as discussed in Remark 2.1.

Lastly, we define Besov functions on a manifold.

Definition 2.14 (Besov Function on Manifold). *Let $\mathcal{M}$ be a compact Riemannian manifold of dimension d. Let $\{(U_i, \phi_i)\}_{i=1}^{C_{\mathcal{M}}}$ be a finite atlas on $\mathcal{M}$ and $\{\rho_i\}_{i=1}^{C_{\mathcal{M}}}$ be a partition of unity on $\mathcal{M}$ such that* $\operatorname{supp}(\rho_i) \subset U_i$. *A function $f : \mathcal{M} \mapsto \mathbb{R}$ is in $B^{\alpha}_{p,q}(\mathcal{M})$ if*

$$\|f\|_{B^{\alpha}_{p,q}(\mathcal{M})} = \sum_{i=1}^{C_{\mathcal{M}}} \|(f\rho_i) \circ \phi_i^{-1}\|_{B^{\alpha}_{p,q}(\mathbb{R}^d)} < \infty.$$

Since ρ_i is supported on U_i, the function $(f\rho_i) \circ \phi_i^{-1}$ is supported on $\phi(U_i)$. We can extend $(f\rho_i) \circ \phi_i^{-1}$ from $\phi(U_i)$ to $\mathbb{R}^d$ by setting the function to be 0 in $\mathbb{R}^d \setminus \phi(U_i)$. The extended function lies in the Besov space $B^{\alpha}_{p,q}(\mathbb{R}^d)$ [29, Chapter 7].

2.3 Neural Network Architecture

2.3.1 Feedforward Neural Network

Feedforward neural networks has attracted many attentions due to its simple compositional form:

$$f(\mathbf{x}) = W_L \cdot \mathrm{ReLU}(W_{L-1} \cdots \mathrm{ReLU}(W_1\mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1}) + \mathbf{b}_L, \tag{2.2}$$

with W_i's and $\mathbf{b}_i$'s being weight matrices and intercepts, respectively. The ReLU activation function computes $\mathrm{ReLU}(a) = \max\{a, 0\}$ and is applied entrywise. We define the following feedforward network architecture:

$$\begin{aligned}\mathrm{FNN}(R, \kappa, L, p, K) = \Big\{ g \;\Big|\; & g \text{ in the form of (Eq. 2.2)}, \\ & \text{with } L \text{ layers and max width } p, \\ & \|g_i\|_\infty \leq R,\; \|W_i\|_\infty \leq \kappa,\; \|\mathbf{b}_i\|_\infty \leq \kappa, \\ & \sum_{j=1}^{L} \|W_i\|_0 + \|\mathbf{b}_i\|_0 \leq K, \text{ for } i = 1, \ldots, L \Big\},\end{aligned} \tag{2.3}$$

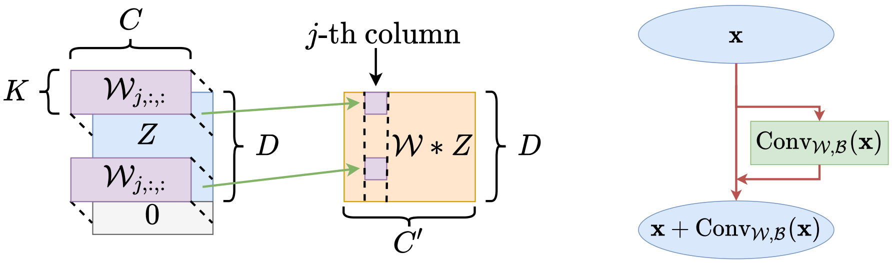
where $\|\cdot\|_0$ denotes the number of nonzero entries in a vector or a matrix.

2.3.2 Convolutional Residual Network

Convolutional Residual Networks (ConvResNets) are widely used in computer vision and natural language processing. We consider one-sided stride-one convolution in our network. Let $\mathcal{W} = \{\mathcal{W}_{j,k,l}\} \in \mathbb{R}^{C' \times K \times C}$ be a filter, where C' is the output channel size, K is the filter size and C is the input channel size. For $Z \in \mathbb{R}^{D \times C}$, the convolution of $\mathcal{W}$ with Z gives $Y = \mathcal{W} * Z \in \mathbb{R}^{D \times C'}$ with

$$Y_{i,j} = \sum_{k=1}^{K} \sum_{l=1}^{C} \mathcal{W}_{j,k,l} Z_{i+k-1,l},$$

where we set $Z_{i+k-1,l} = 0$ for $i+k-1 > D$. See a graphical demonstration in Figure 2.2(a).


C
K
$\mathcal{W}_{j,:,:}$
Z
$\mathcal{W}_{j,:,:}$
0
D
j-th column
$\mathcal{W} * Z$
D
C'
$\mathbf{x}$
$\text{Conv}_{\mathcal{W},\mathcal{B}}(\mathbf{x})$
$\mathbf{x} + \text{Conv}_{\mathcal{W},\mathcal{B}}(\mathbf{x})$

(a) Convolution.

(b) A residual block.

Figure 2.2: (a) Convolution of $\mathcal{W} * Z$, where the input is $Z \in \mathbb{R}^{D \times C}$, and the output is $\mathcal{W} * Z \in \mathbb{R}^{D \times C'}$. Here $\mathcal{W} = \{\mathcal{W}_{j,k,l}\} \in \mathbb{R}^{C' \times K \times C}$ is a filter where C' is the output channel size, K is the filter size and C is the input channel size. $\mathcal{W}_{j,:,:}$ is a $D \times C$ matrix for the j-th output channel. (b) A convolutional residual block.

In this thesis, we study ConvResNets equipped with the ReLU activation function. The ConvResNet we consider consists consecutively of a padding layer, several residual blocks, and finally a fully connected output layer.

Given an input vector $\mathbf{x} \in \mathbb{R}^D$, the network first applies a padding operator $P : \mathbb{R}^D \to \mathbb{R}^{D \times C}$ for some integer $C \geq 1$ such that

$$Z = P(\mathbf{x}) = \begin{bmatrix} \mathbf{x} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{D \times C}.$$

Then the matrix Z is passed through M residual blocks. To ease the notation, we denote the input matrix to the m-th block as Z_m and its output as Z_{m+1} (Consequently, $Z_1 = Z$).

In the m-th block, let $\mathcal{W}_m = \{\mathcal{W}_m^{(1)}, ..., \mathcal{W}_m^{(L_m)}\}$ and $\mathcal{B}_m = \{B_m^{(1)}, ..., B_m^{(L_m)}\}$ be a collection of filters and biases of proper sizes. The m-th residual block maps its input matrix Z_m from $\mathbb{R}^{D\times C}$ to $\mathbb{R}^{D\times C}$ by the operator

$$\mathrm{Conv}_{\mathcal{W}_m,\mathcal{B}_m} + \mathrm{id},$$

where id is the identity mapping (also known as the shortcut connection) and

$$\mathrm{Conv}_{\mathcal{W}_m,\mathcal{B}_m}(Z_m) = \mathrm{ReLU}\Big(\mathcal{W}_m^{(L_m)} * \cdots * \mathrm{ReLU}\left(\mathcal{W}_m^{(1)} * Z_m + B_m^{(1)}\right) \cdots + B_m^{(L_m)}\Big),$$

with ReLU applied entrywise. We denote the mapping from input $\mathbf{x}$ to the output of the M-th residual block as

$$Q(\mathbf{x}) = (\mathrm{Conv}_{\mathcal{W}_M,\mathcal{B}_M} + \mathrm{id}) \circ \cdots \circ (\mathrm{Conv}_{\mathcal{W}_1,\mathcal{B}_1} + \mathrm{id}) \circ P(\mathbf{x}). \tag{2.4}$$

Given (Eq. 2.4), a ConvResNet applies an additional fully connected layer to Q and outputs

$$f(\mathbf{x}) = W \otimes Q(\mathbf{x}) + b,$$

where $W \in \mathbb{R}^{D\times C}$ and $b \in \mathbb{R}$ are a weight matrix and a bias, respectively, and $\otimes$ denotes sum of entrywise product, i.e., $W \otimes Q(\mathbf{x}) = \sum_{i,j} W_{i,j}[Q(\mathbf{x})]_{i,j}$. To this end, we define ConvResNet architecture as

$$\begin{aligned}\mathrm{CRN}(M, L, J, K, \kappa_1, \kappa_2, R) = \Big\{ f \mid f(\mathbf{x}) = W \otimes Q(\mathbf{x}) + b \text{ with } \|W\|_\infty \vee |b| \le \kappa_2,\\ Q(\mathbf{x}) \text{ in the form of (Eq. 2.4) with } M \text{ residual blocks.}\end{aligned}$$

$$\begin{aligned}
&\text{The number of filters per block is bounded by } L;\\
&\text{filter size is bounded by } K;\\
&\text{the number of channels is bounded by } J;\\
&\max_{m,l} \|\mathcal{W}_m^{(l)}\|_\infty \vee \|B_m^{(l)}\|_\infty \leq \kappa_1, \|f\|_\infty \leq R\Big\}.
\end{aligned} \tag{2.5}$$

CHAPTER 3

REPRESENTATION THEORY OF NEURAL NETWORKS

A line of research attempts to explain the empirical success of neural networks through the lens of expressivity – neural networks can effectively approximate various classes of functions. Among existing works, the most well-known results are the universal approximation theorems, see [30, 31, 32, 33, 34, 35]. Specifically, [32] showed that neural networks with one single hidden layer and continuous sigmoidal activations ($\sigma(x)$ is sigmoidal, if $\sigma(x) \to 0$ as $x \to -\infty$, and $\sigma(x) \to 1$ as $x \to \infty$) can approximate continuous functions in a unit cube with arbitrary accuracy. Later, [33] extended the universal approximation theorem to general feed-forward networks with a single hidden layer, while the width of the network has to be exponentially large. Specific approximation rates of shallow networks (with one hidden layer) with smooth activation functions were given in [36] and [37]. Recently, [38] proved the universal approximation theorem for width-bounded deep neural networks, and [39] improved the result with ReLU (Rectified Linear Units) activations, i.e. ReLU$(x) = \max\{0, x\}$. [40] further showed that deep ReLU networks can uniformly approximate functions in Sobolev spaces, while the network size scales exponentially in the approximation error with an exponent depending on the data dimension. Moreover, the network size in [40] matches its lower bound.

The network size considered in applications, however, is significantly smaller than what is predicted by the theory above. Recall from Chapter 1 that the ImageNet consists of RGB images with a resolution of 224×224. The theory above suggests that, to achieve a ϵ uniform approximation error, the number of neurons has to scale as $\epsilon^{-224\times 224\times 3/2}$ [36]. Setting $\epsilon = 0.1$ already gives rise to $10^{224\times 224\times 3/2}$ neurons. However, the AlexNet [1] only consists of 650000 neurons and 60 million parameters to beat the state-of-the-art. To boost the performance on the ImageNet, several more sophisticated network structures were proposed later, such as VGG16 [41] which consists of about 138 million parameters. The size of both networks remains extremely small compared to $10^{224\times 224\times 3/2}$.

To bridge the gap between theory and practice, we exploit data manifold structures and aim to answer a natural question:

Can deep neural networks efficiently approximate functions supported on low dimensional manifolds?

In literature, function approximation on manifolds has been well studied using local polynomials [42] and wavelets [43]. However, studies using neural networks are very limited. Two noticeable works are [44] and [45]. In [44], high order differentiable functions on manifolds are approximated by neural networks with smooth activations, e.g., sigmoid activations and rectified quadratic unit functions ($\sigma(x) = (\max\{0, x\})^2$). These smooth activations, however, are rarely used in the mainstream applications such as computer vision [1, 3, 10]. In [45], a 4-layer network with ReLU activations was proposed to approximate C^2 functions on low dimensional manifolds that have absolutely summable wavelet coefficients. However, this theory does not cover arbitrarily smooth functions, and the analysis is built upon a restrictive assumption – there exists a linear transformation that maps the input data to sparse coordinates, but such transformation is not explicitly given.

In this chapter, we propose a framework to construct deep neural networks with nonsmooth activations to approximate functions supported on a d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$. We prove that, in order to achieve a fixed approximation error, the network size scales exponentially with respect to the intrinsic dimension d, instead of the ambient dimension D. Our framework is flexible: 1). It applies to popular network architectures with nonsmooth activations, e.g., ReLU and leaky ReLU activations; 2). It applies to a wide class of functions, such as Hölder, Sobolev, and Besov classes which are typical examples in nonparametric statistics [13]; 3). It exploits high order smoothness of functions for making the approximation as efficient as possible.

The rest of the chapter is organized as follows: Section 3.1 presents an efficient approximation theory using feedforward neural networks for approximating Hölder functions supported on a low dimensional Riemannian manifold; Section 3.2 extends to Convolutional Residual Networks (ConvResNets) for approximating Besov functions supported on a low dimensional Riemannian manifold; Section 3.3 demonstrates additional continuity of wide and deep neural networks in approximating Sobolev functions; Section 3.4 provides a conclusion and discusses related topics.

3.1 Efficient Approximation of Feedforward Neural Networks

We begin with some assumptions.

Assumption 3.1. *$\mathcal{M}$ is a d-dimensional compact Riemannian manifold isometrically embedded in $\mathbb{R}^D$. There exists a constant B such that for any point $\mathbf{x} \in \mathcal{M}$, we have $|x_i| \leq B$ for all $i = 1, \ldots, D$.*

Assumption 3.2. *The reach of $\mathcal{M}$ is $\tau > 0$.*

Assumption 3.3. *$f : \mathcal{M} \mapsto \mathbb{R}$ belongs to the Hölder space $\mathcal{H}^\alpha(\mathcal{M})$ with a positive index $\alpha > 1$.*

We now state our function approximation result.

Theorem 3.1. *Suppose Assumption 3.1 and 3.2 hold. Given any $\epsilon \in (0, 1)$, there exists a ReLU network structure such that, for any $f : \mathcal{M} \to \mathbb{R}$ satisfying Assumption 3.3, if the weight parameters are properly chosen, the network yields a function $\widehat{f}$ satisfying $\|\widehat{f} - f\|_\infty \leq \epsilon$. Such a network has no more than $c_1(\log \frac{1}{\epsilon} + \log D)$ layers, and at most $c_2(\epsilon^{-\frac{d}{\alpha}} \log \frac{1}{\epsilon} + D \log \frac{1}{\epsilon} + D \log D)$ neurons and weight parameters, where c_1, c_2 depend on d, α, f, τ, and the surface area of $\mathcal{M}$.*

The network structure identified by Theorem 3.1 consists of three sub-networks as shown in Figure 3.1:

- *Chart determination sub-network*, which assigns the input to its corresponding neighborhoods;
- *Taylor approximation sub-network*, which approximates f by polynomials in each neighborhood;
- *Pairing sub-network*, which yields multiplications of the proper pairs of outputs from the chart determination and the Taylor approximation sub-networks.

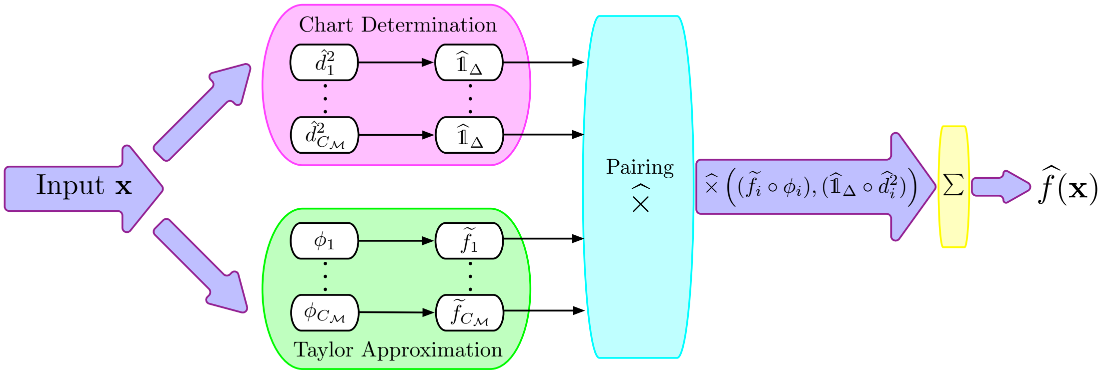


Chart Determination
$\widehat{d}_1^2$
$\widehat{\mathbb{1}}_\Delta$
$\widehat{d}_{C_\mathcal{M}}^2$
$\widehat{\mathbb{1}}_\Delta$
Input $\mathbf{x}$
ϕ_1
$\widetilde{f}_1$
$\phi_{C_\mathcal{M}}$
$\widetilde{f}_{C_\mathcal{M}}$
Taylor Approximation
Pairing
$\widehat{\times}$
$\widehat{\times}\left((\widetilde{f}_i \circ \phi_i), (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2)\right)$
Σ
$\widehat{f}(\mathbf{x})$

Figure 3.1: The ReLU network identified by Theorem 3.1.

Specifically, we partition the manifold as $\mathcal{M} = \bigcup_{i=1}^{C_\mathcal{M}} U_i$, where the U_i's are open sets contained in a Euclidean ball of radius no larger than $\tau/4$. $C_\mathcal{M}$ depends on the reach τ, the surface area of $\mathcal{M}$, and the dimension d (see Section 3.1.1 for an explicit characterization). For each chart, the chart determination sub-network computes an approximation of the indicator function on U_i. The Taylor approximation sub-network provides a local polynomial approximation of f on U_i. Then the pairing sub-network approximates the product for the proper pairs of outputs in the previous two sub-networks. Finally, $\widehat{f}$ is obtained by taking a sum over $C_\mathcal{M}$ outputs from the pairing sub-network.

The size of our ReLU network matches its lower bound up to a logarithmic factor for the approximation of functions in Hölder spaces. Denote $F^{\alpha,d}$ as functions defined on $[0,1]^d$ in the Hölder space $\mathcal{H}^\alpha([0,1]^d)$. We state a lower bound due to [46].

Theorem 3.2. *Fix d and α. Let W be a positive integer and $\kappa : \mathbb{R}^W \mapsto C([0,1]^d)$ be any mapping. Suppose there is a continuous map $\Theta : F^{\alpha,d} \mapsto \mathbb{R}^W$ such that $\|f - \kappa(\Theta(f))\|_\infty \leq$*

ϵ for any $f \in F^{\alpha,d}$. Then $W \geq c\epsilon^{-\frac{d}{\alpha}}$ with c depending on α only.

We take $\mathbb{R}^W$ as the parameter space of a ReLU network, and κ as the network structure. Then to approximate any $f \in F^{\alpha,d}$, the ReLU network has at least $c\epsilon^{-\frac{d}{\alpha}}$ weight parameters. Although Theorem 3.2 holds for functions on $[0,1]^d$, our network size remains in the same order up to a logarithmic factor even when the function is supported on a manifold of dimension d.

3.1.1 Proof – Construction of Network Approximator

This section contains a constructive proof of Theorem 3.1. Before we proceed, we show how to approximate the multiplication operation using ReLU networks. This operation is heavily used in the Taylor approximation sub-network, since Taylor polynomials involve a sum of products. We first show ReLU networks can approximate quadratic functions.

Lemma 3.1 (Proposition 2 in [40]). *The function $f(x) = x^2$ with $x \in [0,1]$ can be approximated by a ReLU network with any error $\epsilon > 0$. The network has depth and the number of neurons and weight parameters no more than $c\log(1/\epsilon)$ with an absolute constant c, and the width of the network is an absolute constant.*

This lemma is proved in Appendix A.1.1. The idea is to approximate quadratic functions using a weighted sum of a series of sawtooth functions. Those sawtooth functions are obtained by compositing the triangular function

$$g(x) = 2\text{ReLU}(x) - 4\text{ReLU}(x - 1/2) + 2\text{ReLU}(x - 1),$$

which can be implemented by a single layer ReLU network.

We then approximate the multiplication operation by invoking the identity $ab = \frac{1}{4}((a+b)^2-(a-b)^2)$ where the two squares can be approximated by ReLU networks in Lemma 3.1.

Corollary 3.1 (Proposition 3 in [40]). *Given a constant $C > 0$ and $\epsilon \in (0, C^2)$, there is a ReLU network which implements a function $\widehat{\times} : \mathbb{R}^2 \mapsto \mathbb{R}$ such that: **1)**. For all inputs x*

and y satisfying $|x| \leq C$ and $|y| \leq C$, we have $|\widehat{\times}(x, y) - xy| \leq \epsilon$; ***2).*** *The depth and the weight parameters of the network is no more than $c \log \frac{C^2}{\epsilon}$ with an absolute constant c.*

The ReLU network in Theorem 3.1 is constructed in the following 5 steps.

Step 1. Construction of an atlas. Denote the open Euclidean ball with center $\mathbf{c}$ and radius r in $\mathbb{R}^D$ by $\mathcal{B}(\mathbf{c}, r)$. For any r, the collection $\{\mathcal{B}(\mathbf{x}, r)\}_{\mathbf{x} \in \mathcal{M}}$ is an open cover of $\mathcal{M}$. Since $\mathcal{M}$ is compact, there exists a finite collection of points $\mathbf{c}_i$ for $i = 1, \ldots, C_{\mathcal{M}}$ such that $\mathcal{M} \subset \bigcup_i \mathcal{B}(\mathbf{c}_i, r)$.

The following lemma says that when the radius r is properly chosen, $U_i = \mathcal{B}(\mathbf{c}_i, r) \cap \mathcal{M}$ is diffeomorphic to $\mathbb{R}^d$.

Lemma 3.2. *Suppose Assumption 3.1 and 3.2 hold and let $r \leq \tau/4$. Then the local neighborhood $U_i = \mathcal{B}(\mathbf{c}_i, r) \cap \mathcal{M}$ is diffeomorphic to $\mathbb{R}^d$. In particular, the orthogonal projection P_i onto the tangent space $T_{\mathbf{c}_i}(\mathcal{M})$ at $\mathbf{c}_i$ is a diffeomorphism.*

The proof is provided in Appendix A.2.1, which utilizes the results in [47]. Therefore, we pick radius $r \leq \tau/4$, and let $\{(U_i, \phi_i)\}_{i=1}^{C_{\mathcal{M}}}$ be an atlas on $\mathcal{M}$ as illustrated in Figure 3.2, where ϕ_i is to be defined in **Step 2**. The number of charts $C_{\mathcal{M}}$ is upper bounded by

$$C_{\mathcal{M}} \leq \left\lceil \frac{SA(\mathcal{M})}{r^d} T_d \right\rceil,$$

where $SA(M)$ is the surface area of $\mathcal{M}$, and T_d is the thickness of the U_i's, which is defined as the average number of U_i's that contain a point on $\mathcal{M}$ (See Eq. (1) in Chapter 2 of [48]).

Remark 3.1. *The thickness T_d scales approximately linear in d. As shown in Eq. (19) in Chapter 2 of [48], there exist coverings with $\frac{d}{e\sqrt{e}} \lesssim T_d \leq d \log d + d \log \log d + 5d$.*

Step 2. Projection with rescaling and translation. We denote the tangent space at $\mathbf{c}_i$ as

$$T_{\mathbf{c}_i}(\mathcal{M}) = \text{span}(\mathbf{v}_{i1}, \ldots, \mathbf{v}_{id}),$$

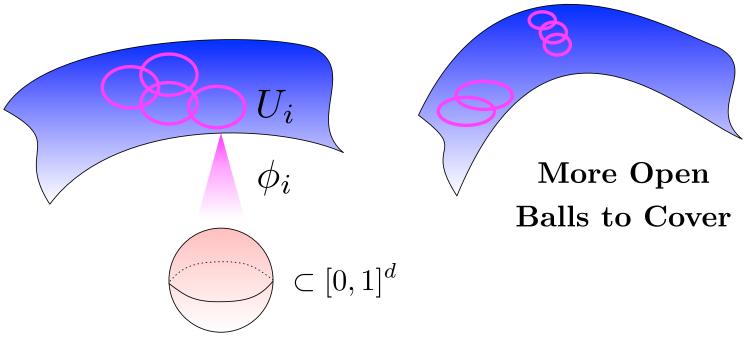


Figure 3.2: Curvature decides the number of charts: smaller reach requires more chart.

where $\{\mathbf{v}_{i1}, \dots, \mathbf{v}_{id}\}$ form an orthonormal basis. We obtain the matrix $V_i = [\mathbf{v}_{i1}, \dots, \mathbf{v}_{id}] \in \mathbb{R}^{D \times d}$ by concatenating the $\mathbf{v}_{ij}$'s as column vectors.

Define

$$\phi_i(\mathbf{x}) = b_i(V_i^\top(\mathbf{x} - \mathbf{c}_i) + \mathbf{u}_i) \in [0,1]^d$$

for any $\mathbf{x} \in U_i$, where $b_i \in (0,1]$ is a scaling factor and $\mathbf{u}_i$ is a translation vector. Since U_i is bounded, we can choose proper b_i and $\mathbf{u}_i$ to guarantee $\phi_i(\mathbf{x}) \in [0,1]^d$. We rescale and translate the projection to ease the notation for the development of local Taylor approximations in **Step 4**. We also remark that each ϕ_i is a linear function, and can be realized by a single layer linear network.

Step 3. Chart determination. This step is to assign a given input $\mathbf{x}$ to the proper charts to which $\mathbf{x}$ belongs. This avoids projecting $\mathbf{x}$ using unmatched charts (i.e., $\mathbf{x} \notin U_j$ for some j) as illustrated in Figure 3.3.

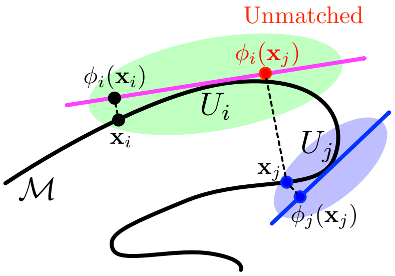


Figure 3.3: Projecting $\mathbf{x}_j$ using a matched chart (blue) (U_j, ϕ_j), and an unmatched chart (green) (U_i, ϕ_i).

An input $\mathbf{x}$ can belong to multiple charts, and the chart determination sub-network determines all these charts. This can be realized by compositing an indicator function and the squared Euclidean distance

$$d_i^2(\mathbf{x}) = \|\mathbf{x} - \mathbf{c}_i\|_2^2 = \sum_{j=1}^{D} (x_j - c_{i,j})^2$$

for $i = 1, \ldots, C_{\mathcal{M}}$. The squared distance $d_i^2(\mathbf{x})$ is a sum of univariate quadratic functions, thus, we can apply Lemma 3.1 to approximate $d_i^2(\mathbf{x})$ by ReLU networks. Denote $\widehat{h}_{\text{sq}}$ as an approximation of the quadratic function x^2 on $[0, 1]$ with an approximation error ν. Then we define

$$\widehat{d}_i^2(\mathbf{x}) = 4B^2 \sum_{j=1}^{D} \widehat{h}_{\text{sq}} \left(\left| \frac{x_j - c_{i,j}}{2B} \right| \right).$$

as an approximation of $d_i^2(\mathbf{x})$. The approximation error is $\|\widehat{d}_i^2 - d_i^2\|_\infty \leq 4B^2 D\nu$, by the triangle inequality. We consider an approximation of the indicator function $\mathbb{1}(x \in [0, r^2])$ as in Figure 3.4:

$$\widehat{\mathbb{1}}_\Delta(a) = \begin{cases} 1 & a \leq r^2 - \Delta + 4B^2 D\nu \\ -\frac{1}{\Delta - 8B^2 D\nu} a + \frac{r^2 - 4B^2 D\nu}{\Delta - 8B^2 D\nu} & a \in [r^2 - \Delta + 4B^2 D\nu, r^2 - 4B^2 D\nu] \\ 0 & a > r^2 - 4B^2 D\nu \end{cases}, \tag{3.1}$$

where Δ ($\Delta \geq 8B^2 D\nu$) will be chosen later according to the accuracy ϵ.

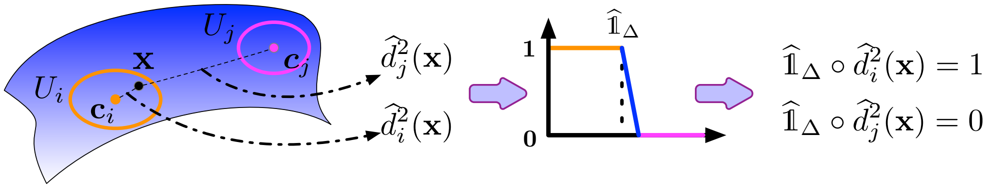

U_j
$\mathbf{x}$
$\boldsymbol{c}_j$
U_i
$\mathbf{c}_i$
$\widehat{d}_j^2(\mathbf{x})$
$\widehat{d}_i^2(\mathbf{x})$
$\widehat{\mathbb{1}}_\Delta$
1
0
$\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2(\mathbf{x}) = 1$
$\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_j^2(\mathbf{x}) = 0$

Figure 3.4: Chart determination utilizes the composition of approximated distance function $\widehat{d}_i^2$ and the indicator function $\widehat{\mathbb{1}}_\Delta$.

To implement $\widehat{\mathbb{1}}_\Delta(a)$, we consider a basic step function $g = 2\text{ReLU}(x - 0.5(r^2 -$

$4B^2D\nu)) - 2\mathrm{ReLU}(x - r^2 + 4B^2D\nu)$. It is straightforward to check

$$
\begin{aligned}
g_k(a) &= \underbrace{g \circ \cdots \circ g}_{k}(a) \\
&= \begin{cases} 0 & a < (1 - 2^{-k})(r^2 - 4B^2D\nu) \\ 2^k(a - r^2 + 4B^2D\nu) + r^2 - 4B^2D\nu & a \in \left[(1 - \frac{1}{2^k})(r^2 - 4B^2D\nu), r^2 - 4B^2D\nu\right] \\ r^2 - 4B^2D\nu & a > r^2 - 4B^2D\nu \end{cases} .
\end{aligned}
$$

Let $\widehat{\mathbb{1}}_\Delta = 1 - \frac{1}{r^2 - 4B^2D\nu} g_k$. It suffices to choose k satisfying $(1 - \frac{1}{2^k})(r^2 - 4B^2D\nu) \geq r^2 - \Delta + 4B^2D\nu$, which yields $k = \left\lceil \log \frac{r^2}{\Delta} \right\rceil$. We use $\widehat{\mathbb{1}}_\Delta \circ \widehat{d_i^2}$ to approximate the indicator function on U_i:

- if $\mathbf{x} \notin U_i$, i.e., $d_i^2(\mathbf{x}) \geq r^2$, we have $\widehat{\mathbb{1}}_\Delta \circ \widehat{d_i^2}(\mathbf{x}) = 0$;
- if $\mathbf{x} \in U_i$ and $d_i^2(\mathbf{x}) \leq r^2 - \Delta$, we have $\widehat{\mathbb{1}}_\Delta \circ \widehat{d_i^2}(\mathbf{x}) = 1$.

We remark that although the approximate indicator function $\widehat{\mathbb{1}}_\Delta$ is a piecewise linear function with two breakpoints, we implement it using a deep neural network to control the range of weight parameters in the network. Otherwise, the parameter upper bound can be as large as $1/\Delta$ due to the steep slope in $\widehat{\mathbb{1}}_\Delta$.

Step 4. Taylor approximation. In each chart (U_i, ϕ_i), we locally approximate f using Taylor polynomials of order n as shown in Figure 3.5. Specifically, we decompose f as

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

where ρ_i is an element in a C^∞ partition of unity on $\mathcal{M}$ which is supported inside U_i. The existence of such a partition of unity is guaranteed by Proposition 2.1. Since $\mathcal{M}$ is a compact smooth manifold and ρ_i is C^∞, f_i preserves the regularity (smoothness) of f such that $f_i \in \mathcal{H}^\alpha(\mathcal{M})$ for $i = 1, \ldots, C_\mathcal{M}$.

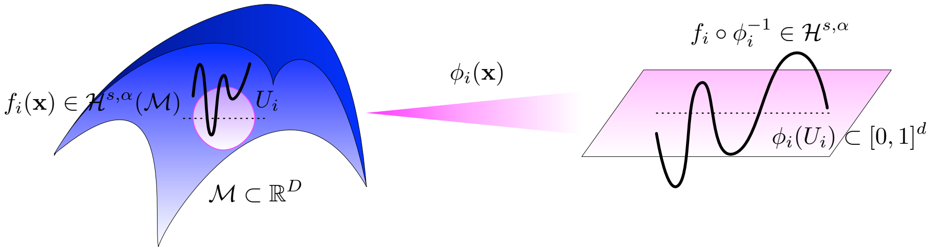

$f_i(\mathbf{x}) \in \mathcal{H}^{s,\alpha}(\mathcal{M})$
U_i
$\mathcal{M} \subset \mathbb{R}^D$
$\phi_i(\mathbf{x})$
$f_i \circ \phi_i^{-1} \in \mathcal{H}^{s,\alpha}$
$\phi_i(U_i) \subset [0,1]^d$

Figure 3.5: Locally approximate f in each chart (U_i, ϕ_i) using Taylor polynomials.

Lemma 3.3. *Suppose Assumption 3.3 holds. For $i = 1, \ldots, C_{\mathcal{M}}$, the function f_i is Hölder continuous on $\mathcal{M}$, in the sense that there exists a Hölder coefficient L_i depending on d, the upper bounds of derivatives of the partition of unity ρ_i and coordinate system ϕ_i, up to order s, such that for any $|\mathbf{s}| = \lfloor \alpha \rfloor$, we have*

$$\left| D^{\mathbf{s}}(f_i \circ \phi_i^{-1})\big|_{\phi_i(\mathbf{x}_1)} - D^{\mathbf{s}}(f_i \circ \phi_i^{-1})\big|_{\phi_i(\mathbf{x}_2)} \right| \leq L_i \left\| \phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2) \right\|_2^{\alpha - \lfloor \alpha \rfloor}, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in U_i.$$

Proof Sketch. We provide a sketch here. More details are deferred to Appendix A.2.2. Without loss of generality, suppose Assumption 3.3 holds with the atlas chosen in **Step 1**. Denote $g_1 = f \circ \phi_i^{-1}$ and $g_2 = \rho_i \circ \phi_i^{-1}$. By the Leibniz rule, we have

$$D^{\mathbf{s}}(f_i \circ \phi_i^{-1}) = D^{\mathbf{s}}(g_1 \times g_2) = \sum_{|\mathbf{p}|+|\mathbf{q}|=\lfloor \alpha \rfloor} \binom{\lfloor \alpha \rfloor}{|\mathbf{p}|} D^{\mathbf{p}} g_1 D^{\mathbf{q}} g_2.$$

Consider each term in the sum: for any $\mathbf{x}_1, \mathbf{x}_2 \in U_i$,

$$\begin{aligned}
&\left| D^{\mathbf{p}} g_1 D^{\mathbf{q}} g_2 |_{\phi_i(\mathbf{x}_1)} - D^{\mathbf{p}} g_1 D^{\mathbf{q}} g_2 |_{\phi_i(\mathbf{x}_2)} \right| \\
\leq & \left| D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_1)) \right| \left| D^{\mathbf{q}} g_2 |_{\phi_i(\mathbf{x}_1)} - D^{\mathbf{q}} g_2 |_{\phi_i(\mathbf{x}_2)} \right| + \left| D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_2)) \right| \left| D^{\mathbf{p}} g_1 |_{\phi_i(\mathbf{x}_1)} - D^{\mathbf{p}} g_1 |_{\phi_i(\mathbf{x}_2)} \right| \\
\leq & \lambda_i \theta_{i,\alpha} \left\| \phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2) \right\|_2^{\alpha - \lfloor \alpha \rfloor} + \mu_i \beta_{i,\alpha} \left\| \phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2) \right\|_2^{\alpha - \lfloor \alpha \rfloor}.
\end{aligned}$$

Here λ_i and μ_i are uniform upper bounds on the derivatives of g_1 and g_2 with order up to s, respectively. The quantities $\theta_{i,\alpha}$ and $\beta_{i,\alpha}$ in the last inequality above is chosen as follows:

by the mean value theorem, we have

$$\begin{aligned}\left|D^{\mathbf{q}}g_2|_{\phi_i(\mathbf{x}_1)} - D^{\mathbf{q}}g_2|_{\phi_i(\mathbf{x}_2)}\right| &\leq \sqrt{d}\mu_i \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2 \\ &= \sqrt{d}\mu_i \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2^{1-\alpha+\lfloor\alpha\rfloor} \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2^{\alpha-\lfloor\alpha\rfloor} \\ &\leq \sqrt{d}\mu_i(2r)^{1-\alpha+\lfloor\alpha\rfloor} \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2^{\alpha-\lfloor\alpha\rfloor},\end{aligned}$$

where the last inequality is due to the fact that $\left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2 \leq b_i \left\|V_i\right\| \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|_2 \leq 2r$. Then we set $\theta_{i,\alpha} = \sqrt{d}\mu_i(2r)^{1-\alpha+\lfloor\alpha\rfloor}$ and by a similar argument, we set $\beta_{i,\alpha} = \sqrt{d}\lambda_i(2r)^{1-\alpha+\lfloor\alpha\rfloor}$. We complete the proof by taking $L_i = 2^{\lfloor\alpha\rfloor+1}\sqrt{d}\lambda_i\mu_i(2r)^{1-\alpha+\lfloor\alpha\rfloor}$. □

Lemma 3.3 is crucial for the error estimation in the local approximation of $f_i \circ \phi_i^{-1}$ by Taylor polynomials. This error estimate is given in the following theorem, where some of the proof techniques are from Theorem 1 in [40].

Theorem 3.3. *Let $f_i = f\rho_i$ as in **Step 4**. For any $\delta \in (0, 1)$, there exists a ReLU network structure that, if the weight parameters are properly chosen, the network yields an approximation of $f_i \circ \phi_i^{-1}$ uniformly with an L_∞ error δ. Such a network has*

1. *no more than $c_1\left(\log\frac{1}{\delta} + 1\right)$ layers, with width bounded by $c_2\delta^{-d/\alpha}$,*
2. *at most $c_3\delta^{-\frac{d}{\alpha}}\left(\log\frac{1}{\delta} + 1\right)$ neurons and weight parameters, with the range of weight parameters bounded by $\kappa = c_4 \max\{1, \sqrt{d}\}$,*

where c_1, c_2, c_3 depend on α, d, τ, and the upper bound of derivatives of $f_i \circ \phi_i^{-1}$ up to order $\lfloor\alpha\rfloor$, and c_4 depends on the upper bound of the derivatives of ρ_i's up to order $\lfloor\alpha\rfloor$.

Proof Sketch. The detailed proof is provided in Appendix A.2.3. The proof consists of two steps:

1. Approximate $f_i \circ \phi_i^{-1}$ using a weighted sum of Taylor polynomials;
2. Implement the weighted sum of Taylor polynomials using ReLU networks.

Specifically, we set up a uniform grid and divide $[0,1]^d$ into small cubes, and then approximate $f_i \circ \phi_i^{-1}$ by its $\lfloor \alpha \rfloor$-th order Taylor polynomial in each cube. To implement such polynomials by ReLU networks, we recursively apply the multiplication $\widehat{\times}$ operator in Corollary 3.1, since these polynomials are sums of the products of different variables. □

Step 5. Estimating the total error. We have collected all the ingredients to implement the entire ReLU network to approximate f on $\mathcal{M}$. Recall that the network structure consists of 3 main sub-networks as demonstrated in Figure 3.1. Let $\widehat{\times}$ be an approximation to the multiplication operator in the pairing sub-network with error η. Accordingly, the function given by the whole network is

$$\widetilde{f} = \sum_{i=1}^{C_{\mathcal{M}}} \widehat{\times}(\widehat{f}_i, \widehat{\mathbb{1}}_\Delta \circ \widehat{d_i^2}) \quad \text{with } \widehat{f}_i = \widetilde{f}_i \circ \phi_i,$$

where $\widetilde{f}_i$ is the approximation of $f_i \circ \phi_i^{-1}$ using Taylor polynomials in Theorem 3.3. The total error can be decomposed into three components according to Lemma 3.4 below. We denote $\mathbb{1}(\mathbf{x} \in U_i)$ as the indicator function of U_i. Let the approximation errors of the multiplication operation $\widehat{\times}$ and the local Taylor polynomial in Theorem 3.3 be η and δ, respectively.

Lemma 3.4. *For any $i = 1, \ldots, C_{\mathcal{M}}$, we have $\|\widetilde{f} - f\|_\infty \leq \sum_{i=1}^{C_{\mathcal{M}}}(A_{i,1} + A_{i,2} + A_{i,3})$, where*

$$\begin{aligned}
A_{i,1} &= \left\| \widehat{\times}(\widehat{f}_i, \widehat{\mathbb{1}}_\Delta \circ \widehat{d_i^2}) - \widehat{f}_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d_i^2}) \right\|_\infty \leq \eta, \\
A_{i,2} &= \left\| \widehat{f}_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d_i^2}) - f_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d_i^2}) \right\|_\infty \leq \delta, \\
A_{i,3} &= \left\| f_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d_i^2}) - f_i \times \mathbb{1}(\mathbf{x} \in U_i) \right\|_\infty \leq \frac{c(\pi+1)}{r(1-r/\tau)}\Delta \quad \textit{for some constant } c.
\end{aligned}$$

Lemma 3.4 is proved in Appendix A.2.4. In order to achieve an ϵ total approximation error, i.e., $\|f - \widetilde{f}\|_\infty \leq \epsilon$, we need to control the errors in the three sub-networks. In other words, we need to decide ν for $\widehat{d_i^2}$, Δ for $\widehat{\mathbb{1}}_\Delta$, δ for $\widetilde{f}_i$, and η for $\widehat{\times}$. Note that $A_{i,1}$ is the error from the pairing sub-network, $A_{i,2}$ is the approximation error in the Taylor approximation sub-network, and $A_{i,3}$ is the error from the chart determination sub-network. The error bounds on $A_{i,1}, A_{i,2}$ are straightforward from the constructions of $\widehat{\times}$ and $\widehat{f}_i$. The estimate of $A_{i,3}$ involves some technical analysis since $\|\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2 - \mathbb{1}(\mathbf{x} \in U_i)\|_\infty = 1$. Note that we have

$$\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2(\mathbf{x}) - \mathbb{1}(\mathbf{x} \in U_i) = 0$$

whenever $\|\mathbf{x} - \mathbf{c}_i\|_2^2 < r^2 - \Delta$ or $\|\mathbf{x} - \mathbf{c}_i\|_2^2 > r^2$. Therefore, we only need to prove that $|f_i(\mathbf{x})|$ is sufficiently small in the shell region

$$\mathcal{K}_i = \{\mathbf{x} \in \mathcal{M} : r^2 - \Delta \leq \|\mathbf{x} - \mathbf{c}_i\|_2^2 \leq r^2\}.$$

We bound the maximum of f_i on $\mathcal{K}_i$ using a first-order Taylor expansion. Since f_i vanishes at the boundary of U_i due to the partition of unity ρ_i, we can show that $\sup_{\mathbf{x} \in \mathcal{K}_i} |f_i(\mathbf{x})|$ is proportional to the width Δ of $\mathcal{K}_i$. In particular, there exists a constant c depending on f_i's and ϕ_i's such that

$$\max_{\mathbf{x} \in \mathcal{K}_i} |f_i(\mathbf{x})| \leq \frac{c(\pi + 1)}{r(1 - r/\tau)}\Delta \quad \text{for any} \quad i = 1, \ldots, C_{\mathcal{M}}. \tag{3.2}$$

Then (Eq. 3.2) immediately implies the upper bound on $A_{i,3}$. The formal statement of (Eq. 3.2) and its proof are deferred to Lemma A.1 and Appendix A.2.5.

Given Lemma 3.4, we choose

$$\eta = \delta = \frac{\epsilon}{3C_{\mathcal{M}}} \quad \text{and} \quad \Delta = \frac{r(1 - r/\tau)\epsilon}{3c(\pi + 1)C_{\mathcal{M}}} \tag{3.3}$$

so that the approximation error is bounded by ϵ. Moreover, we choose

$$\nu = \frac{\Delta}{16B^2D} \tag{3.4}$$

to guarantee $\Delta > 8B^2 D\nu$ so that the definition of $\widehat{\mathbb{1}}_\Delta$ is valid.

Finally we quantify the size of the ReLU network. Recall that the chart determination sub-network has $c_1 \log \frac{1}{\nu}$ layers, the Taylor approximation sub-network has $c_2 \log \frac{1}{\delta}$ layers, and the pairing sub-network has $c_3 \log \frac{1}{\eta}$ layers. Here c_2 depends on d, α, f, and c_1, c_3 are absolute constants. Combining these with (Eq. 3.3) and (Eq. 3.4) yields the depth in Theorem 3.1. By a similar argument, we can obtain the number of neurons and weight parameters. A detailed analysis is given in Appendix A.2.6.

3.2 Efficient Approximation of ConvResNets

Feedforward neural networks have attracted many theoretical studies due to its simplicity. In practice, Convolutional Residual Networks (ConvResNets) give rise to the state-of-the-art performance in wide applications. In this section, we build upon the framework developed in Section 3.1 and study the universal approximation properties of ConvResNets.

We shift our focus to Besov functions as a generalization of Hölder functions.

Assumption 3.4. *Let $0 < p, q \leq \infty$, $d/p + 1 \leq \alpha < \infty$. Assume $f \in B^\alpha_{p,q}(\mathcal{M})$ and $\|f\|_{B^\alpha_{p,q}(\mathcal{M})} \leq c_0$ from a constant $c_0 > 0$. Additionally, we assume $\|f\|_{L^\infty} \leq C$ for a constant $C > 0$.*

Our universal approximation guarantee of ConvResNets for Besov functions on $\mathcal{M}$ is summarized in the following (Proof can be found in [49]). We recall the ConvResNet architecture $\mathrm{CRN}(\dots)$ from (Eq. 2.5).

Theorem 3.4. *Assume Assumption 3.1 and 3.2 hold. For any function f satisfying Assumption 3.4, any $\epsilon \in (0,1)$, and positive integer $K \in [2, D]$, there is a ConvResNet architecture $\mathrm{CRN}(M, L, J, K, \kappa_1, \kappa_2, \infty)$ such that if the weight parameters of this ConvResNet are properly chosen, the network yields a function $\widehat{f} \in \mathrm{CRN}(M, L, J, K, \kappa_1, \kappa_2, \infty)$ satisfying*

$$\|\widehat{f} - f\|_{L^\infty} \leq \epsilon. \tag{3.5}$$

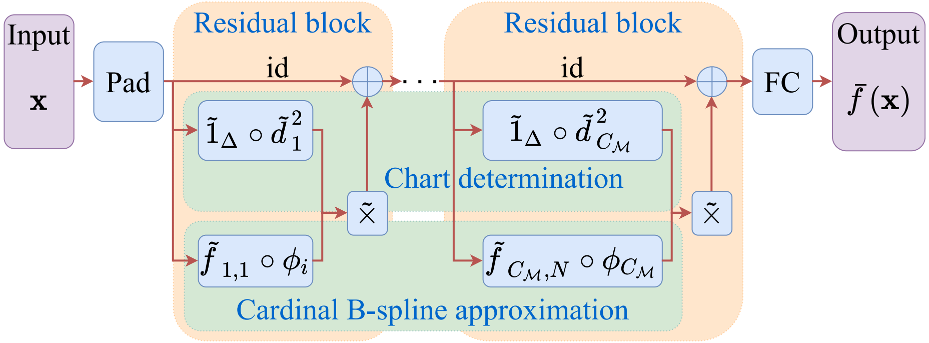

Input
$\mathbf{x}$
Pad
Residual block
id
$\tilde{\mathbb{1}}_\Delta \circ \tilde{d}_1^2$
$\tilde{f}_{1,1} \circ \phi_i$
$\tilde{\times}$
Chart determination
Cardinal B-spline approximation
Residual block
id
$\tilde{\mathbb{1}}_\Delta \circ \tilde{d}_{C_\mathcal{M}}^2$
$\tilde{f}_{C_\mathcal{M},N} \circ \phi_{C_\mathcal{M}}$
$\tilde{\times}$
FC
Output
$\bar{f}(\mathbf{x})$

Figure 3.6: The ConvResNet in Theorem 3.4 contains a padding layer, M residual blocks, and a fully connected (FC) layer.

Such a network architecture has

$$M = O\left(\epsilon^{-d/\alpha}\right), \; L = O(\log(1/\epsilon)), \; J = O(1), \; \kappa_1 = O(1), \; \log \kappa_2 = O(\log^2(1/\epsilon)).$$

The constant hidden in $O(\cdot)$ depend on d, $D \log D$, α, $\frac{2d}{\alpha p - d}$, p, q, c_0, τ and the surface area of $\mathcal{M}$. In particular, the constant depends on $D \log D$ linearly.

The architecture of the ConvResNet in Theorem 3.4 is illustrated in Figure 3.6. It has the following properties:

- The network has a fixed filter size and a fixed number of channels.
- There is no cardinality constraint (number of nonzero weight parameters).
- The network size depends on the intrinsic dimension d, and only weakly depends on D.

Theorem 3.4 can be compared with [50] on the approximation theory for Besov functions in $\mathbb{R}^D$ by FNNs as follows: (1) To universally approximate Besov functions in $\mathbb{R}^D$ with ϵ error, the FNN constructed in [50] requires $O\left(\log(1/\epsilon)\right)$ depth, $O\left(\epsilon^{-D/\alpha}\right)$ width and $O\left(\epsilon^{-D/\alpha}\log(1/\epsilon)\right)$ nonzero parameters. By exploiting the manifold model, our network size depends on the intrinsic dimension d and weakly depends on D. (2) The ConvResNet in Theorem 3.4 does not require any cardinality constraint, while such a constraint is needed in [50].

3.3 Approximation with Smoothness Constraints

3.3.1 Benefits of Overparameterized Neural Networks

Deep neural networks of enormous sizes have achieved remarkable success in various applications. Some well-known examples include ViT-Huge of 632 million parameters [51], BERT-Large of 336 million parameters [52], and the gigantic GPT-3 of 175 billion parameters [53]. In addition to outstanding testing accuracy, there has been evidence that large neural networks favor smoothness and yield good robustness [54, 55].

Among vast literature on explaining the success of neural networks, universal approximation theories analyze how well neural networks can represent complex data models (see literature in related work section). These works focus on approximating a target function in terms of its function value (i.e., in function L_∞ norm). However, other important properties, espcifically the smoothness of the neural networks, are less investigated. A few early results provide asymptotic results on two-layer networks with smooth activation for approximating both function value and derivatives [56, 57]. Recently, [58, 59] established nonasymptotic approximation theory of feedforward networks in terms of Sobolev norms.

In real-world applications, on the other hand, practitioners empirically demonstrated a close tie between the smoothness of a trained neural network to its adversarial robustness [60, 61, 62, 63]. The intuition behind is relatively clear. Consider, for instance, adding some adversarial perturbation to an input. A network of small (local) Lipschitz constant produces less deviation to the original output, and therefore, is often resilient to adversarial attackes. On the contrary, a network that is vulnerable to adversarial attacks usually has a large Lipschitz constant. Over the years, many computational methods are proposed and extensively tested in experiments for promoting network smoothness [64, 54, 63, 65]. Apart from these explicit training methodologies, the size of a network is also recognized as a critical factor to its generalization and robustness [66, 54, 67]. Yet, theoretical understanding is largely missing.

In this section, we investigate universal approximation ability of neural networks with smoothness guarantees. We consider ConvResNet with ReLU activation as an example. We measure the approximation error of ConvResNet in terms of not only the function value, but also higher order smoothness. Specifically, suppose given a target function f belonging to a Sobolev space in a D-dimensional hypercube. We provide an approximation error estimate in terms of Sobolev norm as a function of the size of ConvResNet. We also extend our theory to functions supported on a d-dimensional Riemannian manifold ($d \ll D$). All of the proofs can be found in [68].

3.3.2 $W^{s,p}$-approximation in Euclidean Space

Consider a Sobolev function class defined on a unit hypercube $(0,1)^D$. We aim to use convolutional residual networks for approximating functions in the target class in terms of the $W^{s,p}$ norm. Here p is a positive integer and s can vary in $[0,1]$; in particular, $s = 0$ corresponds to function value approximation, and $s = 1$ resembles the result in previous sections. We formally define our target function class as a Sobolev norm ball.

Assumption 3.5. *Let $\alpha \geq 2, 1 \leq p \leq +\infty$ be integers. Assume the target function f satisfies*

$$f \in W^{\alpha,p}\left((0,1)^D\right) \quad \textit{and} \quad \|f\|_{W^{\alpha,p}((0,1)^D)} \leq 1.$$

We set the norm ball of radius 1 for the sake of simplicity, while the results in the sequel hold for any constant radius. We also let $\alpha \geq 2$ for techincal convenience. In the following theorem, we show that ConvResNets can approximate any functions in a Sobolev norm ball in terms of $W^{s,p}$ norm ($s \leq 1$). The approximation error is obtained as a function of the network configuration.

Theorem 3.5. *For any positive integers $K \in [2, D]$, $\widetilde{M}$, and $\widetilde{J} > 0$, we choose*

$$L = O(\log(\widetilde{M}\widetilde{J})),\ J = O(\widetilde{J}),\ \kappa_1 = O((\widetilde{M}\widetilde{J})^{1/D}),\ \kappa_2 = O((\widetilde{M}\widetilde{J})^{1/D}),\ M = O(\widetilde{M}).$$

Then given $s \in [0, 1]$, the ConvResNet architecture $\mathrm{CRN}(M, L, J, K, \kappa_1, \kappa_2)$ can approximate any function f satisfying Assumption 3.5, i.e., there exists $\widehat{f} \in \mathrm{CRN}(M, L, J, K, \kappa_1, \kappa_2, \infty)$ with

$$\|\widehat{f} - f\|_{W^{s,p}((0,1)^D)} \leq C_1(\widetilde{M}\widetilde{J})^{-\frac{\alpha - s}{D}}$$

for some constant C_1 depending on D, α, p.

Theorem 3.5 says that the approximation power of ConvResNet amplifies as its width and depth increase. To better interpret the result, we choose $s = 1$ and $p = \infty$, which corresponds to simultaneously approximating function value and first-order derivatives.

Corollary 3.2. *In the setup of Theorem 3.5, taking $s = 1$ and $p = \infty$, the ConvResNet architecture $\mathrm{CRN}(M, L, J, K, \kappa_1, \kappa_2, \infty)$ can approximate any f satisfying Assumption 3.5 up to first-order, i.e., there exists $\widehat{f} \in \mathrm{CRN}(M, L, J, K, \kappa_1, \kappa_2, \infty)$ with*

$$\left\|\widehat{f} - f\right\|_\infty \leq C_2(\widetilde{M}\widetilde{J})^{-\frac{\alpha-1}{D}} \quad \textit{and} \quad \sup_i \left\|\frac{\partial \widehat{f}}{\partial x_i} - \frac{\partial f}{\partial x_i}\right\|_\infty \leq C_2(\widetilde{M}\widetilde{J})^{-\frac{\alpha-1}{D}},$$

where the constant C_2 depends on D and α. In particular, we have Lipschitz continuity bound

$$\left\|\widehat{f}\right\|_{\mathrm{Lip}} \leq 1 + C_2\sqrt{D}(\widetilde{M}\widetilde{J})^{-\frac{\alpha-1}{D}}.$$

Theorem 3.5 and Corollary 3.2 have rich implications.

Large network for smooth approximation. Taking $s = 0$ in Theorem 3.5 recovers function approximation in terms of L^∞ norm. The corresponding approximation error scales as

$O((\widetilde{M}\widetilde{J})^{-\frac{\alpha}{D}})$. A quick comparison to Corollary 3.2 indicates that in order to additionally capture the first-order information of a target function, large network is needed to achieve the same function value error bound.

Arbitrary width and depth. [58, 59] provide approximation guarantees of feedforward networks in terms of $W^{s,p}$ norm. Despite different network architectures, we remark that our theory covers general networks with arbitrary width and depth. More specifically, for a given approximation error ϵ, [58] set the network depth and width as $O(\log 1/\epsilon)$ and $O(\epsilon^{-D/(\alpha-s)})$, respectively. Yet in our result, we only need to ensure $\widetilde{M}\widetilde{J} = O(\epsilon^{-D/(\alpha-s)})$, which does not require any scaling relation between $\widetilde{M}$ and $\widetilde{J}$.

3.3.3 $W^{s,p}$-approximation on Manifold

Theorem 3.5 indicates a curse of data dimensionality: When data dimension D is large, such as image data, Theorem 3.5 converges extremely slowly and becomes less attractive. Motivated by applications, we model data as a low-dimensional Riemannian manifold $\mathcal{M}$. We will show that ConvResNet is still adaptable to manifold structures, even we impose smoothness constraints on approximation. Analogous to the Euclidean case, we consider a Sobolev norm ball on a manifold.

Assumption 3.6. *Let $\alpha \geq 2$ be an integer. Assume the target function f satisfies*

$$f \in W^{\alpha,\infty}(\mathcal{M}) \quad \textit{and} \quad \|f\|_{W^{\alpha,\infty}(\mathcal{M})} \leq 1.$$

We now present a counterpart of Theorem 3.5, showing an efficient approximation of functions in a Sobolev norm ball on $\mathcal{M}$.

Theorem 3.6. *For any positive integers $K \in [2, D]$, $\widetilde{M}$, and $\widetilde{J} > 0$, we choose*

$$L = O(\log(\widetilde{M}\widetilde{J})) + D,\ J = O(D\widetilde{J}),\ \kappa_1 = O((\widetilde{M}\widetilde{J})^{1/d}),$$
$$\kappa_2 = O((\widetilde{M}\widetilde{J})^{1/d}),\ M = O(\widetilde{M}).$$

Then given $k \in \{0, 1\}$, the ConvResNet architecture $\mathrm{CRN}(M, L, J, K, \kappa_1, \kappa_2, \infty)$ can approximate any function f satisfying Assumption 3.6, i.e., there exists $\widehat{f} \in \mathcal{C}(M, L, J, K, \kappa_1, \kappa_2, \infty)$ with

$$\|\widehat{f} - f\|_{W^{k,\infty}(\mathcal{M})} \leq C_3(\widetilde{M}\widetilde{J})^{-\frac{\alpha-k}{d}},$$

where constant C_3 depends on d, α, B, τ, and the surface area of $\mathcal{M}$.

As can be seen, the approximation error decays at a rate only depending on intrinsic data dimension d, which is a significant improvement over Theorem 3.5 given $d \ll D$. We also note that the size of ConvResNet has a weak dependence on D, yet it is inevitable due to the residual connection preserves input dimensionality.

3.4 Conclusion and Discussion

In this chapter, we develop efficient function approximation theories of feedforward neural networks and convolutional residual networks using the ReLU activation. We show that these network architectures enjoy fast rate in approximating Hölder, Sobolev, and Besov functions. We also prove wide and deep networks can not only approximate function value, but first-order derivatives. We discuss related topics and future directions.

ReLU activations We consider neural networks with ReLU activations for a practical concern — ReLU activations are widely used in deep networks. Moreover, ReLU networks are easier to train compared with sigmoid or hyperbolic tangent activations, which are known for their notorious vanishing gradient problem [69, 70].

Low-dimensional Manifolds The low dimensional manifold model plays a vital role to reduce the network size. As shown in Theorem 3.2, to approximate functions in $F^{n,D}$ with accuracy ϵ, the minimal number of weight parameters is $O(\epsilon^{-\frac{D}{n}})$. This lower bound is huge, and can not be improved without low dimensional structures of data.

Existence vs. Learnability and Generalization Our Theorem 3.1 shows the existence of a ReLU network structure that gives efficient approximations of functions on low dimensional manifolds, if the weight parameters are properly chosen. In practice, it is observed that larger neural networks are easier to train and yield better generalization performances [71, 72, 73]. This is referred to as overparameterization. Establishing the connection between learnability and generalization is an important future direction.

Convolutional Filters Convolutional neural networks (CNNs, [1]) are widely used in computer vision, language modeling, etc. Empirical results reveal that different convolutional filters can capture various patterns in images, e.g., edge detection filters. An interesting question is whether convolutional filters serve as charts in our framework.

Equivalent Networks The ReLU network identified in Theorem 3.1 and ConvResNet in Theorem 3.4 are capable of approximating same functions. Several other network structures can also yield the same function. It is interesting to investigate whether these network structures also possess the universal approximation property and whether different architectures exhibit advantages in different scenarios.

CHAPTER 4
NONPARAMETRIC REGRESSION/CLASSIFICATION USING NEURAL NETWORKS

In approximation theories, we constructively show the existence of a network for approximating target functions supported on manifolds. Such results provide valuable guidelines for choosing proper network architectures depending on problem regularity. A followup question naturally arises as “Can we establish sample complexity bounds of these networks in various learning problems?”. In Chapter 4, Chapter 5, and Chapter 6, we devote to studying statistical applications using neural networks. In this chapter, we consider nonparametric regression and classification problems [74, 13, 14] using neural networks in exploitation of low-dimensional geometric structures of data; the corresponding results are presented in Section 4.1 and Section 4.2, respectively. Section 4.3 concludes the chapter.

4.1 Nonparametric Regression

In nonparametric regression, the goal is to recover the regression function f_0 supported on a manifold $\mathcal{M}$ using samples $S_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $\mathbf{x} \in \mathcal{M}$ and $y \in \mathbb{R}$. The $\mathbf{x}_i$'s are i.i.d. sampled from a distribution $\mathcal{D}_x$ on $\mathcal{M}$, and the response y_i satisfies

$$y_i = f_0(\mathbf{x}_i) + \xi_i,$$

where ξ_i's are i.i.d. sub-Gaussian noise independent of $\mathbf{x}_i$'s.

We use multi-layer ReLU (Rectified Linear Unit) neural networks to recover f_0. We denote $\mathcal{F}$ as a class of neural networks with bounded weight parameters and bounded output:

$$\mathcal{F}(R, \kappa, L, p, K) = \mathrm{FNN}(R, \kappa, L, p, K) \quad \text{(defined in (Eq. 2.3))}. \tag{4.1}$$

To obtain an estimator $\widehat{f} \in \mathcal{F}(R, \kappa, L, p, K)$ of f_0, we minimize the empirical quadratic risk

$$\widehat{f}_n = \underset{f \in \mathcal{F}(R,\kappa,L,p,K)}{\operatorname{argmin}} \widehat{\mathcal{R}}_n(f) = \underset{f \in \mathcal{F}(R,\kappa,L,p,K)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(f(\mathbf{x}_i) - y_i\right)^2. \tag{4.2}$$

The subscript n emphasizes that the estimator is obtained using n pairs of samples.

Related work Nonparametric regression has been widely studied in statistics. A variety of methods has been proposed to estimate the regression function, including kernel methods, wavelets, splines, and local polynomials [75, 76, 77, 14, 13]. Nonetheless, there is limited study on regression using deep ReLU networks until recently. The earliest works focused on neural networks with a single hidden layer and smooth activations (e.g., sigmoidal and sinusoidal functions, [78, 79]). Later results achieved the minimax lower bound for the mean squared error in the order of $O(n^{-\frac{2s}{2s+D}})$ up to a logarithmic factor for C^s functions in $\mathbb{R}^D$ [80, 81, 82, 83]. Theories for deep ReLU networks were developed in [84], where the estimate matches the minimax lower bound up to a logarithmic factor for Hölder functions. Extensions to more general function spaces, such as Besov spaces, can be found in [50] and results for classification problems can be found in [85, 86].

The rate of convergence in the results above cannot fully explain the success of deep learning due to the curse of the data dimension with a large D. Fortunately, many real-world data sets exhibit low-dimensional geometric structures. It has been demonstrated that, some classical methods are adaptive to the low-dimensional structures of data sets, and perform as well as if the low-dimensional structures were known. Results in this direction include local linear regression [87, 88], multiscale polynomial regression [89], k-nearest neighbor

[90], kernel regression [91], and Bayesian Gaussian process regression [92], where optimal rates depending on the intrinsic dimension were proved for functions having the second order of continuity [87], globally Lipschitz functions [90], and Hölder functions with Hölder index no more than 1 [91].

Recently, several independent works [93, 94, 95] justified the adaptability of deep neural networks to the low-dimensional data structures. [93] considered function approximation and regression of Hölder functions on a low-dimensional manifold, which is similar to the setup in this paper. The proofs in [93] and this paper both utilize a collection of charts to map each point on $\mathcal{M}$ into a local coordinate in $\mathbb{R}^d$, and then approximate functions in $\mathbb{R}^d$. There are two differences in the detailed proof: (1) In exploitation of a positive reach property of $\mathcal{M}$, we construct local coordinates on the manifold given by orthogonal projections onto the tangent spaces, while [93] assumed the existence of smooth local coordinates; (2) A main novelty of our work is to explicitly construct a chart determination sub-network which assigns each data point to its proper chart. In [93], the chart determination is realized by the partition of unity. In order to approximate functions in $\mathcal{H}^\alpha(\mathcal{M})$, [93] required a uniform upper bound on the derivatives of each coordinate map and each function in the partition of unity, up to order $\alpha D/d$. Our proof does not rely on such regularity conditions depending on the ambient dimension D. To describe the intrinsic dimensionality of data, [94] applied the notion of Minkowski dimension, which can be defined for a broader class of sets without smoothness restrictions. The intrinsic dimension of manifolds and the Minkowski dimension are different notions for low-dimensional sets, and one does not naturally imply the other. [93] and [94] established a $O(n^{-\frac{2\alpha}{2\alpha+d}})$ convergence rate of the mean squared error for learning functions in $\mathcal{H}^\alpha(\mathcal{M})$, where d is the manifold dimension in [93] and Minkowski dimension in [94], respectively. Recently [95] studied the approximation and regression error of ReLU neural networks for a class of functions in the form of $f(\mathbf{x}) = g(\pi_{\mathcal{M}}(\mathbf{x}))$, where $\mathbf{x}$ is near the low-dimensional manifold $\mathcal{M}$, $\pi_{\mathcal{M}}$ is a projection onto $\mathcal{M}$, and g is a Hölder function on $\mathcal{M}$.

4.1.1 Statistical Estimation Guarantee

We characterize the convergence rate for the estimation of f_0.

Theorem 4.1. *Suppose Assumption 3.1 and 3.2 hold. Let $\widehat{f}_n$ be the minimizer of empirical risk* (Eq. 4.2) *with the network class $\mathcal{F}(R, \kappa, L, p, K)$ properly designed such that*

$$L = \widetilde{O}\left(\frac{\alpha}{2\alpha+d}\log n\right), \quad p = \widetilde{O}\left(n^{\frac{d}{2\alpha+d}}\right), \quad K = \widetilde{O}\left(\frac{\alpha}{2\alpha+d} n^{\frac{d}{2\alpha+d}}\log n\right),$$
$$R = \|f_0\|_{\infty}, \quad \textit{and} \quad \kappa = O(\max\{1, B, \sqrt{d}, \tau^2\}).$$

Then we have

$$\mathbb{E}\left[\int_{\mathcal{M}} \left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 d\mathcal{D}_x(\mathbf{x})\right] \leq c(R^2+\sigma^2)\left(n^{-\frac{2\alpha}{2\alpha+d}} + \frac{D}{n}\right)\log^3 n,$$

where the expectation is taken over the training samples S_n, and c is a constant depending on $\log D$*, d, α, τ, B, the surface area of $\mathcal{M}$, and the upper bounds of derivatives of the coordinate systems ϕ_i's and partition of unity ρ_i's, up to order $\lfloor \alpha \rfloor$.*

Our theory implies that, in order to estimate an α-Hölder function up to an ϵ-error, the sample complexity is $n \gtrsim \epsilon^{-\frac{2\alpha+d}{\alpha}}$ up to a log factor. This sample complexity depends on the intrinsic dimension d, and thus largely improves on existing theories of nonparametric regression using neural networks, where the sample complexity scales as $\widetilde{O}(\epsilon^{-\frac{2\alpha+D}{\alpha}})$ [80, 81, 82, 83, 84]. Our result partially explains the success of deep ReLU neural networks in tackling high-dimensional data with low-dimensional geometric structures.

Theorem 4.1 is established by a bias-variance trade-off. We decompose the mean squared error to a squared bias term and a variance term. The bias is quantified by Theorem 3.1, and the variance term is proportional to the network size. A detailed proof of Theorem 4.1 is provided in Subsection 4.1.2. Here are some remarks:

1. The network class in Theorem 4.1 is sparsely connected, i.e. $K = O(Lp)$, while

densely connected networks satisfy $K = O(Lp^2)$.

2. The network class $\mathcal{F}(R, \kappa, L, p, K)$ has outputs uniformly bounded by R. Such a requirement can be achieved by appending an additional clipping layer to the end of the network structure, i.e.,

$$g(a) = \max\{-R, \min\{a, R\}\} = \text{ReLU}(a - R) - \text{ReLU}(a + R) - R.$$

3. Each weight parameter in our network class is bounded by a constant κ only depending on the curvature τ, the range B of the manifold $\mathcal{M}$, and the manifold dimension d. Such a boundedness condition is crucial to our theory and can be computationally realized by normalization after each step of the stochastic gradient descent.

4.1.2 Proof – Bias-variance Tradeoff

To prove Theorem 4.1, we decompose the mean squared error of the estimator $\widehat{f}_n$ into a squared bias term and a variance term. We bound the bias and variance separately, where the bias is tackled using the approximation theory (Theorem 3.1), and the variance is bounded using the metric entropy arguments [96, 13]. We begin with an oracle-type decomposition of the L^2 risk, in which we introduce the empirical L^2 risk as the intermediate term:

$$\begin{aligned}
&\mathbb{E}\left[\int_{\mathcal{M}} \left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 d\mathcal{D}_x(\mathbf{x})\right] \\
&= \underbrace{2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2\right]}_{T_1} \\
&\quad + \underbrace{\mathbb{E}\left[\int_{\mathcal{M}} \left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 d\mathcal{D}_x(\mathbf{x})\right] - 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2\right]}_{T_2},
\end{aligned}$$

where T_1 reflects the squared bias of using neural networks for estimating f_0 and T_2 is the variance term.

Bias Characterization – Bounding T_1

Since T_1 is the empirical L_2 risk of $\widehat{f}_n$ evaluated on the samples S_n, we relate T_1 to the empirical risk (Eq. 4.2) by rewriting $f_0(\mathbf{x}_i) = y_i - \xi_i$. Substituting into T_1, we derive the following decomposition,

$$\begin{aligned}
T_1 &= 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i) - y_i + \xi_i)^2\right] \\
&\overset{(i)}{=} 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left[(\widehat{f}_n(\mathbf{x}_i) - y_i)^2 + 2\xi_i\widehat{f}_n(\mathbf{x}_i) - \xi_i^2\right]\right] \\
&= 2\mathbb{E}\left[\inf_{f\in\mathcal{F}(R,\kappa,L,p,K)}\frac{1}{n}\sum_{i=1}^{n}\left[(f(\mathbf{x}_i) - y_i)^2 - \xi_i^2 + 2\xi_i\widehat{f}_n(\mathbf{x}_i)\right]\right] \\
&\overset{(ii)}{\leq} \underbrace{2\inf_{f\in\mathcal{F}(R,\kappa,L,p,K)}\int_{\mathcal{M}}(f(\mathbf{x}) - f_0(\mathbf{x}))^2 d\mathcal{D}_x(\mathbf{x})}_{(A)} + \underbrace{4\,\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}_n(\mathbf{x}_i)\right]}_{(B)}.
\end{aligned} \qquad (4.3)$$

Equality (i) is obtained by expanding the square, where the cross term $\mathbb{E}[\xi_i y_i] = \mathbb{E}[\xi_i(f_0(\mathbf{x}_i) + \xi_i)] = \mathbb{E}[\xi_i^2]$ due to the independence between $\mathbf{x}_i$ and ξ_i. Inequality (ii) invokes the Jensen's inequalty to pass the expectation. To obtain term (A), we expand $(f(\mathbf{x}_i) - y_i)^2 = (f(\mathbf{x}_i) - f_0(\mathbf{x}_i) - \xi_i)^2$, and observe the cancellation of $-\xi_i^2$. Note that term (A) is the squared approximation error of neural networks, and we will tackle it later using Theorem 3.1. We bound term (B) by quantifying the complexity of the network class $\mathcal{F}(R, \kappa, L, p, K)$. A precise upper bound of T_1 is given in the following lemma, whose proof follows a similar argument in [84, Lemma 4].

Lemma 4.1. *Fix the neural network class $\mathcal{F}(R, \kappa, L, p, K)$. For any constant $\delta \in (0, 2R)$,* *we have*

$$T_1 \leq 4 \inf_{f \in \mathcal{F}(R,\kappa,L,p,K)} \int_{\mathcal{M}} (f(\mathbf{x}) - f_0(\mathbf{x}))^2 d\mathcal{D}_x(\mathbf{x})$$
$$+ 48\sigma^2 \frac{\log \mathcal{N}(\delta, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty) + 2}{n}$$
$$+ (8\sqrt{6}\sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty) + 2}{n}} + 8)\sigma\delta,$$

where $\mathcal{N}(\delta, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty)$ denotes the δ-covering number of $\mathcal{F}(R,\kappa,L,p,K)$ with respect to the ℓ_∞ norm, i.e., there exists a discretization of $\mathcal{F}(R,\kappa,L,p,K)$ into $\mathcal{N}(\delta, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty)$ distinct elements, such that for any $f \in \mathcal{F}$, there is $\bar{f}$ in the discretization satisfying $\|\bar{f} - f\|_\infty \leq \epsilon$.

Proof Sketch. Given the derivation in (Eq. 4.3), we need to bound term (B). We discretize the neural network class $\mathcal{F}(R,\kappa,L,p,K)$ as $\{f_i^*\}_{i=1}^{\mathcal{N}(\delta,\mathcal{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)}$. By the definition of covering, there exists f^* such that $\|\widehat{f}_n - f^*\|_\infty \leq \delta$. Denoting $\|f - f_0\|_n = \frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2$, we cast (B) into

$$\begin{aligned}
(B) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \xi_i(\widehat{f}_n(\mathbf{x}_i) - f^*(\mathbf{x}_i) + f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))\right] \\
&\overset{(i)}{\leq} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))\right] + \delta\sigma \\
&= \mathbb{E}\left[\frac{\|f^* - f_0\|_n}{\sqrt{n}} \frac{\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\,\|f^* - f_0\|_n}\right] + \delta\sigma \\
&\overset{(ii)}{\leq} \sqrt{2}\mathbb{E}\left[\frac{\|\widehat{f}_n - f_0\|_n + \delta}{\sqrt{n}}\left|\frac{\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\,\|f^* - f_0\|_n}\right|\right] + \delta\sigma,
\end{aligned}$$

where (i) follows from Hölder's inequality and (ii) is obtained by some algebraic manipulation. To break the dependence between f^* and the samples, we replace f^* by any f_j^* in the δ-covering and observe that $\left|\frac{\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f^* - f_0\|_n}\right| \leq \max_j \left|\frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^* - f_0\|_n}\right|$. Applying the Cauchy-Schwarz inequality, we can show

$$(B) \leq \sqrt{2} \left(\sqrt{\frac{1}{n} \mathbb{E} \left[\|\widehat{f}_n - f_0\|_n^2 \right]} + \frac{\delta}{\sqrt{n}} \right) \sqrt{\mathbb{E} \left[\max_j \ z_j^2 \right]} + \delta\sigma,$$

where $z_j = \left| \frac{\sum_{i=1}^n \xi_i (f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f^* - f_0\|_n} \right|$. Given $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we note that z_j is a sub-Gaussian random variable with parameter σ (i.e., its variance is bounded by σ^2). It is well established in the existing literature on empirical processes [96] that the maximum of a collection of squared sub-Gaussian random variables satisfies

$$\mathbb{E} \left[\max_j \ z_j^2 \mid \mathbf{x}_1, \ldots, \mathbf{x}_n \right] \leq 3\sigma^2 \log \mathcal{N}(\delta, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty) + 6\sigma^2.$$

Substituting the above inequality into (B) and combining (A) and (B), we have

$$\begin{aligned} T_1 &= 2\mathbb{E} \left[\|\widehat{f}_n - f_0\|_n^2 \right] \\ &\leq 2 \inf_{f \in \mathcal{F}(R,\kappa,L,p,K)} \mathbb{E} \left[(f(\mathbf{x}) - f_0(\mathbf{x}))^2 \right] + 4\delta\sigma \\ &\quad + 4\sqrt{6}\sigma \left(\sqrt{\mathbb{E} \left[\|\widehat{f}_n - f_0\|_n^2 \right]} + \delta \right) \sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty) + 2}{n}}. \end{aligned}$$

Some manipulation gives rise to the desired result

$$\begin{aligned} T_1 \leq 4 &\inf_{f \in \mathcal{F}(R,\kappa,L,p,K)} \int_{\mathcal{M}} (f(\mathbf{x}) - f_0(\mathbf{x}))^2 d\mathcal{D}_x(\mathbf{x}) \\ &+ 48\sigma^2 \frac{\log \mathcal{N}(\delta, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty) + 2}{n} \\ &+ (8\sqrt{6}\sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty) + 2}{n}} + 8)\sigma\delta. \end{aligned}$$

See proof details in Appendix B.1.1. □

Variance Characterization – Bounding T_2

We observe that T_2 is the difference between the population L_2 risk of $\widehat{f}_n$ and its empirical counterpart. However, bounding such a difference is distinct from conventional concentration results due to the scaling factor 2 before the empirical risk. In particular, we split the empirical risk evenly into two parts, and bound one part using its higher-order moment (fourth moment). Using Bernstein-type inequality allows us to establish a $1/n$ convergence rate of T_2; the corresponding upper bound is presented in the following lemma.

Lemma 4.2. *For any constant $\delta \in (0, 2R)$, T_2 satisfies*

$$T_2 \leq \frac{104R^2}{3n} \log \mathcal{N}(\delta/4R, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty) + \left(4 + \frac{1}{2R}\right)\delta.$$

Proof Sketch. The detailed proof is deferred to Appendix B.1.2. For notational simplicity, we denote $\widehat{g}(\mathbf{x}) = (\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x}))^2$ and $\|\widehat{g}\|_\infty \leq 4R^2$. Applying the inequality $\int_{\mathcal{M}} \widehat{g}^2 d\mathcal{D}_x(\mathbf{x}) \leq 4R^2 \int_{\mathcal{M}} \widehat{g} d\mathcal{D}_x(\mathbf{x})$ [78], we rewrite T_2 as

$$\begin{aligned} T_2 &= \mathbb{E}\left[\int_{\mathcal{M}} \widehat{g}(\mathbf{x}) d\mathcal{D}_x(\mathbf{x}) - \frac{2}{n}\sum_{i=1}^n \widehat{g}(\mathbf{x}_i)\right] \\ &= 2\mathbb{E}\left[\int_{\mathcal{M}} \widehat{g}(\mathbf{x}) d\mathcal{D}_x(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n \widehat{g}(\mathbf{x}_i) - \frac{1}{2}\int_{\mathcal{M}} \widehat{g}(\mathbf{x}) d\mathcal{D}_x(\mathbf{x})\right] \\ &\leq 2\mathbb{E}\left[\int_{\mathcal{M}} \widehat{g}(\mathbf{x}) d\mathcal{D}_x(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n \widehat{g}(\mathbf{x}_i) - \frac{1}{8R^2}\int_{\mathcal{M}} \widehat{g}^2(\mathbf{x}) d\mathcal{D}_x(\mathbf{x})\right]. \end{aligned}$$

We now utilize ghost samples of $\mathbf{x}$ to bound T_2, which is a common technique in existing literature on nonparametric statistics [96, 13]. Specifically, let $\bar{\mathbf{x}}_i$'s be independent replications of $\mathbf{x}_i$'s. We bound T_2 as

$$\begin{aligned} T_2 &\leq 2\mathbb{E}\left[\sup_{g \in \mathcal{G}} \int_{\mathcal{M}} g(\mathbf{x}) d\mathcal{D}_x(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n g(\mathbf{x}_i) - \frac{1}{8R^2}\int_{\mathcal{M}} g^2(\mathbf{x}) d\mathcal{D}_x(\mathbf{x})\right] \\ &\leq 2\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n (g(\bar{\mathbf{x}}_i) - g(\mathbf{x}_i)) - \frac{1}{16R^2}\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}}}\left[g^2(\mathbf{x}) + g^2(\bar{\mathbf{x}})\right]\right], \end{aligned}$$

where $\mathcal{G} = \{g = (f - f_0)^2 \mid f \in \mathcal{F}(R, \kappa, L, p, K)\}$. We use the shorthand $\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}}}[\cdot]$ to denote the double integral $\int_{\mathcal{M}} \int_{\mathcal{M}} \cdot d\mathcal{D}_x(\mathbf{x}) d\mathcal{D}_x(\bar{\mathbf{x}})$ with respect to the joint distribution of $(\mathbf{x}, \bar{\mathbf{x}})$. The last inequality holds due to Jensen's inequality. Note here $g^2(\mathbf{x}) + g^2(\bar{\mathbf{x}})$ contributes as the variance term of $g(\bar{\mathbf{x}}_i) - g(\mathbf{x}_i)$, which yields a fast convergence of T_2 as n grows.

Similar to bounding T_1, we discretize the function space $\mathcal{G}$ using a δ-covering denoted by $\mathcal{G}^*$. This allows us to replace the supremum by the maximum over a finite set:

$$T_2 \leq 2\mathbb{E}_{\bar{\mathbf{x}},\mathbf{x}} \left[\sup_{g^* \in \mathcal{G}^*} \frac{1}{n} \sum_{i=1}^{n} (g^*(\bar{\mathbf{x}}_i) - g^*(\mathbf{x}_i)) - \frac{1}{16R^2} \mathbb{E}_{\mathbf{x},\bar{\mathbf{x}}} \left[(g^*)^2(\mathbf{x}) + (g^*)^2(\bar{\mathbf{x}}) \right] \right] \\ + \left(4 + \frac{1}{2R}\right) \delta.$$

We can bound the above maximum by the Bernstein's inequality, which yields

$$T_2 \leq \frac{104R^2}{3n} \log \mathcal{N}(\delta, \mathcal{G}, \|\cdot\|_\infty) + \left(4 + \frac{1}{2R}\right) \delta.$$

The last step is to relate the covering number of $\mathcal{G}$ to that of $\mathcal{F}(R, \kappa, L, p, K)$. Specifically, consider any $g_1, g_2 \in \mathcal{G}$ with $g_1 = (f_1 - f_0)^2$ and $g_2 = (f_2 - f_0)^2$, respectively. We can derive

$$\|g_1 - g_2\|_\infty = \sup_{\mathbf{x} \in \mathcal{M}} |f_1(\mathbf{x}) - f_2(\mathbf{x})| \, |f_1(\mathbf{x}) + f_2(\mathbf{x}) - 2f_0(\mathbf{x})| \leq 4R \|f_1 - f_2\|_\infty.$$

Therefore, the inequality $\mathcal{N}(\delta, \mathcal{G}, \|\cdot\|_\infty) \leq \mathcal{N}(\delta/4R, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty)$ holds, which implies

$$T_2 \leq \frac{104R^2}{3n} \log \mathcal{N}(\delta/4R, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty) + \left(4 + \frac{1}{2R}\right) \delta.$$

The proof is complete. □

Covering Number of Neural Networks

The upper bounds of T_1 and T_2 in Lemma 4.1 and 4.2 both depend on the covering number of the network class $\mathcal{F}(R, \kappa, L, p, K)$. In this section, we provide an upper bound on the covering number $\mathcal{N}(\delta, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty)$ for a given a resolution $\delta > 0$. Since each weight parameter in the network is bounded by a constant κ, we construct a covering by partitioning the range of each weight parameter into a uniform grid. By choosing a proper grid size, we show the following lemma.

Lemma 4.3. *Given $\delta > 0$, the covering number of neural network class $\mathcal{F}(R, \kappa, L, p, K)$ satisfies*

$$\mathcal{N}(\delta, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty) \leq \left(\frac{2L^2(pB+2)\kappa^L p^{L+1}}{\delta}\right)^K. \tag{4.4}$$

Proof Sketch. Consider $f, f' \in \mathcal{F}(R, \kappa, L, p, K)$ with each weight parameter differing at most h. By an induction on the number of layers in the network, we show that the ℓ_∞ norm of the difference $f - f'$ scales as

$$\|f - f'\|_\infty \leq hL(pB+2)(\kappa p)^{L-1}.$$

As a result, to achieve a δ-covering, it suffices to choose h such that $hL(pB+2)(\kappa p)^{L-1} = \delta$. Moreover, there are $\binom{Lp^2}{K} \leq (Lp^2)^K$ different choices of K non-zero entries out of Lp^2 weight parameters. Therefore, the covering number is bounded by

$$\mathcal{N}(\delta, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty) \leq \left(Lp^2\right)^K \left(\frac{2\kappa}{h}\right)^K \leq \left(\frac{2L^2(pB+2)\kappa^L p^{L+1}}{\delta}\right)^K.$$

The detailed proof is provided in Appendix B.1.3. □

Putting Together and Tradeoff

We are ready to finish the proof of Theorem 4.1. Combining the upper bounds of T_1 in Lemma 4.1 and T_2 in Lemma 4.2 together and substituting the covering number (Eq. 4.4), we obtain

$$\begin{aligned}\mathbb{E}\left[\int_{\mathcal{M}}\left(\widehat{f}_n(\mathbf{x})-f_0(\mathbf{x})\right)^2 d\mathcal{D}_x(\mathbf{x})\right] \leq\ & 4\inf_{f\in\mathcal{F}(R,\kappa,L,p,K)}\int_{\mathcal{M}}(f(\mathbf{x})-f_0(\mathbf{x}))^2 d\mathcal{D}_x(\mathbf{x})\\ &+48\sigma^2\frac{\log\mathcal{N}(\delta,\mathcal{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)+2}{n}\\ &+8\sqrt{6}\sqrt{\frac{\log\mathcal{N}(\delta,\mathcal{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)+2}{n}}\sigma\delta\\ &+\frac{104R^2}{3n}\log\mathcal{N}(\delta/4R,\mathcal{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)\\ &+\left(4+\frac{1}{2R}+8\sigma\right)\delta.\end{aligned}$$

It suffices to choose $\delta = 1/n$, which gives rise to

$$\begin{aligned}\mathbb{E}\left[\int_{\mathcal{M}}\left(\widehat{f}_n(\mathbf{x})-f_0(\mathbf{x})\right)^2 d\mathcal{D}_x(\mathbf{x})\right] \leq\ & 4\inf_{f\in\mathcal{F}(R,\kappa,L,p,K)}\int_{\mathcal{M}}(f(\mathbf{x})-f_0(\mathbf{x}))^2 d\mathcal{D}_x(\mathbf{x})\\ &+\widetilde{O}\left(\frac{R^2+\sigma^2}{n}KL\log(R\kappa Lpn)+\frac{\sigma^2}{n}\right), \qquad (4.5)\end{aligned}$$

where we also plug in the covering number upper bound in Lemma Eq. 4.4. We further set the approximation error as ϵ, i.e., $\inf_{f\in\mathcal{F}(R,\kappa,L,p,K)}\|f(\mathbf{x})-f_0(\mathbf{x})\|_\infty \leq \epsilon$. Theorem 3.1 suggests that we choose $L=\widetilde{O}(\log\frac{1}{\epsilon})$, $p=\widetilde{O}(\epsilon^{-\frac{d}{\alpha}})$, and $K=\widetilde{O}\left(\epsilon^{-\frac{d}{\alpha}}\log\frac{1}{\epsilon}+D\log\frac{1}{\epsilon}\right)$. Substituting L, p, and K into (Eq. 4.5), we have

$$\mathbb{E}\left[\int_{\mathcal{M}}\left(\widehat{f}_n(\mathbf{x})-f_0(\mathbf{x})\right)^2 d\mathcal{D}_x(\mathbf{x})\right]=\widetilde{O}\left(\epsilon^2+\frac{R^2+\sigma^2}{n}\left(\epsilon^{-\frac{d}{\alpha}}+D\right)\log^3\frac{1}{\epsilon}+\frac{1}{n}\right).$$

To balance the error terms, we pick ϵ satisfying $\epsilon^2=\frac{1}{n}\epsilon^{-\frac{d}{\alpha}}$, which gives $\epsilon=n^{-\frac{\alpha}{d+2\alpha}}$. The proof of Theorem 4.1 is complete by plugging in $\epsilon=n^{-\frac{\alpha}{d+2\alpha}}$ and rearranging the terms.

4.2 Nonparametric Classification

This section studies binary classification on a smooth manifold. Different from regression problems, in classification, we use neural networks for approximating likelihood functions. Moreover, to measure the performance of a classifier, we focus on the excess risk by competing with the optimal Bayes classifier. We impose the following data assumption.

Assumption 4.1. *Assume the given data set $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathcal{M}$ and $y_i \in \{-1, 1\}$ is the label, are i.i.d samples from a probability measure $(\mathbf{x}, y) \sim \mu$.*

Denote $\eta(\mathbf{x}) = \mathbb{E}(\mathbb{1}\{y = 1\} \mid \mathbf{x})$ as the probability that the label of $\mathbf{x}$ is 1 where $\mathbb{1}\{\cdot\} = 1$ if $\{\cdot\}$ is true and is 0 otherwise.

Let $\mathcal{Q}$ be the class of all functions mapping $\mathcal{M}$ to $\{-1, 1\}$. The Bayes classifier, which minimizes the misclassification error, is defined as

$$f^* = \operatorname*{argmin}_{Q \in \mathcal{Q}} \mathbb{E}(\mathbb{1}\{Q(\mathbf{x}) \neq y\}).$$

It can be shown that $f^* = \operatorname{sign}(\eta - 1/2)$. Given a classifier f, the excess risk is defined as

$$\mathcal{E}(f, f^*) = \mathcal{E}(f) - \mathcal{E}(f^*) \tag{4.6}$$

with $\mathcal{E}(f) = \mathbb{E}\left[\frac{1}{2}(1 - yf(\mathbf{x}))\right]$ being the misclassification risk of f.

While it is natural to define optimal classifier as the minimizer of $\mathcal{E}(f)$, $\mathcal{E}(f)$ is not differentiable and is NP hard to minimize. Instead, surrogate loss is considered whose minimizer has the same sign as f^* for all $\mathbf{x}$. One popular choice is the logistic loss defined as $\phi(z) = \log(1 + \exp(-z))$. The logistic risk $\mathcal{E}_\phi(f)$ of a classifier f and its empirical risk $\mathcal{E}_{\phi,n}(f)$ are defined as

$$\mathcal{E}_\phi(f) = \mathbb{E}(\phi(yf)), \qquad \mathcal{E}_{\phi,n}(f) = \frac{1}{n}\sum_{i=1}^n \phi(y_i f(\mathbf{x}_i)). \tag{4.7}$$

Denote the minimizer of $\mathcal{E}_\phi(f)$ and $\mathcal{E}_{\phi,n}(f)$ by

$$f_\phi^* = \operatorname*{argmin}_f \mathcal{E}_\phi(f), \qquad \widehat{f}_{\phi,n} = \operatorname*{argmin}_f \mathcal{E}_{\phi,n}(f).$$

One has $f_\phi^* = \log \frac{\eta(\mathbf{x})}{1-\eta(\mathbf{x})}$ and $\mathrm{sign}(f_\phi^*) = f^*$. The logistic excess risk of a classifier f is defined as

$$\mathcal{E}_\phi(f, f_\phi^*) = \mathcal{E}_\phi(f) - \mathcal{E}_\phi(f_\phi^*). \tag{4.8}$$

Related Work Statistical theories for binary classification by FNNs are established with the hinge loss [86, 97] and the logistic loss [85]. Among these works, [97] uses a parametric model given by a teacher-student network. The nonparametric results in [86, 85] are cursed by the data dimension, and therefore require a large number of samples for high-dimensional data.

Binary classification by CNNs has been studied in [98, 99, 100, 101]. Image binary classification is studied in [98, 99] in which the probability function is assumed to be in a hierarchical max-pooling model class. ResNet-type classifiers are considered in [100, 101] while the generalization error is not given explicitly.

4.2.1 Excess Risk Bound

We use ConvResNets to learn a classifier on a smooth manifold by minimizing the logistic loss. In the following theorem, we establish an upper bound on the excess risk of the learned classifier (A proof can be found in [49]).

Theorem 4.2. *Suppose Assumption 3.1, 3.2 and 4.1 hold. Assume* $0 < p, q \leq \infty$, $0 < \alpha < \infty$, $\alpha > d/p + 1$ *and* $\eta \in U(B^\alpha_{p,q}(\mathcal{M}))$. *For any* $2 \leq K \leq D$ *and a ConvResNet architecture* $\mathcal{C}^{(n)}$ *defined as*

$$\mathcal{C}^{(n)} = \left\{ \bar{f} \mid \bar{f} = \bar{g}_2 \circ \bar{h} \circ \bar{g}_1 \circ \bar{\eta} \textit{ where } \bar{\eta} \in \mathrm{CRN}\left(M_1, L, p, K, \kappa_1, \infty, \infty\right), \right.$$

$$\bar{g}_1 \in \mathrm{CRN}\left(1,4,8,1,\kappa_2,\infty,\infty\right),\ \bar{h} \in \mathrm{CRN}\left(M_2,L,p,1,\kappa_1,\infty,\infty\right),$$

$$\bar{g}_2 \in \mathcal{C}\left(1,3,8,1,\kappa_3,1,R\right) \textit{ with}$$

$$M_1 = O\left(n^{\frac{2d}{\alpha+2(\alpha\vee d)}}\right),\ M_2 = O\left(n^{\frac{2\alpha}{\alpha+2(\alpha\vee d)}}\right),\ L = O(\log(n)),\ p = O(1),$$

$$\kappa_1 = O(1),\ \log\kappa_2 = O(\log^2 n),\ \kappa_3 = O(\log n),\ R = O(\log n).\}$$

Let $\bar{f}_{\phi,n}$ be the minimizer of the empirical risk in (Eq. 4.7) among functions in $\mathcal{C}^{(n)}$. We have

$$\mathbb{E}(\mathcal{E}_\phi(\bar{f}, f_\phi^*)) \leq Cn^{-\frac{\alpha}{2\alpha+2(\alpha\vee d)}}\log^4 n$$

for some constant C. The constants hidden in $O(\cdot)$ depend on $d, \log D, \alpha, \frac{2d}{\alpha p-d}, \tau$ and the surface area of $\mathcal{M}$.

The ConvResNet $\mathcal{C}^n$ in Theorem 4.2 consists of four sub-ConvResNets: $\bar{\eta},\ \bar{g}_1,\ \bar{h}$ and $\bar{g}_2$. $\bar{\eta}$ is a network estimating the probability function η and $\bar{g}_1$ is a function which truncates $\bar{\eta}$ to some range. $\bar{h}$ approximates the function $\frac{\log z}{\log(1-z)}$ and $\bar{g}_2$ truncates the output of $\bar{h}$ to some range. Recall that

$$f_\phi^* = \frac{\log\eta}{\log(1-\eta)}. \tag{4.9}$$

$\bar{h}$ approximates the operation on the right-hand side of (Eq. 4.9).

Theorem 4.2 shows that if the architecture of the ConvResNet is properly chosen, the minimizer of the empirical logistic risk in (Eq. 4.7) has the excess risk in the order of $n^{-\frac{\alpha}{2\alpha+2(\alpha\vee d)}}\log^3 n$. The exponent in the rate only depends on the smoothness of the probability function η and the intrinsic dimension d, not the ambient dimension D. In the case that $s > d$, we get rate $O\left(n^{-\frac{1}{4}}\log^3 n\right)$. Compared with [85, Theorem 4] in which the authors made assumptions on the smoothness of the decision boundary of the Bayes classifier, the margin ([85, Assumption (M)]) and the behavior of f_ϕ^* ([85, Assumption (E)]), in

Theorem 4.2, only assumption on the smoothness of η is made.

4.3 Conclusion

In this chapter, we establish sample complexity bounds using neural networks for nonparametric regression/classification, when data are sampled from a low-dimensional Riemannian manifold. We demonstrate that properly chosen neural networks can circumvent the curse of data ambient dimensionality.

CHAPTER 5

DISTRIBUTION ESTIMATION OF GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GANs, [2]) utilize two neural networks competing with each other to generate new samples with the same distribution as the training data. They have been successful in many applications including producing photorealistic images, improving astronomical images, and modding video games [102, 103, 104, 105, 106, 107, 108].

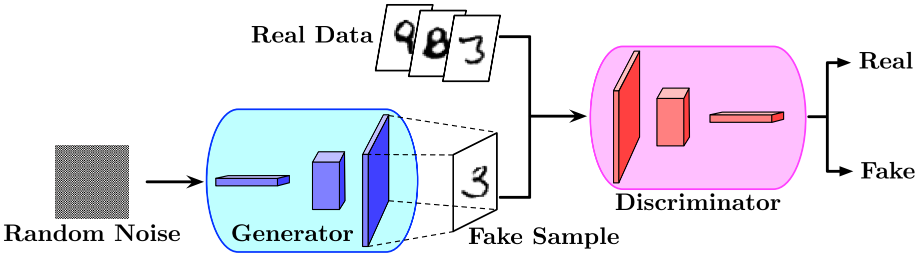


Figure 5.1: The architecture of GANs.

From the perspective of statistics, GANs have stood out as an important unsupervised method for learning target data distributions. Different from explicit distribution estimators, such as the kernel density estimator, GANs implicitly learn the data distribution and act as samplers to generate new fake samples mimicking the data distribution (see Figure 5.1).

To estimate a data distribution μ, GANs solve the following minimax optimization problem

$$(g^*, f^*) \in \operatorname*{argmin}_{g \in \mathcal{G}} \max_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{z} \sim \rho}[f(g(\mathbf{z}))] - \mathbb{E}_{\mathbf{x} \sim \mu}[f(\mathbf{x})], \tag{5.1}$$

where $\mathcal{G}$ denotes a class of generators, $\mathcal{F}$ denotes a symmetric class (if $f \in \mathcal{F}$, then $-f \in \mathcal{F}$) of discriminators, and $\mathbf{z}$ follows some easy-to-sample distribution ρ, e.g., a uniform distribution. The estimator of μ is given by a pushforward distribution of ρ under g^*.

The inner maximization problem of (Eq. 5.1) is an Integral Probability Metric (IPM, [109]), which quantifies the discrepancy between two distributions μ and ν w.r.t. the symmetric function class $\mathcal{F}$:

$$d_{\mathcal{F}}(\mu,\nu) = \sup_{f\in\mathcal{F}} \mathbb{E}_{\mathbf{x}\sim\mu}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{y}\sim\nu}[f(\mathbf{y})].$$

Accordingly, GANs essentially minimize an IPM between the generated distribution and the data distribution. IPM unifies many standard discrepancy metrics. For example, when $\mathcal{F}$ is taken to be all 1-Lipschitz functions, $d_{\mathcal{F}}(\cdot,\cdot)$ is the Wasserstein-1 distance $W_1(\cdot,\cdot)$; when $\mathcal{F}$ is the class of all indicator functions, $d_{\mathcal{F}}(\cdot,\cdot)$ is the total variation distance; when $\mathcal{F}$ is taken as neural networks, $d_{\mathcal{F}}(\cdot,\cdot)$ is the so-called "neural net distance" [110].

In practical GANs, the generator and discriminator classes $\mathcal{G}$ and $\mathcal{F}$ are parametrized by neural networks. We denote $\mathcal{G} = \mathcal{G}_{\text{NN}}$ and $\mathcal{F} = \mathcal{F}_{\text{NN}}$ to emphasize such a parameterization. In this chapter, we focus on using feedforward ReLU networks, since it has wide applications [111, 70, 112] and can ease the notorious vanishing gradient issue during training, which commonly arises with sigmoid or hyperbolic tangent activations [70, 69]

When n samples of the data distribution μ are given, denoted as $\{x_i\}_{i=1}^n$, one can replace μ in (Eq. 5.1) by its empirical counterpart $\widehat{\mu}_n$, and (Eq. 5.1) becomes

$$(g_\theta^*, f_\omega^*) \in \operatorname*{argmin}_{g_\theta\in\mathcal{G}_{\text{NN}}} \max_{f_\omega\in\mathcal{F}_{\text{NN}}} \mathbb{E}_{\mathbf{z}\sim\rho}[f_\omega(g_\theta(\mathbf{z}))] - \frac{1}{n}\sum_{i=1}^n f_\omega(\mathbf{x}_i), \tag{5.2}$$

where θ and ω are parameters in the generator and discriminator networks, respectively. The empirical estimator of μ given by GANs is the pushforward distribution of ρ under g_θ^*, denoted by $(g_\theta^*)_\sharp\rho$.

In contrast to the prevalence of GANs in applications, there are very limited works on the theoretical properties of GANs [110, 113, 114, 115, 116]. This chapter focuses on the following fundamental questions from a theoretical point of view:

- *(Q1).* What types of distributions can be approximated by a deep neural network generator?

- *(Q2).* If the distribution can be approximated, what is the statistical rate of estimation using GANs?
- *(Q3).* If further there are unknown low-dimensional structures in the data distribution, can GANs capture the low-dimensional data structure and enjoy a fast rate of estimation?

5.1 Results in A Nutshell

Results in Euclidean space To address *(Q1)* and *(Q2)*, we show that, if the generator and discriminator network architectures are properly chosen, GANs can learn distributions with Hölder densities supported on a convex domain. Specifically, we consider a data distribution μ supported on a compact convex subset $\mathcal{X} \subset \mathbb{R}^D$, where D is the data dimension. We assume μ has an α-Hölder density with respect to Lebesgue measure in $\mathbb{R}^D$ and the density is lower bounded away from 0 on $\mathcal{X}$.

Our generator and discriminator network architectures are explicitly chosen – we specify the width and depth of the network, total number of neurons, and total number of weight parameters (details are provided in Section 5.3). Roughly speaking, the generator needs to be flexible enough to approximately transform an easy-to-sample distribution to the data distribution, and the discriminator is powerful enough to distinguish the generated distribution from the data distribution.

Let g_θ^* be the optimal solution of (Eq. 5.2), and then $(g_\theta^*)_\sharp \rho$ is the generated data distribution as an estimation of μ. Our main result can be summarized as, for any $\beta \geq 1$, if the generator and discriminator network architectures are properly chosen, then

$$\mathbb{E}\left[d_{\mathcal{H}^\beta}\left((g_\theta^*)_\sharp \rho, \mu\right)\right] = \widetilde{O}\left(n^{-\frac{\beta}{2\beta+D}} \log^2 n\right), \tag{5.3}$$

where the expectation is taken over the randomness of samples and $\widetilde{O}$ hides polynomial factors in β, D. It shows that the β-Hölder IPM between the generated distribution and the data distribution converges at a rate depending on the Hölder index β and dimension D. When $\beta = 1$, our theory implies that GANs can estimate any distribution with a Hölder density under the Wasserstein-1 distance. A comparison to closely related works is provided in Section 5.5.

In our analysis, we decompose the distribution estimation error into a statistical error and an approximation error by an oracle inequality. A key step is to properly choose the generator network architecture to control the approximation error. Specifically, the generator architecture allows an accurate approximation to a data transformation T such that $T_{\sharp}\rho = \mu$. The existence of such a transformation T is guaranteed by optimal transport theory [117], and holds universally for all the data distributions with Hölder densities.

Results in low-dimensional linear subspace Moreover, we provide a positive answer to *(Q3)* by considering data distributions with low-dimensional linear structures. Specifically, we assume the data support $\mathcal{X} \subset \mathbb{R}^D$ is a compact subset of a q-dimensional linear subspace. Let columns of $A \in \mathbb{R}^{D \times q}$ denote a set of orthonormal basis of the q-dimensional linear subspace. We assume the pushforward $A_{\sharp}^{\top}\mu$ of data distribution has a density function p_μ defined in $\mathbb{R}^q$, and p_μ is α-Hölder continuous and lower bounded away from 0 on its support. We leverage the data geometric structures and generate samples by transforming an easy-to-sample distribution ρ in $\mathbb{R}^q$. With a proper choice of the generator and discriminator network architectures, the statistical error of GANs converges at a fast rate

$$\mathbb{E}\left[W_1\left((g_\theta^*)_{\sharp}\rho, \mu\right)\right] = \widetilde{O}\left(n^{-\frac{1}{2+q}} \log^2 n\right). \tag{5.4}$$

By taking $\beta = 1$ in (Eq. 5.3), we note that (Eq. 5.4) enjoys a faster statistical convergence in the Wasserstein-1 distance, since the exponent only depends on the intrinsic dimension q. Meanwhile, (Eq. 5.4) indicates that GANs can circumvent the curse of ambient dimensionality when data are supported on a low-dimensional subspace. A detailed comparison with existing works is given in Section 5.5.

From a technical point of view, a key challenge in obtaining the fast rate in (Eq. 5.4) is to prove that the generator can capture the unknown linear structure in data. We achieve this by introducing a learnable linear projection layer in the generator, and pairing it with an "anti-projection" layer in the discriminator. We show (see Lemma 5.9) that by optimizing (Eq. 5.2), the linear projection layer in generator accurately recovers the linear subspace of data.

The rest of the chapter is organized as follows: Section 5.2 briefly introduces IPM and optimal transport theory. Section 5.3 presents the statistical guarantees of GANs for learning data distributions with a Hölder density. Section 5.4 extends the statistical theory to low-dimensional data, and shows that GANs can adapt to the intrinsic structures in data. Section 5.5 comapres with existing works. Section 5.6 proves main results. Section 5.7 concludes the chapter and discusses related topics.

5.2 IPM and Optimal Transport

In order to measure the performance of GANs in estimating target distribution μ, we adopt the Integral Probability Metric (IPM) with respect to Hölder discriminative functions. In particular, suppose GAN generates a fake distribution ν. For any $\beta \geq 1$, we denote

$$d_{\mathcal{H}^\beta}(\mu, \nu) = \sup_{f \in \mathcal{H}^\beta} \mathbb{E}_{\mathbf{x} \sim \mu}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{y} \sim \nu}[f(\mathbf{y})].$$

Remark 5.1. *It is convenient to restrict $\mathcal{H}^\beta$ in IPM $d_{\mathcal{H}^\beta}$ to have a bounded radius. Specifically, for any $f \in \mathcal{H}^\beta$, we assume $\|f\|_{\mathcal{H}^\beta} \leq C$ for some constant C. Otherwise, we can simply rescale f while maintaining the discriminative power of the IPM. In addition, since IPMs are translation invariant, meaning that discriminative functions f and $f + c$ for some constant c are equivalent. Therefore, we also assume $f(\mathbf{0}) = 0$ for simplicity.*

In the special case of $\beta = 1$, $d_{\mathcal{H}^\beta}(\cdot, \cdot)$ shares the same discriminative power as Wasserstein-

1 distance, which can be defined using the dual formulation,

$$W_1(\mu, \nu) = \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \mathbb{E}_{\mathbf{x} \sim \mu}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{y} \sim \nu}[f(\mathbf{y})].$$

In the right-hand side above, $\|f\|_{\mathrm{Lip}}$ denotes the Lipschitz coefficient of f. It can be checked that Lipschitz functions are Hölder continuous with Hölder index equal to 1. Therefore, $W_1(\cdot, \cdot)$ is equivalent to $d_{\mathcal{H}^1}(\cdot, \cdot)$.

GANs are closely related to Optimal Transport (OT, [118, 119, 120, 121]), as the generator essentially learns a pushforward mapping of an easy-to-sample distribution. A typical problem in OT is the following: Let $\mathcal{X}, \mathcal{Z}$ be subsets of $\mathbb{R}^D$. Given two probability spaces $(\mathcal{X}, \mu)$ and $(\mathcal{Z}, \rho)$, OT aims to find a transformation $T : \mathcal{Z} \mapsto \mathcal{X}$, such that $T(\mathbf{z}) \sim \mu$ for $\mathbf{z} \sim \rho$. In general, the transformation T may neither exist nor be unique. Fortunately, in the case that μ and ρ have Hölder densities p_μ and p_ρ, respectively, the Monge map ensures the existence of a Hölder transformation T^*, when $\mathcal{X}$ is convex. In particular, the Monge map T^* is the solution to the following optimization problem:

$$T^* \in \underset{T}{\operatorname{argmin}} \;\; \mathbb{E}_{\mathbf{z} \sim \rho}\left[\ell(\mathbf{z}, T(\mathbf{z}))\right], \quad \text{subject to} \quad T_\sharp \rho = \mu, \tag{5.5}$$

where ℓ is a cost function. (Eq. 5.5) is known as the Monge problem. When $\mathcal{X}$ is convex and the cost function is quadratic, the solution to (Eq. 5.5) satisfies the Monge-Ampère equation [122]. The regularity of T^* was proved in [123, 124, 125] and [126, 127] independently. Their main result is summarized in the following lemma.

Lemma 5.1 ([123])**.** *Suppose μ and ρ both have α-Hölder densities, and the support $\mathcal{X}$ is convex. Then there exists a transformation $T^* : \mathcal{Z} \mapsto \mathcal{X}$ such that $T^*_\sharp \rho = \mu$. Moreover, this transformation T^* belongs to the Hölder class $\mathcal{H}^{\alpha+1}(\mathcal{Z})$.*

We will see later that Lemma 5.1 provides important guidelines for choosing proper generator networks in distribution estimation.

5.3 Distribution Estimation in Euclidean Space

We consider a data distribution μ supported on a convex subset $\mathcal{X} \subset \mathbb{R}^D$ and assume that μ has a density function p_μ with respect to the Lebesgue measure in $\mathbb{R}^D$. GANs seek to estimate the data distribution μ by transforming some easy-to-sample distribution ρ supported on domain $\mathcal{Z} \subset \mathbb{R}^D$, such as a uniform distribution. Our main results provide statistical guarantees of GANs for the estimation of μ, based on the following assumptions.

Assumption 5.1. *The domains $\mathcal{X}$ and $\mathcal{Z}$ are compact, and $\mathcal{X}$ is convex. There exists a constant $B > 0$ such that for any $\mathbf{x} \in \mathcal{X}$ or $\mathbf{x} \in \mathcal{Z}$, $\|\mathbf{x}\|_\infty \leq B$.*

Assumption 5.2. *Given a Hölder index $\alpha > 0$, the density function p_μ of μ (w.r.t. Lebesgue measure in $\mathbb{R}^D$) belongs to the Hölder class $\mathcal{H}^\alpha(\mathcal{X})$ with $\|p_\mu\|_{\mathcal{H}^\alpha(\mathcal{X})} \leq C$ for some constant $C > 0$. Meanwhile, p_μ is lower bounded, i.e.,*

$$\inf_{\mathbf{x} \in \mathcal{X}} p_\mu(\mathbf{x}) \geq \tau$$

for some constant $\tau > 0$.

Assumption 5.3. *The easy-to-sample distribution ρ has a C^∞ (smooth) density function p_ρ.*

Hölder regularity is commonly used in literature on smooth density estimation [74, 14]. In the remaining of the paper, we occasionally omit the domain in Hölder spaces when it is clear from the context. The condition of p_μ being lower bounded is a common technical assumption in the optimal transport theory [128, 125]. This condition and the convexity of $\mathcal{X}$ guarantee that, there exists a Hölder transformation T such that $T_\sharp \rho = \mu$ (see Lemma 5.1). Besides, Assumption 5.3 is always satisfied, since ρ is often taken as a uniform distribution.

Given Assumption 5.1 – 5.3, we set the generator network architecture as

$$\mathcal{G}_{\rm NN}(R,\kappa,L,p,J)=\mathrm{FNN}(R,\kappa,L,p,J)\quad\text{with input output dimension } d_{\rm in}=d_{\rm out}=d.$$

and the discriminator network architecture as

$$\mathcal{F}_{\rm NN}(\bar{R},\bar{\kappa},\bar{L},\bar{p},\bar{J})=\mathrm{FNN}(\bar{R},\bar{\kappa},\bar{L},\bar{p},\bar{J})\quad\text{with input output dimension } d_{\rm in}=d, d_{\rm out}=1.$$

We first show a properly chosen generator network can universally approximate data distributions with a Hölder density.

Theorem 5.1 (Distribution approximation theory)**.** *For any data distribution $(\mathcal{X},\mu)$ and easy-to-sample distribution $(\mathcal{Z},\rho)$ satisfying Assumption 5.1 - 5.3, there exists an $(\alpha+1)$-Hölder continuous transformation $T:\mathbb{R}^D\to\mathbb{R}^D$ such that $T_\sharp\rho=\mu$. Moreover, given any $\epsilon\in(0,1)$, there exists a generator network with configuration*

$$\begin{gathered}L=O(\log(1/\epsilon)),\quad p=O(D\epsilon^{-\frac{D}{\alpha+1}}),\quad J=O(D\epsilon^{-\frac{D}{\alpha+1}}\log(1/\epsilon)),\\ R=B,\quad \kappa=\max\{C,B\},\end{gathered}\tag{5.6}$$

such that, if the weight parameters of this network are properly chosen, then it yields a transformation g_θ satisfying

$$\max_{\mathbf{z}\in\mathcal{Z}}\|g_\theta(\mathbf{z})-T(\mathbf{z})\|_\infty\le\epsilon\quad\textit{and}\quad W_1((g_\theta)_\sharp\rho,\mu)\le\sqrt{D}\epsilon.$$

In Theorem 5.1, the existence of a transformation T is guaranteed by optimal transport theory (Lemma 5.1). Furthermore, we explicitly choose a generator network architecture to approximately realize T, such that the easy-to-sample distribution is approximately transformed to the data distribution.

Our statistical result is the following finite-sample estimation error bound in terms of the Hölder IPM between $(g_\theta^*)_\sharp\rho$ and μ, where g_θ^* is the optimal solution of GANs in (Eq. 5.2).

We use $O(\cdot)$ to hide constant factors depending on B, C, α, and β; $\widetilde{O}(\cdot)$ further hides polynomial factors of D and logarithmic factors of n.

Theorem 5.2 (Statistical estimation theory)**.** *Suppose Assumption 5.1 – 5.3 hold. For any $\beta \geq 1$, choose $\epsilon = n^{-\frac{\beta}{2\beta+D}}$ in Theorem 5.1 for the generator network and*

$$\bar{L} = O\left(\frac{\beta}{2\beta + D}\log n\right), \quad \bar{p} = O\left(n^{\frac{D}{2\beta+D}}\right), \quad \bar{J} = O\left(\frac{\beta}{2\beta + D} n^{\frac{D}{2\beta+D}}\log n\right),$$
$$\bar{R} = C, \quad \bar{\kappa} = C,$$

for the discriminator network. Then it holds

$$\mathbb{E}\left[d_{\mathcal{H}^\beta}((g_\theta^*)_\sharp \rho, \mu)\right] = \widetilde{O}\left(n^{-\frac{\beta}{2\beta+D}}\log^2 n\right). \tag{5.7}$$

Theorem 5.2 demonstrates that GANs can effectively learn data distributions, with a convergence rate depending on the smoothness of the function class in IPM and the dimension D.

In the case that only m samples from the easy-to-sample distribution ρ are collected, GANs solve the following empirical minimax problem

$$\min_{g_\theta \in \mathcal{G}_{\mathrm{NN}}} \max_{f_\omega \in \mathcal{F}_{\mathrm{NN}}} \frac{1}{m}\sum_{i=1}^{m} f_\omega(g_\theta(\mathbf{z}_i)) - \frac{1}{n}\sum_{j=1}^{n} f_\omega(\mathbf{x}_j). \tag{5.8}$$

We denote $(g_\theta^{*,m}, f_\omega^{*,m})$ as the optimal solution of (Eq. 5.8). We show in the following corollary that GANs retain similar statistical guarantees for distribution estimation with finite generated samples.

Corollary 5.1. *Suppose Assumption 5.1 – 5.3 hold and $m \geq n$. We choose*

$$L = O\left(\frac{\alpha + 1}{2(\alpha+1) + D}\log m\right), \quad p = O\left(Dm^{\frac{D}{2(\alpha+1)+D}}\right),$$
$$J = O\left(\frac{D(\alpha+1)}{2(\alpha+1) + D} m^{\frac{D}{2(\alpha+1)+D}}\log m\right), \quad R = B, \quad \kappa = \max\{C, B\},$$

for the generator network and the same architecture as in Theorem 5.2 for the discriminator network. Then it holds

$$\mathbb{E}\left[d_{\mathcal{H}^\beta}((g_\theta^{*,m})_\sharp \rho, \mu)\right] = \widetilde{O}\left(n^{-\frac{\beta}{2\beta+D}} + m^{-\frac{\alpha+1}{2(\alpha+1)+D}}\right).$$

Here $\widetilde{O}$ also hides a logarithmic factors m. As it is often cheap to obtain a large amount of samples from ρ, the convergence rate in Corollary 5.1 is dominated by $n^{-\frac{\beta}{2\beta+D}}$ whenever $m \geq n^{\frac{\beta}{\alpha+1}\frac{2(\alpha+1)+D}{2\beta+D}\vee 1}$.

Theorem 5.2 and Corollary 5.1 suggest that GANs suffer from the curse of data dimensionality. However, such an exponential dependence on the dimension d is inevitable without further assumptions on the data, as indicated by the minimax optimal rate of distribution estimation: To estimate a distribution μ with a $\mathcal{H}^\alpha(\mathcal{X})$ density, the minimax optimal rate under the $\mathcal{H}^\beta$ IPM loss satisfies

$$\inf_{\widetilde{\mu}_n} \sup_{\mu \in \mathcal{H}^\alpha} \mathbb{E}\left[d_{\mathcal{H}^\beta}(\widetilde{\mu}_n, \mu)\right] \gtrsim n^{-\frac{\alpha+\beta}{2\alpha+D}} + n^{-\frac{1}{2}},$$

where $\widetilde{\mu}_n$ is any estimator of μ based on n data points [114, 129].

5.4 Distribution Estimation in Low-dimensional Linear Subspace

In this section, we prove that GANs are adaptive to unknown low-dimensional linear structures in data. We consider the data domain $\mathcal{X} \subset \mathbb{R}^D$ being a compact subset of a q-dimensional linear subspace with $q \ll D$. Our analysis holds for general $q \leq D$, while $q \approx D$ is less of interest as practical data sets are often low-dimensional with intrinsic dimension much smaller than ambient dimension [18, 19, 130].

Assumption 5.4. *The data domain $\mathcal{X}$ is compact, i.e., there exists a constant $B > 0$ such that for any $\mathbf{x} \in \mathcal{X}$, $\|\mathbf{x}\|_\infty \leq B$. Moreover, $\mathcal{X}$ is a convex subset of a q-dimensional linear subspace in $\mathbb{R}^d$, and the span of $\mathcal{X}$ is the q-dimensional subspace.*

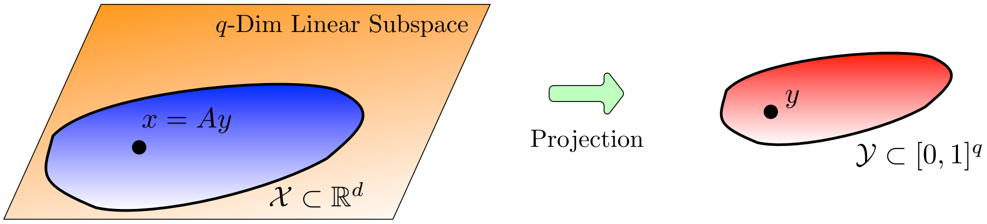

q-Dim Linear Subspace
$x = Ay$
$\mathcal{X} \subset \mathbb{R}^d$
Projection
y
$\mathcal{Y} \subset [0,1]^q$

Figure 5.2: Low-dimensional linear structures in $\mathcal{X}$.

Under Assumption 5.4, a data point $\mathbf{x} \in \mathcal{X}$ can be represented as $A\mathbf{y}$, where $\mathbf{y} \in \mathbb{R}^q$ and $A \in \mathbb{R}^{D\times q}$ is a linear transformation (See graphical illustration in Figure 5.2). The following lemma formally justifies the existence of the linear transformation A.

Lemma 5.2. *Suppose Assumption 5.4 holds. Consider a matrix $A \in \mathbb{R}^{D\times q}$ with columns being an orthonormal basis of the q-dimensional linear subspace. Then it holds that $\mathcal{Y} = A^\top \mathcal{X} = \{A^\top \mathbf{x} : \mathbf{x} \in \mathcal{X}\}$ is a compact and convex subset of $\mathbb{R}^q$, and $A\mathcal{Y} = \mathcal{X}$.*

The proof is deferred to Appendix C.4. The projected domain $\mathcal{Y}$ captures the intrinsic geometric structures in $\mathcal{X}$. More importantly, using transformation A allows us to define smoothness of the target data distribution. Specifically, we consider a data distribution μ supported on $\mathcal{X}$. Since $\mathcal{X}$ is a low-dimensional space, μ does not have a well defined density function with respect to the Lebesgue measure in $\mathbb{R}^D$. Thanks to Lemma 5.2, the pushforward distribution $A^\top_\sharp \mu$ has a well-defined density function. Accordingly, we make the following data distribution assumption.

Assumption 5.5. *Without loss of generality, we assume $\mathcal{Y} \subset [0,1]^q$. Given a Hölder index $\alpha > 0$, the density function p_μ of $A^\top_\sharp \mu$ belongs to $\mathcal{H}^\alpha(\mathcal{Y})$ with a bounded Hölder norm $\|p_\mu\|_{\mathcal{H}^\alpha(\mathcal{Y})} \le C$ for some constant $C > 0$, and $p_\mu \ge \tau > 0$ on $\mathcal{Y}$ for some constant τ.*

We assume $\mathcal{Y} \subset [0,1]^q$ for convenience. Otherwise, we can rescale input space $\mathcal{X}$ by a constant c, so that the projected space $\mathcal{Y} \subset [0,1]^q$. Since $\mathcal{X}$ is compact, the constant c is bounded and will not undermine the statistical rate of convergence.

To generate samples mimicking data distribution μ, we consider transforming a q-dimensional easy-to-sample distribution ρ supported on $[0,1]^q$ to leverage the structural assumption in domain $\mathcal{X}$. We define the generator network architecture $\mathcal{G}_{\text{NN}}^{\text{ld}}(R, \kappa, L, p, J)$ as

$$\begin{aligned} \mathcal{G}_{\text{NN}}^{\text{ld}}(R, \kappa, L, p, J) = \big\{ & U \circ g : U \in \mathbb{R}^{d \times q} \text{ with orthonormal columns and} \\ & g \in \text{FNN}(R, \kappa, L, p, J) \text{ with } d_{\text{in}} = q, d_{\text{out}} = d \big\}. \end{aligned} \tag{5.9}$$

Note that $U \in \mathbb{R}^{D \times q}$ lifts the transformed easy-to-sample distribution $(g_\theta)_\sharp \rho$ to $\mathbb{R}^D$. We expect U to extract the linear structures in data, while g_θ approximates an optimal transport plan for transforming ρ to $A_\sharp^\top \mu$.

In correspondence with the generator, we define the discriminator network architecture $\mathcal{F}_{\text{NN}}^{\text{ld}}(\bar{R}, \bar{\kappa}, \bar{L}, \bar{p}, \bar{J}, \bar{\gamma})$ as

$$\begin{aligned} \mathcal{F}_{\text{NN}}^{\text{ld}}(\bar{R}, \bar{\kappa}, \bar{L}, \bar{p}, \bar{J}, \bar{\gamma}) = \big\{ & f \circ V^\top : V \in \mathbb{R}^{d \times q} \text{ with } \|V\|_2 \leq 1, \\ & f \in \text{FNN}(\bar{R}, \bar{\kappa}, \bar{L}, \bar{p}, \bar{J}) \text{ with } d_{\text{in}} = q, d_{\text{out}} = 1, \text{ and} \\ & |f(\mathbf{x}) - f(\mathbf{y})| \leq \bar{\gamma} \|\mathbf{x} - \mathbf{y}\|_\infty \text{ for } \mathbf{x}, \mathbf{y} \in [0,1]^q \big\}. \end{aligned} \tag{5.10}$$

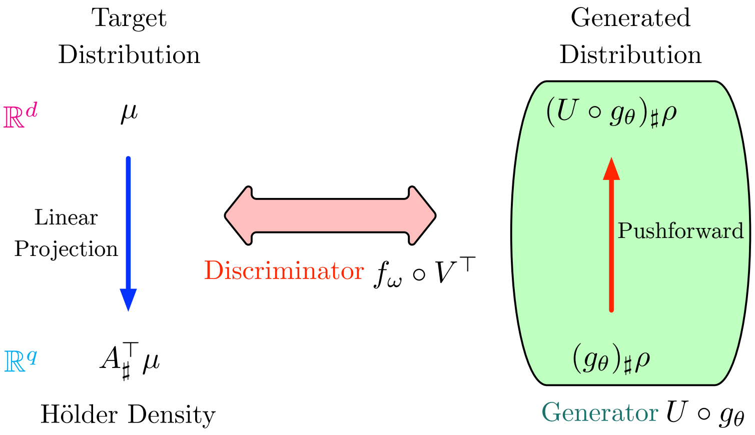


Figure 5.3: Learning data distribution μ with unknown linear structures using generator in (Eq. 5.9) and discriminator in (Eq. 5.10).

Matrix V is chosen to "couple" with the linear structures learned by the generator ("anti-projection") and f_ω will approximate Lipschitz functions in $\mathbb{R}^q$ for approximating Wasserstein distance. We remark that an appropriate choice of Lipschitz coefficient $\bar{\gamma}$ on f_ω will not undermine the approximation power of $\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}$ as confirmed in Lemma 5.8. Meanwhile, the Lipschitz constraint of discriminator ensures that the generator can accurately capture the linear structures in data. In practice, such a Lipschitz regularity is often enforced by computational heuristics [131, 132, 133].

With proper configurations of network classes (Eq. 5.9) and (Eq. 5.10), we train GANs using (Eq. 5.2) (see Figure 5.3 for illustration) and denote the optimizer as $(U^*, g_\theta^*, V^*, f_\omega^*)$, i.e.,

$$(U^*, g_\theta^*, V^*, f_\omega^*) \in \underset{f_\omega \circ V^\top \in \mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}{\operatorname{argmin}} \max_{U \circ g_\theta \in \mathcal{G}_{\mathrm{NN}}^{\mathrm{ld}}} \mathbb{E}_{\mathbf{z} \sim \rho} \left[(f_\omega \circ V^\top) \circ (U \circ g_\theta)(\mathbf{z}) \right] \\ - \frac{1}{n} \sum_{i=1}^{n} (f_\omega \circ V^\top)(\mathbf{x}_i).$$

The following theorem establishes a fast statistical rate of convergence of $(U^* \circ g_\theta^*)_\sharp \rho$ to data distribution μ.

Theorem 5.3. *Suppose Assumption 5.4 and 5.5 hold. We choose*

$$R = B, \quad \kappa = \max\{B, C\}, \quad L = O\left(\frac{\alpha}{2\alpha + q} \log n\right),$$
$$p = O\left(qn^{\frac{q\alpha}{(\alpha+1)(2\alpha+q)}} \vee D\right), \quad J = O\left(Dq + \frac{\alpha}{2\alpha + q} n^{\frac{q\alpha}{(\alpha+1)(2\alpha+q)}} \log n\right).$$

for the generator $\mathcal{G}_{\mathrm{NN}}^{\mathrm{ld}}(R, \kappa, L, p, J)$ *in* (Eq. 5.9) *and*

$$\bar{R} = C, \quad \bar{\kappa} = C, \quad \bar{\gamma} = 10q, \quad \bar{L} = O\left(\frac{1}{2+q} \log n\right),$$
$$\bar{p} = O\left(n^{q/(2+q)} \vee D\right), \quad \bar{J} = O\left(Dq + \frac{1}{2+q} n^{q/(2+q)} \log n\right).$$

for the discriminator $\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}(\bar{R}, \bar{\kappa}, \bar{L}, \bar{p}, \bar{J}, \bar{\gamma})$ *in* (Eq. 5.10). *Then it holds*

$$\mathbb{E}\left[W_1((U^* \circ g_\theta^*)_\sharp \rho, \mu)\right] = \widetilde{O}\left(n^{-\frac{1}{2+q}} \log^2 n\right).$$

Compared to Theorem 5.2, we observe that the sizes of generator and discriminator in Theorem 5.3 only weakly depend on D. Meanwhile, the rate of convergence is fast as the exponent only depends on q. This result provides important understandings of why GANs can circumvent the curse of dimensionality in real-world applications, since low-dimensional intrinsic structures are often seen in real-world data sets. Nonetheless, linear structures in Assumption 5.4 is largely simplified, as it is rare the case that real-world data lie in a subset of a low-dimensional linear subspace. At the same time, real data are often contaminated with observational noise and concentrate only near a low-dimensional manifold.

Theorem 5.3 demonstrates that GANs with properly chosen generator and discriminator are adaptive to the unknown linear structures in data. Since data are concentrated on a linear subspace, one may advocate PCA-like methods for estimating the linear structure first and then learn the data distribution on a projected subspace. However, such a method requires two-step learning and is rarely used in practical GANs. In fact, GANs simultaneously capture the linear structure and learning the target data distribution via optimizing the empirical risk (Eq. 5.2).

A major difficulty in establishing Theorem 5.3 is proving GANs can capture unknown linear structures in data. We exploit the optimality of (U^*, g_θ^*) to prove that $\|U^* - A\|_{\mathrm{F}}$ is small, i.e., the column spaces of U^* and the ground truth matrix A match closely. In particular, the mismatch $\|U^* - A\|_{\mathrm{F}}$ depends on the approximation power of the generator and discriminator (see Lemma 5.9). Built upon this crucial ingredient, the remaining analysis focuses on tackling the projected Wasserstein distance with respect to the data transformation A (see [134, 135] for applications of projected Wasserstein distance in two-sample test). In this way, we circumvent the curse of ambient dimensionality.

5.5 Comparison with Existing Literature

• **Distribution approximation using deep generative models**. Using generator to accurately approximate the data distribution is of essential importance in understanding the statistical properties of GANs. [113] considered data distribution being exactly realized by an invertible generator, i.e., all the weight matrices and activation functions are invertible. Such an invertibility requires the width of the generator to be the same as input data dimension D. Existing literature has shown that such narrow networks lack approximation ability [38, 136]. In fact, to ensure universal approximation for Lebesgue-integrable functions and L^p functions in $\mathbb{R}^D$, the weakest width requirement needs to be $D+4$ and $D+1$, respectively. Our work, in contrast, allows the generator to be wide and expressive for any data distribution with Hölder densities.

$\star$ *Approximating empirical distribution using neural networks*. After the release of an early version of the manuscript, the authors were aware of a concurrent work studying distribution approximation using generative networks. Specifically, [137] established universal approximation abilities of neural network generators for approximating sub-Gaussian data distributions. They proved the existence of a properly chosen generator architecture for achieving an ϵ approximation error of data distribution in Wasserstein-1 distance. Our Theorem 5.1 shares a similar conclusion to [137] for data distributions with Hölder densities. However, the analysis in [137] is very different and relies on memorizing discretized data distribution using neural networks. More recently, [138] showed that GANs can approximate any data distribution (in any dimension) by transforming an absolutely continuous distribution. The idea is to memorize the empirical data distribution using ReLU networks. Nonetheless, the designed generator may not be able to generate new samples (different from the training data), which cannot explain the success of GANs in practice.

• **Statistical properties of GANs**. Statistical guarantees of generative models for distribution estimation has been studied in several works. We compare with existing works in

Table 5.1 and provide more details in following context.

Table 5.1: A comparison to closely related works in problem setups and statistical results. NN stands for neural networks and '—' indicates no specific choice is given. Weak metric refers to "neural net distance" in [110] and strong metric refers to IPMs with non-parametric discriminative function classes, e.g., using 1-Lipschitz discriminative functions corresponds to the Wasserstein-1 distance.

	Generator	Discriminator	Distribution	Metric
Generalization error bound				
[110, 139, 140]	NN	NN	General (Euclidean)	Weak
[113, 141, 114]	Invertible NN	NN	Realizable by invertible NN generators (Eulidean and low-d)	Strong
[142]	C^s	—	C^s pushforward of sub-Gaussian distributions (Eulidean and low-d)	Sinkhorn
[116]	—	Hölder	General (Euclidean and low-d)	Strong
Statistical estimation bound				
[115]	NN	Lipschitz	C^s pushforward of uniform distributions (low-d)	Strong
Ours	NN	NN	Having Hölder densities (Euclidean and low-d)	Strong

$\star$ *Generalization bound of GANs.* [110] studied the generalization error of GANs. Lemma 1 in [110] shows that GANs cannot generalize under the Wasserstein distance and the Jensen-Shannon divergence unless the sample size is $\widetilde{O}(\epsilon^{-\text{poly}(D)})$, where ϵ is the generalization gap. Alternatively, they defined a surrogate metric called "neural net distance" $d_{\mathcal{F}_{\text{NN}}}(\cdot,\cdot)$, where $\mathcal{F}_{\text{NN}}$ is the class of discriminator networks. They proved that GANs generalize under the neural net distance, with sample complexity of $\widetilde{O}(\epsilon^{-2})$. This result has two limitations: 1). The sample complexity depends on some unknown parameters of the discriminator network class (e.g., the Lipschitz constant of discriminators with respect to parameters); 2). A small neural net distance does not necessarily imply that two distributions are close [110, Corollary 3.2], which in turn can not answer *(Q1)* firmly. Our results are explicit in the network architectures, and provide a statistical convergence of GANs under the Wasserstein distance.

Some follow-up works attempted to address the first limitation in [110]. [139] explicitly quantified the Lipschitz constant and the covering number of the discriminator network. They improved the generalization bound in [110] with the technique in [143]. Whereas the bound has an exponential dependence on the depth of the discriminator. [140] proved a tighter generalization bound under spectral normalization applied to the discriminator, where the bound has a polynomial dependence on the size of the discriminator. These generalization theories rely on the assumption that the generator can approximate the data distribution well with respect to the neural net distance, nonetheless, the existence of such a generator is unknown.

[113] tackled the second limitation in [110], and studied the estimation error of GANs under the Wasserstein distance for a special class of distributions implemented by a generator, while the discriminator is designed to guarantee zero bias (or approximation error). Specifically, [113] showed that for certain generator classes, there exist corresponding discriminator classes with a strong discriminative power against the generator. Particular examples include two-layer ReLU network discriminators (half spaces) for distinguishing Gaussian distributions/mixture of Gaussians, and $(L+2)$-layer discriminators for $(L+1)$-layer invertible generators. In these examples, if the data distribution can be exactly implemented by some generator, then the neural net distance can provably approximate the Wasserstein distance. Consequently, GANs can generalize under the Wasserstein distance. As mentioned earlier, these results require an invertibility assumption on the generator.

Concurrent with [113], [114] studied the estimation error of GANs under the Sobolev IPMs. [114] considered both nonparametric and parametric settings. In the nonparametric setting, the generator and discriminator network architectures are not explicitly chosen, so the bias of the distribution estimation remains unknown. As a result, the bound cannot provide an explicit sample complexity for distribution estimation. Their parametric results are very similar to [113], which requires the same invertibility assumptions and the data distribution needs to be exactly implementable by the generator.

$\star$ *Generative distribution estimation under IPMs.* Recently, several works studied distribution estimation under certain discrepancy measures using generative models, when data exhibit low-dimensional structures [142, 115, 116]. The distribution estimation framework is

$$g^* \in \operatorname*{argmin}_{g \in \mathcal{G}} \texttt{discrepancy}(g_\sharp \rho, \mu)$$

and the corresponding statistical rate of estimation is free of the curse of data ambient dimensionality. Specifically, in [142], the generative models are assumed to be continuously differentiable up to order s. By simultaneously optimize the choice of latent distribution ρ and generative model g, they proved that the Sinkhorn divergence between the generated distribution and data distribution converges only depending on data intrinsic dimension. [115] consider data being generated by a ground truth pushfowrad mapping applied to latent samples from a low-dimensional unit cube. Using Lipschitz generator, they proved that the generalization bound in terms of Wassesrstein-1 distance converges only depending on the dimension of the latent space. More recently, [116] established a generalization bound in terms of Hölder IPMs for generative models and the bound converges depending on data intrinsic dimension. Nonetheless, how well the generator can represent the data distribution remains unclear. All of the aforementioned results rely on training the generative model by minimizing certain discrepancy metric, e.g., Wasserstein-1 distance and Sinkhorn divergence. There is no explicit discriminator network involved, nor there exists any straightforward method to parameterize the discrepancy metric by a discriminator network with performance guarantees. It is worth mentioning that [144] considered estimating low-dimensional singular distributions using deep generative models. They adopted a likelihood approach, which is different from GANs. In this regard, these existing works does not directly apply to GANs and cannot precisely evaluate the distribution estimation power of GANs.

$\star$ *Density estimation under IPMs.* There is also a line of works considering nonparametric density estimation under IPMs [145, 146]. [145] studied the minimax error under Sobolev IPMs. Later, [146] generalized the minimax result to Besov IPMs for estimating distributions with Besov densities. Yet our work is different from these works. Specifically, the distribution estimation framework in [145, 146] is

$$\min_{\nu \in \mathcal{P}} \max_{f_\omega \in \mathcal{F}_{\mathrm{NN}}} \int f_\omega(\mathbf{y})\nu(\mathbf{y})dy - \frac{1}{n}\sum_{i=1}^{n} f_\omega(\mathbf{x}_i), \tag{5.11}$$

where $\nu : \mathcal{X} \mapsto \mathbb{R}$ is a density function and $\mathcal{P}$ denotes a class of density estimators, such as the wavelet-thresholding estimator in [146]. Compared to our framework in (Eq. 5.2), we consider the push-forward structure in GANs, where the generator g_θ is a multidimensional mapping. In contrast, (Eq. 5.11) considers density estimators, where $\nu : \mathcal{X} \mapsto \mathbb{R}$ is some density function parameterized by a neural network – involving NO generator architecture which transforms the easy-to-sample distribution to the data distribution. Moreover, to evaluate the integral in (Eq. 5.11), one needs to exactly know the feature space $\mathcal{X}$, and efficiently sample from $\mathcal{X}$. Consequently, [145, 146] only apply to $\mathcal{X} = [0, 1]^d$. Besides, only estimating the density function requires extensive extra efforts to sample from it, e.g., using Monte Carlo simulation, due to the lack of the push-forward structure. However, our theories are applicable to push-forward GANs, and allow an efficient sampling of generated (fake) data.

5.6 Proof Outline

We provide proofs of Theorem 5.1 and Theorem 5.2. The developed analytical framework will also be adopted for proving Theorem 5.3 with additional treatments on low-dimensional structures in Section 5.6.3.

5.6.1 Proof of Distribution Approximation Theory

Theorem 5.1 is obtained by combining Lemma 5.1 and Theorem 3.1 (take Euclidean space itself as a manifold). Under Assumption 5.1 – 5.3, Lemma 5.1 ensures the existence of a $\mathcal{H}^{\alpha+1}(\mathcal{Z})$ data transformation T such that $T_\sharp\rho = \mu$. The remaining step is to choose a proper generator network for approximating T.

If the latent space $\mathcal{Z} \subset [0,1]^D$, we can directly apply Theorem 3.1 for constructing the generator. Otherwise, if $\mathcal{Z} \subset [-B,B]^D$, we define a linear scaling function $\phi(\mathbf{z}) = (\mathbf{z} + B\mathbf{1})/(2B) \in [0,1]^D$ for any $\mathbf{z} \in \mathcal{Z}$, where $\mathbf{1}$ denotes a vector of 1's. For the data transformation T, we rewrite it as $T \circ \phi^{-1}(\phi(\cdot))$ so that it suffices to approximate $T \circ \phi^{-1}$ supported on $[0,1]^D$. $T \circ \phi^{-1}$ retains the same Hölder smoothness as T, since ϕ is invertible and linear. To this end, without loss of generality, we focus on $\mathcal{Z} \subset [0,1]^D$.

Our generator network architecture is constructed in the following way. By denoting $T = [T_1, \ldots, T_D]^\top$ with $T_i : \mathcal{Z} \to \mathbb{R}$ for $i = 1, \ldots, D$, we approximate each coordinate mapping T_i using Theorem 3.1. For a given error $\epsilon \in (0,1)$, T_i can be approximated by a ReLU network with $O\left(\log \frac{1}{\epsilon}\right)$ layers and $O\big(\delta^{-\frac{D}{\alpha+1}} \log \frac{1}{\epsilon}\big)$ neurons and weight parameters. Thus, mapping T can be approximated by D such networks and we denote as g_θ. Further, the distribution approximation error is

$$\begin{aligned} W_1((g_\theta)_\sharp\rho, \mu) &= \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \mathbb{E}_{\mathbf{z}\sim\rho}[f(g_\theta(\mathbf{z}))] - \mathbb{E}_{\mathbf{x}\sim\mu}[f(\mathbf{x})] \\ &\leq \mathbb{E}_{\mathbf{z}\sim\rho} \left\|g_\theta(\mathbf{z}) - T(\mathbf{z})\right\|_2 \\ &\leq \sqrt{D}\epsilon. \end{aligned}$$

5.6.2 Proof of Statistical Estimation Theory

We prove an oracle inequality for establishing Theorem 5.2, which decomposes the distribution estimation error into the generator approximation error $\mathcal{E}_1$, the discriminator approximation error $\mathcal{E}_2$, and the statistical error $\mathcal{E}_3$.

Lemma 5.3. *Let $\mathcal{H}^\beta(\mathcal{X})$ be the Hölder function class defined on $\mathcal{X}$ with Hölder index $\beta \geq 1$. Define $\mathcal{H}^\beta_\infty(\mathcal{X}) = \{f \in \mathcal{H}^\beta(\mathcal{X}) : |f(\mathbf{x}) - f(\mathbf{y})| \leq \|\mathbf{x} - \mathbf{y}\|_\infty\}$. Then it holds*

$$d_{\mathcal{H}^\beta}((g^*_\theta)_\sharp \rho, \mu) \leq \mathcal{E}_1 + 4\mathcal{E}_2 + \mathcal{E}_3,$$

where $\mathcal{E}_1 = \inf_{g_\theta \in \mathcal{G}_{\mathrm{NN}}} d_{\mathcal{H}^\beta_\infty}((g_\theta)_\sharp \rho, \mu)$, $\mathcal{E}_2 = \sup_{f \in \mathcal{H}^\beta} \inf_{f_\omega \in \mathcal{F}_{\mathrm{NN}}} \|f - f_\omega\|_\infty$, and $\mathcal{E}_3 = d_{\mathcal{H}^\beta}(\mu, \widehat{\mu}_n) + d_{\mathcal{F}_{\mathrm{NN}}}(\mu, \widehat{\mu}_n)$.

The proof is provided in Appendix C.1.1. We next bound each error term separately. $\mathcal{E}_1$ and $\mathcal{E}_2$ can be controlled by proper choices of the generator and discriminator architectures. $\mathcal{E}_3$ can be controlled based on empirical process [96, 13].

• **Bounding Generator Approximation Error $\mathcal{E}_1$.** We answer this question: Given $\epsilon_1 \in (0, 1)$, how can we properly choose $\mathcal{G}_{\mathrm{NN}}$ to guarantee $\mathcal{E}_1 \leq \epsilon_1$? Later, we will pick ϵ_1 based on the sample size n, and Hölder indexes β and α.

Lemma 5.4. *Given $\epsilon_1 \in (0, 1)$, there exists a ReLU network architecture $\mathcal{G}_{NN}(R, \kappa, L, p, K)$ with parameters given by* (Eq. 5.6) *with $\epsilon = \epsilon_1$ such that, for any data distribution $(\mathcal{X}, \mu)$ and easy-to-sample distribution $(\mathcal{Z}, \rho)$ satisfying Assumptions 5.1 – 5.3, if the weight parameters of this network are properly chosen, then it yields a transformation g_θ satisfying $d_{\mathcal{H}^\beta_\infty}((g_\theta)_\sharp \rho, \mu) \leq \epsilon_1$.*

The proof is provided in Appendix C.1.2.

• **Bounding Discriminator Approximation Error $\mathcal{E}_2$.** Analogous to the generator, we pre-define an error $\epsilon_2 \in (0, 1)$, and determine the discriminator architecture.

The discriminator is expected to approximate any function $f \in \mathcal{H}^\beta(\mathcal{X})$. We have the following result.

Lemma 5.5. *Given any $\epsilon_2 \in (0, 1)$, there exists a ReLU network architecture $\mathcal{F}_{NN}(\bar{R}, \bar{\kappa}, \bar{L}, \bar{p}, \bar{J})$ with*

$$\bar{L} = O\big(\log(1/\epsilon_2)\big), \quad \bar{p} = O\big(\epsilon_2^{-D/\beta}\big), \quad \bar{J} = O\big(\epsilon_2^{-D/\beta} \log(1/\epsilon_2)\big),$$

$$\bar{R} = C, \quad \bar{\kappa} = C,$$

such that, for any discriminative function $f \in \mathcal{H}^\beta(\mathcal{X})$, if the weight parameters are properly chosen, this network architecture yields a function f_ω satisfying $\|f_\omega - f\|_\infty \leq \epsilon_2$.

The proof is provided in Appendix C.1.3.

• **Bounding Statistical Error** $\mathcal{E}_3$. The statistical error term is essentially the concentration of empirical data distribution $\widehat{\mu}_n$ to its population counterpart. Given a symmetric function class $\mathcal{F}$, we show $\mathbb{E}\left[d_{\mathcal{F}}(\widehat{\mu}_n, \mu)\right]$ scales with the complexity of the function class $\mathcal{F}$.

Lemma 5.6. *For a symmetric function class $\mathcal{F}$ with $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq M$ for a constant M, we have*

$$\mathbb{E}\left[d_{\mathcal{F}}(\widehat{\mu}_n, \mu)\right] \leq 2 \inf_{0<\delta<M} \Big(2\delta + \frac{12}{\sqrt{n}} \int_\delta^M \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)} d\epsilon\Big),$$

where $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ denotes the ϵ-covering number of $\mathcal{F}$ with respect to the L_∞ norm.

The proof is provided in Appendix C.1.4. Now we need to find the covering number of Hölder class and that of the discriminator network. Classical result shows that the δ-covering number of $\mathcal{H}^\beta$ satisfies $\log \mathcal{N}(\delta, \mathcal{H}^\beta, \|\cdot\|_\infty) \leq C(1/\delta)^{\frac{D}{\beta} \vee 2}$ [147].

On the other hand, Lemma 4.3 quantifies the covering number of $\mathcal{F}_{\text{NN}}$:

$$\mathcal{N}\left(\delta, \mathcal{F}_{\text{NN}}(\bar{R}, \bar{\kappa}, \bar{L}, \bar{p}, \bar{K}), \|\cdot\|_\infty\right) \leq \left(\frac{2\bar{L}^2(\bar{p}B+2)(\bar{\kappa}\bar{p})^{\bar{L}+1}}{\delta}\right)^{\bar{J}}.$$

Combining Lemma 5.6 and the covering numbers, the statistical error can be bounded by

$$\begin{aligned} &\mathbb{E}\left[d_{\mathcal{H}^\beta}(\widehat{\mu}_n, \mu) + d_{\mathcal{F}_{\text{NN}}}(\mu, \widehat{\mu}_n)\right] \\ &\leq 4 \inf_{\delta_1 \in (0, C)} \left(\delta_1 + \frac{6}{\sqrt{n}} \int_{\delta_1}^C \sqrt{\log \mathcal{N}(\epsilon, \mathcal{H}^\beta, \|\cdot\|_\infty)} d\epsilon\right) \end{aligned}$$

$$
\begin{aligned}
&+ 4 \inf_{\delta_2 \in (0, C)} \left(\delta_2 + \frac{6}{\sqrt{n}} \int_{\delta_2}^{C} \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}_{\mathrm{NN}}, \|\cdot\|_\infty)} d\epsilon \right) \\
&\overset{(i)}{\leq} 4 \inf_{\delta_1 \in (0, C)} \left(\delta_1 + \frac{6}{\sqrt{n}} \int_{\delta_1}^{C} \sqrt{c \left(\frac{1}{\epsilon} \right)^{\left(\frac{D}{\beta} \vee 2 \right)}} d\epsilon \right) \\
&+ 4 \inf_{\delta_2 \in (0, C)} \left(\delta_2 + \frac{6}{\sqrt{n}} \int_{\delta_2}^{C} \sqrt{\bar{J} \log \frac{\bar{L}(\bar{p}B + 2)(\bar{\kappa}\bar{p})^{\bar{L}}}{\epsilon}} d\epsilon \right).
\end{aligned}
$$

We find that the first infimum in step (i) is attained at $\delta_1 = n^{-\frac{\beta}{D}}$. It suffices to take $\delta_2 = \frac{1}{n}$ in the second infimum. By omitting constants and polynomial dependence on β, we derive

$$
\mathbb{E}\left[d_{\mathcal{H}^\beta}(\widehat{\mu}_n, \mu) + d_{\mathcal{F}_{\mathrm{NN}}}(\mu, \widehat{\mu}_n) \right] = \widetilde{O}\left(\frac{1}{n} + n^{-\frac{\beta}{D}} + \frac{1}{\sqrt{n}} \sqrt{\bar{J}\bar{L} \log\left(n\bar{L}\bar{p}\right)} \right).
$$

• **Balancing the Approximation Error and Statistical Error**. Combining the previous three ingredients, by invoking the oracle inequality (Lemma 5.3), we can establish

$$
\begin{aligned}
\mathbb{E}\left[d_{\mathcal{H}^\beta}((g_\theta^*)_\sharp \rho, \mu) \right] &= \widetilde{O}\left(\epsilon_1 + \epsilon_2 + \frac{1}{n} + n^{-\frac{\beta}{D}} + \sqrt{\frac{\bar{J}\bar{L} \log\left(n\bar{L}\bar{p}\right)}{n}} \right) \\
&= \widetilde{O}\left(\epsilon_1 + \epsilon_2 + \frac{1}{n} + n^{-\frac{\beta}{D}} + \sqrt{\frac{\epsilon_2^{-\frac{D}{\beta}} \log \frac{1}{\epsilon_2} \log\left(n\epsilon_2^{-\frac{D}{\beta}}\right)}{n}} \right).
\end{aligned}
$$

We choose $\epsilon_1 = n^{-\frac{\beta}{2\beta + D}}$, and ϵ_2 satisfying $\epsilon_2 = n^{-\frac{1}{2}} \epsilon_2^{-\frac{D}{2\beta}}$, i.e., $\epsilon_2 = n^{-\frac{\beta}{2\beta+D}}$. This yields (Eq. 5.7).

When given finite generated fake samples, we need an extra concentration argument. This is tackled by an alternative oracle inequality (Eq. 5.12) shown in below. The rest of the proof utilizes the same argument in Theorem 5.2.

Proof of Corollary 5.1. We show an alternative oracle inequality for finite generated samples as follows. Inequality (Eq. C.1) in the proof of Lemma 5.3 yields

$$d_{\mathcal{H}^\beta}((g_\theta^{*,m})_\sharp\rho,\mu) \leq d_{\mathcal{F}_{\text{NN}}}((g_\theta^{*,m})_\sharp\rho,\widehat{\mu}_n) + 2\sup_{f\in\mathcal{H}^\beta}\inf_{f_\omega\in\mathcal{F}_{\text{NN}}}\|f-f_\omega\|_\infty \\ + d_{\mathcal{H}^\beta}(\widehat{\mu}_n,\mu).$$

We further expand the first term on the right-hand side above as

$$d_{\mathcal{F}_{\text{NN}}}((g_\theta^{*,m})_\sharp\rho,\widehat{\mu}_n) \leq d_{\mathcal{F}_{\text{NN}}}((g_\theta^{*,m})_\sharp\rho,(g_\theta^{*,m})_\sharp\widehat{\rho}_m) + d_{\mathcal{F}_{\text{NN}}}((g_\theta^{*,m})_\sharp\widehat{\rho}_m,\widehat{\mu}_n).$$

By the optimality of $g_\theta^{*,m}$, for any $g_\theta \in \mathcal{G}_{\text{NN}}$, we have

$$\begin{aligned}
&d_{\mathcal{F}_{\text{NN}}}((g_\theta^{*,m})_\sharp\widehat{\rho}_m,\widehat{\mu}_n)\\
\leq\ & d_{\mathcal{F}_{\text{NN}}}((g_\theta)_\sharp\widehat{\rho}_m,\widehat{\mu}_n)\\
\leq\ & d_{\mathcal{F}_{\text{NN}}}((g_\theta)_\sharp\widehat{\rho}_m,(g_\theta)_\sharp\rho) + d_{\mathcal{F}_{\text{NN}}}((g_\theta)_\sharp\rho,\mu) + d_{\mathcal{F}_{\text{NN}}}(\mu,\widehat{\mu}_n)\\
\leq\ & d_{\mathcal{F}_{\text{NN}}}((g_\theta)_\sharp\rho,\mu) + \sup_{g_\theta\in\mathcal{G}_{\text{NN}}} d_{\mathcal{F}_{\text{NN}}}((g_\theta)_\sharp\widehat{\rho}_m,(g_\theta)_\sharp\rho) + d_{\mathcal{F}_{\text{NN}}}(\mu,\widehat{\mu}_n)\\
\leq\ & d_{\mathcal{H}^\beta_\infty}((g_\theta)_\sharp\rho,\mu) + 2\sup_{f\in\mathcal{H}^\beta}\inf_{f_\omega\in\mathcal{F}_{\text{NN}}}\|f-f_\omega\|_\infty + 2d_{\mathcal{F}_{\text{NN}}}(\mu,\widehat{\mu}_n)\\
&+ \sup_{g_\theta\in\mathcal{G}_{\text{NN}}} d_{\mathcal{F}_{\text{NN}}}((g_\theta)_\sharp\widehat{\rho}_m,(g_\theta)_\sharp\rho),
\end{aligned}$$

where the last inequality follows the same argument in the proof of Lemma 5.6. Combining all the inequalities together, we have

$$\begin{aligned}
&d_{\mathcal{H}^\beta}((g_\theta^{*,m})_\sharp\rho,\mu)\\
\leq\ & \inf_{g_\theta\in\mathcal{G}_{\text{NN}}} d_{\mathcal{H}^\beta_\infty}((g_\theta)_\sharp\rho,\mu) + 4\sup_{f\in\mathcal{H}^\beta}\inf_{f_\omega\in\mathcal{F}_{\text{NN}}}\|f-f_\omega\|_\infty + 2d_{\mathcal{F}_{\text{NN}}}(\mu,\widehat{\mu}_n)\\
&+ d_{\mathcal{H}^\beta}(\widehat{\mu}_n,\mu) + \sup_{g_\theta\in\mathcal{G}_{\text{NN}}} d_{\mathcal{F}_{\text{NN}}}((g_\theta)_\sharp\widehat{\rho}_m,(g_\theta)_\sharp\rho) + d_{\mathcal{F}_{\text{NN}}}((g_\theta^{*,m})_\sharp\rho,(g_\theta^{*,m})_\sharp\widehat{\rho}_m).
\end{aligned} \tag{5.12}$$

Given the proof of Theorem 5.2, we only need to bound the extra statistical error terms

$$\sup_{g_\theta \in \mathcal{G}_{\rm NN}} d_{\mathcal{F}_{\rm NN}}((g_\theta)_\sharp \widehat{\rho}_m, (g_\theta)_\sharp \rho) \quad \text{and} \quad d_{\mathcal{F}_{\rm NN}}((g_\theta^{*,m})_\sharp \rho, (g_\theta^{*,m})_\sharp \widehat{\rho}_m).$$

In fact, 5.6 and Lemma 4.3 together imply

$$\begin{aligned}\sup_{g_\theta \in \mathcal{G}_{\rm NN}} d_{\mathcal{F}_{\rm NN}}((g_\theta)_\sharp \widehat{\rho}_m, (g_\theta)_\sharp \rho) &= \widetilde{O}\left(\frac{1}{\sqrt{m}}\sqrt{\bar{J}\bar{L}\log\left(m\bar{L}\bar{p}\right) + JL\log\left(mLp\right)}\right), \\ d_{\mathcal{F}_{\rm NN}}((g_\theta^{*,m})_\sharp \rho, (g_\theta^{*,m})_\sharp \widehat{\rho}_m) &= \widetilde{O}\left(\frac{1}{\sqrt{m}}\sqrt{\bar{J}\bar{L}\log\left(m\bar{L}\bar{p}\right)}\right),\end{aligned}$$

where the first inequality is obtained by taking $\mathcal{F} = \mathcal{F}_{\rm NN} \circ \mathcal{G}_{\rm NN}$ in Lemma 5.6, and its covering number is upper bounded by the product of the covering numbers of $\mathcal{F}_{\rm NN}$ and $\mathcal{G}_{\rm NN}$. Putting together, the estimation error $d_{\mathcal{H}^\beta}((g_\theta^{*,m})_\sharp \rho, \mu)$ can be bounded analogously to Theorem 5.2 as

$$\begin{aligned}&\mathbb{E}\left[d_{\mathcal{H}^\beta}((g_\theta^{*,m})_\sharp \rho, \mu)\right] \\ &= \widetilde{O}\left(\epsilon_1 + \epsilon_2 + \frac{1}{n} + \frac{1}{m} + n^{-\frac{\beta}{D}} + \sqrt{\frac{\epsilon_2^{-\frac{D}{\beta}}}{n}} + \sqrt{\frac{\epsilon_1^{-\frac{D}{\alpha+1}} + \epsilon_2^{-\frac{D}{\beta}}}{m}}\right).\end{aligned}$$

It suffices to choose $\epsilon_2 = n^{-\frac{\beta}{2\beta+D}}$ and $\epsilon_1 = m^{-\frac{\alpha+1}{2(\alpha+1)+D}}$, which yields

$$\mathbb{E}\left[d_{\mathcal{H}^\beta}((g_\theta^{*,m})_\sharp \rho, \mu)\right] = \widetilde{O}\left(n^{-\frac{\beta}{2\beta+D}} + m^{-\frac{\alpha+1}{2(\alpha+1)+D}} + \sqrt{\frac{n^{\frac{D}{2\beta+D}}}{m}}\right).$$

In the case of $m \geq n$, we have $\sqrt{\frac{n^{\frac{D}{2\beta+D}}}{m}} \leq n^{-\frac{\beta}{2\beta+D}}$. The proof is complete. □

5.6.3 Proof of Statistical Theory in Low-dimensional Space

The proof idea follows that of Theorem 5.2, with extra attentions to the exploitation of low-dimensional structures in data. We first slightly modify the oracle inequality in Lemma 5.3

to decompose the distribution estimation error.

Lemma 5.7. *Let* $(U^*, g_\theta^*, V^*, f_\omega^*)$ *be the global optimizer of* (Eq. 5.2). *The following error decomposition holds,*

$$
\begin{aligned}
& W_1((U^* \circ g_\theta^*)_\sharp \rho, \mu) \\
\leq\ & \underbrace{\inf_{g_\theta : A \circ g_\theta \in \mathcal{G}_{\mathrm{NN}}^{\mathrm{ld}}} \left\| A \circ g_\theta - A \circ T^{\mathrm{ld}} \right\|_\infty}_{\text{generator approximation error}} + \underbrace{W_1(\widehat{\mu}_n, \mu) + d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}(\widehat{\mu}_n, \mu)}_{\text{statistical error}} \\
& + \underbrace{\sup_{f \in \mathrm{Lip}_1(\mathbb{R}^D)} \inf_{f_\omega \circ V^\top \in \mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \left\| f \circ U^* - f_\omega \circ V^\top U^* \right\|_\infty + \left\| f \circ A - f_\omega \circ V^\top A \right\|_\infty}_{\text{discriminator approximation error (HARD)}} \qquad (5.13) \\
& + \underbrace{2 \sup_{f \in \mathrm{Lip}_1(\mathbb{R}^D)} \inf_{f_\omega \circ V^\top \in \mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \left\| f \circ A - f_\omega \circ V^\top A \right\|_\infty}_{\text{discriminator approximation error (EASY)}} .
\end{aligned}
$$

The proof is provided in Appendix C.2.1. In the sequel, we bound error terms in (Eq. 5.13) respectively. The generator approximation error can be reduced to approximating T^{ld}. By some manipulation on the intrinsic structures of data distribution, we expect that the statistical error scales with the subspace dimension q. The main difficulty stems from bounding the discriminator approximation error. A quick comparison to Lemma 5.3 indicates that the (EASY) error term may be bounded similarly as in Theorem 5.2. In contrast, the (HARD) error term involves simultaneously approximating the discriminative function projected into the column space of U^* and A. In general, such an approximation error is hardly small unless U^*, A share approximately the same column space. Fortunately, this is indeed the case as shown in Lemma 5.9 so that the (HARD) error term can be controlled.

• **Bounding Generator Approximation Error**. Suppose that we require the generator approximation error to be bounded by $\epsilon_1 > 0$, i.e.,

$$
\inf_{g : U \circ g_\omega \in \mathcal{G}_{\mathrm{NN}}^{\mathrm{ld}}} \left\| U \circ g_\omega - A \circ T^{\mathrm{ld}} \right\|_\infty \leq \epsilon_1 .
$$

It suffices to choose a proper generator architecture $\mathcal{G}_{\mathrm{NN}}^{\mathrm{ld}}$ such that there exists $\widetilde{g}_\omega$ satisfying $\left\|\widetilde{g}_\omega - T^{\mathrm{ld}}\right\|_\infty \leq \epsilon_1/q$. To see this, we take $U = A$ and substitute $\widetilde{g}_\omega$ into the generator approximation error,

$$\begin{aligned}\left\|A \circ \widetilde{g}_\omega - A \circ T^{\mathrm{ld}}\right\|_\infty &= \left\|\sum_{j=1}^{q} A_{:,j}[\widetilde{g}_\omega - T^{\mathrm{ld}}]_j\right\|_\infty \\ &\leq \sum_{j=1}^{q} \left\|A_{:,j}\right\|_\infty \left\|\widetilde{g}_\omega - T^{\mathrm{ld}}\right\|_\infty \\ &\leq \epsilon_1,\end{aligned}$$

where the last inequality holds since A has orthonormal columns. We can apply Theorem 3.1 and Lemma 5.4 for choosing proper network configuration of $\mathcal{G}_{\mathrm{NN}}^{\mathrm{ld}}$ to ensure the existence of $\widetilde{g}_\omega$. We recall that T^{ld} is a $\mathcal{H}^{\alpha+1}$ continuous mapping in $\mathbb{R}^q$ by Lemma 5.1. Therefore, the resulting network architecture has the following configuration

$$\begin{gathered} R = B, \quad \kappa = O(1), \quad L = O\left(\log \frac{1}{\epsilon_1}\right), \\ p = O\left(q\epsilon_1^{-\frac{q}{\alpha+1}}\right), \quad K = O\left(Dq + \epsilon_1^{-\frac{q}{\alpha+1}} \log \frac{1}{\epsilon_1}\right). \end{gathered} \tag{5.14}$$

We will choose ϵ_1 later in the last step of the proof to balance all the error terms.

• **Bounding Discriminator Approximation Error**. We first consider the (EASY) error term. Suppose that we require the (EASY) discriminator approximation error to be bounded by $\epsilon_2 > 0$. We check that once $f : \mathbb{R}^D \mapsto \mathbb{R}$ is 1-Lipschitz and A has orthonormal columns, then $f \circ A : \mathbb{R}^q \mapsto \mathbb{R}$ is also 1-Lipschitz. To see this, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$, we have

$$|f(A\mathbf{x}) - f(A\mathbf{y})| \leq \|A\mathbf{x} - A\mathbf{y}\|_2 \leq \|A\|_2 \|\mathbf{x} - \mathbf{y}\|_2 = \|\mathbf{x} - \mathbf{y}\|_2 .$$

By taking $V = A$ in the (EASY) term, it suffices to ensure that f_ω can approximate any 1-Lipschitz function in a compact subset of $[0, 1]^q$. Due to the additional $\bar{\gamma}$-Lipschitz continuity constraint in (Eq. 5.10), we need a stronger universal approximation theory of the discriminator. The following lemma shows that ReLU neural networks can accurately approximating 1-Lipschitz functions in L^∞-norm, while the Lipschitz continuity of the network remains independent of the approximation error.

Lemma 5.8. *For any $\epsilon_2 \in (0,1)$, there exists a ReLU network architecture $\mathcal{F}(\bar{R}, \bar{\kappa}, \bar{L}, \bar{p}, \bar{J})$, such that for any target 1-Lipschitz function f defined on $[0,1]^q$ with $f(0) = 0$, the architecture yields an approximation $\widehat{f}$ satisfying $\|f - \widehat{f}\|_\infty \le \epsilon_2$. Moreover, the Lipschitz continuity of $\widehat{f}$ is bounded by*

$$\left|\widehat{f}(\mathbf{x}) - \widehat{f}(\mathbf{y})\right| \le 10q \left\|\mathbf{x} - \mathbf{y}\right\|_\infty \quad \textit{for any} \quad \mathbf{x}, \mathbf{y} \in [0,1]^q.$$

The configuration of $\mathcal{F}$ is

$$\bar{R} = \sqrt{q}, \quad \bar{\kappa} = O(1), \quad \bar{L} = O\left(\log 1/\epsilon_2 + q\right),$$
$$\bar{p} = O\left(\epsilon_2^{-q}\right), \quad \bar{J} = O\left(\epsilon_2^{-q}(\log 1/\epsilon_2 + q)\right).$$

The proof is defered to Appendix C.2.2. Lemma 5.8 improves the approximation guarantee in Theorem 3.1 with the additional Lipschitz continuity characterization, while the newtork size shares the same order of magnitude when specializing Theorem 3.1 to $d = q$ and $\beta = 1$. We take $\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}(\bar{R}, \bar{\kappa}, \bar{L}, \bar{p}, \bar{J}, \bar{\gamma})$ with $\bar{\gamma} = 10q$ and all the other parameters the same as in Lemma 5.8. Since the (EASY) error term is invariant with respect to translations on f, we can always assume $f(\mathbf{0}) = 0$ without loss of generality. It then holds

$$\text{(EASY) Error Term} \le 2\epsilon_2.$$

We next bound the (HARD) term. Recall that we need the column spaces of U^* and A to be approximately identical for controlling this error. Thanks to the choice of both the generator and discriminator class, we can show that the column spans of U^* and A match up to some error.

Lemma 5.9. *Given $\epsilon_1, \epsilon_2 \in (0, 1)$. Suppose Assumption 5.5 and 5.4 hold. Let the generator $\mathcal{G}_{\mathrm{NN}}^{\mathrm{ld}}$ be chosen as* (Eq. 5.14) *and discriminator $\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}$ be chosen as in Lemma 5.8 with $\bar{\gamma} = 10q$. For the global optimizer (U^*, g_θ^*), it holds*

$$\begin{aligned}
&\|U^* - A\|_{\mathrm{F}}^2 \\
&\quad \le 4q \left(1 + 4\sqrt{q}\left(\min_i \mathbb{E}_{\mathbf{z}\sim\rho}\left[T_i^{\mathrm{ld}}(\mathbf{z})\right]\right)^{-1} \mathbb{E}_{\mathbf{z}\sim\rho}\left[\|g_\theta^*(\mathbf{z})\|_2\right]\right)^2 \\
&\quad\quad \cdot \left(\min_i \mathbb{E}_{\mathbf{z}\sim\rho}\left[T_i^{\mathrm{ld}}(\mathbf{z})\right]\right)^{-2} \epsilon^2,
\end{aligned}$$

where $\epsilon = 10q\epsilon_1 + 3\epsilon_2$.

The full proof is deferred to Appendix C.2.3. We remark that $\mathbb{E}_{\mathbf{z}\sim\rho}[T_i^{\mathrm{ld}}(\mathbf{z})]$ is always lower bounded by a positive constant τ for any $i = 1, \ldots, q$, since its density is positive on the support by Assumption 5.5. To establish Lemma 5.9, we leverage the optimality of U^*, g_θ^* and the corresponding discriminator network. We show by contraction that if the column spaces of U^* and A do not match closely, there exists a discriminator network capable of distinguishing the generated distribution and data distribution.

Given Lemma 5.9, we are ready to derive an upper bound for the (HARD) discriminator approximation error term.

$$\begin{aligned}
&\sup_{f\in\mathrm{Lip}_1(\mathbb{R}^D)} \inf_{f_\omega\circ V^\top\in\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \left\|f\circ U^* - f_\omega\circ V^\top U^*\right\|_\infty + \left\|f\circ A - f_\omega\circ V^\top A\right\|_\infty \\
\overset{(i)}{\le}\; &\sup_{f\in\mathrm{Lip}_1(\mathbb{R}^D)} \inf_{f_\omega\circ A^\top\in\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \left\|f\circ U^* - f_\omega\circ A^\top U^*\right\|_\infty + \left\|f\circ A - f_\omega\right\|_\infty \\
\overset{(ii)}{\le}\; &\sup_{f\in\mathrm{Lip}_1(\mathbb{R}^D)} \inf_{f_\omega\circ A^\top\in\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \left\|f\circ U^* - f\circ A\right\|_\infty + \left\|f_\omega - f_\omega\circ A^\top U^*)\right\|_\infty \\
&\qquad\qquad + 2\left\|f\circ A - f_\omega\right\|_\infty,
\end{aligned} \tag{5.15}$$

where (i) is obtained by taking $V = A$, and inequality (ii) is obtained by the triangle inequality

$$\begin{aligned}\left\|f \circ U^* - f_\omega \circ A^\top U^*\right\|_\infty \leq \|f \circ U^* - f \circ A\|_\infty + \|f \circ A - f_\omega\|_\infty \\ + \left\|f_\omega - f_\omega \circ A^\top U^*\right\|_\infty .\end{aligned}$$

The first term on the right-hand side of (Eq. 5.15) can be bounded using the Lipschitz continuity of f, i.e.,

$$\begin{aligned}&\|f \circ U^* - f \circ A\|_\infty \\ &\leq \sup_{\mathbf{x} \in [0,1]^q} \|U^* - A^*\|_2 \|\mathbf{x}\|_2 \\ &\leq 2q\big(1 + 4\sqrt{q}\left(\min_i \mathbb{E}_{\mathbf{z}\sim\rho}\left[T_i^{\mathrm{ld}}(\mathbf{z})\right]\right)^{-1} \mathbb{E}_{\mathbf{z}\sim\rho}\left[\|g_\theta^*(\mathbf{z})\|_2\right]\big) \max_i \mathbb{E}_{\mathbf{z}\sim\rho}^{-1}\left[T_i^{\mathrm{ld}}(\mathbf{z})\right] \epsilon.\end{aligned}$$

A similar argument applies to

$$\begin{aligned}&\left\|f_\omega - f_\omega \circ A^\top U^*\right\|_\infty \\ &\leq \sup_{\mathbf{x} \in [0,1]^q} 10q \left\|I - A^\top U^*\right\|_2 \|\mathbf{x}\|_2 \\ &\leq \sup_{\mathbf{x} \in [0,1]^q} 10q^{3/2} \|A - U^*\|_2 \\ &\leq 20q^2\big(1 + 4\sqrt{q}\left(\min_i \mathbb{E}_{\mathbf{z}\sim\rho}\left[T_i^{\mathrm{ld}}(\mathbf{z})\right]\right)^{-1} \mathbb{E}_{\mathbf{z}\sim\rho}\left[\|g_\theta^*(\mathbf{z})\|_2\right]\big) \max_i \mathbb{E}_{\mathbf{z}\sim\rho}^{-1}\left[T_i^{\mathrm{ld}}(\mathbf{z})\right] \epsilon.\end{aligned}$$

The last term in the right-hand side of (Eq. 5.15) is the discriminator approximation error, which is bounded by ϵ_2. As a result, the (HARD) error term is upper bounded by

$$\begin{aligned}&\text{(HARD) Error Term} \\ &\leq 2q\big(1 + 4\sqrt{q}\left(\min_i \mathbb{E}_{\mathbf{z}\sim\rho}\left[T_i^{\mathrm{ld}}(\mathbf{z})\right]\right)^{-1} \mathbb{E}_{\mathbf{z}\sim\rho}\left[\|g_\theta^*(\mathbf{z})\|_2\right]\big) \max_i \mathbb{E}_{\mathbf{z}\sim\rho}^{-1}\left[T_i^{\mathrm{ld}}(\mathbf{z})\right] \epsilon \\ &+ 20q^2\big(1 + 4\sqrt{q}\left(\min_i \mathbb{E}_{\mathbf{z}\sim\rho}\left[T_i^{\mathrm{ld}}(\mathbf{z})\right]\right)^{-1} \mathbb{E}_{\mathbf{z}\sim\rho}\left[\|g_\theta^*(\mathbf{z})\|_2\right]\big) \max_i \mathbb{E}_{\mathbf{z}\sim\rho}^{-1}\left[T_i^{\mathrm{ld}}(\mathbf{z})\right] \epsilon \\ &+ 2\epsilon_2\end{aligned}$$

$$= O(\epsilon_2 + q^3\epsilon),$$

where the last step is obtained by $\|g_\theta^*(\mathbf{z})\|_2 \leq \sqrt{q}$ due to $g_\theta^*(\mathbf{z}) \in [0,1]^q$.

• **Bounding Statistical Error**. Similar to the statistical error in Lemma 5.3, we can bound it via finite-sample concentration. Yet we can pursue a faster convergence rate here by rewriting the data distribution as a pushforward of a low-dimensional distribution.

Lemma 5.10. *Suppose Assumption 5.4 and 5.5 hold. Statistical error terms in Lemma 5.7 are bounded by*

$$W_1(\widehat{\mu}_n, \mu) = O\left(n^{-1/q}\log n\right),$$
$$d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}(\widehat{\mu}_n, \mu) = O\left(\frac{1}{n} + \frac{1}{\sqrt{n}}\sqrt{\bar{J}\bar{L}\log(\bar{L}\bar{p}\bar{\kappa}n)}\right).$$

From Lemma 5.10, we observe that the statistical error $W_1(\widehat{\mu}_n, \mu)$ only depends on dimension q. To make sense the result, we rewrite the data distribution $\mu = A_\sharp(A_\sharp^\top \mu)$. In this way, we can translate the concentration of $\widehat{\mu}_n$ to μ in $\mathbb{R}^D$ into a counterpart in $\mathbb{R}^q$. Recall that $A_\sharp^\top \mu$ is a distribution with a $\mathcal{H}^\alpha(\mathbb{R}^q)$ density by Assumption 5.5. Threfore, we can apply Lemma 5.6 to complete the proof. See detailed arguments in Appendix C.2.4.

• **Balancing the Approximation Error and Statistical Error**. We collect all the error terms in the oracle inequality of Lemma 5.7 and choose optimal scalings on ϵ_1 and ϵ_2. We list all the error upper bounds in the following for a quick reference.

1. Generator approximation error $O(\epsilon_1)$.
2. Statistical error $O\left(n^{-1/q}\log n + \frac{1}{\sqrt{n}}\sqrt{\bar{J}\bar{L}\log(\bar{L}\bar{p}\bar{\kappa}n)}\right)$.
3. (EASY) discriminator approximation error $O(\epsilon_2)$.
4. (HARD) discriminator approximation error $O(\epsilon_2 + q^3\epsilon)$.

Summing up four error bounds above yields

$$W_1((U^* \circ g_\theta^*)_\sharp \rho, \mu) = O\left(\epsilon_1 + \epsilon_2 + n^{-1/q} \log n + q^3 \epsilon + \frac{1}{\sqrt{n}} \sqrt{\bar{J}\bar{L} \log(\bar{L}\bar{p}\bar{\kappa}n)}\right).$$

Substituting the configuration of $\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}$ in Lemma 5.8 into the last display, we set $\epsilon_1 = \epsilon_2 = n^{-\frac{1}{2+q}}$. By collecting terms, we derive

$$W_1\left((U^* \circ g_\theta^*)_\sharp \rho, \mu\right) = \widetilde{O}\left(n^{-\frac{1}{2+q}} \log^2 n\right).$$

The corresponding configurations of generator $\mathcal{G}_{\mathrm{NN}}^{\mathrm{ld}}$ and discriminator $\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}$ is obtained by substituting ϵ_1 and ϵ_2 in (Eq. 5.14) and Lemma 5.8, respectively.

5.7 Conclusion and Discussion

We establish statistical convergence of distribution estimation using GANs. Specifically, with proper generator and discriminator network architecture, we show GANs are consistent estimator of data distribution in terms of the Wasserstein distance. Moreover, when data have intrinsic low-dimensional linear structures, we show GANs can capture the unknown linear structure and enjoy a faster statistical rate of estimation, which is free of the curse of dimensionality. Compared to existing works, our theory exploits the pushforward structure of GANs and network architectures are explicitly given without invertibility constraints. In the sequel, we discuss several related topics and future directions.

Convolutional filters and residual connections Convolutional filters [1] are widely used in GANs for image generating and processing. Empirical results show that convolutional filters can learn hidden representations aligned with various patterns in images [148, 149], e.g., textures and skeletons. An interesting question is to understand how convolutional filters capture the aforementioned low-dimensional structures in data sets.

Smoothness of data distributions and regularized distribution Estimation Theorem 5.2 indicates a convergence rate independent of the smoothness of the data distribution. The reason behind is that the empirical data distribution $\widehat{\mu}_n$ cannot inherit the same smoothness as the underlying data distribution. This limitation exists in all previous works [141, 145, 146]. It is interesting to investigate whether GANs can achieve a faster convergence rate (e.g., attain the minimax optimal rate).

From a theoretical perspective, [114] suggested first obtaining a smooth kernel estimator from $\widetilde{\mu}_n$, and then replacing $\widehat{\mu}_n$ by $\widetilde{\mu}_n$ to train GANs. In practice, kernel smoothing is hardly used in GANs. Instead, regularization (e.g., entropy regularization) and normalization (e.g., spectral normalization and batch-normalization) are widely applied as implicit regularizers to promote the smoothness of the learned distribution. Several empirical studies of GANs suggest that divergence-based and mutual information-based regularization can stabilize the training and improve the performance [150, 151] of GANs. We leave the studies on statistical properties of regularized GANs for future investigation.

Computational concerns Our statistical results hold for global optimizer of (Eq. 5.2), whereas solving (Eq. 5.2) is often difficult. In practice, it is observed that larger neural networks are easier to train and yield better statistical performance [72, 152, 153, 154, 155, 156, 157, 158, 159]. This is referred to as overparameterization. Establishing a connection between computation and statistical properties of GANs is an important direction.

CHAPTER 6
OFFLINE DOUBLY-ROBUST POLICY LEARNING USING NEURAL NETWORKS

6.1 Personalized Offline Policy Learning

Causal inference studies the causal connection between actions and rewards, which has wide applications in healthcare [160, 161], digital advertising [162], product recommendation [163], and policy formulation [164]. For example in healthcare, each patient can be characterized by a set of covariates (also called features), and the actions are a set of treatments. Each patient has the corresponding reactions, or rewards, to different treatments. Causal inference enables one to personalize the treatment to each patient to maximize the total rewards. Such a personalized decision-making rule is referred to as a policy, which is a map from the covariate set to the action set. In off-policy learning, a batch of observational data is given, which typically consists of a covariate, the action taken (e.g. medical treatments and recommendations), and the observed reward. In this chapter, we are interested in learning an optimal policy that targets personalized treatments or services to different individuals based on the logged data. This is also known as the optimal treatment assignment in literature [165, 166].

Conventional causal inference methods often rely on parametric models [161, 167, 168, 169], which can introduce a large bias when the real model is not in the assumed parametric form. Many nonparametric methods are proposed [170, 169, 171, 172, 173, 174, 175, 176, 177, 178, 179], while the statistical theories often suffer from the curse of dimensionality. Recently neural networks became a popular modeling tool for causal inference. Many results have shown that neural networks outperform conventional nonparametric approaches, especially when the learning task involves high-dimensional complex data. For example,

[180] proposed to discover causal and anticausal features in images from ImageNet using a 20-layer residual network. [181] used recurrent neural networks to study the causality between group forming loans and the funding time on an online non-profit financial platform. Other examples can be found in diverse areas, including climate analysis, medical diagnosis, cognitive science, and online recommendations [182, 183, 184, 185, 186].

Despite the great progress of causal inference, there is still a huge gap between theory and practice. In casual inference, many existing theories on nonparametric or neural networks approaches are asymptotic, and suffer from the curse of dimensionality. Specifically, to achieve an ϵ accuracy, the sample complexity needs to grow in the order of ϵ^{-D}, where D is the covariate dimension. Such theories can not explain the empirical success when D is large. For example, in [180], the RGB images in ImageNet are of resolution $3 \times 224 \times 224$. To obtain a 0.1 error, the sample complexity needs to scale like $10^{-3\times224\times224}$, which well exceeds the training size of $99,309$. Besides, the curse of dimensionality is inevitable unless additional data structures are considered. [187] proved that, for binary policy learning problems, the sample complexity obtained by the optimal algorithm still grows exponentially in the covariate dimension D in the order of ϵ^{-D}.

This chapter establishes statistical guarantees of policy learning in causal inference using neural networks. We consider the doubly robust method (see Section 6.2 for details), and use deep ReLU neural networks to parameterize the policy class, the propensity score, and the conditional expected reward. We summarize our contributions as follows.

1. We establish nonasymptotic regret bounds for doubly robust policy learning in the finite-action scenario (Theorem 6.1 and Theorem 6.2). Our nonasymptotic theory optimally balances bias and variance under general regularity conditions (see a detailed discussion after Theorem 6.1). We highlight that we develop novel policy approximation theory using neural networks, which can efficiently approximate deterministic policies that are not continuous with respect to its input covariate (Theorem 6.2).

2. We leverage low-dimensional intrinsic structures in covariates. The obtained nonasymp-

totic regret bound of learned policy converges at a fast rate dependent on covariate intrinsic dimension, instead of the covariate ambient dimension D (Theorem 6.1 – Theorem 6.3). This partially explains the success of deep learning-based causal inference in high-dimensional applications.

3. We consider a discretization method and propose new analysis for policy learning in the continuous-action setting. A nonasymptotic regret bound is established (Theorem 6.3), where we carefully balance the discretization error in addition to the bias and variance in the finite-action scenario.

6.1.1 Related Work

In off-policy learning, one line of research learns the optimal policy by evaluating the expected reward of candidate policies and then finding the policy with the largest expected reward. The procedure of evaluating a target policy from the given data is called off-policy evaluation, which has been intensively studied in literature. The simplest way to evaluate a policy is the direct method which estimates the empirical reward of the target policy from collected data [188]. The direct method is unbiased if one specifies the reward model correctly. However, model specification is a difficult task in practice. Another method is the inverse propensity weighting [189, 168], which uses the importance weighting to correct the mismatch between the propensity scores of the target policy and the data collection policy. This method is unbiased if the data collection policy can be exactly estimated, yet it has a large variance especially when some actions are rarely observed. A more robust method is the doubly robust method [190, 167, 191], which integrates the direct method and the inverse propensity weighting. This method is unbiased if the reward model is correctly specified or the data collection policy is known.

The aforementioned methods have been used in [169, 171, 192, 193] for off-policy learning. [169] used the inverse propensity weighting, and [192] and [171] used the doubly robust method to learn the optimal policy with binary actions. In [193], an algorithm based on decision trees was proposed to learn the optimal policy with multiple actions using the doubly robust method. [194] proposed a balanced method which minimizes the worst-case conditional mean squared error to evaluate and learn the optimal policy with multiple actions.

Another line of research learns the optimal policy without evaluating policies. In [195, 196], the authors transformed the policy learning task with binary actions into a classification problem. Other works on off-policy learning include [197, 198], and [199].

Most of the aforementioned works provide asymptotic regret bounds with finite actions, which are valid when the number of samples goes to infinity. A nonasymptotic bound was derived in [169], but this work requires that the propensity score is known and the algorithm only works for policy learning with binary actions. Meanwhile, off-policy learning with continuous actions has not been addressed until recently [200, 201, 202, 203]. [201] developed a semi-parametric off-policy learning algorithm, which relies on the special form of the reward function. Then a regret bound was derived, while explicit dependency of the bound on the number of samples is not given. Although their algorithm can be applied when the reward model is misspecified, in this case the regret between the learned policy and the optimal one (not restricted to the semi-parametric model) is unclear. [200] applied a kernel method to extend the inverse propensity weighting and the doubly robust method to the continuous-action setting. Incorporated with the kernel method, the framework proposed in [204] can be used to learn policies with continuous actions. However, such an extension is only discussed in [204], without explicit statistical guarantees provided. [203] proposed a kernel based nonparametric double debiased machine learning estimator for causal effects with continuous actions. These works on continuous actions did not provide a nonasymptotic regret bound with an explicit dependency on the number of samples.

The rest of the chapter is organized as follows: Section 6.2 presents the doubly robust estimation framework; Section 6.3 states our regret bounds of the learned policy; Section 6.4 concludes the chapter and discusses related topics.

6.2 Doubly-robust Learning Framework

We introduce a two-stage policy learning scheme using neural networks. Suppose we receive n i.i.d. triples $\{(\mathbf{x}_i, \mathbf{a}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathcal{M}$ denotes a covariate independently sampled from an unknown distribution on $\mathcal{M}$, $\mathbf{a}_i \in \mathcal{A}$ denotes the action taken, and $y_i \in \mathbb{R}$ is the observed reward. To incorporate the low-dimensional geometric structures of the covariates, we assume $\mathcal{M}$ is a d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$. The action space $\mathcal{A}$ can be either finite or continuous. For each covariate and action pair $(\mathbf{x}, \mathbf{a})$, there is an associated random reward. We adopt the unconfoundedness assumption to simplify the model, which is commonly used in existing literature on causal inference [205, 193].

Assumption 6.1 (Unconfoundedness)**.** *The reward is independent of* $\mathbf{a}$ *conditioned on* $\mathbf{x}$.

To interpret Assumption 6.1, we first consider a finite action space $\mathcal{A} = \{A_1, \dots, A_{|\mathcal{A}|}\}$, where A_j is a one-hot vector, i.e. $A_j = [0, \dots, 0, 1, 0, \dots, 0]^\top$ with 1 appearing at the j-th position. Given the covariate $\mathbf{x}$, there is a reward $\{Y(\mathbf{x}, A_1), \dots, Y(\mathbf{x}, A_{|\mathcal{A}|})\}$ for each action, where the randomness of $Y(\mathbf{x}, A_j)$ depends on $\mathbf{x}$ and the noise in the reward. The observed reward y_i is a realization of $Y(\mathbf{x}_i, A_j)$ with $\mathbf{a}_i = A_j$.

6.2.1 Policy Learning with Discrete Action

When the action space is finite, a policy $\pi : \mathcal{M} \to \Delta^{|\mathcal{A}|}$ maps a covariate on $\mathcal{M}$ to a vector on the $|\mathcal{A}|$-dimensional simplex

$$\Delta^{|\mathcal{A}|} = \left\{ \mathbf{z} \in \mathbb{R}^{|\mathcal{A}|} : z_i \geq 0 \text{ and } \sum_i z_i = 1 \right\}.$$

The j-th entry of $\pi(\mathbf{x})$ denotes the probability of choosing the action A_j given $\mathbf{x}$. A policy in the interior of the simplex is called a randomized policy. If $\pi(\mathbf{x})$ is a one-hot vector, it is called a deterministic policy. The expected reward of deploying a policy π is

$$Q(\pi) = \mathbb{E}[Y(\pi(\mathbf{x}))] = \mathbb{E}\left[\left\langle [Y(\mathbf{x}, A_1), \ldots, Y(\mathbf{x}, A_{|\mathcal{A}|})]^\top, \pi(\mathbf{x}) \right\rangle\right]. \tag{6.1}$$

We investigate the doubly robust approach [190, 168, 191] for policy learning, which consists of two stages. After receiving the training data, we split them into two groups

$$\mathcal{S}_1 = \{(\mathbf{x}_i, \mathbf{a}_i, y_i)\}_{i=1}^{n_1} \quad \text{and} \quad \mathcal{S}_2 = \{(\mathbf{x}_i, \mathbf{a}_i, y_i)\}_{i=n_1+1}^{n}. \tag{6.2}$$

We denote $n_2 = n - n_1$ and choose n_1, n_2 to be proportional to n such that n_1/n is a constant. In the first stage, we solve nonparametric regression problems using $\mathcal{S}_1$ to estimate two important functions — the propensity score and the conditional expected reward. For any action A_j, the propensity score

$$e_{A_j}(\mathbf{x}) = \mathbb{P}(\mathbf{a} = A_j \mid \mathbf{x})$$

quantifies the probability of choosing A_j given the covariate $\mathbf{x}$, and the expected reward of choosing A_j is

$$\mu_{A_j}(\mathbf{x}) = \mathbb{E}[Y(\mathbf{x}, A_j) \mid \mathbf{x}].$$

Substituting the definition above into (Eq. 6.1), we can write

$$Q(\pi) = \mathbb{E}\left\langle [\mu_{A_1}(\mathbf{x}), ..., \mu_{A_{|\mathcal{A}|}}(\mathbf{x})]^\top, \pi(\mathbf{x}) \right\rangle = \int_{\mathcal{M}} \left\langle [\mu_{A_1}(\mathbf{x}), ..., \mu_{A_{|\mathcal{A}|}}(\mathbf{x})]^\top, \pi(\mathbf{x}) \right\rangle d\mathbf{x}. \tag{6.3}$$

In the second stage, we learn a policy using $\mathcal{S}_2$ based on our estimated e_{A_j}'s and μ_{A_j}'s, which only requires that either the e_{A_j}'s or the μ_{A_j}'s are accurately estimated.

- **Stage 1: Estimating μ_{A_j} and e_{A_j}.** For each action A_j, we use a neural network to

Algorithm 1 A two-stage algorithm for doubly-robust off-policy learning with discrete actions.

```
**Input:** Collected data $\{(\mathbf{x}_i, \mathbf{a}_i, y_i)\}_{i=1}^n$. Network architectures $\mathcal{F}_{\rm NN}, \mathcal{G}_{\rm NN}$ and $\Pi_{\rm NN}$.
**Stage 1: Estimating $\mu_{A_j}$ and $e_{A_j}$.**
For each action $A_j$,
    • For each action $A_j$, estimate the reward function $\mu_{A_j}$ by minimizing (Eq. 6.4),
    • Estimate $e_{A_j}$'s by solving (Eq. 6.5) – (Eq. 6.6).
**Stage 2: Policy Learning.**
    • Learn the optimal policy by solving (Eq. 6.7) – (Eq. 6.8).
**Output:** Learned policy $\widehat{\pi}_{\rm DR}$.
```

estimate the reward function μ_{A_j} by minimizing the following empirical quadratic loss

$$\widehat{\mu}_{A_j}(\mathbf{x}) = \mathop{\mathrm{argmin}}_{f\in\mathcal{F}_{\rm NN}} \frac{1}{n_{A_j}} \sum_{i=1}^{n_1} (y_i - f(\mathbf{x}_i))^2 \mathbb{1}\{\mathbf{a}_i = A_j\} \quad \text{with} \quad n_{A_j} = \sum_{i=1}^{n_1} \mathbb{1}\{\mathbf{a}_i = A_j\}, \tag{6.4}$$

where $\mathcal{F}_{\rm NN} : \mathcal{M} \to \mathbb{R}$ is a properly chosen network class defined in Lemma 6.1.

An estimator of the propensity score e_{A_j} is obtained by minimizing the multinomial logistic loss. Let $\mathcal{G}_{\rm NN} : \mathcal{M} \mapsto \mathbb{R}^{|\mathcal{A}|-1}$ be a properly chosen network class defined in Lemma 6.1. We obtain $\widehat{e}_{A_j}(\mathbf{x})$ via

$$\widehat{g}(\mathbf{x}) = \mathop{\mathrm{argmin}}_{g\in\mathcal{G}_{\rm NN}} \frac{1}{n_1} \sum_{i=1}^{n_1} -[g(\mathbf{x}_i)^\top, 1]\mathbf{a}_i + \log\Big(1 + \sum_{j=1}^{|\mathcal{A}|-1} \exp([g(\mathbf{x}_i)]_j)\Big), \tag{6.5}$$

$$\widehat{e}_{A_j}(\mathbf{x}) = \frac{\exp([\widehat{g}(\mathbf{x})]_j)}{1 + \sum_{j=1}^{|\mathcal{A}|-1} \exp([\widehat{g}(\mathbf{x})]_j)} \text{ for } j \leq |\mathcal{A}| - 1, \text{ and } \widehat{e}_{A_{|\mathcal{A}|}}(\mathbf{x}) = \frac{1}{1 + \sum_{j=1}^{|\mathcal{A}|-1} \exp([\widehat{g}(\mathbf{x})]_j)}. \tag{6.6}$$

Here $[g]_j$ denotes the j-th entry, and $[g^\top, 1] \in \mathbb{R}^{|\mathcal{A}|}$ is obtained by augmenting g by 1.

• **Stage 2: Policy Learning**. Given $\widehat{\mu}_{A_j}$ and $\widehat{e}_{A_j}$, we learn an optimal policy by maximizing a doubly robust empirical reward:

$$\widehat{Q}(\pi) := \frac{1}{n_2} \sum_{i=n_1+1}^{n} \pi(\mathbf{x}_i)^\top \widehat{\Gamma}_i$$
$$\text{with} \quad \widehat{\Gamma}_i = \frac{y_i - \widehat{\mu}_{\mathbf{a}_i}(\mathbf{x}_i)}{\widehat{e}_{\mathbf{a}_i}(\mathbf{x}_i)} \cdot \mathbf{a}_i + [\widehat{\mu}_{A_1}(\mathbf{x}_i), \ldots, \widehat{\mu}_{A_{|\mathcal{A}|}}(\mathbf{x}_i)]^\top \in \mathbb{R}^{|\mathcal{A}|}. \tag{6.7}$$

A doubly robust optimal policy is learned by

$$\widehat{\pi}_{\mathrm{DR}} = \underset{\pi \in \Pi_{\mathrm{NN}}}{\operatorname{argmax}} \ \widehat{Q}(\pi), \tag{6.8}$$

where Π_{NN} is a properly chosen network class (see Section 6.3 for the configurations of Π_{NN}, e.g., (Eq. 6.19) and (Eq. 6.20)). The doubly robust reward $\widehat{Q}$ can tolerate a relatively large estimation error in either $\widehat{\mu}_{A_j}$ or $\widehat{e}_{A_j}$ (see the discussion after Theorem 6.1). Our algorithm is summarized in Algorithm 1.

In the proposed two-stage method, one only needs to estimate the reward functions and propensity scores, and then learn the optimal policy. Although the covariates are defined on a low-dimensional manifold, we do not need to explicitly learn the manifold. Instead, neural networks are adaptive to the low-dimensional structure and the manifold is learned implicitly during the estimation process.

6.2.2 Policy Learning with Continuous Action

Continuous actions, e.g. doses of drugs, often arise in applications, but there are limited studies on policy learning with continuous actions. In this chapter, we consider the continuous action space $\mathcal{A} = [0, 1]$ and use $a \in \mathcal{A}$ to denote an action. When the random action a takes the value $A \in [0, 1]$, we denote $Y(\mathbf{x}, A)$ as its random reward. The propensity score and conditional expected reward are defined analogously to the finite action case:

$$e(\mathbf{x}, A) = \frac{d}{dA}\mathbb{P}(a \leq A, A \in \mathcal{A} \mid \mathbf{x}) \quad \text{and} \quad \mu(\mathbf{x}, A) = \mathbb{E}[Y(\mathbf{x}, A) \mid \mathbf{x}].$$

Note that $e(\mathbf{x}, A)$ is a probability density function.

In this scenario, we can learn an optimal policy by replicating the two-stage scheme with a discretization technique on the continuous action space. Specifically, we uniformly partition the action space $\mathcal{A}$ into V sub-intervals and denote $I_j = [(j-1)/V, j/V]$ for $j = 1, \dots V$. Accordingly, we define the discretized propensity score and conditional expected reward for the sub-interval I_j as

$$e_{I_j}(\mathbf{x}) = \mathbb{P}(a \in I_j \mid \mathbf{x}) \quad \text{and} \quad \mu_{I_j}(\mathbf{x}) = \mathbb{E}[Y(\mathbf{x}, a)\mathbb{1}\{a \in I_j\} \mid \mathbf{x}]/e_{I_j}. \tag{6.9}$$

After the discretization on the action space, we identify all the actions a belonging to a single sub-interval I_j as the midpoint $A_j = (2j-1)/2V$ of I_j and equips A_j with the average expected reward μ_{I_j}. After discretization, we resemble the setup in the finite-action scenario, and then apply the aforementioned two-stage doubly robust approach to learn a discretized policy concentrated on the A_j's. In the first stage, we obtain $\widehat{\mu}_{I_j}$ and $\widehat{e}_{I_j}$ as estimators of μ_{I_j} and e_{I_j}, respectively. In the second stage, we use neural networks for policy learning by maximizing the discretized doubly robust empirical reward. Specifically, we define $I(a_i) = I_j$ for $a_i \in I_j$ which maps the continuous action to the corresponding discretized sub-interval. For $a_i \in I_j$, we denote $\mathbf{a}_i \in \{0,1\}^V$ as the one-hot vector with the j-th element being 1, which encodes the action a_i. The discretized doubly robust empirical reward is defined as

$$\begin{aligned} &\widehat{Q}^{(\mathrm{D})}(\pi) = \frac{1}{n_2} \sum_{i=n_1+1}^{n} \left\langle \widehat{\Gamma}_i^{(\mathrm{D})}, \pi(\mathbf{x}_i) \right\rangle \\ &\text{with } \widehat{\Gamma}_i^{(\mathrm{D})} = \frac{y_i - \widehat{\mu}_{I(a_i)}(\mathbf{x}_i)}{\widehat{e}_{I(a_i)}(\mathbf{x}_i)} \cdot \mathbf{a}_i + [\widehat{\mu}_{I_1}(\mathbf{x}_i), \dots, \widehat{\mu}_{I_V}(\mathbf{x}_i)]^\top, \end{aligned} \tag{6.10}$$

where the superscript (D) denotes the discretized quantities. We learn an optimal policy by solving the following maximization problem:

$$\widehat{\pi}_{\text{C-DR}} = \underset{\pi \in \Pi_{\mathrm{NN}}}{\operatorname{argmax}} \ \widehat{Q}^{(\mathrm{D})}(\pi) \tag{6.11}$$

where Π_{NN} is a properly chosen neural network. See Section Subsection 6.3.2 for more details of the learning procedure, a proper choice of V, and the statistical guarantees of the learned policy.

6.3 Policy Regret Bound

Our main results are nonasymptotic regret bounds (see Definition 6.1) on the policy learned by the two-stage scheme in Section 6.2, when the covariates are concentrated on a low-dimensional manifold.

The regret of a policy π against a reference policy $\bar{\pi}$ is defined as the difference between their respective expected rewards. The formal definition is given as follows.

Definition 6.1. *Let $\bar{\pi}$ be a fixed reference policy. For any policy π, the regret of π against $\bar{\pi}$ is*

$$R(\bar{\pi}, \pi) = Q(\bar{\pi}) - Q(\pi).$$

Here $Q(\pi)$ is the expected reward either in the finite-action scenario defined in (Eq. 6.1) or the continuous-action scenario which is defined later in (Eq. 6.26). We consider two reference policies: 1) the optimal Hölder policy that maximizes the expected reward; 2) the unconstrained optimal policy that maximizes the expected reward. We establish high probability bounds on the regret of the learned policy for both discrete actions (Section 6.3.1) and continuous actions (Section 6.3.2).

6.3.1 Regret with Discrete Action

Our theory is based on the following assumptions, including some standard assumptions on the smoothness of the propensity score and the reward.

Assumption 6.A.2. *The propensity score and random reward satisfy:*

(i) Overlap: $e_{A_j}(\mathbf{x}) \geq \eta$ *for* $j = 1, \ldots, |\mathcal{A}|$, *where* $\eta > 0$ *is a constant;*

(ii) Bounded Reward: $Y(\mathbf{x}, A_j)$ *is bounded and has a bounded variance, i.e., we have* $\sup_{\mathbf{x} \in \mathcal{M}} |Y(\mathbf{x}, A_j)| \leq M_1$ *and* $\mathrm{Var}[Y(\mathbf{x}, A_j)|\mathbf{x}] \leq \sigma^2$ *for any* $j = 1, \ldots, |\mathcal{A}|$, *where* $M_1 > 0$ *and* $\sigma > 0$ *are constants.*

Assumption 6.A.2 is a standard assumption for statistical guarantees of all learning approaches using the inverse propensity score [205, 206, 193]. Assumption 6.A.2 implies that expected reward μ_{A_j} is bounded since $|\mu_{A_j}(\mathbf{x})| \leq \mathbb{E}[|Y(\mathbf{x}, A_j)| \mid \mathbf{x}] \leq M_1$ for every $\mathbf{x} \in \mathcal{M}$.

Assumption 6.A.3. *Assume covariate* $\mathbf{x}$ *lies on* $\mathcal{M}$. *Given a Hölder index* $\alpha \geq 1$, *we further assume* $\mu_{A_j}(\mathbf{x}) \in \mathcal{H}^\alpha(\mathcal{M})$ *and* $e_{A_j}(\mathbf{x}) \in \mathcal{H}^\alpha(\mathcal{M})$ *for* $j = 1, \ldots, |\mathcal{A}|$. *Moreover, for a fixed* C^∞ *atlas of* $\mathcal{M}$, *there exists* $M_2 > 0$ *such that*

$$\max_j \left\|\mu_{A_j}\right\|_{\mathcal{H}^\alpha} \leq M_2 \quad \textit{and} \quad \max_j \|\log e_{A_j}\|_{\mathcal{H}^\alpha} \leq M_2.$$

Thanks to Assumption 6.A.2 *(i)*, $e_{A_j} \in \mathcal{H}^\alpha$ implies $\log e_{A_j} \in \mathcal{H}^\alpha$ (see Lemma D.10 in Appendix D.4). In the first part of Assumption 6.A.3, all covariates are located on $\mathcal{M}$. Since all of μ_{A_j}'s and e_{A_j}'s are functions of $\mathbf{x}$, they are defined on the same manifold $\mathcal{M}$. The second part of Assumption 6.A.3 characterizes the smoothness of these functions. Now we are ready to derive the following estimation bounds for μ_{A_j} and e_{A_j} using nonparametric regression techniques [14]. To simplify the notation, we denote

$$M = \max\{1, M_1, 2M_2, -\log \eta\}. \tag{6.12}$$

Estimation Bounds of $\mu_{A_j}(\mathbf{x})$ and $e_{A_j}(\mathbf{x})$

By choosing networks

$$\mathcal{F}_{\text{NN}} = \mathcal{F}(L_1, p_1, K_1, \kappa_1, R_1) \quad \text{and} \quad \mathcal{G}_{\text{NN}} = \mathcal{F}(L_2, p_2, K_2, \kappa_2, R_2) \tag{6.13}$$

to estimate μ_{A_j} and e_{A_j} in (Eq. 6.4) and (Eq. 6.6), respectively, we prove the following estimation error bounds for the estimators $\widehat{\mu}_{A_j}$ and $\widehat{e}_{A_j}$ (Lemma 6.1 is proved in Appendix D.3.1). We use $O(\cdot)$ to hide absolute constants and polynomial factors of α, Hölder norm, $\log D$, d, τ, $|\mathcal{A}|$, and the surface area of $\mathcal{M}$.

Lemma 6.1. *Suppose Assumption 3.1, 3.2, 6.1, and 6.A.2 – 6.A.3 hold. We choose*

$$\begin{gathered} L_1 = O(\log \eta n_1), \quad p_1 = O\big((\eta n_1)^{\frac{d}{2\alpha+d}}\big), \quad K_1 = O\big((\eta n_1)^{\frac{d}{2\alpha+d}} \log \eta n_1\big), \\ \kappa_1 = \max\{B, M, \sqrt{d}, \tau^2\}, \quad R_1 = M \end{gathered} \tag{6.14}$$

for $\mathcal{F}_{\text{NN}}$ and

$$\begin{gathered} L_2 = O(\log n_1), \quad p_2 = O\big(|\mathcal{A}|^{\frac{2\alpha}{2\alpha+d}} n_1^{\frac{d}{2\alpha+d}}\big), \quad K_2 = O\big(|\mathcal{A}|^{\frac{2\alpha}{2\alpha+d}} n_1^{\frac{d}{2\alpha+d}} \log n_1\big), \\ \kappa_2 = \max\{B, M, \sqrt{d}, \tau^2\}, \quad R_2 = M, \end{gathered} \tag{6.15}$$

for $\mathcal{G}_{\text{NN}}$ in (Eq. 6.13). *Then for any $j = 1, \dots, |\mathcal{A}|$, we have*

$$\mathbb{E}_{\mathcal{S}_1}\left[\left\|\widehat{\mu}_{A_j} - \mu_{A_j}\right\|_{L^2}^2\right] \leq C_1(M^2 + \sigma^2)(\eta n_1)^{-\frac{2\alpha}{2\alpha+d}} \log^3(\eta n_1), \tag{6.16}$$

$$\mathbb{E}_{\mathcal{S}_1}\left[\left\|\widehat{e}_{A_j} - e_{A_j}\right\|_{L^2}^2\right] \leq C_2 M^2 |\mathcal{A}|^{\frac{4\alpha+d}{2\alpha+d}} n_1^{-\frac{2\alpha}{2\alpha+d}} \log^3 n_1, \tag{6.17}$$

where C_1, C_2 depend on $\log D$, B, τ and the surface area of $\mathcal{M}$.

In (Eq. 6.16) and (Eq. 6.17), the expectation is taken with respect to $\mathcal{S}_1$ defined in (Eq. 6.2). Lemma 6.1 provides performance guarantees of neural networks to solve regression problems (Eq. 6.4) and (Eq. 6.6) in order to estimate μ_{A_j} and e_{A_j}. When the covariates

$\mathbf{x}$ are on a manifold, we prove that the estimation errors converge at a fast rate in which the exponent only depends on the intrinsic dimension d instead of the ambient dimension D. Such a fast convergence is consistent with [93, 207] and indicates the adaptability of neural networks to data intrinsic structures. Note that the network architectures in (Eq. 6.14) and (Eq. 6.15) only require the knowledge of generic information of the data manifold: manifold dimension d, reach τ, and range B. Local geometries of the manifold is automatically captured by the network during empirical risk minimization in **Stage 1** of the doubly robust learning method.

Regret Bound of Learned Policy versus Constrained Oracle Policy

Our first main result is a regret bound of $\widehat{\pi}_{\mathrm{DR}}$ obtained in (Eq. 6.8) against the oracle policy in a Hölder policy class:

$$\pi_\beta^* = \operatorname*{argmax}_{\pi \in \Pi_{\mathcal{H}^\beta}} \mathbb{E}[Q(\pi(\mathbf{x}))],$$

where the Hölder policy class $\Pi_{\mathcal{H}^\beta}$ is defined as

$$\Pi_{\mathcal{H}^\beta} = \left\{ \text{Softmax}[\nu_1(\mathbf{x}), \dots, \nu_{|\mathcal{A}|}(\mathbf{x})]^\top : \nu_j \in \mathcal{H}^\beta(\mathcal{M}) \ \text{ and } \|\nu_j\|_{\mathcal{H}^\beta} \leq M \text{ for } j = 1, \dots, |\mathcal{A}| \right\}. \tag{6.18}$$

Accordingly, we pick the neural network policy class as

$$\Pi_{\mathrm{NN}}^{|\mathcal{A}|} = \{\text{Softmax}(f) \text{ with } f : \mathcal{M} \to \mathbb{R}^{|\mathcal{A}|} \in \mathcal{F}(L_\Pi, p_\Pi, K_\Pi, \kappa_\Pi, R_\Pi)\}. \tag{6.19}$$

Our first theorem shows that $\widehat{\pi}_{\mathrm{DR}}$ is a consistent estimator of the oracle Hölder policy π_β^* as long as the network parameters $L_\Pi, p_\Pi, K_\Pi, \kappa_\Pi, R_\Pi$ are properly chosen.

Theorem 6.1. *Suppose Assumption 3.1, 3.2, 6.1, and 6.A.2 – 6.A.3 hold. Under the setup* *in Lemma 6.1, if the network parameters of* $\Pi_{\rm NN}^{|\mathcal{A}|}$ *are chosen with*

$$\begin{gathered} L_\Pi = O(\log n), \quad p_\Pi = O\big(|\mathcal{A}| n^{\frac{d}{2\beta+d}}\big), \quad K_\Pi = O\big(|\mathcal{A}| n^{\frac{d}{2\beta+d}} \log n\big), \\ \kappa_\Pi = \max\{B, M, \sqrt{d}, \tau^2\}, \quad R_\Pi = M, \end{gathered} \tag{6.20}$$

then with probability no less than $1 - C_1 |\mathcal{A}| n^{-\frac{\beta}{2\beta+d}}$ *over the randomness of data* $\mathcal{S}_1$ *and* $\mathcal{S}_2$*, the following bound holds*

$$\begin{aligned} R(\pi_\beta^*, \widehat{\pi}_{\rm DR}) \leq{} & C|\mathcal{A}|^2 n^{-\frac{\beta}{2\beta+d}} \log^{3/2} n \\ & + \eta^{-1}|\mathcal{A}| \sqrt{\frac{1}{n_2}\sum_{i=n_1+1}^{n} \big(\widehat{\mu}_{A_j}(\mathbf{x}_i) - \mu_{A_j}(\mathbf{x}_i)\big)^2} \sqrt{\frac{1}{n_2}\sum_{i=n_1+1}^{n} \big(\widehat{e}_{A_j}(\mathbf{x}_i) - e_{A_j}(\mathbf{x}_i)\big)^2}, \end{aligned} \tag{6.21}$$

where $C_1 > 0$ *is an absolute constant and* C *depends on* $\log D$*,* d*,* B*,* M*,* τ*,* η*,* β*, and the surface area of* $\mathcal{M}$*.*

Theorem 6.1 is proved in Appendix D.2.1. Theorem 6.1 corroborates the doubly robust property of $\widehat{\pi}_{\rm DR}$. The regret of $\widehat{\pi}_{\rm DR}$ is not sensitive to the individual estimation error of either $\widehat{\mu}_{A_j}$ or $\widehat{e}_{A_j}$, since the bound depends on the product of the estimation errors. Combining Theorem 6.1 and Lemma 6.1 yields the following corollary (see proof in Appendix D.2.2).

Corollary 6.1. *Suppose Assumption 3.1, 3.2, 6.1, and 6.A.2 – 6.A.3 hold. If the network structures are chosen as in Lemma 6.1 and Theorem 6.1, the following regret bound holds with probability no less than* $1 - C_1 n^{-\frac{\alpha\wedge\beta}{2(\alpha\wedge\beta)+d}} \log^3 n$

$$R(\pi_\beta^*, \widehat{\pi}_{\rm DR}) \leq C|\mathcal{A}|^{\frac{16\alpha+7d}{2(2\alpha+d)}} n^{-\frac{\alpha\wedge\beta}{2(\alpha\wedge\beta)+d}} \log^{3/2} n \tag{6.22}$$

where C_1 *is an absolute constant, and* C *depends on* $\log D$*,* d*,* B*,* M*,* σ*,* τ*,* η*,* α*,* β*, and the surface area of* $\mathcal{M}$*.*

In comparison with existing works, our theory has several advantages:

- We allow competing the learned policy $\widehat{\pi}_{\mathrm{DR}}$ with the best oracle policy, while several closely related works [206, 193] only consider competing with the best in-class policy. Specifically, our learnable policy class $\Pi_{\mathrm{NN}}^{|\mathcal{A}|}$ is chosen depending on the sample size n and regularity of oracle policies. In Theorem 6.1, we established a regret bound on $R(\pi_\beta^*, \widehat{\pi}_{\mathrm{DR}})$. Note that $\pi_\beta^* \in \Pi_{\mathcal{H}^\beta}$ does not necessarily belong to $\Pi_{\mathrm{NN}}^{|\mathcal{A}|}$. In contrast, [193] only applies to a comparison of $\widehat{\pi}_{\mathrm{DR}}$ to the best policy $\pi_{\mathrm{NN}}^* \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}$, with $\Pi_{\mathrm{NN}}^{|\mathcal{A}|}$ being given a priori. Figure 6.1 demonstrates such a difference.

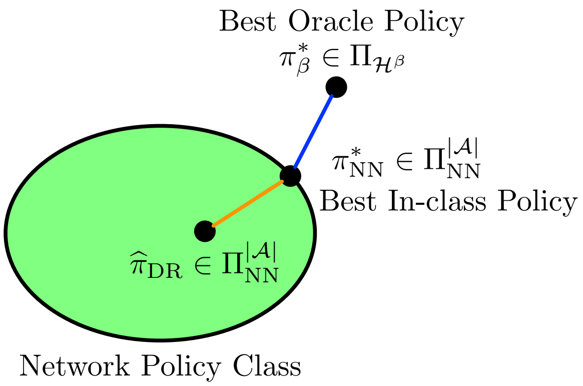

Best Oracle Policy
$\pi_\beta^* \in \Pi_{\mathcal{H}^\beta}$
$\pi_{\mathrm{NN}}^* \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}$
Best In-class Policy
$\widehat{\pi}_{\mathrm{DR}} \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}$
Network Policy Class

Figure 6.1: Learned policy $\widehat{\pi}_{\mathrm{DR}}$ competes with best in-class policy π_{NN}^* and best oracle policy π_β^*. [193] analyzed the regret between $\widehat{\pi}_{\mathrm{DR}}$ and π_{NN}^*, while our analysis applies to $\widehat{\pi}_{\mathrm{DR}}$ compared to π_β^*.

- By considering the low-dimensional geometric structures of the covariates, we obtain a fast rate depending on the intrinsic dimension d. Our theory partially justifies the success of off-policy learning by neural networks for high-dimensional data with low-dimensional structures.

- Our assumptions on the propensity score and expected reward are weak in the sense that the Hölder index $\alpha \geq 1$ can be arbitrary. In [206] and [193], the Hölder index α of the propensity score and expected reward needs to satisfy $2\alpha > D$. This condition is hard to satisfy when the covariates are high-dimensional, unless μ_{A_j}'s and e_{A_j}'s are super smooth with bounded high-order derivatives. Moreover, our theory is obtained by optimally choosing the policy network class $\Pi_{\mathrm{NN}}^{|\mathcal{A}|}$, yet [193, Assumption 3] requires a pre-fixed policy class with certain bounded complexity.

- Our theory is nonasymptotic, while most existing works focus on asymptotic analysis [193, 192, 206].

Remark 6.1. *In Lemma 6.1 and Theorem 6.1, we consider the weight parameters in network classes being uniformly bounded. Such a condition is often implicitly implemented in practice, such as using weight decay (ℓ_2-norm penalty), weight normalization [208, 209] and weight clipping. From a theoretical point of view, the boundedness condition is only imposed for technical convenience to control the complexity of the network class (see the covering number in Lemma D.5). In fact, we can remove such a condition and obtain the same nonasymptotic policy learning guarantees up to a logarithmic factor in n, when the propensity score and expected reward functions are α-Hölder continuous on manifold $\mathcal{M}$ with $\alpha \geq 1$ being an integer. We provide a detailed analysis in Appendix D.1.*

Regret Bound of Learned Policy versus Unconstrained Optimal Policy

We have shown that neural networks can accurately learn an oracle Hölder policy in Corollary 6.1. In this section, we enlarge the oracle policy class to capture all possible policies, including highly nonsmooth polices, e.g., deterministic policies. We show that neural networks can still achieve a small regret, due to their strong expressive power. The relationship between the Hölder policy class, neural network policy class, and unconstrained policy class is depicted in Figure 6.2.

The unconstrained optimal policy is defined as

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\ Q(\pi).$$

To establish the regret bound of $\widehat{\pi}_{\mathrm{DR}}$ in (Eq. 6.8) against π^*, we need the following assumption on the μ_{A_j}'s.

Assumption 6.A.4 (Noise Condition)**.** *Let $q \geq 1$ and denote $j^*(\mathbf{x}) = \operatorname{argmax}_j \mu_{A_j}(\mathbf{x})$.*

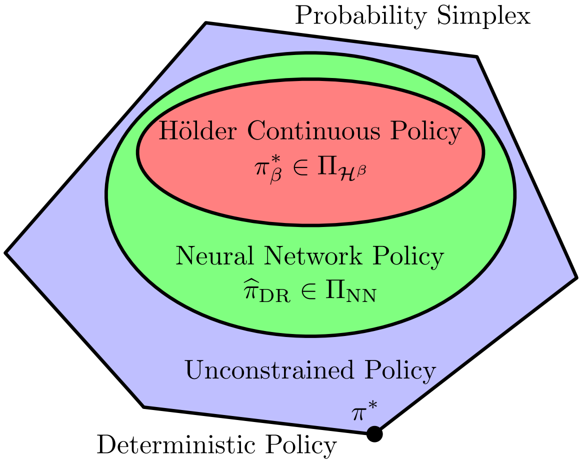


Figure 6.2: The unconstrained policy class is the whole probability simplex with vertices being deterministic polices. The inclusion relation of the neural network policy class and the Hölder policy class indicates that for any Hölder continuous policy, there is an approximation given by a neural network policy.

There exists $c > 0$*, such that*

$$\mathbb{P}\Big[\big|\mu_{A_{j^*(\mathbf{x})}}(\mathbf{x}) - \max_{j \neq j^*(\mathbf{x})} \mu_{A_j}(\mathbf{x})\big| \leq Mt\Big] \leq ct^q, \quad \textit{for any } t \in (0,1).$$

Assumption 6.A.4 implies that, with high probability, there exists an optimal action whose expected reward is larger than those of others by a positive margin. This is an analogue of Tsybakov low-noise condition [210] in multi-class classification problems, which appears similarly in [211]. A similar noisy condition for policy learning with binary actions are proposed in [169]. Assumption 6.A.4 is a generalization of that in [169] to multiple actions. We illustrate the noise condition in a binary-action scenario in Figure 6.3.

We utilize a temperature parameter H in the Softmax layer of the neural network to better learn the unconstrained optimal policy. Under the Hölder continuity in Assumption 6.A.3, there exists a deterministic optimal policy π^*, i.e., $\pi^*(\mathbf{x}) = A_{j^*(\mathbf{x})}$, which is a one-hot vector. In contrast, the output of the Softmax function is a randomized policy (i.e., a vector in the interior of the simplex), unless the output of the neural network is positive infinity. Accordingly, we adopt the Softmax function with a tunable temperature parameter

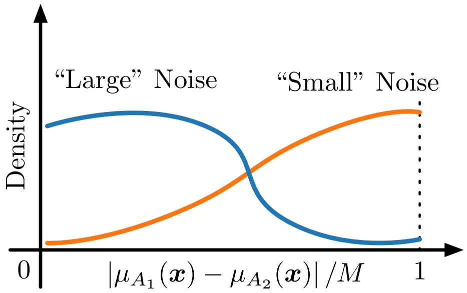


Figure 6.3: Noise condition in a binary-action scenario. "Large" noise corresponds to high densities when $|\mu_{A_1}(\mathbf{x}) - \mu_{A_2}(\mathbf{x})|$ is small; "Small" noise corresponds to low densities near the origin.

H to push the learned policy to a one-hot vector. This idea has given many empirical successes in reinforcement learning [212, 213]. Specifically, we set

$$\Pi^{|\mathcal{A}|}_{\mathrm{NN}(H)} = \{\mathrm{Softmax}_H(f) \text{ with } f : \mathcal{M} \mapsto \mathbb{R}^{|\mathcal{A}|} \in \mathcal{F}(L_\Pi, p_\Pi, K_\Pi, \kappa_\Pi, R_\Pi)\}, \tag{6.23}$$

where $[\mathrm{Softmax}_H(f)]_i = \frac{\exp(f_i/H)}{\sum_j \exp(f_j/H)}$. A small temperature H will push the output of $\Pi^{|\mathcal{A}|}_{\mathrm{NN}(H)}$ towards a one-hot vector, which can better approximate the deterministic policy π^*.

Our main result is the following regret bound of $\widehat{\pi}_{\mathrm{DR}}$ (see proof in Appendix D.2.3).

Theorem 6.2. *Suppose Assumption 3.1, 3.2, 6.1, and 6.A.2 – 6.A.4 hold. Assume the network structures defined in Lemma 6.1 are used to estimate the μ_{A_j}'s and the e_{A_j}'s. If the network parameters of $\Pi^{|\mathcal{A}|}_{\mathrm{NN}(H)}$ are chosen with*

$$L_\Pi = O(\log n),\ p_\Pi = O\big(|\mathcal{A}| n^{\frac{d}{2\alpha+d}}\big),\ K_\Pi = O\big(|\mathcal{A}| n^{\frac{d}{2\alpha+d}} \log n\big),$$
$$\kappa_\Pi = \max\{B, M, \sqrt{d}, \tau^2, 1/H\},\ R_\Pi = M,$$

then the following bound holds with probability no less than $1 - C_1 n^{-\frac{\alpha}{2\alpha+d}} \log^3 n$,

$$R(\pi^*, \widehat{\pi}_{\mathrm{DR}}) \leq \underbrace{C|\mathcal{A}|^{\frac{16\alpha+7d}{2(2\alpha+d)}} n^{-\frac{\alpha}{2\alpha+d}} \log^{3/2} n \log^{1/2}(1/H)}_{\mathcal{T}_1}$$

$$+ \underbrace{\min_{t\in(0,1)} 2cMt^q + M|\mathcal{A}|^2 \exp\left[\left(-Mt + 2n^{-\frac{\alpha}{2\alpha+d}}\right)/H\right]}_{\mathcal{T}_2}, \tag{6.24}$$

where C_1 is an absolute constant, and C depends on $\log D$, d, B, M, σ, τ, η, α, and the surface area of $\mathcal{M}$.

We observe that our choice of the policy network $\Pi^{|\mathcal{A}|}_{\mathrm{NN}(H)}$ is adaptive to the smoothness in the propensity score and expected reward. However, the policy network $\Pi^{|\mathcal{A}|}_{\mathrm{NN}(H)}$ is constructed for approximating a deterministic optimal policy π^* (akin to a classifier) that is not smooth with respect to its input covariate. Indeed, we propose novel policy approximation technique beyond existing function approximation theories of neural networks on manifolds [214, 93]. Specifically, we compute $\mathrm{Softmax}_H[\widetilde{\mu}_{A_1}, \dots, \widetilde{\mu}_{A_{|\mathcal{A}|}}]^\top$ to approximate π^*, where $\widetilde{\mu}$ is an approximation to the expected reward function μ. There are two advantages. On the one hand, $\mathrm{Softmax}_H[\widetilde{\mu}_{A_1}, \dots, \widetilde{\mu}_{A_{|\mathcal{A}|}}]^\top$ maintains the preference to the optimal action returned by π^*, thanks to the noise condition in Assumption 6.A.4. On the other hand, by Assumption 6.A.3, we can efficiently approximate the expected reward functions using neural networks as they are α-Hölder smooth.

The regret $R(\pi^*, \widehat{\pi}_{\mathrm{DR}})$ consists of two parts: a variance term $\mathcal{T}_1$ and a bias term $\mathcal{T}_2$. When the temperature H is fixed, the variance $\mathcal{T}_1$ converges at the rate $n^{-\frac{\alpha}{2\alpha+d}}$, while the bias $\mathcal{T}_2$ does not vanish. This is because $\widehat{\pi}_{\mathrm{DR}}$ is a random policy as the output of a softmax function, while π^* is deterministic as a one-hot vector under Assumption 6.A.4. Furthermore, $\widehat{\pi}_{\mathrm{DR}}$ is asymptotically consistent with π^* when $H \to 0$. If we choose $H = n^{-\frac{2\alpha}{2\alpha+d}}$ and $t = 2n^{-\frac{\alpha}{2\alpha+d}}$, then $\mathcal{T}_2$ converges at the rate $n^{-\frac{q\alpha}{2\alpha+d}}$ and $\mathcal{T}_1$ converges at the rate $n^{-\frac{\alpha}{2\alpha+d}}$. We have the following corollary.

Corollary 6.2. *Suppose Assumption 3.1, 3.2, 6.1, and 6.A.2 – 6.A.4 hold. In the setup of*

Theorem 6.2, setting $H = n^{-\frac{2\alpha}{2\alpha+d}}$ *and* $t = 2n^{-\frac{\alpha}{2\alpha+d}}$ *gives rise to*

$$R(\pi^*, \widehat{\pi}_{\mathrm{DR}}) \leq C|\mathcal{A}|^{\frac{16\alpha+7d}{2(2\alpha+d)}} n^{-\frac{(1\wedge q)\alpha}{2\alpha+d}} \log^2 n \tag{6.25}$$

with probability no less than $1 - C_1 n^{-\frac{\alpha}{2\alpha+d}} \log^3 n$*, where* C_1 *is an absolute constant, and* C *depends on* $\log D$*,* d*,* B*,* M*,* σ*,* τ*,* η*,* α*, and the surface area of* $\mathcal{M}$*.*

6.3.2 Regret with Continuous Action

Our analysis can be extended to the continuous-action scenario. For simplicity, we let the action space be a unit interval, i.e., $\mathcal{A} = [0, 1]$. In such a continuous-action scenario, a policy $\pi(\mathbf{x}, \cdot)$, either randomized or deterministic, is a probability distribution on $[0, 1]$ for each covariate $\mathbf{x} \in \mathcal{M}$. The expected reward of the policy π is defined as

$$Q(\pi) = \int_{\mathcal{M}} \int_0^1 \mu(\mathbf{x}, A)\pi(\mathbf{x}, A)dAd\mathbb{P}(\mathbf{x}), \tag{6.26}$$

where $\mathbb{P}$ is the marginal distribution of covariate $\mathbf{x}$.

As mentioned in Section 6.2.2, we tackle the continuous-action scenario using a discretization technique on the action space. This is motivated by practical applications where continuous objects are often quantized. The action space $\mathcal{A}$ is uniformly partitioned into V sub-intervals $I_j = [(j-1)/V, j/V]$ for $j = 1, \ldots, V$, where V is to be determined in Theorem 6.3. The discretized version of the propensity score and the expected reward on I_j are defined in (Eq. 6.9).

We also consider discretized policies on $\mathcal{A}$. In particular, we identify all the actions belonging to a single sub-interval I_j as its midpoint $A_j = \frac{2j-1}{2V}$. A discretized policy is defined as

$$\pi^{(\mathrm{D})}(\mathbf{x}, A) = \sum_{j=1}^{V} p_j(\mathbf{x})\delta_{A_j}(A), \tag{6.27}$$

where δ_{A_j} is the Dirac delta function at A_j and $p_j(\mathbf{x})$ denotes the probability of choosing action A_j, which satisifes $\sum_{j=1}^{V} p_j(\mathbf{x}) = 1$. In fact, $\pi^{(\mathrm{D})}(\mathbf{x}, \cdot)$ can be interpreted as a vector in the V-dimensional simplex, since it is only supported on V discretized actions. For simplicity, we denote vector $\pi^{(\mathrm{D})}(\mathbf{x}) = [p_1(\mathbf{x}), \ldots, p_V(\mathbf{x})]^\top$ with $p_j(\mathbf{x})$ representing the probability of choosing the action A_j, as an equivalent notation of $\pi^{(\mathrm{D})}(\mathbf{x}, \cdot)$.

For the discretized policy in (Eq. 6.27), the discretized expected reward is defined as

$$Q^{(\mathrm{D})}(\pi^{(\mathrm{D})}) = \int_{\mathcal{M}} \left\langle [\mu_{I_1}(\mathbf{x}), \ldots, \mu_{I_V}(\mathbf{x})]^\top, \pi^{(\mathrm{D})}(\mathbf{x}) \right\rangle d\mathbb{P}(\mathbf{x}). \tag{6.28}$$

We observe the analogy between (Eq. 6.28) and (Eq. 6.3) — the discrete conditional reward μ_{A_j} is replaced by the discretized conditional reward μ_{I_j}, and the number of discrete actions $|\mathcal{A}|$ becomes the number of discretized actions V. On the other hand, the expected reward of a discretized policy is

$$\begin{aligned} Q(\pi^{(\mathrm{D})}) &= \int_{\mathcal{M}} \int_0^1 \mu(\mathbf{x}, A) \sum_{j=1}^{V} \pi_j(\mathbf{x}) \delta_{A_j}(A) dA d\mathbb{P}(\mathbf{x}) \\ &= \int_{\mathcal{M}} \left\langle [\mu(\mathbf{x}, A_1), \ldots, \mu(\mathbf{x}, A_V)]^\top, \pi^{(\mathrm{D})}(\mathbf{x}) \right\rangle d\mathbb{P}(\mathbf{x}). \end{aligned} \tag{6.29}$$

The following lemma shows that if the $\mu(\mathbf{x}, A)$ is Lipschitz in A uniformly for any $\mathbf{x} \in \mathcal{M}$, $Q^{(\mathrm{D})}(\pi^{(\mathrm{D})})$ is close to $Q(\pi^{(\mathrm{D})})$ when V is large (see proof in Appendix D.3.2).

Lemma 6.2. *Assume there exists a constant $L_\mu > 0$ such that*

$$\sup_{\mathbf{x} \in \mathcal{M}} |\mu(\mathbf{x}, A) - \mu(\mathbf{x}, \widetilde{A})| \leq L_\mu |A - \widetilde{A}| \quad \text{for any} \quad A, \widetilde{A} \in [0, 1].$$

Then for any discretized policy $\pi^{(\mathrm{D})}$ given in (Eq. 6.27)*, we have*

$$|Q(\pi^{(\mathrm{D})}) - Q^{(\mathrm{D})}(\pi^{(\mathrm{D})})| \leq L_\mu / V.$$

We remark a key difference between (Eq. 6.28) and (Eq. 6.29). To evaluate (Eq. 6.29), one needs to accurately estimate the $\mu(\mathbf{x}, A_j)$'s, which requires the action A_j to be repeatedly observed. However, this is prohibitive in the continuous-action scenario, since an action A_j is observed with probability 0. In contrast, (Eq. 6.28) relies on the average expected reward on a sub-interval, which can be estimated using standard nonparametric methods in the following Section 6.3.2. Moreover, thanks to Lemma 6.2, we can well approximate $Q(\pi^{(\mathrm{D})})$ by $Q^{(\mathrm{D})}(\pi^{(\mathrm{D})})$ up to a small discretization error. This is crucial to establish the regret bound in Theorem 6.3.

Doubly Robust Policy Learning with Continuous Actions

After discretization, we can apply the doubly robust framework to learn an optimal discretized policy.

In the first stage, we estimate the μ_{I_j}'s and e_{I_j}'s. In the sequel, we use the plain font a_i to denote the observed action of the i-th sample. The bold font $\mathbf{a}_i \in \{0,1\}^V$ denotes the one-hot vector with the j-th element being 1, if $a_i \in I_j$. Similar to (Eq. 6.4) – (Eq. 6.6) in the finite-action case, we obtain estimators of the μ_{I_j}'s and e_{I_j}'s by minimizing the following empirical risks:

$$\widehat{\mu}_{I_j}(\mathbf{x}) = \underset{f \in \mathcal{F}_{\mathrm{NN}}}{\operatorname{argmin}} \ \frac{1}{n_{I_j}} \sum_{i=1}^{n_1} (y_i - f(\mathbf{x}_i))^2 \mathbb{1}\{a_i \in I_j\} \quad \text{with} \quad n_{I_j} = \sum_{i=1}^{n_1} \mathbb{1}\{a_i \in I_j\}, \tag{6.30}$$

and

$$\widehat{g}(\mathbf{x}) = \underset{g \in \mathcal{G}_{\mathrm{NN}}}{\operatorname{argmin}} \ \frac{1}{n_1} \sum_{i=1}^{n_1} - \left\langle [g(\mathbf{x}_i)^\top, 1]^\top, \mathbf{a}_i \right\rangle + \log\Big(1 + \sum_{j=1}^{V-1} \exp([g(\mathbf{x}_i)]_j)\Big), \tag{6.31}$$

$$\widehat{e}_{I_j}(\mathbf{x}) = \frac{\exp([\widehat{g}(\mathbf{x})]_j)}{1 + \sum_{j=1}^{V-1} \exp([\widehat{g}(\mathbf{x})]_j)} \ \text{for } j \leq V-1, \ \text{and} \ \widehat{e}_{I_V}(\mathbf{x}) = \frac{1}{1 + \sum_{j=1}^{V-1} \exp([\widehat{g}(\mathbf{x})]_j)}, \tag{6.32}$$

where $\mathcal{F}_{\mathrm{NN}} : \mathcal{M} \to \mathbb{R}$ and $\mathcal{G}_{\mathrm{NN}} : \mathcal{M} \to \mathbb{R}^{V-1}$ are neural networks.

Algorithm 2 A two–stage algorithm for doubly-robust off-policy learning with continuous actions.

```
Input: Collected data $\{(\mathbf{x}_i, \mathbf{a}_i, y_i)\}_{i=1}^n$. Network architectures $\mathcal{F}_{\mathrm{NN}}, \mathcal{G}_{\mathrm{NN}}$ and $\Pi^V_{\mathrm{NN}(H)}$.
Stage 1: Estimating $\mu_{A_j}$ and $e_{A_j}$.
For each action $A_j$,
    • For each action $A_j$, estimate the reward function $\mu_{A_j}$ by solving (Eq. 6.30),
    • Estimate $e_{A_j}$'s by solving (Eq. 6.31) – (Eq. 6.32).
Stage 2: Policy Learning.
    • Learn the optimal policy by solving (Eq. 6.33).
Output: Learned policy $\widehat{\pi}_{\mathrm{DR}}$.
```

Remark 6.2. *Although our method suggests partitioning the continuous action space into V sub-intervals, we do not require training V independent neural networks for estimating the expected reward and propensity score functions. In practice, the networks for estimating expected reward functions often share a large amount of parameters. The shared part is expected to extract useful data embeddings, while the remaining network layers adapt to individual target expected reward functions. In addition, to estimate the propensity score, we directly train a single multi-dimensional input-output neural network.*

In the second stage, we learn an optimal discretized policy using the $\widehat{\mu}_{I_j}$'s and $\widehat{e}_{I_j}$'s. Recall that we define a mapping $I(a_i) := I_j$ for $a_i \in I_j$ to index which sub-interval a_i belongs to. We use neural networks to learn a discretized policy by maximizing the doubly robust empirical reward in (Eq. 6.10) and (Eq. 6.11), i.e.,

$$\widehat{\pi}_{\text{C-DR}} = \operatorname*{argmax}_{\pi \in \Pi^V_{\mathrm{NN}(H)}} \widehat{Q}^{(\mathrm{D})}(\pi), \tag{6.33}$$

where $\Pi^V_{\mathrm{NN}(H)}$ represents the network class in (Eq. 6.10), which is defined as (Eq. 6.23). We emphasize that $\widehat{\pi}_{\text{C-DR}}$ is a discretized policy and the output $\widehat{\pi}_{\text{C-DR}}(\mathbf{x})$ is a V-dimensional vector in the simplex. Our algorithm for learning policies with continuous actions is summarized in Algorithm 2.

Regret Bound of Learned Discretized Policy

We begin with several assumptions, which are the continuous counterparts of Assumptions 6.A.2 – 6.A.4 in the finite-action scenario.

Assumption 6.B.2. *The propensity score and random reward satisfy:*

(i) Overlap: $e(\mathbf{x}, A) \geq \eta$ *for any* $A \in \mathcal{A}$*, where* $\eta > 0$ *is a constant.*

(ii) Bounded Reward: $|Y(\mathbf{x}, A)| \leq M_1$ *for any* $(\mathbf{x}, A) \in \mathcal{M} \times [0, 1]$*, where* $M_1 > 0$ *is a constant.*

Assumption 6.B.3. *Given a Hölder index* $\alpha \geq 1$*, we have both the expected reward* $\mu(\cdot, A) \in \mathcal{H}^\alpha(\mathcal{M})$ *and the propensity score* $e(\cdot, A) \in \mathcal{H}^\alpha(\mathcal{M})$ *for any fixed action* A*. Moreover, the Hölder norms of* $\mu(\cdot, A)$ *and* $e(\cdot, A)$ *are uniformly bounded for any* $A \in [0, 1]$*, i.e.,*

$$\sup_{A \in [0,1]} \|\mu(\cdot, A)\|_{\mathcal{H}^\alpha} \leq M_2, \quad \textit{and} \quad \sup_{A \in [0,1]} \|e(\cdot, A)\|_{\mathcal{H}^\alpha} \leq M_2$$

for some constant $M_2 > 0$*. Furthermore, there exists a constant* $M_3 > 0$ *such that*

$$\sup_{\mathbf{x} \in \mathcal{M}} |\mu(\mathbf{x}, A_1) - \mu(\mathbf{x}, A_2)| \leq M_3 |A_1 - A_2| \quad \textit{for any } A_1, A_2 \in [0, 1].$$

There also exists a constant $M_4 > 0$ *such that*

$$\left\| \log \left(\int_{I_1} e(\mathbf{x}, A) dA \Big/ \int_{I_2} e(\mathbf{x}, A) dA \right) \right\|_{\mathcal{H}^\alpha} \leq M_4$$

for any intervals $I_1, I_2 \subset [0, 1]$ *of the same length.*

Let $p_I(\mathbf{x}) = \int_I e(\mathbf{x}, A) dA$ denote the probability of choosing actions in I given $\mathbf{x}$. In Assumption 6.B.3, the condition $e(\cdot, A) \in \mathcal{H}^\alpha(\mathcal{M})$ for any given A implies $p_I \in \mathcal{H}^\alpha(\mathcal{M})$ (see Lemma D.8). Combining this and Assumption 6.B.2 *(ii)* of $e(\mathbf{x}, A) \geq \eta > 0$, one deduces that $\log(p_{I_1}(\mathbf{x})/p_{I_2}(\mathbf{x}))$ belongs to $\mathcal{H}^\alpha(\mathcal{M})$ with a bounded Hölder norm. See

Lemma D.9 – D.10 in Appendix D.4 for a formal justification. For simplicity, we denote

$$M = \max\{1, M_1, 2M_2, M_3, M_4, -\log \eta\}. \tag{6.34}$$

Assumption 6.B.4 (Continuous Noise Condition)**.** *The following two conditions hold:*

(i) *For each fixed $\mathbf{x} \in \mathcal{M}$, $\mu(\mathbf{x}, A)$ is unimodal with respect to A: there exists a unique optimal action $A^*(\mathbf{x}) \in \mathcal{A}$ such that $\mu(\mathbf{x}, A^*(\mathbf{x})) = \max_{A \in \mathcal{A}} \mu(\mathbf{x}, A)$.*

(ii) *There exist constants $q \geq 1$ and $c > 0$, such that*

$$\mathbb{P}\left[\mu(\mathbf{x}, A^*(\mathbf{x})) - \mu(\mathbf{x}, A) \leq Mt \quad \textit{given} \quad |A - A^*(\mathbf{x})| \geq \gamma\right] \leq ct^q(1-\gamma),$$

holds for any $t \in (0,1)$ and any $\gamma \in (0,1)$, where $\mathbb{P}$ denotes the marginal distribution on $\mathbf{x}$.

Assumption 6.B.4 generalizes the noise condition for finite actions in Assumption 6.A.4, to the continuous-action scenario. Assmption 6.B.4 *(i)* assures the uniqueness of the optimal action given each covariate. Assumption 6.B.4 *(ii)* means that, with high probability, there is a gap between the reward at the optimal action and the rewards in its neighbors. Such a noise condition is important to derive the regret bound of the learned policy against the unconstrained optimal policy.

We establish a regret bound of $\widehat{\pi}_{\text{C-DR}}$ against the unconstrained optimal (deterministic) policy

$$\pi_{\text{C}}^* = \underset{\pi}{\operatorname{argmax}}\ Q(\pi)$$

where $Q(\pi)$ is defined in (Eq. 6.26). Due to Assumption 6.B.4 *(i)*, π_{C}^* is deterministic with $\pi_{\text{C}}^*(\mathbf{x}, \cdot) = \delta_{A^*(\mathbf{x})}(\cdot)$. The following theorem establishes the regret bound of $\widehat{\pi}_{\text{C-DR}}$ against π_{C}^*.

Theorem 6.3. *Suppose Assumption 3.1, 3.2, 6.1, and 6.B.2 – 6.B.4 hold. Set* $\mathcal{F}_{\rm NN} = \mathcal{F}(L_1, p_1, K_1, \kappa_1, R_1)$ *in* (Eq. 6.30) *and* $\mathcal{G}_{\rm NN} = \mathcal{F}(L_2, p_2, K_2, \kappa_2, R_2)$ *in* (Eq. 6.31) *with*

$$\begin{gathered} L_1 = O(\log n),\ p_1 = O\left(\eta^{\frac{d}{2\alpha+d}} n^{\frac{12\alpha d+7d^2}{7(2\alpha+d)^2}}\right),\ K_1 = O\left(\eta^{\frac{d}{2\alpha+d}} n^{\frac{12\alpha d+7d^2}{7(2\alpha+d)^2}} \log n\right), \\ \kappa_1 = \max\{B, M, \sqrt{d}, \tau^2\},\ R_1 = M, \end{gathered} \tag{6.35}$$

and

$$\begin{gathered} L_2 = O(\log n),\ p_2 = O\left(n^{\frac{4\alpha^2+14\alpha d+7d^2}{7(2\alpha+d)^2}}\right),\ K_2 = O\left(n^{\frac{4\alpha^2+14\alpha d+7d^2}{7(2\alpha+d)^2}} \log n\right), \\ \kappa_2 = \max\{B, M, \sqrt{d}, \tau^2\},\ R_2 = M. \end{gathered} \tag{6.36}$$

If the network parameters in $\Pi^V_{{\rm NN}(H)}$ *are chosen as*

$$\begin{gathered} L_\Pi = O(\log n),\ p_\Pi = O\left(n^{\frac{2\alpha+7d}{7(2\alpha+d)}}\right),\ K_\Pi = O\left(n^{\frac{2\alpha+7d}{7(2\alpha+d)}} \log n\right), \\ \kappa_\Pi = \max\{B, M, \sqrt{d}, \tau^2\},\ R_\Pi = M, \end{gathered} \tag{6.37}$$

and we set $V = O\left(n^{2\alpha/7(2\alpha+d)}\right)$*, the following bound holds with probability no less than* $1 - C_1 n^{-\frac{6\alpha^2+5\alpha d}{7(2\alpha+d)^2}} \log^3 n$

$$\begin{aligned} R(\pi^*_{\rm C}, \widehat{\pi}_{\rm C-DR}) \leq{}& C n^{-\frac{2\alpha}{7(2\alpha+d)}} \log^{3/2} n \log^{1/2} 1/H \\ &+ \left[2cMt^q + Mn^{\frac{4\alpha}{7(2\alpha+d)}} \exp\left(-\left(Mt - 4Mn^{-\frac{2\alpha}{7(2\alpha+d)}}\right)/H\right)\right] \end{aligned} \tag{6.38}$$

for any $t \in \left(2(1+1/M)n^{-\frac{2\alpha}{7(2\alpha+d)}}, 1\right)$*, where* C_1 *is an absolute constant and* C *depends on* $\log D,\ d,\ B,\ M,\ \sigma,\ \tau,\ \eta,\ \alpha$*, and the surface area of* $\mathcal{M}$*.*

Theorem 6.3 is proved in Appendix D.2.4. Rewriting $V = O\left(n^{\frac{1}{7\left(1+\frac{d}{2\alpha}\right)}}\right)$, Theorem 6.3 suggests that

- When d and α are fixed, as sample size n grows, we form finer discretization (large V) to solve the problem with higher accuracy.

- If the intrinsic dimension d is large or smoothness index α is small, the problem is of high complexity. Therefore, when the sample size n is fixed, the discretization is coarse (small V), to ensure a good estimation using sufficient samples in each sub-interval.

This indeed matches intuition: It is an indication of increased complexity if either the intrinsic dimension d is large or smoothness index α is small. Therefore, we need to be conservative in discretization as a compromise of limited training sample budget.

Similar to (Eq. 6.24) in Theorem 6.2, the bound in Theorem 6.3 contains a variance term (the first term) and a bias term (the second term in square brackets). In particular, the bias term consists of problem reduction error, neural network approximation error and the discretization error. The network approximation error is the same one as that in Theorem 6.2. The problem reduction error is due to a nontrivial reduction of the policy learning problem with continuous actions to a policy learning problem with discrete actions. The discretization error is due to the discretization of the action domain and the difference between μ_{I_j}'s and μ_{A_j}'s. See the term I_3 and III_3 in (Eq. D.52) for details. In our proof, new techniques are developed to bound the additional error terms. Note that the variance converges in the rate of $n^{-\frac{2\alpha}{7(2\alpha+d)}}$. For a fixed H , the bias term does not vanish as n goes to infinity. If we set $H = n^{-\frac{2\alpha}{7(2\alpha+d)}}$ and $t = 4(1+M+1/M)n^{-\frac{\alpha}{7(2\alpha+d)}}$, the bias term converges in the rate of $n^{-\frac{2q\alpha}{7(2\alpha+d)}}$. Under this choice, the behavior of $R(\pi^*_{\mathrm{C}}, \widehat{\pi}_{\mathrm{C-DR}})$ is summarized in the following corollary.

Corollary 6.3. *Suppose Assumption 3.1, 3.2, 6.1, and 6.B.2 – 6.B.4 hold. In the setup of Theorem 6.3, setting* $H = n^{-\frac{2\alpha}{7(2\alpha+d)}}$ *and* $t = 4(1+M+1/M)n^{-\frac{\alpha}{7(2\alpha+d)}}$ *gives rise to*

$$R(\pi^*_{\mathrm{C}}, \widehat{\pi}_{\mathrm{C-DR}}) \leq C n^{-\frac{2\alpha}{7(2(q\wedge 1)\alpha+d)}} \log^2 n \tag{6.39}$$

with probability no less than $1 - C_1 n^{-\frac{6\alpha^2+5\alpha d}{7(2\alpha+d)^2}} \log^3 n$*, where* C_1 *is an absolute constant and* C *depends on* $\log D$, $d,\ B,\ M,\ \sigma,\ \tau,\ \eta,\ \alpha$*, and the surface area of* $\mathcal{M}$*.*

There are limited theoretical guarantees for causal inference with continuous actions. [172] proposed a doubly robust method to estimate continuous treatment effects. The asymptotic behavior of the method was analyzed while the policy learning problem was not addressed. To our knowledge, Theorem 6.3 and Corollary 6.3 is the first finite-sample performance guarantee of policy learning with continuous actions.

6.4 Conclusion and Discussion

This chapter establishes statistical guarantee for doubly robust off-policy learning by neural networks. The covariate is assumed to be on a low-dimensional manifold. Non-asymptotic regret bounds for the learned policy are proved in the finite-action scenario and in the continuous-action scenario. Our results show that when the covariates exhibit low dimensional-structures, neural networks provide a fast convergence rate whose exponent depends on the intrinsic dimension of the manifold instead of the ambient dimension. Our results partially justify the success of neural networks in causal inference with high-dimensional covariates.

We finally provide some discussions in connection with the existing literature.

Sample Complexity Lower Bound without Low-dimensional Structures [187] established a lower bound of the sample complexity for policy evaluation (or treatment effect estimation), when the covariates are in $\mathbb{R}^D$ and do not have low-dimensional structures. Specifically, they assume that both the initial policy and reward functions belong to a Hölder space. The sample complexity needs to be at least exponential in the dimension D. This result shows that the rate can not be improved unless additional assumptions are made. By assuming that the covariates are on a d-dimensional manifold, our sample complexity only depends on the intrinsic dimension d. We remark that [187] studied the Hölder space with a Hölder index $\alpha \in (0, 1]$, while we focus on the case of $\alpha \geq 1$. In the case that $\alpha = 1$, if we have $q \geq 1$ (in Assumption 6.A.4), Corollary 6.2 gives the convergence rate

$O\left(n^{-\frac{1}{2+d}}\log^3 n\right)$. This rate is better than the minimax rate $O\left(n^{-\frac{1}{2+D}}\right)$ from [187] thanks to the low-dimensional structures of the covariates.

Nonconvex Optimization of Deep Neural Networks Our theoretical guarantees hold for the global optimum of (Eq. 6.4)-(Eq. 6.8). However, solving these optimization can be difficult in practice. Some recent empirical and theoretical results have shown that large neural networks help to ease the optimization without sacrificing statistical efficiency [157, 158]. This is also referred to as an overparameterization phenomenon. We will leave it for future investigation.

Inference on the Optimal Expected Value Function Our analysis provides nonasymptotic regret bounds of learned policies, which can provide a rough confidence interval of the optimal expected value function. For example, Corollary 6.1 implies

$$\begin{aligned}&\mathbb{E}[Q(\pi^*_\beta)]\\&\quad\in\left[Q(\widehat{\pi}_{\mathrm{DR}})-C|\mathcal{A}|^{\frac{16\alpha+7d}{2(2\alpha+d)}}n^{-\frac{\alpha\wedge\beta}{2(\alpha\wedge\beta)+d}}\log^2 n,\ Q(\widehat{\pi}_{\mathrm{DR}})+C|\mathcal{A}|^{\frac{16\alpha+7d}{2(2\alpha+d)}}n^{-\frac{\alpha\wedge\beta}{2(\alpha\wedge\beta)+d}}\log^2 n\right]\end{aligned}$$

with probability $1-\widetilde{O}(n^{-\frac{\alpha\wedge\beta}{2(\alpha\wedge\beta)+d}})$. Nonetheless, the constant C is not exactly given.

It is yet relatively difficult to establish precise asymptotic confidence intervals for the optimal expected value function. The optimal value function is obtained by optimizing policy among a properly chosen class Π_{NN} (neural network policy class). Therefore, we need to study the asymptotic distribution of $\sup_{\pi\in\Pi_{\mathrm{NN}}}\mathbb{E}[Q(\pi(\mathbf{x}))]$, which can be viewed as the supremum of a stochastic process indexed by $\pi\in\Pi_{\mathrm{NN}}$. As can be seen, this is much more difficult compared to policy evaluation problems, where the target policy is fixed. We suspect that Stein's method [215, 216] can provide an initial analytical framework, however, a rigorous analysis is beyond the scope of the chapter.

CHAPTER 7
CONCLUDING REMARKS

In this thesis, we show deep neural networks are adaptive to data intrinsic structures in function approximation and statistical applications. Specifically, when data concentrate on a low-dimensional Riemannian manifold, neural networks are capable of efficiently approximating Hölder, Sobolev, and Besov functions. The size of the network grows depending on the manifold dimension. Besides approximating a target function in terms of function value, we extend approximation guarantees to first-order derivatives.

Turning towards statistical applications, we show that by choosing network architectures according to our approximation theories, neural networks can circumvent the curse of data ambient dimensionality. In particular, for nonparametric regression/classification, distribution estimation, and causal inference problems, we obtain fast statistical guarantees. The convergence rate weakly depends on the data ambient dimension. These results partially explain the remarkable success of neural networks in practice.

7.1 Future Directions

Besides future directions presented in the end of each previous chapter, we discuss additional future topics.

Robust function approximation and generalization guarantees Existing works on function approximation theories focus on how well a neural network can approximate target function value under certain metrics, e.g., function L^∞ norm. However, such approximation theories may fail to control the Lipschitz continuity of the network, and the obtained neural network can be highly zigzagging. This leaves a significant vacancy in theory on understanding whether neural network can well approximate the target function beyond zero-th order. I aim to establish robust function approximation theories, showing that over-parameterized neural networks not only recover function value, but also replicate target function's Lipschitz continuity.

Furthermore, in practice, Lipschitzness or even higher-order smoothness is often induced by computational heuristics, e.g., weight decay and spectral normalization, and the trained networks yield better generalization performance and are robust against adversarial attacks. I plan to study whether popular optimization algorithms can optimize the network architecture suggested by the robust function approximation theory and bias towards the desired network. This will build upon my preliminary understandings of algorithmic behaviors of first-order methods in nonconvex problems [217, 218, 219]. I foresee these results possess intimate connections to max-margin classification, adversarial training, and transfer learning.

Deep reinforcement learning Deep reinforcement learning creates astounding AI game players like AlphaGo to convincingly beat top professional human players. In multi-player games, deep reinforcement learning is also competitive or even better than human beings. However, the success of deep reinforcement learning lacks theoretical understanding. Many existing works of RL are established under simplified assumptions, e.g., linear function approximation in reward and transition kernels. I plan to study neural network-parameterized policy learning and evaluation problems in general online/off-line RL. This will leverage the efficient function approximation theories and statistical analyzing tools. In addition, I expect neural networks can adapt to geometric structures in the state-action space, and circumvent the curse of data dimensionality.

In addition, for multi-agent environment, the joint state-action space grows exponentially in the number of agents (curse of many agents). I tackle such a challenge by utilizing the rotational invariance among agents [220], which results in a mean-field formulation of multi-agent RL problems. I am interested in exploring other structural interactions between agents, e.g., interaction according to a sparsely connected graph, for potential mitigation of the curse of many agents.

Transfer learning Transfer learning is widely used in modern deep learning applications, including natural language processing and computer vision. Adapting large pre-trained models to downstream tasks gives rise to the state-of-the-art performance. More interestingly, empirical results suggest that fine-tuning a small fraction of the pre-trained model or adding simple adaptation layers already yield superior performance, and therefore, the adaptation is both computation and sample efficient. Despite empirical successes, limited theory has been developed to understand the power of pre-trained models and their ability of adaptation. I plan to study transfer learning through the function approximation and statistical perspective. In particular, I aim to investigate the following questions:

- Under what conditions, fine-tuning a small fraction of parameters in a large network allows rich function approximation?

- Under what conditions, adding simple adaptation layers allows rich function approximation?

- From a statistical perspective, does adapting a pre-trained model gives rise to better sample complexity in comparison to training from scratch?

The first two questions are closely related to perturbation analysis of neural networks. Different from conventional studies, the perturbation is now added on the model parameters. Therefore, I expect to utilize tools from constructive function approximation as well as develop new ones. The third question targets at a statistical measure of useful information extracted from the pre-training stage. The information can be viewed as increasing the "effective" sample complexity of downstream tasks. Nonetheless, a systematic measure of the information is very open.

Appendices

APPENDIX A
OMITTED PROOFS IN CHAPTER 3

A.1 Proofs of Preliminary Results in Section 3.1

A.1.1 Proof of Lemma 3.1

Proof. We partition the interval $[0, 1]$ uniformly into 2^N subintervals $I_k = [\frac{k}{2^N}, \frac{k+1}{2^N}]$ for $k = 0, \ldots, 2^N - 1$. We approximate $f(x) = x^2$ on these subintervals by a linear interpolation

$$\widehat{f}_k = \frac{2k+1}{2^N}\left(x - \frac{k}{2^N}\right) + \frac{k^2}{2^{2N}}, \quad \text{for } x \in I_k.$$

It is straightforward to check that $\widehat{f}_k$ meets f at the endpoints $\frac{k}{2^N}, \frac{k+1}{2^N}$ of I_k.

We evaluate the approximation error of $\widehat{f}_k$ on the interval I_k:

$$\begin{aligned}
\max_{x \in I_k} \left| f(x) - \widehat{f}_k(x) \right| &= \max_{x \in I_k} \left| x^2 - \frac{2k+1}{2^N} x + \frac{k^2 + k}{2^{2N}} \right| \\
&= \max_{x \in I_k} \left| \left(x - \frac{2k+1}{2^{N+1}} \right)^2 - \frac{1}{2^{2N+2}} \right| \\
&= \frac{1}{2^{2N+2}}.
\end{aligned}$$

Note that this approximation error does not depend on k. Thus, in order to achieve an ϵ approximation error, we only need

$$\frac{1}{2^{2N+2}} \leq \epsilon \implies N \geq \frac{\log \frac{1}{\epsilon}}{2 \log 2} - 1.$$

Since $2 \log 2 > 1$, we let $N = \lceil \log \frac{1}{\epsilon} \rceil$ and denote $f_N = \sum_{k=0}^{2^N - 1} \widehat{f}_k \mathbb{1}\{x \in I_k\}$. We compute the increment from f_{N-1} to f_N for $x \in \left[\frac{k}{2^{N-1}}, \frac{k+1}{2^{N-1}}\right]$ as

$$f_{N-1} - f_N = \begin{cases} \frac{k^2}{2^{2(N-1)}} + \frac{2k+1}{2^{N-1}}\left(x - \frac{k}{2^{N-1}}\right) - \frac{k^2}{2^{2(N-1)}} - \frac{4k+1}{2^N}\left(x - \frac{k}{2^{N-1}}\right), & x \in \left[\frac{k}{2^{N-1}}, \frac{2k+1}{2^N}\right) \\ \frac{k^2}{2^{2(N-1)}} + \frac{2k+1}{2^{N-1}}\left(x - \frac{k}{2^{N-1}}\right) - \frac{(2k+1)^2}{2^{2N}} - \frac{4k+3}{2^N}\left(x - \frac{2k+1}{2^N}\right), & x \in \left[\frac{2k+1}{2^N}, \frac{k+1}{2^{N-1}}\right) \end{cases}$$

$$= \begin{cases} \frac{1}{2^N}x - \frac{k}{2^{2N-1}}, & x \in \left[\frac{k}{2^{N-1}}, \frac{2k+1}{2^N}\right) \\ -\frac{1}{2^N}x + \frac{k+1}{2^{2N-1}}, & x \in \left[\frac{2k+1}{2^N}, \frac{k+1}{2^{N-1}}\right) \end{cases}.$$

We observe that $f_{N-1} - f_N$ is a triangular function on $\left[\frac{k}{2^{N-1}}, \frac{k+1}{2^{N-1}}\right]$. The maximum is $\frac{1}{2^{2N}}$ independent of k attained at $x = \frac{2k+1}{2^N}$. The minimum is 0 attained at the endpoints $\frac{k}{2^{N-1}}, \frac{k+1}{2^{N-1}}$. To implement f_N, we consider a triangular function representable by a one-layer ReLU network:

$$g(x) = 2\sigma(x) - 4\sigma(x - 0.5) + 2\sigma(x - 1).$$

Denote by $g_m = g \circ g \circ \cdots \circ g$ the composition of totally m functions g. Observe that g_m is a sawtooth function with 2^{m-1} peaks at $\frac{2k+1}{2^m}$ for $k = 0, \ldots, 2^{m-1} - 1$, and we have $g_m\left(\frac{2k+1}{2^m}\right) = 1$ for $k = 0, \ldots, 2^{m-1} - 1$. Then we have $f_{N-1} - f_N = \frac{1}{2^{2N}} g_N$. By induction, we have

$$\begin{aligned} f_N &= f_{N-1} - \frac{1}{2^{2N}} g_N \\ &= f_{N-2} - \frac{1}{2^{2N}} g_N - \frac{1}{2^{2N-2}} g_{N-1} \\ &= \cdots \\ &= x - \sum_{k=1}^{N} \frac{1}{2^{2k}} g_k. \end{aligned}$$

Therefore, f_N can be implemented by a ReLU network of depth $\left\lceil \log \frac{1}{\epsilon} \right\rceil \leq \log \frac{1}{\epsilon} + 1$. Meanwhile, each layer consists of at most 3 neurons. Hence, the total number of neurons and weight parameters is no more than $c \log \frac{1}{\epsilon}$ for an absolute constant c. □

A.1.2 Proof of Corollary 3.1

Proof. Let $\widehat{f_\delta}$ be an approximation of the quadratic function on $[0,1]$ with error $\delta \in (0,1)$. We set

$$\widehat{\times}(x,y) = C^2 \left(\widehat{f_\delta}\left(\frac{|x+y|}{2C}\right) - \widehat{f_\delta}\left(\frac{|x-y|}{2C}\right)\right).$$

Now we determine δ. We bound the error of $\widehat{\times}$

$$\begin{aligned}\left|\widehat{\times}(x,y) - xy\right| &= C^2 \left| \widehat{f_\delta}\left(\frac{|x+y|}{2C}\right) - \frac{|x+y|^2}{4C^2} - \widehat{f_\delta}\left(\frac{|x-y|}{2C}\right) + \frac{|x-y|^2}{4C^2}\right| \\ &\leq C^2 \left| \widehat{f_\delta}\left(\frac{|x+y|}{2C}\right) - \frac{|x+y|^2}{4C^2}\right| + \left|\widehat{f_\delta}\left(\frac{|x-y|}{2C}\right) - \frac{|x-y|^2}{4C^2}\right| \\ &\leq 2C^2\delta.\end{aligned}$$

Thus, we pick $\delta = \frac{\epsilon}{2C^2}$ to ensure $\left|\widehat{\times}(x,y) - xy\right| \leq \epsilon$ for any inputs x and y. As shown in Lemma 3.1, we can implement $\widehat{f_\delta}$ using a ReLU network of depth at most $c' \log \frac{1}{\delta} = c \log \frac{C^2}{\epsilon}$ with absolute constants c', c. The proof is complete. □

A.2 Proof of Approximation Theory of Feedforward Network (Theorem 3.1)

This section consists of the detailed proofs of Lemma 3.2, Lemma 3.3, local approximation theory Theorem 3.3, error decomposition in Lemma 3.4 and a technical Lemma A.1 for bounding the error, as well as the configuration of the desired ReLU network class for universally approximating Hölder functions. For notational simplicity, we let $s = \lfloor \alpha \rfloor$ and reload $\alpha = \alpha - \lfloor \alpha \rfloor$.

A.2.1 Proof of Lemma 3.2

Proof. We first show P_i defined on U_i is a homeomorphism, which implies (U_i, P_i) is a chart on the manifold. Then by Proposition 6.10 in [22], we conclude P_i is a diffeomorphism.

To show P_i is a homeomorphism on U_i, we only need to show P_i has a continuous inverse. By Lemma 5.4 in [47], the derivative of P_i is nonsingular in U_i. The inverse function theorem implies that P_i is locally invertible in an open neighborhood $\mathcal{B}(\mathbf{c}_i, c\tau) \bigcap \mathcal{M}$ for some constant $c > 0$. In the following, we show by contradiction that the constant $c \geq 1/4$. Suppose not, there exist distinct points $\mathbf{a}, \mathbf{b} \in U_i$ such that $\mathsf{P}_i(\mathbf{a}) = \mathsf{P}_i(\mathbf{b})$ with $\|\mathbf{a} - \mathbf{c}_i\|_2 < \tau/4$ and $\|\mathbf{b} - \mathbf{c}_i\|_2 < \tau/4$. Using the triangle inequality, we obtain $\|\mathbf{a} - \mathbf{b}\|_2 < \tau/2$. Applying Proposition 6.3 in [47], we derive

$$d_{\mathcal{M}}(\mathbf{a}, \mathbf{b}) < \tau \quad \text{and} \quad d_{\mathcal{M}}(\mathbf{a}, \mathbf{c}_i) < \tau(1 - \sqrt{2}/2)$$
$$\text{with} \quad d_{\mathcal{M}}(\cdot, \cdot) \quad \text{being the geodesic distance.}$$

Furthermore, using Proposition 6.2 in [47], we lower bound the angle between the tangent spaces $T_{\mathbf{c}_i}(\mathcal{M})$ and $T_{\mathbf{a}}(\mathcal{M})$ by

$$\cos\left(\angle(T_{\mathbf{a}}(\mathcal{M}), T_{\mathbf{c}_i}(\mathcal{M}))\right) \triangleq \min_{\mathbf{u} \in T_{\mathbf{a}}(\mathcal{M})} \max_{\mathbf{v} \in T_{\mathbf{c}_i}(\mathcal{M})} |\langle \mathbf{u}, \mathbf{v} \rangle| \geq 1 - \frac{1}{\tau} d_{\mathcal{M}}(\mathbf{a}, \mathbf{c}_i) > \sqrt{2}/2. \tag{A.1}$$

On the other hand, we consider a unit speed geodesic $\gamma(t)$ starting from $\mathbf{a}$ and ending at $\mathbf{b}$, with $\gamma(0) = \mathbf{a}$, $\gamma(d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})) = \mathbf{b}$, and $\|\dot{\gamma}\|_2 = 1$. Integration by parts yields

$$\begin{aligned} \mathbf{b} - \mathbf{a} &= \gamma(d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})) - \gamma(0) \\ &= \int_0^{d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})} \dot{\gamma}(t) dt \\ &= \dot{\gamma}(0) d_{\mathcal{M}}(\mathbf{a}, \mathbf{b}) + \int_0^{d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})} \int_0^t \ddot{\gamma}(s) ds dt. \end{aligned}$$

Rearranging terms gives rise to

$$\|\mathbf{b} - \mathbf{a} - \dot{\gamma}(0) d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})\|_2 \leq \int_0^{d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})} \int_0^t \|\ddot{\gamma}(s)\|_2 \, ds dt \leq \frac{d_{\mathcal{M}}^2(\mathbf{a}, \mathbf{b})}{2\tau}, \tag{A.2}$$

where the last inequality follows from Proposition 6.1 in [47]. Dividing (Eq. A.2) by

$d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})$ and plugging in $d_{\mathcal{M}}(\mathbf{a}, \mathbf{b}) \leq \tau$, we have

$$\left\| \frac{\mathbf{b} - \mathbf{a}}{d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})} - \dot{\gamma}(0) \right\|_2 < \frac{1}{2}.$$

For any unit vector $\mathbf{v} \in T_{\mathbf{c}_i}(\mathcal{M})$, we evaluate the inner product

$$\begin{aligned} |\langle \dot{\gamma}(0), \mathbf{v} \rangle| &\leq \left| \left\langle \dot{\gamma}(0) - \frac{\mathbf{b} - \mathbf{a}}{d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})}, \mathbf{v} \right\rangle \right| + \left| \left\langle \frac{\mathbf{b} - \mathbf{a}}{d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})}, \mathbf{v} \right\rangle \right| \\ &\overset{(i)}{=} \left| \left\langle \dot{\gamma}(0) - \frac{\mathbf{b} - \mathbf{a}}{d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})}, \mathbf{v} \right\rangle \right| \\ &\leq \left\| \frac{\mathbf{b} - \mathbf{a}}{d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})} - \dot{\gamma}(0) \right\|_2 \\ &< \frac{1}{2}, \end{aligned} \tag{A.3}$$

where $\left| \left\langle \frac{\mathbf{b}-\mathbf{a}}{d_{\mathcal{M}}(\mathbf{a},\mathbf{b})}, \mathbf{v} \right\rangle \right| = 0$ in equality (i), since $\mathsf{P}_i(\mathbf{a}) = \mathsf{P}_i(\mathbf{b})$ by our assertion. Combining (Eq. A.1) and (Eq. A.3), we obtain

$$\frac{\sqrt{2}}{2} < \cos\left(\angle(T_{\mathbf{a}}(\mathcal{M}), T_{\mathbf{c}_i}(\mathcal{M}))\right) \leq \max_{\mathbf{v} \in T_{\mathbf{c}_i}(\mathcal{M})} |\langle \dot{\gamma}(0), \mathbf{v} \rangle| < \frac{1}{2},$$

which is a contradiction. Therefore, we conclude that P_i is injective, and hence invertible on the local neighborhood $\mathcal{B}(\mathbf{c}_i, \tau/4) \bigcap \mathcal{M}$. The continuity of P_i follows from its definition, and the inverse map of a continuous map is also continuous. Therefore, P_i is a homeomorphism on $\mathcal{B}(\mathbf{c}_i, r) \bigcap \mathcal{M}$ for $r \leq \tau/4$.

The last step is to show P_i is also a diffeomorphism. We leverage the following proposition.

Proposition A.1 (Proposition 6.10 in [22]). *If (U, ϕ) is a chart on a manifold $\mathcal{M}$, then the coordinate map $\phi : U \mapsto \phi(U)$ is a diffeomorphism.*

Since P_i is a homeomorphism, we deduce that (U_i, P_i) is a chart of $\mathcal{M}$. Applying Proposition A.1, we conclude that P_i is a diffeomorphism. □

A.2.2 Proof of Lemma 3.3

Proof. Recall that we choose local coordinate neighborhood U_i in **Step 1** in Section 3.1.1. Let P_i be the projection onto the tangent space $T_{\mathbf{c}_i}(\mathcal{M})$. Then $\{(U_i, \mathsf{P}_i)\}$ is an atlas of $\mathcal{M}$. Without loss of generality, we assume that $\{(U_i, \mathsf{P}_i)\}$ verifies the Hölder condition in Definition 2.12. Now we rewrite $f_i \circ \phi_i^{-1}$ as

$$\underbrace{(f \circ \phi_i^{-1})}_{g_1} \times \underbrace{(\rho_i \circ \phi_i^{-1})}_{g_2}. \tag{A.4}$$

By the definition of the partition of unity, we know g_2 is C^∞. This implies that g_2 is $(s+1)$ continuously differentiable. Since supp(ρ_i) is compact, the k-th derivative of g_2 is uniformly bounded by $\lambda_{i,k}$ for any $k \leq s+1$. Let $\lambda_i = \max_{k \leq n+1} \lambda_{i,k}$. We have for any $|\mathbf{n}| \leq n$ and $\mathbf{x}_1, \mathbf{x}_2 \in U_i$,

$$\begin{aligned}|D^{\mathbf{n}} g_2(\phi_i(\mathbf{x}_1)) - D^{\mathbf{n}} g_2(\phi_i(\mathbf{x}_2))| &\leq \sqrt{d}\lambda_i \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2 \\ &\leq \sqrt{d}\lambda_i b_i^{1-\alpha} \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|_2^{1-\alpha} \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2^{\alpha}.\end{aligned}$$

The last inequality follows from $\phi_i(\mathbf{x}) = b_i(V_i^\top(\mathbf{x} - \mathbf{c}_i) + \mathbf{u}_i)$ and $\|V_i\|_2 = 1$. Observe that U_i is bounded, hence, we have $\|\mathbf{x}_1 - \mathbf{x}_2\|_2^{1-\alpha} \leq (2r)^{1-\alpha}$. Absorbing $\|\mathbf{x}_1 - \mathbf{x}_2\|_2^{1-\alpha}$ into $\sqrt{d}\lambda_i b_i^{1-\alpha}$, we have the derivative of g_2 is Hölder continuous. We denote $\beta_{i,\alpha} = \sqrt{d}\lambda_i b_i^{1-\alpha}(2r)^{1-\alpha} \leq \sqrt{d}\lambda_i(2r)^{1-\alpha}$. Similarly, g_1 is C^{s-1} by Assumption 3.3. Then there exists a constant μ_i such that the k-th derivative of g_1 is uniformly bounded by μ_i for any $k \leq n-1$. These derivatives are also Hölder continuous with coefficient $\theta_{i,\alpha} \leq \sqrt{d}\mu_i(2r)^{1-\alpha}$.

By the Leibniz rule, for any $|\mathbf{s}| = s$, we expand the s-th derivative of $f_i \circ \phi_i^{-1}$ as

$$D^{\mathbf{s}}(g_1 \times g_2) = \sum_{|\mathbf{p}|+|\mathbf{q}|=s} \binom{s}{|\mathbf{p}|} D^{\mathbf{p}} g_1 D^{\mathbf{q}} g_2.$$

Consider each summand in the above right-hand side. For any $\mathbf{x}_1, \mathbf{x}_2 \in U_i$, we derive

$$\begin{aligned}
&\left|D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_1)) D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_1)) - D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_2)) D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_2))\right| \\
=&\big|D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_1)) D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_1)) - D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_1)) D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_2)) \\
&\quad + D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_1)) D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_2)) - D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_2)) D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_2))\big| \\
\leq&|D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_1))||D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_1)) - D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_2))| \\
&\quad + |D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_2))||D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_1)) - D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_2))| \\
\leq&\mu_i \theta_{i,\alpha} \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2^{\alpha} + \lambda_i \beta_{i,\alpha} \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2^{\alpha} \\
\leq&2\sqrt{d}\mu_i \lambda_i (2r)^{1-\alpha} \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2^{\alpha}.
\end{aligned}$$

Observe that there are totally 2^s summands in the right hand side of (Eq. A.4). Therefore, for any $\mathbf{x}_1, \mathbf{x}_2 \in U_i$ and $|\mathbf{s}| = s$, we have

$$\left|D^{\mathbf{s}}(f_i \circ \phi_i^{-1})\big|_{\phi_i(\mathbf{x}_1)} - D^{\mathbf{s}}(f_i \circ \phi_i^{-1})\big|_{\phi_i(\mathbf{x}_2)}\right| \leq 2^{s+1}\sqrt{d}\mu_i \lambda_i (2r)^{1-\alpha} \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2^{\alpha}.$$

□

A.2.3 Proof of Taylor Polynomial Approximation

Proof of Theorem 3.3. The proof consists of two steps. We first approximate $f_i \circ \phi_i^{-1}$ by a Taylor polynomial, and then implement the Taylor polynomial using a ReLU network. To ease the analysis, we extend $f_i \circ \phi_i^{-1}$ to the whole cube $[0,1]^d$ by assigning $f_i \circ \phi_i^{-1}(\mathbf{x}) = 0$ for $\phi_i(\mathbf{x}) \in [0,1]^d \setminus \phi_i(U_i)$. It is straightforward to check that this extension preserves the regularity of $f_i \circ \phi_i^{-1}$, since f_i vanishes on the complement of the compact set $\operatorname{supp}(\rho_i) \subset U_i$. For notational simplicity, we denote $f_i^{\phi} = f_i \circ \phi_i^{-1}$ with the extension. Accordingly, Lemma 3.3 can be extended to the whole cube $[0,1]^d$ without changing its proof, i.e., for any $\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^d$ and $|\mathbf{s}| = s$, we have

$$\left| D^{\mathbf{s}} f_i^{\phi}\big|_{\mathbf{x}_1} - D^{\mathbf{s}} f_i^{\phi}\big|_{\mathbf{x}_2} \right| \leq 2^{s+1}\sqrt{d}\mu_i \lambda_i (2r)^{1-\alpha} \left\| \mathbf{x}_1 - \mathbf{x}_2 \right\|_2^{\alpha}. \tag{A.5}$$

Step 1. We define a trapezoid function

$$\psi(x) = \begin{cases} 1 & |x| < 1 \\ 2 - |x| & 1 \leq |x| \leq 2 \\ 0 & |x| > 2 \end{cases}.$$

Note that we have $\|\psi\|_\infty = 1$. Let N be a positive integer, we form a uniform grid on $[0,1]^d$ by dividing each coordinate into N subintervals. We then consider a partition of unity on these grid defined by

$$\zeta_{\mathbf{m}}(\mathbf{x}) = \prod_{k=1}^{d} \psi\left(3N\left(x_k - \frac{m_k}{N}\right)\right).$$

We can check that $\sum_{\mathbf{m}} \zeta_{\mathbf{m}}(\mathbf{x}) = 1$ as in Figure A.1.

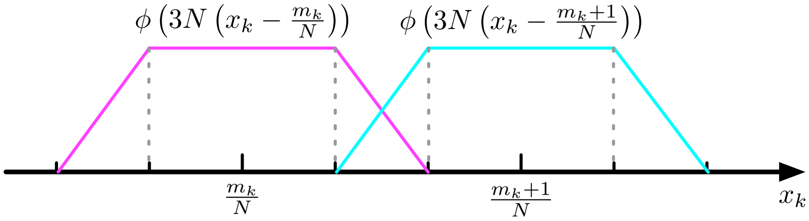

$\phi\left(3N\left(x_k - \frac{m_k}{N}\right)\right)$
$\phi\left(3N\left(x_k - \frac{m_k+1}{N}\right)\right)$
$\frac{m_k}{N}$
$\frac{m_k+1}{N}$
x_k

Figure A.1: Illustration of the construction of $\zeta_{\mathbf{m}}$ on the k-th coordinate.

We also observe

$$\text{supp}(\zeta_{\mathbf{m}}) = \left\{\mathbf{x} : \left|x_k - \frac{m_k}{N}\right| \leq \frac{2}{3N}, k = 1, \dots, d\right\} \subset \left\{\mathbf{x} : \left|x_k - \frac{m_k}{N}\right| \leq \frac{1}{N}, k = 1, \dots, d\right\}.$$

We use the slightly enlarged support set of length $2/N$ to simplify the constant computation.

Now we construct a Taylor polynomial of degree s for approximating f_i^ϕ at $\frac{\mathbf{m}}{N}$:

$$P_{\mathbf{m}}(\mathbf{x}) = \sum_{|\mathbf{s}|\le s} \frac{D^{\mathbf{s}} f_i^\phi}{\mathbf{s}!}\bigg|_{\mathbf{x}=\frac{\mathbf{m}}{N}} \left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{\mathbf{s}}.$$

Define $\bar{f}_i = \sum_{\mathbf{m}\in\{0,\dots,N\}^d} \zeta_{\mathbf{m}} P_{\mathbf{m}}$. We bound the approximation error $\left\|\bar{f}_i - f_i^\phi\right\|_\infty$:

$$\begin{aligned}
\max_{\mathbf{x}\in[0,1]^d} \left|\bar{f}_i(\mathbf{x}) - f_i^\phi(\mathbf{x})\right| &= \max_{\mathbf{x}} \left|\sum_{\mathbf{m}} \phi_{\mathbf{m}}(\mathbf{x})(P_{\mathbf{m}}(\mathbf{x}) - f_i^\phi(\mathbf{x}))\right| \\
&\le \max_{\mathbf{x}} \sum_{\mathbf{m}:\left|x_k - \frac{m_k}{N}\right|\le\frac{1}{N}} \left|P_{\mathbf{m}}(\mathbf{x}) - f_i^\phi(\mathbf{x})\right| \\
&\le \max_{\mathbf{x}} 2^d \max_{\mathbf{m}:\left|x_k - \frac{m_k}{N}\right|\le\frac{1}{N}} \left|P_{\mathbf{m}}(\mathbf{x}) - f_i^\phi(\mathbf{x})\right| \\
&\overset{(i)}{\le} \max_{\mathbf{x}} \frac{2^d d^s}{s!}\left(\frac{1}{N}\right)^s \max_{|\mathbf{s}|=s} \left|D^{\mathbf{s}} f_i^\phi\big|_{\frac{\mathbf{m}}{N}} - D^{\mathbf{s}} f_i^\phi\big|_{\mathbf{y}}\right| \\
&\overset{(ii)}{\le} \max_{\mathbf{x}} \frac{2^d d^s}{s!}\left(\frac{1}{N}\right)^s 2^{s+1}\sqrt{d}\mu_i\lambda_i(2r)^{1-\alpha}\left\|\frac{\mathbf{m}}{N} - \mathbf{x}\right\|_2^\alpha \\
&\le \sqrt{d}\mu_i\lambda_i(2r)^{1-\alpha}\frac{2^{d+s+1}d^{s+\alpha/2}}{s!}\left(\frac{1}{N}\right)^{s+\alpha}.
\end{aligned}$$

Here $\mathbf{y}$ is the linear interpolation of $\frac{\mathbf{m}}{N}$ and $\mathbf{x}$, determined by the Taylor remainder, and inequality (i) follows from the Taylor expansion of f_i^ϕ around $\mathbf{m}/N$. Note that only s-th order derivative remains in step (i) and there are at most d^s terms. Inequality (ii) is obtained by the Hölder continuity in the inequality (Eq. A.5).

By setting

$$\sqrt{d}\mu_i\lambda_i(2r)^{1-\alpha}\frac{2^{d+s+1}d^{s+\alpha/2}}{s!}\left(\frac{1}{N}\right)^{s+\alpha} \le \frac{\delta}{2},$$

we get $N \ge \left(\frac{\sqrt{d}\mu_i\lambda_i(2r)^{1-\alpha}2^{d+s+2}d^{s+\alpha/2}}{\delta s!}\right)^{\frac{1}{s+\alpha}}$. Accordingly, the approximation error is bounded by $\|\bar{f}_i - f_i^\phi\|_\infty \le \frac{\delta}{2}$.

Step 2. We next implement $\widetilde{f}_i$ by a ReLU network that approximates $\bar{f}_i$ up to an error

$\frac{\delta}{2}$. We denote

$$P_{\mathbf{m}}(\mathbf{x}) = \sum_{|\mathbf{s}| \leq s} a_{\mathbf{m},\mathbf{s}} \left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{\mathbf{s}},$$

where $a_{\mathbf{m},\mathbf{s}} = \frac{D^{\mathbf{s}} f_i^{\phi}}{\mathbf{s}!}\Big|_{\mathbf{x}=\frac{\mathbf{m}}{N}}$. Then we rewrite $\bar{f}_i$ as

$$\bar{f}_i(\mathbf{x}) = \sum_{\mathbf{m} \in \{0,\ldots,N\}^d} \sum_{|\mathbf{s}| \leq s} a_{\mathbf{m},\mathbf{s}} \zeta_{\mathbf{m}}(\mathbf{x}) \left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{\mathbf{s}}. \tag{A.6}$$

Note that (Eq. A.6) is a linear combination of products $\zeta_{\mathbf{m}} \left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{\mathbf{s}}$. Each product involves at most $d + n$ univariate terms: d terms for $\zeta_{\mathbf{m}}$ and n terms for $\left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{\mathbf{s}}$. We recursively apply Corollary 3.1 to implement the product. Specifically, let $\widehat{\times}_\epsilon$ be the approximation of the product operator in Corollary 3.1 with error ϵ, which will be chosen later. Consider the following chain application of $\widehat{\times}_\epsilon$:

$$\widetilde{f}_{\mathbf{m},\mathbf{s}}(\mathbf{x}) = \widehat{\times}_\epsilon \left(\psi(3Nx_1 - 3m_1), \widehat{\times}_\epsilon \left(\ldots, \widehat{\times}_\epsilon(\psi(3N_d x_d - m_d), \widehat{\times}_\epsilon(x_1 - \frac{m_1}{N}, \ldots))\right)\right).$$

Now we estimate the error of the above approximation. Note that we have $|\psi(3Nx_k - 3m_k)| \leq 1$ and $\left|x_k - \frac{m_k}{N}\right| \leq 1$ for all $k \in \{1, \ldots, d\}$ and $\mathbf{x} \in [0,1]^d$. We then have

$$\begin{aligned}
&\left|\widetilde{f}_{\mathbf{m},\mathbf{s}}(\mathbf{x}) - \zeta_{\mathbf{m}} \left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{\mathbf{s}}\right| \\
=& \left|\widehat{\times}_\epsilon\Big(\psi(3Nx_1 - 3m_1), \widehat{\times}_\epsilon\big(\ldots, \widehat{\times}_\epsilon(x_1 - \frac{m_1}{N}, \ldots)\big)\Big) - \zeta_{\mathbf{m}} \left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{\mathbf{s}}\right| \\
\leq& \left|\widehat{\times}_\epsilon \left(\psi(3Nx_1 - 3m_1), \widehat{\times}_\epsilon(\psi(3Nx_2 - 3m_2), \ldots)\right)\right. \\
&\qquad \left. - \psi(3N_1 - 3m_1)\widehat{\times}_\epsilon(\psi(3Nx_2 - 3m_2), \ldots)\right| \\
&\quad + |\psi(3Nx_1 - m_1)| \left|\widehat{\times}_\epsilon(\psi(3Nx_2 - 3m_2), \ldots) - \psi(3Nx_2 - 3m_2)\widehat{\times}_\epsilon(\ldots)\right| \\
&\quad + \ldots \\
\leq& (s + d)\epsilon.
\end{aligned}$$

Moreover, we have $\widetilde{f}_{\mathbf{m},\mathbf{s}}(\mathbf{x}) = \zeta_{\mathbf{m}}\left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{\mathbf{s}} = 0$, if $\mathbf{x} \notin \operatorname{supp}(\zeta_{\mathbf{m}})$. Now we define

$$\widetilde{f}_i = \sum_{\mathbf{m}\in\{0,\ldots,N\}^d} \sum_{|\mathbf{s}|\leq s} a_{\mathbf{m},\mathbf{s}} \widetilde{f}_{\mathbf{m},\mathbf{s}}.$$

The approximation error is bounded by

$$\begin{aligned}
\max_{\mathbf{x}} \left|\widetilde{f}_i(\mathbf{x}) - \bar{f}_i(\mathbf{x})\right| &= \left| \sum_{\mathbf{m}\in\{0,\ldots,N\}^d} \sum_{|\mathbf{s}|\leq n} a_{\mathbf{m},\mathbf{s}} \left(\widetilde{f}_{\mathbf{m},\mathbf{n}}(\mathbf{x}) - \zeta_{\mathbf{m}}\left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{\mathbf{s}}\right)\right| \\
&\leq \max_{\mathbf{x}} \lambda_i \mu_i 2^{d+s+1} \max_{\mathbf{m}:\mathbf{x}\in\operatorname{supp}(\zeta_{\mathbf{m}})} \sum_{|\mathbf{s}|\leq s} \left|\widetilde{f}_{\mathbf{m},\mathbf{s}}(\mathbf{x}) - \zeta_{\mathbf{m}}\left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{\mathbf{s}}\right| \\
&\leq \lambda_i \mu_i 2^{d+s+1} d^s (d+s)\epsilon.
\end{aligned}$$

We choose $\epsilon = \frac{\delta}{\lambda_i \mu_i 2^{d+s+2} d^s (d+s)}$, so that $\|\bar{f}_i - \widetilde{f}_i\|_\infty \leq \frac{\delta}{2}$. Thus, we eventually have $\|\widetilde{f}_i - f_i^{\phi}\|_\infty \leq \delta$. Now we compute the depth and computational units for implement $\widetilde{f}_i$. $\widetilde{f}_i$ can be implemented by a collection of parallel sub-networks that compute each $\widetilde{f}_{\mathbf{m},\mathbf{s}}$. The total number of parallel sub-networks is bounded by $d^s(N+1)^d$. For each sub-network, we observe that ψ can be exactly implemented by a single layer ReLU network, i.e., $\psi(x) = \mathrm{ReLU}(x+2) - \mathrm{ReLU}(x+1) - \mathrm{ReLU}(x-1) + \mathrm{ReLU}(x-2)$. Corollary 3.1 shows that $\widehat{\times}_\epsilon$ can be implemented by a depth $c_1 \log \frac{1}{\epsilon}$ ReLU network. Therefore, the whole network for implementing $\widetilde{f}_i$ has no more than $c_1'\left(\log\frac{1}{\epsilon} + 1\right)$ layers with width bounded by $O(d^s(N+1)^d)$ and $c_1' d^s (N+1)^d \left(\log\frac{1}{\epsilon} + 1\right)$ neurons and weight parameters. With $\epsilon = \frac{\delta}{\lambda_i \mu_i 2^{d+s+2} d^s (d+s)}$ and $N = \left\lceil \left(\frac{\mu_i \lambda_i (2r)^{1-\alpha} 2^{d+s+2} d^{s+\alpha/2}}{\delta s!}\right)^{\frac{1}{s+\alpha}} \right\rceil$, we obtain that the whole network has no more than $L = c_1 \log\frac{1}{\delta}$ layers, with width bounded by $p = c_2 \delta^{-\frac{d}{s+\alpha}}$, and at most $K = c_2 \delta^{-\frac{d}{s+\alpha}} \left(\log\frac{1}{\delta} + 1\right)$ neurons and weight parameters, for constants c_1, c_2, c_3 depending on d, s, τ, and upper bound of derivatives of $f_i \circ \phi_i^{-1}$, up to order s. Lastly, from (Eq. A.6), we see each parameter has a range bounded by the upper bound of derivatives of $f_i \circ \phi_i^{-1}$ up to order s – scales as $\sqrt{d}$ as in (Eq. A.5). □

A.2.4 Proof of Lemma 3.4

Proof. We expand the estimation error as

$$\begin{aligned}
&\left\|\widehat{f}-f\right\|_\infty \\
&=\left\|\sum_{i=1}^{C_{\mathcal{M}}}\widehat{\times}(\widehat{f}_i,\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)-f\right\|_\infty \\
&=\left\|\sum_{i=1}^{C_{\mathcal{M}}}\widehat{\times}(\widehat{f}_i,\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)-f\rho_i\mathbb{1}(\mathbf{x}\in U_i)\right\|_\infty \\
&\leq\sum_{i=1}^{C_{\mathcal{M}}}\left\|\widehat{\times}(\widehat{f}_i,\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)-f_i\mathbb{1}(\mathbf{x}\in U_i)\right\|_\infty \\
&\leq\sum_{i=1}^{C_{\mathcal{M}}}\Big\|\widehat{\times}(\widehat{f}_i,\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)-\widehat{f}_i\cdot(\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)+\widehat{f}_i\cdot(\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)-f_i\cdot(\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2) \\
&\quad+f_i\cdot(\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)-f_i\cdot\mathbb{1}(\mathbf{x}\in U_i)\Big\|_\infty \\
&\leq\sum_{i=1}^{C_{\mathcal{M}}}\underbrace{\left\|\widehat{\times}(\widehat{f}_i,\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)-\widehat{f}_i\times(\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)\right\|_\infty}_{A_{i,1}}+\underbrace{\left\|\widehat{f}_i\times(\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)-f_i\times(\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)\right\|_\infty}_{A_{i,2}} \\
&\qquad+\underbrace{\left\|f_i\times(\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)-f_i\times\mathbb{1}(\mathbf{x}\in U_i)\right\|_\infty}_{A_{i,3}}.
\end{aligned}$$

The first two terms $A_{i,1}, A_{i,2}$ are straightforward to handle, since by the construction we have

$$\begin{aligned}
A_{i,1}&=\left\|\widehat{\times}(\widehat{f}_i,\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)-\widehat{f}_i\cdot(\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)\right\|_\infty\leq\eta,\quad\text{and}\\
A_{i,2}&=\left\|\widehat{f}_i\times(\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)-f_i\cdot(\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)\right\|_\infty\leq\left\|\widehat{f}_i-f_i\right\|_\infty\left\|\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2\right\|_\infty\leq\delta.
\end{aligned}$$

By Lemma A.1, we have $\max_{\mathbf{x}\in\mathcal{K}_i}|f_i(\mathbf{x})|\leq\frac{c(\pi+1)}{r(1-r/\tau)}\Delta$ for a constant c depending on f_i. Then we bound $A_{i,3}$ as

$$A_{i,3}=\left\|f_i\times(\widehat{\mathbb{1}}_\Delta\circ\widehat{d}_i^2)-f_i\times\mathbb{1}(\mathbf{x}\in U_i)\right\|_\infty\leq\max_{\mathbf{x}\in\mathcal{K}_i}|f_i(\mathbf{x})|\leq\frac{c(\pi+1)}{r(1-r/\tau)}\Delta.$$

□

A.2.5 Helper Lemma for Bounding $A_{i,3}$ and Its Proof

Lemma A.1. *For any $i = 1, \dots, C_{\mathcal{M}}$, denote*

$$\mathcal{K}_i = \{\mathbf{x} \in \mathcal{M} : r^2 - \Delta \leq \|\mathbf{x} - \mathbf{c}_i\|_2^2 \leq r^2\}.$$

Then there exists a constant c depending on the upper bounds of the first derivatives of the partition of unity ρ_i's and coordinate system ϕ_i's such that

$$\max_{\mathbf{x} \in \mathcal{K}_i} |f_i(\mathbf{x})| \leq \frac{c(\pi + 1)}{r(1 - r/\tau)} \Delta.$$

Proof. We extend $f_i \circ \phi_i^{-1}$ to the whole cube $[0,1]^d$ as in the proof of Theorem Theorem 3.3. We also have $f_i(\mathbf{x}) = 0$ for $\|\mathbf{x} - \mathbf{c}_i\|_2 = r$. By the first order Taylor expansion, for any $\mathbf{x}, \mathbf{y} \in U_i$, we have

$$\begin{aligned}
|f_i(\mathbf{x}) - f_i(\mathbf{y})| &= \left|f_i \circ \phi_i^{-1}(\phi_i(\mathbf{x})) - f_i \circ \phi_i^{-1}(\phi_i(\mathbf{y}))\right| \\
&\leq \left\|\nabla(f_i \circ \phi_i^{-1})(\mathbf{z})\right\|_2 \|\phi_i(\mathbf{x}) - \phi_i(\mathbf{y})\|_2 \\
&\leq \left\|\nabla(f_i \circ \phi_i^{-1})(\mathbf{z})\right\|_2 b_i \|V_i\|_2 \|\mathbf{x} - \mathbf{y}\|_2 ,
\end{aligned}$$

where $\mathbf{z}$ is a linear interpolation of $\phi_i(\mathbf{x})$ and $\phi_i(\mathbf{y})$ satisfying the mean value theorem. Since $f_i \circ \phi_i^{-1}$ is C^s in $[0,1]^d$, the first derivative is uniformly bounded, i.e., $\left\|\nabla f_i \circ \phi_i^{-1}(\mathbf{z})\right\|_2 \leq \alpha_i$ for any $\mathbf{z} \in [0,1]^d$. Let $\mathbf{y} \in U_i$ satisfying $f_i(\mathbf{y}) = 0$. In order to bound the function value for any $\mathbf{x} \in \mathcal{K}_i$, we only need to bound the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$. More specifically, for any $\mathbf{x} \in \mathcal{K}_i$, we need to show that there exists $\mathbf{y} \in U_i$ satisfying $f_i(\mathbf{y}) = 0$, such that $\|\mathbf{x} - \mathbf{y}\|_2$ is sufficiently small.

Before continuing with the proof, we introduce some notations. Let $\gamma(t)$ be a geodesic on $\mathcal{M}$ parameterized by the arc length. In the following context, we use $\dot{\gamma}$ and $\ddot{\gamma}$ to denote the first and second derivatives of γ with respect to t. By the definition of geodesic, we have $\|\dot{\gamma}(t)\|_2 = 1$ (unit speed) and $\ddot{\gamma}(t) \perp \dot{\gamma}(t)$.

Without loss of generality, we shift $\mathbf{c}_i$ to $\mathbf{0}$. We consider a geodesic starting from $\mathbf{x}$ with initial "velocity" $\dot{\gamma}(0) = \mathbf{v}$ in the tangent space of $\mathcal{M}$ at $\mathbf{x}$. To utilize polar coordinate, we define two auxiliary quantities: $\ell(t) = \|\gamma(t)\|_2$ and $\theta(t) = \arccos \frac{\gamma(t)^\top \dot{\gamma}(t)}{\|\gamma(t)\|_2} \in [0, \pi]$. As can be seen in Figure Figure A.2, ℓ and θ have clear geometrical interpretations: ℓ is the radial distance from the center $\mathbf{c}_i$, and θ is the angle between the velocity and $\gamma(t)$.

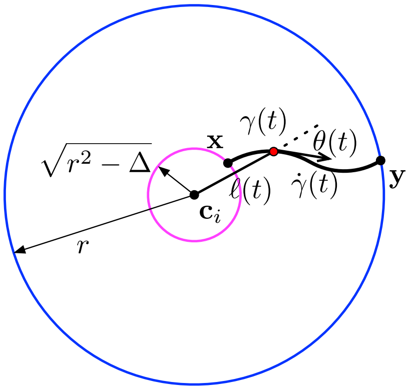


Figure A.2: Illustration of ℓ and θ along a parametric curve γ.

Suppose $\mathbf{y} = \gamma(T)$, we need to upper bound T. Note that $\ell(T) - \ell(0) \leq r - \sqrt{r^2 - \Delta} \leq \Delta / r$. Moreover, observe that the derivative of ℓ is $\dot{\ell}(t) = \cos\theta(t)$, since γ has unit speed. It suffices to find a lower bound on $\dot{\ell}(t) = \cos\theta(t)$ so that $T \leq \frac{\Delta}{r \inf_t \dot{\ell}(t)}$.

We immediately have the second derivative of ℓ as $\ddot{\ell}(t) = -\sin\theta(t)\dot{\theta}(t)$. Meanwhile, using the equation $\ell(t) = \sqrt{\gamma(t)^\top \gamma(t)}$, we also have

$$\ddot{\ell}(t) = \frac{\left(\ddot{\gamma}(t)^\top \gamma(t) + \dot{\gamma}(t)^\top \dot{\gamma}(t)\right) \sqrt{\gamma(t)^\top \gamma(t)} - \left(\gamma(t)^\top \dot{\gamma}(t)\right)^2 / \sqrt{\gamma(t)^\top \gamma(t)}}{\gamma(t)^\top \gamma(t)}. \tag{A.7}$$

Note that by definition, we have $\dot{\gamma}(t)^\top \dot{\gamma}(t) = 1$ and $\gamma(t)^\top \dot{\gamma}(t) = \cos\theta(t) \sqrt{\gamma(t)^\top \gamma(t)}$. Plugging into (Eq. A.7), we can derive

$$\ddot{\ell}(t) = \frac{1 + \ddot{\gamma}(t)^\top \gamma(t) - \cos^2\theta(t)}{\ell(t)} = \frac{\sin^2\theta(t) + \ddot{\gamma}(t)^\top \gamma(t)}{\ell(t)}. \tag{A.8}$$

Now we find a lower bound on $\ddot{\gamma}(t)^\top \gamma(t)$. Specifically, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}\ddot{\gamma}(t)^\top \gamma(t) &\geq - \left\|\ddot{\gamma}(t)\right\|_2 \left\|\gamma(t)\right\|_2 \left|\cos \angle\left(\ddot{\gamma}(t), \gamma(t)\right)\right| \\ &\geq -\frac{r}{\tau} \left|\cos \angle\left(\ddot{\gamma}(t), \gamma(t)\right)\right|.\end{aligned}$$

The last inequality follows from $\left\|\ddot{\gamma}(t)\right\|_2 \leq \frac{1}{\tau}$ [47] and $\left\|\gamma(t)\right\|_2 \leq r$. We now need to bound $\angle(\ddot{\gamma}(t), \gamma(t))$, given $\angle\left(\gamma(t), \dot{\gamma}(t)\right) = \theta(t)$ and $\ddot{\gamma}(t) \perp \dot{\gamma}(t)$. Consider the following optimization problem,

$$\begin{aligned} \min \quad & a^\top x, \\ \text{subject to} \quad & x^\top x = 1, \\ & b^\top x = 0. \end{aligned} \tag{A.9}$$

By assigning $a = \frac{\gamma(t)}{\left\|\gamma(t)\right\|_2}$ and $b = \frac{\dot{\gamma}(t)}{\left\|\dot{\gamma}(t)\right\|_2}$, the optimal objective value is exactly the minimum of $\cos \angle\left(\ddot{\gamma}(t), \gamma\right)$. Additionally, we can find the maximum of $\cos \angle\left(\ddot{\gamma}(t), \gamma\right)$ by replacing the minimization in (Eq. A.9) by maximization. We solve (Eq. A.9) by the Lagrangian method. More precisely, let

$$\mathcal{L}(x, \lambda, \mu) = -a^\top x + \lambda(x^\top x - 1) + \mu(b^\top x).$$

We have the optimal solution x^* satisfying $\nabla_x \mathcal{L} = 0$, which implies $x^* = \frac{1}{2\lambda^*}(a - \mu^* b)$ with μ^* and λ^* being the optimal dual variable. By the primal feasibility, we have $\mu^* = a^\top b$ and $\lambda^* = -\frac{1}{2}\sqrt{1 - (a^\top b)^2}$. Therefore, the optimal objective value is $-\sqrt{1 - (a^\top b)^2}$. Similarly, the maximum is $\sqrt{1 - (a^\top b)^2}$. Note that $a^\top b = \cos \theta(t)$, we then get

$$\ddot{\gamma}(t)^\top \gamma(t) \geq -\frac{r}{\tau} \sin \theta(t).$$

Substituting into (Eq. A.8), we have the following lower bound

$$\ddot{\ell}(t) = \frac{\sin\theta^2(t) + \ddot{\gamma}(t)^\top \gamma(t)}{\ell(t)} \geq \frac{1}{\ell(t)} \left(\sin^2\theta(t) - \frac{r}{\tau}\sin\theta(t) \right).$$

Now combining with $\ddot{\ell}(t) = -\sin\theta(t)\dot{\theta}(t)$, we can derive

$$\dot{\theta}(t) \leq -\frac{1}{\ell(t)} \left(\sin\theta(t) - \frac{r}{\tau} \right). \tag{A.10}$$

Inequality (Eq. A.10) has an important implication: When $\sin\theta(t) > \frac{r}{\tau}$, as t increasing, $\theta(t)$ is monotone decreasing until $\sin\theta(t') = \frac{r}{\tau}$ for some $t' = t$. Thus, we distinguish two cases depending on the value of $\theta(0)$. Indeed, we only need to consider $\theta(0) \in [0, \pi/2]$. The reason behind is that if $\theta(0) \in (\pi/2, \pi]$, we only need to set the initial velocity in the opposite direction.

Case 1: $\theta(0) \in \left[0, \arcsin\frac{r}{\tau}\right]$. We claim that $\theta(t) \in \left[0, \arcsin\frac{r}{\tau}\right]$ for all $t \leq T$. In fact, suppose there exists some $t_1 \leq T$ such that $\theta(t_1) > \arcsin\frac{r}{\tau}$. By the continuity of θ, there exists $t_0 < t_1$, such that $\theta(t_0) = \arcsin\frac{r}{\tau}$ and $\theta(t) \geq \arcsin\frac{r}{\tau}$ for $t \in [t_0, t_1]$. This already gives us a contradiction:

$$\theta(t_0) < \theta(t_1) = \theta(t_0) + \underbrace{\int_{t_0}^{t_1} \dot{\theta}(t)dt}_{\leq 0} \leq \theta(t_0).$$

Therefore, we have $\dot{\ell}(t) \geq \cos\arcsin\frac{r}{\tau} = \sqrt{1 - \frac{r^2}{\tau^2}}$, and thus $T \leq \frac{\Delta}{r\sqrt{1-\frac{r^2}{\tau^2}}}$.

Case 2: $\theta(0) \in \left(\arcsin\frac{r}{\tau}, \pi/2\right]$. It is enough to show that $\theta(0)$ can be bounded sufficiently away from $\pi/2$. Let $\gamma_{\mathbf{c},\mathbf{x}} \subset \mathcal{M}$ be a geodesic from $\mathbf{c}_i$ to $\mathbf{x}$. We analogously define $\theta_{\mathbf{c},\mathbf{x}}$ and $\ell_{\mathbf{c},\mathbf{x}}$ as for the geodesic from $\mathbf{x}$ to $\mathbf{y}$. Let $T_{r/2} = \sup\{t : \ell_{\mathbf{c},\mathbf{x}}(t) \leq r/2 - \Delta/r\}$, and denote $\mathbf{z} = \gamma_{\mathbf{c},\mathbf{x}}(T_{r/2})$. We must have $\theta_{\mathbf{c},\mathbf{x}}(T_{r/2}) \in [0, \pi/2]$ and $\ell_{\mathbf{c},\mathbf{x}}(T_{r/2}) = r/2 - \Delta/r$, otherwise there exists $T'_{r/2} > T_{r/2}$ satisfying $\ell_{\mathbf{c},\mathbf{x}}(T'_{r/2}) \leq r/2$. Denote $T_{\mathbf{x}}$ satisfying

$\mathbf{x} = \gamma_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}})$. We bound $\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}})$ as follows,

$$\begin{aligned}
\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) &= \theta_{\mathbf{c},\mathbf{x}}(T_{r/2}) + \int_{T_{r/2}}^{T_{\mathbf{x}}} \dot{\theta}_{\mathbf{c},\mathbf{x}}(t) dt \\
&\le \frac{\pi}{2} - \int_{T_{r/2}}^{T_{\mathbf{x}}} \frac{1}{\ell_{\mathbf{c},\mathbf{x}}(t)} \left(\sin\theta_{\mathbf{c},\mathbf{x}}(t) - \frac{r}{\tau} \right) dt.
\end{aligned}$$

If there exists some $t \in (T_{r/2}, T_{\mathbf{x}}]$ such that $\sin\theta_{\mathbf{c},\mathbf{x}}(t) \le \frac{r}{\tau}$, by the previous reasoning, we have $\sin\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) \le \frac{r}{\tau}$. Thus, we only need to handle the case when $\sin\theta_{\mathbf{c},\mathbf{x}}(t) > \frac{r}{\tau}$ for all $t \in (T_{r/2}, T_{\mathbf{x}}]$. In this case, $\theta_{\mathbf{c},\mathbf{x}}(t)$ is monotone decreasing, hence we further have

$$\begin{aligned}
\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) &\le \frac{\pi}{2} - \int_{T_{r/2}}^{T_{\mathbf{x}}} \frac{1}{\ell_{\mathbf{c},\mathbf{x}}(t)} \left(\sin\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) - \frac{r}{\tau} \right) dt \\
&\le \frac{\pi}{2} - (T_{\mathbf{x}} - T_{r/2}) \frac{1}{r} \left(\sin\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) - \frac{r}{\tau} \right) \\
&\le \frac{\pi}{2} - \frac{1}{2} \left(\sin\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) - \frac{r}{\tau} \right).
\end{aligned}$$

The last inequality follows from $T_{\mathbf{x}} - T_{r/2} \ge r/2$. Using the fact, $\sin x \ge \frac{2}{\pi} x$, we can derive

$$\begin{aligned}
& \theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) \le \frac{\pi}{2} - \frac{1}{2} \left(\frac{2}{\pi} \theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) - \frac{r}{\tau} \right) \\
\implies \quad & \theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) \le \frac{\pi}{2} \left(\frac{\pi + r/\tau}{\pi + 1} \right).
\end{aligned}$$

We can then set $\theta(0) = \theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}})$, and thus

$$\begin{aligned}
\cos\theta(0) \ge \cos\left(\frac{\pi}{2} \frac{\pi + r/\tau}{\pi + 1} \right) &= \cos\left(\frac{\pi}{2} \left(1 - \frac{1 - r/\tau}{\pi + 1} \right) \right) \\
&= \sin\left(\frac{\pi}{2} \frac{1 - r/\tau}{\pi + 1} \right) \\
&\ge \frac{1 - r/\tau}{\pi + 1}.
\end{aligned}$$

Therefore, we have $T \le \frac{\Delta}{r\cos\theta(0)} \le \frac{\pi+1}{r(1-r/\tau)} \Delta$. By the choice of $r \le \tau/4$, we immediately have $\frac{\tau}{\sqrt{\tau^2-r^2}} < \frac{\pi+1}{1-r/\tau}$. Hence, combining case 1 and case 2, we conclude

$$T \leq \frac{\pi+1}{r(1-r/\tau)}\Delta.$$

Therefore, the function value $f(\mathbf{x})$ on $\mathcal{K}_i$ is bounded by $\alpha_i \frac{\pi+1}{r(1-r/\tau)}\Delta$. It suffices to set $c = \max_i \alpha_i b_i \left\|V_i\right\|_2$, and we complete the proof. □

A.2.6 Characterization of the Size of the ReLU Network

Proof. We evenly split the error ϵ into 3 parts for $A_{i,1}$, $A_{i,2}$, and $A_{i,3}$, respectively. We pick $\eta = \frac{\epsilon}{3C_{\mathcal{M}}}$ so that $\sum_{i=1}^{C_{\mathcal{M}}} A_{i,1} \leq \frac{\epsilon}{3}$. The same argument yields $\delta = \frac{\epsilon}{3C_{\mathcal{M}}}$. Analogously, we can choose $\Delta = \frac{r(1-r/\tau)\epsilon}{3c(\pi+1)C_{\mathcal{M}}}$. Finally, we pick $\nu = \frac{\Delta}{16B^2D}$ so that $8B^2D\nu < \Delta$.

Now we compute the number of layers, width, the number of neurons and weight parameters, and the range of each weight parameter in the ReLU network identified by Theorem 3.1.

1. For the chart determination sub-network, $\widehat{\mathbb{1}}_{\Delta}$ can be implemented by a ReLU network with $\left\lceil \log \frac{r^2}{\Delta} \right\rceil$ layers and 2 neurons in each layer. The weight parameters in the network is bounded by $O(\max\{\tau^2, 1\})$. The approximation of the distance function $\widehat{d_i^2}$ can be implemented by a network of depth $O\left(\log \frac{1}{\nu}\right)$, width bounded by a constant, and the number of neurons and weight parameters is at most $O\left(\log \frac{1}{\nu}\right)$. Each weight parameter is bounded by B. Plugging in our choice of ν and Δ, we have the depth is no greater than $c_1\left(\log \frac{1}{\epsilon} + \log D\right)$ with c_1 depending on d, f, τ, and the surface area of $\mathcal{M}$. The number of neurons and weight parameters is also $c_1'\left(\log \frac{1}{\epsilon} + \log D\right)$ except for a different constant. Note that there are D parallel networks computing $\widehat{d_i^2}$ for $i = 1, \ldots, C_{\mathcal{M}}$. Hence, the total number of neurons and weight parameters is $c_1' C_{\mathcal{M}} D\left(\log \frac{1}{\epsilon} + \log D\right)$ with c_1' depending on d, f, τ, and the surface area of $\mathcal{M}$. As can be seen, the width of the chart-determination network is bounded by $O(C_{\mathcal{M}}D)$, and the weight parameter is bounded by $O(\max\{1, \tau^2, B\})$.

2. For the Taylor polynomial sub-network, ϕ_i can be implemented by a linear network with at most Dd weight parameters. To implement each $\widehat{f}_i$, we need a ReLU network of depth $c_4 \log \frac{1}{\delta}$. The number of neurons and weight parameters is $c_4' \delta^{-\frac{d}{s+\alpha}} \log \frac{1}{\delta}$, and the width is bounded by $c_4'' \delta^{-\frac{d}{s+\alpha}}$. Here c_4, c_4', c_4'' depend on $s, d, \tau, f_i \circ \phi_i^{-1}$. In addition, all the weight parameters are bounded by the upper bound of the derivatives of $f_i \circ \phi_i^{-1}$ up to order s (which scales as $\sqrt{d}$ as in Lemma 3.3). Substituting $\delta = \frac{\epsilon}{3C_{\mathcal{M}}}$, we get the depth is $c_2 \log \frac{1}{\epsilon}$ and the number of neurons and weight parameters is $c_2' \epsilon^{-\frac{d}{s+\alpha}} \log \frac{1}{\epsilon}$. There are totally $C_{\mathcal{M}}$ parallel $\widehat{f}_i$'s, hence the width is further bounded by $c_2'' C_{\mathcal{M}} \epsilon^{-\frac{d}{s+\alpha}}$. Meanwhile, the total number of neurons and weight parameters is $c_2' C_{\mathcal{M}} \epsilon^{-\frac{d}{s+\alpha}} \log \frac{1}{\epsilon}$. Here constants c_2' and c_2'' depend on $d, s, f_i \circ \phi_i^{-1}, \tau$, and the surface area of $\mathcal{M}$.

3. For the product sub-network, the analysis is similar to the chart determination sub-network. The depth is $O\left(\log \frac{1}{\eta}\right)$, the width is bounded by a constant, he number of neurons and weight parameters is $O\left(\log \frac{1}{\eta}\right)$, and all the weight parameters are bounded by a constant. The choice of η yields that the depth is $c_3 \log \frac{1}{\epsilon}$, and the number of neurons and weight parameters is $c_3' \log \frac{1}{\epsilon}$. There are $C_{\mathcal{M}}$ parallel pairs of outputs from the chart determination and the Taylor polynomial sub-networks. Hence, the total number of weight parameters is $c_3' C_{\mathcal{M}} \log \frac{1}{\epsilon}$ with c_3' depending on d, τ, and the surface area of $\mathcal{M}$.

Combining these 3 sub-networks, and redefining the constants c_1, c_2, c_3 and c_4 in the sequel, we obtain that the depth of the full network is $L = c_1 \left(\log \frac{1}{\epsilon} + \log D\right)$ for some constant c_1 depending on d, s, τ, and the surface area of $\mathcal{M}$. The depth of the neural network is bounded by $p = c_2(\epsilon^{-\frac{d}{s+\alpha}} + D)$ with c_2 depending on d, s, τ, the surface area of $\mathcal{M}$, and the upper bounds on derivatives of ϕ_i's and ρ_i's, up to order s. The total number of neurons and weight parameters is $K = c_3 \left(\epsilon^{-\frac{d}{s+\alpha}} \log \frac{1}{\epsilon} + D \log \frac{1}{\epsilon} + D \log D\right)$ for some constant c_3 depending on d, s, f, τ, and the surface area of $\mathcal{M}$. Lastly, all the weight parameters in the network is bounded by $c_4 \max\{1, \tau^2, B, \sqrt{d}\}$ with c_4 depends on the upper bound of derivatives of ρ_i's up to order s. $\square$

APPENDIX B

OMITTED PROOFS IN CHAPTER 4

B.1 Proof of Nonparametric Regression (Theorem 4.1)

This section consists of the detailed proofs, in Appendix B.1.1, Appendix B.1.2 and Appendix B.1.3, respectively, for upper bounding bias in Lemma 4.1, upper bounding variance in Lemma 4.2, and upper bounding covering number in Lemma 4.3. Lastly, the statistical bound in Theorem 4.1 is established in Appendix B.1.4 by choosing a proper approximation error and covering accuracy via the bias-variance trade-off argument.

B.1.1 Proof of Lemma 4.1

Proof. T_1 essentially reflects the bias of estimating f_0:

$$\begin{aligned}
T_1 &= \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i) - \xi_i + \xi_i)^2\right] \\
&= \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i) - \xi_i)^2 + 2\xi_i(\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i) - \xi_i) + \xi_i^2\right] \\
&\overset{(i)}{=} \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i) - \xi_i)^2 + 2\xi_i\widehat{f}_n(\mathbf{x}_i) - \xi_i^2\right] \\
&= \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i) - y_i)^2 + 2\xi_i\widehat{f}_n(\mathbf{x}_i) - \xi_i^2\right] \\
&= \frac{2}{n}\mathbb{E}\left[\inf_{f\in\mathcal{F}(R,\kappa,L,p,K)}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2 + 2\xi_i\widehat{f}_n(\mathbf{x}_i) - \xi_i^2\right] \\
&\overset{(ii)}{\leq} 2\inf_{f\in\mathcal{F}(R,\kappa,L,p,K)}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - f_0(\mathbf{x}_i) - \xi_i)^2 - \xi_i^2\right] + \mathbb{E}\left[\frac{4}{n}\sum_{i=1}^{n}\xi_i\widehat{f}_n(\mathbf{x}_i)\right] \\
&= 2\inf_{f\in\mathcal{F}(R,\kappa,L,p,K)}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2 - 2\xi_i(f(\mathbf{x}_i) - f_0(\mathbf{x}_i))\right] + \mathbb{E}\left[\frac{4}{n}\sum_{i=1}^{n}\xi_i\widehat{f}_n(\mathbf{x}_i)\right]
\end{aligned}$$

$$= 2 \inf_{f \in \mathcal{F}(R,\kappa,L,p,K)} \int_{\mathcal{M}} (f(\mathbf{x}) - f_0(\mathbf{x}))^2 d\mathcal{D}_x(\mathbf{x}) + \mathbb{E}\left[\frac{4}{n}\sum_{i=1}^{n} \xi_i \widehat{f}_n(\mathbf{x}_i)\right], \tag{B.1}$$

where (i) follows from $\mathbb{E}[\xi_i f_0(\mathbf{x}_i)] = 0$ due to the independence between ξ_i and $\mathbf{x}$, and (ii) follows from Jensen's inequality. Now we need to bound $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \xi_i \widehat{f}_n(\mathbf{x}_i)\right]$. We discretize the class $\mathcal{F}(R,\kappa,L,p,K)$ into $\mathcal{F}^*(R,\kappa,L,p,K) = \{f_i^*\}_{i=1}^{\mathcal{N}(\delta,\mathcal{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)}$, where $\mathcal{N}(\delta,\mathcal{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)$ denotes the δ-covering number with respect to the ℓ_∞ norm. Accordingly, there exists f^* such that $\|f^* - \widehat{f}_n\|_\infty \leq \delta$. Denote $\|\widehat{f}_n - f_0\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2$. Then we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \xi_i \widehat{f}_n(\mathbf{x}_i)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \xi_i(\widehat{f}_n(\mathbf{x}_i) - f^*(\mathbf{x}_i) + f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))\right] \\
&\overset{(i)}{\leq} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))\right] + \delta\sigma \\
&= \mathbb{E}\left[\frac{\|f^* - f_0\|_n}{\sqrt{n}} \frac{\sum_{i=1}^{n} \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\,\|f^* - f_0\|_n}\right] + \delta\sigma \\
&\overset{(ii)}{\leq} \sqrt{2}\mathbb{E}\left[\frac{\|\widehat{f}_n - f_0\|_n + \delta}{\sqrt{n}} \left|\frac{\sum_{i=1}^{n} \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\,\|f^* - f_0\|_n}\right|\right] + \delta\sigma. \quad \text{(B.2)}
\end{aligned}$$

Here (i) is obtained by applying Hölder's inequality to $\xi_i(\widehat{f}_n(\mathbf{x}_i) - f^*(\mathbf{x}_i))$ and invoking the Jensen's inequality:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \xi_i(\widehat{f}_n(\mathbf{x}_i) - f^*(\mathbf{x}_i))\right] &\leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} |\xi_i| \left\|f^* - \widehat{f}_n\right\|_\infty\right] \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[|\xi_i|]\delta \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \sqrt{\mathbb{E}[|\xi_i|^2]}\delta \\
&\leq \delta\sigma.
\end{aligned}$$

Step (ii) holds, since by invoking the inequality $2ab \leq a^2 + b^2$, we have

$$\begin{aligned}
\|f^* - f_0\|_n &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f^*(\mathbf{x}_i) - \widehat{f}_n(\mathbf{x}_i) + \widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2} \\
&\leq \sqrt{\frac{2}{n}\sum_{i=1}^{n}(f^*(\mathbf{x}_i) - \widehat{f}_n(\mathbf{x}_i))^2 + (\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2} \\
&\leq \sqrt{\frac{2}{n}\sum_{i=1}^{n}\left[\delta^2 + \widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2\right]} \\
&\leq \sqrt{2}\left\|\widehat{f}_n - f_0\right\|_n + \sqrt{2}\delta.
\end{aligned}$$

To bound the expectation term in (Eq. B.2), we first break the dependence between f^* and the samples $(\mathbf{x}_i, y_i)$. In detail, we replace f^* by any f_j^* in the δ-covering, and observe that $\frac{\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f^* - f_0\|_n} \leq \max_j \frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^* - f_0\|_n}$. For notational simplicity, we denote $z_j = \frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^* - f_0\|_n}$. Applying Cauchy-Schwarz inequality, we cast the expectation term in (Eq. B.2) as

$$\begin{aligned}
&\mathbb{E}\left[\frac{\|\widehat{f}_n - f_0\|_n + \delta}{\sqrt{n}}\left|\frac{\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\,\|f^* - f_0\|_n}\right|\right] \\
&\leq \mathbb{E}\left[\frac{\|\widehat{f}_n - f_0\|_n + \delta}{\sqrt{n}}\max_j |z_j|\right] \\
&= \mathbb{E}\left[\frac{\|\widehat{f}_n - f_0\|_n}{\sqrt{n}}\max_j |z_j| + \frac{\delta}{\sqrt{n}}\max_j |z_j|\right] \\
&\leq \mathbb{E}\left[\left(\sqrt{\frac{1}{n}\mathbb{E}\left[\|\widehat{f}_n - f_0\|_n^2\right]} + \frac{\delta}{\sqrt{n}}\right)\sqrt{\mathbb{E}\left[\max_j \ z_j^2\right]}\right]. \qquad \text{(B.3)}
\end{aligned}$$

For given $\mathbf{x}_1, \ldots, \mathbf{x}_n$, each term $\frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^* - f_0\|_n}$ is sub-guassian with parameter σ. Consequently, the last inequality (Eq. B.3) involves the maximum of a collection of squared sub-Gaussian random variables z_j^2. Indeed, z_j^2 is sub-exponential for each j. We can bound it using the moment generating function: For any $t > 0$, we have

$$\begin{aligned}
\mathbb{E}\left[\max_j \ z_j^2 \mid \mathbf{x}_1, \ldots, \mathbf{x}_n\right] &= \frac{1}{t} \log \exp\left(t\mathbb{E}[\max_j \ z_j^2 \mid \mathbf{x}_1, \ldots, \mathbf{x}_n]\right) \\
&\overset{(i)}{\le} \frac{1}{t} \log \mathbb{E}\left[\exp\left(t \max_j \ z_j^2\right) | \mathbf{x}_1, \ldots, \mathbf{x}_n\right] \\
&\le \frac{1}{t} \log \mathbb{E}\left[\sum_j \exp\left(t z_j^2\right) | \mathbf{x}_1, \ldots, \mathbf{x}_n\right] \\
&\le \frac{1}{t} \log \mathcal{N}(\delta, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty) \\
&\quad + \frac{1}{t} \log \mathbb{E}[\exp(t z_1^2) | \mathbf{x}_1, \ldots, \mathbf{x}_n]. \qquad \text{(B.4)}
\end{aligned}$$

Since z_1 is σ^2-sub-Gaussian given $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we derive

$$\begin{aligned}
\mathbb{E}[\exp(t z_1^2) | \mathbf{x}_1, \ldots, \mathbf{x}_n] &= 1 + \sum_{p=1}^{\infty} \frac{t^p \mathbb{E}[z_1^{2p} | \mathbf{x}_1, \ldots, \mathbf{x}_n]}{p!} \\
&= 1 + \sum_{p=1}^{\infty} \left[\frac{t^p}{p!} \int_0^\infty \mathbb{P}(|z_1| \ge u^{1/2p}) du\right] \\
&\le 1 + 2 \sum_{p=1}^{\infty} \left[\frac{t^p}{p!} \int_0^\infty \exp\left(-\frac{u^{1/p}}{2\sigma^2}\right) du\right] \\
&= 1 + 2 \sum_{p=1}^{\infty} (2t\sigma^2)^p.
\end{aligned}$$

Taking $t = (3\sigma^2)^{-1}$ and substituting into (Eq. B.4), we deduce $\mathbb{E}\left[\max_j \ z_j^2 \mid \mathbf{x}_1, \ldots, \mathbf{x}_n\right]$ is bounded by

$$\begin{aligned}
\mathbb{E}\left[\max_j \ z_j^2 \mid \mathbf{x}_1, \ldots, \mathbf{x}_n\right] &\le 3\sigma^2 \log \mathcal{N}(\delta, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty) + 3\sigma^2 \log 5 \\
&\le 3\sigma^2 \log \mathcal{N}(\delta, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty) + 6\sigma^2. \qquad \text{(B.5)}
\end{aligned}$$

Combining (Eq. B.5), (Eq. B.3), (Eq. B.2), and substituting back into (Eq. B.1), we obtain the following implicit error estimation on T_1:

$$
\begin{aligned}
T_1 &= 2\mathbb{E}\left[\|\widehat{f}_n - f_0\|_n^2\right] \\
&\leq 2 \inf_{f \in \mathcal{F}(R,\kappa,L,p,K)} \int_{\mathcal{M}} (f(\mathbf{x}) - f_0(\mathbf{x}))^2 d\mathcal{D}_x(\mathbf{x}) + 4\delta\sigma \\
&\quad + 4\sqrt{6}\sigma \left(\sqrt{\mathbb{E}\left[\|\widehat{f}_n - f_0\|_n^2\right]} + \delta \right) \sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty) + 2}{n}}.
\end{aligned}
$$

We denote $v = \sqrt{\mathbb{E}\left[\|\widehat{f}_n - f_0\|_n^2\right]}$. Then the above implicit bound on T_1 implies

$$
\begin{aligned}
& v^2 \leq b + 2av && \text{(B.6)} \\
\text{with} \quad & a = \sqrt{6}\sigma \sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty) + 2}{n}}, \\
& b = \inf_{f \in \mathcal{F}(R,\kappa,L,p,K)} \int_{\mathcal{M}} (f(\mathbf{x}) - f_0(\mathbf{x}))^2 d\mathcal{D}_x(\mathbf{x}) \\
& \qquad + \left(2\sqrt{6} \sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty) + 2}{n}} + 2 \right) \sigma\delta.
\end{aligned}
$$

Rearranging (Eq. B.6) for $a, b > 0$, we deduce $(v-a)^2 \leq b + a^2$. Some manipulation then yields $v^2 \leq 4a^2 + 2b$, which implies

$$
\begin{aligned}
T_1 = 2v^2 \leq 4 &\inf_{f \in \mathcal{F}(R,\kappa,L,p,K)} \int_{\mathcal{M}} (f(\mathbf{x}) - f_0(\mathbf{x}))^2 d\mathcal{D}_x(\mathbf{x}) \\
&+ 48\sigma^2 \frac{\log \mathcal{N}(\delta, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty) + 2}{n} \\
&+ \left(8\sqrt{6} \sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty) + 2}{n}} + 8 \right) \sigma\delta.
\end{aligned}
$$

The proof is complete. □

B.1.2 Proof of Lemma 4.2

Proof. Recall that we denote $\widehat{g}(\mathbf{x}) = (\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x}))^2$. We rewrite T_2 as

$$\begin{aligned}
T_2 &= \mathbb{E}\left[\int_{\mathcal{M}} \widehat{g}(\mathbf{x}) d\mathcal{D}_x(\mathbf{x}) - \frac{2}{n}\sum_{i=1}^{n} \widehat{g}(\mathbf{x}_i)\right] \\
&= 2\mathbb{E}\left[\int_{\mathcal{M}} \widehat{g}(\mathbf{x}) d\mathcal{D}_x(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} \widehat{g}(\mathbf{x}_i) - \frac{1}{2}\int_{\mathcal{M}} \widehat{g}(\mathbf{x}) d\mathcal{D}_x(\mathbf{x})\right] \\
&\leq 2\mathbb{E}\left[\int_{\mathcal{M}} \widehat{g}(\mathbf{x}) d\mathcal{D}_x(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} \widehat{g}(\mathbf{x}_i) - \frac{1}{8R^2}\int_{\mathcal{M}} \widehat{g}^2(\mathbf{x}) d\mathcal{D}_x(\mathbf{x})\right].
\end{aligned}$$

We lower bound $\int_{\mathcal{M}} \widehat{g}(\mathbf{x}) d\mathcal{D}_x(\mathbf{x})$ by its second moment:

$$\begin{aligned}
\int_{\mathcal{M}} \widehat{g}(\mathbf{x}) d\mathcal{D}_x(\mathbf{x}) &= \int_{\mathcal{M}} \left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^4 d\mathcal{D}_x(\mathbf{x}) \\
&= \int_{\mathcal{M}} \left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 \widehat{g}(\mathbf{x}) d\mathcal{D}_x(\mathbf{x}) \\
&\leq \int_{\mathcal{M}} 4R^2 \widehat{g}(\mathbf{x}) d\mathcal{D}_x(\mathbf{x}).
\end{aligned}$$

The last inequality follows from $\left|\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right| \leq 2R$. Now we cast T_2 into

$$T_2 \leq 2\mathbb{E}\left[\int_{\mathcal{M}} \widehat{g}(\mathbf{x}) d\mathcal{D}_x(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} \widehat{g}(\mathbf{x}_i) - \frac{1}{8R^2}\int_{\mathcal{M}} \widehat{g}^2(\mathbf{x}) d\mathcal{D}_x(\mathbf{x})\right]. \tag{B.7}$$

Introducing the second moment allows us to establish a fast convergence of T_2. Specifically, we denote $\bar{\mathbf{x}}_i$'s as independent copies of $\mathbf{x}_i$'s following the same distribution. We also denote

$$\mathcal{G} = \left\{g(\mathbf{x}) = (f(\mathbf{x}) - f_0(\mathbf{x}))^2 \ \middle|\ f \in \mathcal{F}(R, \kappa, L, p, K)\right\}$$

as the function class induced by $\mathcal{F}(R, \kappa, L, p, K)$. Then we upper bound (Eq. B.7) as

$$T_2 \leq 2\mathbb{E}\left[\sup_{g \in \mathcal{G}} \left(\int_{\mathcal{M}} g(\bar{\mathbf{x}}) d\mathcal{D}_x(\bar{\mathbf{x}}) - \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) - \frac{1}{8R^2}\int_{\mathcal{M}} g^2(\mathbf{x}) d\mathcal{D}_x(\mathbf{x})\right)\right]$$

$$\overset{(i)}{\leq} 2\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}}}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}(g(\bar{\mathbf{x}}_i)-g(\mathbf{x}_i))-\frac{1}{16R^2}\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}}}[g^2(\bar{\mathbf{x}})+g^2(\mathbf{x})]\right], \tag{B.8}$$

where (i) follows from Jensen's inequality and shorthand $\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}}}[\cdot]$ denotes the expectation (double integral $\int_{\mathcal{M}}\int_{\mathcal{M}}\cdot d\mathcal{D}_x(\mathbf{x})d\mathcal{D}_x(\bar{\mathbf{x}})$) with respect to the joint distribution of $(\mathbf{x},\bar{\mathbf{x}})$.

We discretize $\mathcal{G}$ with respect to the ℓ_∞ norm. The δ-covering number is denoted as $\mathcal{N}(\delta,\mathcal{G},\|\cdot\|_\infty)$ and the elements in the covering is denoted as $\mathcal{G}^*=\{g_i^*\}_{i=1}^{\mathcal{N}(\delta,\mathcal{G},\|\cdot\|_\infty)}$, that is, for any $g\in\mathcal{G}$, there exists a g^* satisfying $\|g-g^*\|_\infty\leq\delta$.

We replace $g\in\mathcal{G}$ by $g^*\in\mathcal{G}^*$ in bounding T_2, which then boils down to deriving concentration results on a finite concept class. Specifically, for g^* satisfying $\|g-g^*\|_\infty\leq\delta$, we have

$$\begin{aligned} g(\bar{\mathbf{x}}_i)-g(\mathbf{x}_i) &= g(\bar{\mathbf{x}}_i)-g^*(\bar{\mathbf{x}}_i)+g^*(\bar{\mathbf{x}}_i)-g^*(\mathbf{x}_i)+g^*(\mathbf{x}_i)-g(\mathbf{x}_i)\\ &\leq g^*(\bar{\mathbf{x}}_i)-g^*(\mathbf{x}_i)+2\delta. \end{aligned}$$

We also have

$$\begin{aligned} & g^2(\bar{\mathbf{x}})+g^2(\mathbf{x})\\ &= \left[g^2(\bar{\mathbf{x}})-(g^*)^2(\bar{\mathbf{x}})\right]+\left[(g^*)^2(\bar{\mathbf{x}})+(g^*)^2(\mathbf{x})\right]-\left[(g^*)^2(\mathbf{x})-g^2(\mathbf{x})\right]\\ &= (g^*)^2(\bar{\mathbf{x}})+(g^*)^2(\mathbf{x})+(g(\bar{\mathbf{x}})-g^*(\bar{\mathbf{x}}))(g(\bar{\mathbf{x}})+g^*(\bar{\mathbf{x}}))+(g^*(\mathbf{x})-g(\mathbf{x}))(g^*(\mathbf{x})+g(\mathbf{x}))\\ &\geq (g^*)^2(\bar{\mathbf{x}})+(g^*)^2(\mathbf{x})-|g(\bar{\mathbf{x}})-g^*(\bar{\mathbf{x}})|\,|g(\bar{\mathbf{x}})+g^*(\bar{\mathbf{x}})|-|g^*(\mathbf{x})-g(\mathbf{x})|\,|g^*(\mathbf{x})+g(\mathbf{x})|\\ &\geq (g^*)^2(\bar{\mathbf{x}})+(g^*)^2(\mathbf{x})-2R\delta-2R\delta. \end{aligned}$$

Plugging the above two items into (Eq. B.8), we upper bound T_2 as

$$\begin{aligned} T_2 &\leq 2\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}}}\left[\sup_{g^*\in\mathcal{G}^*}\frac{1}{n}\sum_{i=1}^{n}\left(g^*(\bar{\mathbf{x}}_i)-g^*(\mathbf{x}_i)\right)-\frac{1}{16R^2}\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}}}[(g^*)^2(\bar{\mathbf{x}})+(g^*)^2(\mathbf{x})]\right]+\left(4+\frac{1}{2R}\right)\delta\\ &= 2\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}}}\left[\max_{j}\frac{1}{n}\sum_{i=1}^{n}\left(g_j^*(\bar{\mathbf{x}}_i)-g_j^*(\mathbf{x}_i)\right)-\frac{1}{16R^2}\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}}}[(g_j^*)^2(\bar{\mathbf{x}})+(g_j^*)^2(\mathbf{x})]\right]+\left(4+\frac{1}{2R}\right)\delta. \end{aligned}$$

Denote $h_j(i) = g_j^*(\bar{\mathbf{x}}_i) - g_j^*(\mathbf{x}_i)$. By symmetry, it is straightforward to see $\mathbb{E}[h_j(i)] = 0$. The variance of $h_j(i)$ is computed as

$$\mathrm{Var}[h_j(i)] = \mathbb{E}\left[h_j^2(i)\right] = \mathbb{E}\left[\left(g_j^*(\bar{\mathbf{x}}_i) - g_j^*(\mathbf{x}_i)\right)^2\right] \overset{(i)}{\leq} 2\mathbb{E}\left[(g_j^*)^2(\bar{\mathbf{x}}_i) + (g_j^*)^2(\mathbf{x}_i)\right].$$

The last inequality (i) utilizes the identity $(a-b)^2 \leq 2(a^2+b^2)$. Therefore, we derive the following upper bound for T_2,

$$T_2 \leq 2\mathbb{E}\left[\max_j \frac{1}{n}\sum_{i=1}^n h_j(i) - \frac{1}{32R^2}\frac{1}{n}\sum_{i=1}^n \mathrm{Var}[h_j(i)]\right] + \left(4 + \frac{1}{2R}\right)\delta.$$

We invoke the moment generating function to bound T_2. Note that we have $\|h_j\|_\infty \leq (2R)^2$. Then by Taylor expansion, for $0 < t/n < \frac{3}{4R^2}$ and any j, we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{t}{n}h_j(i)\right)\right] &= \mathbb{E}\left[1 + \frac{t}{n}h_j(i) + \sum_{k=2}^\infty \frac{(t/n)^k h_j^k(i)}{k!}\right] \\
&\leq \mathbb{E}\left[1 + \frac{t}{n}h_j(i) + \sum_{k=2}^\infty \frac{(t/n)^k h_j^2(i)(4R^2)^{k-2}}{2 \times 3^{k-2}}\right] \\
&= \mathbb{E}\left[1 + \frac{t}{n}h_j(i) + \frac{(t/n)^2 h_j^2(i)}{2}\sum_{k=2}^\infty \frac{(t/n)^{k-2}(4R^2)^{k-2}}{3^{k-2}}\right] \\
&= \mathbb{E}\left[1 + \frac{t}{n}h_j(i) + \frac{(t/n)^2 h_j^2(i)}{2}\frac{1}{1 - 4tR^2/(3n)}\right] \\
&= 1 + (t/n)^2\,\mathrm{Var}[h_j(i)]\frac{1}{2 - 8tR^2/(3n)} \\
&\overset{(i)}{\leq} \exp\left(\mathrm{Var}[h_j(i)]\frac{3(t/n)^2}{6 - 8tR^2/n}\right). \qquad \text{(B.9)}
\end{aligned}$$

Step (i) follows from the fact $1 + x \leq \exp(x)$ for $x \geq 0$. Given (Eq. B.9), we proceed to bound T_2. To ease the presentation, we temporarily neglect $\left(4 + \frac{1}{2R}\right)\delta$ term and denote $T_2' = T_2 - \left(4 + \frac{1}{2R}\right)\delta$. Then for $0 < t/n < \frac{3}{4R^2}$, we have

$$\exp\left(t\frac{T_2'}{2}\right) = \exp\left(t\mathbb{E}\left[\max_j \frac{1}{n}\sum_{i=1}^n h_j(i) - \frac{1}{32R^2}\frac{1}{n}\sum_{i=1}^n \mathrm{Var}[h_j(i)]\right]\right)$$

$$
\begin{aligned}
&\overset{(i)}{\leq} \mathbb{E}\left[\exp\left(t\max_j \frac{1}{n}\sum_{i=1}^n h_j(i) - \frac{1}{32R^2}\frac{1}{n}\sum_{i=1}^n \mathrm{Var}[h_j(i)]\right)\right] \\
&\leq \mathbb{E}\left[\sum_j \exp\left(\frac{t}{n}\sum_{i=1}^n h_j(i) - \frac{1}{32R^2}\frac{t}{n}\sum_{i=1}^n \mathrm{Var}[h_j(i)]\right)\right] \\
&\overset{(ii)}{\leq} \mathbb{E}\left[\sum_j \exp\left(\sum_{i=1}^n \mathrm{Var}[h_j(i)]\frac{3(t/n)^2}{6-8tR^2/n} - \frac{1}{32R^2}\frac{t}{n}\mathrm{Var}[h_j(i)]\right)\right] \\
&= \mathbb{E}\left[\sum_j \exp\left(\sum_{i=1}^n \frac{t}{n}\mathrm{Var}[h_j(i)]\left(\frac{3t/n}{6-8tR^2/n} - \frac{1}{32R^2}\right)\right)\right].
\end{aligned}
$$

Step (i) follows from Jensen's inequality, and step (ii) invokes (Eq. B.9) for each $h(i)$. We now choose t so that $\frac{3t/n}{6-8tR^2/n} - \frac{1}{32R^2} = 0$, which yields $t = \frac{3n}{52R^2} < \frac{3n}{4R^2}$. Substituting our choice of t into $\exp(tT_2'/2)$, we have

$$
t\frac{T_2'}{2} \leq \log\sum_j \exp(0) \Longrightarrow T_2' \leq \frac{2}{t}\log\mathcal{N}(\delta, \mathcal{G}, \|\cdot\|_\infty) = \frac{104R^2}{3n}\log\mathcal{N}(\delta, \mathcal{G}, \|\cdot\|_\infty).
$$

To complete the proof, we relate the covering number of $\mathcal{G}$ to that of $\mathcal{F}(R, \kappa, L, p, K)$. Consider any $g_1, g_2 \in \mathcal{G}$ with $g_1 = (f_1 - f_0)^2$ and $g_2 = (f_2 - f_0)^2$, respectively, for $f_1, f_2 \in \mathcal{F}(R, \kappa, L, p, K)$. We can derive

$$
\begin{aligned}
\|g_1 - g_2\|_\infty &= \sup_{\mathbf{x}} \left|(f_1(\mathbf{x}) - f_0(\mathbf{x}))^2 - (f_2(\mathbf{x}) - f_0(\mathbf{x}))^2\right| \\
&= \sup_{\mathbf{x}} |f_1(\mathbf{x}) - f_2(\mathbf{x})|\, |f_1(\mathbf{x}) + f_2(\mathbf{x}) - 2f_0(\mathbf{x})| \\
&\leq 4R\|f_1 - f_2\|_\infty.
\end{aligned}
$$

The above characterization immediately implies $\mathcal{N}(\delta, \mathcal{G}, \|\cdot\|_\infty) \leq \mathcal{N}(\delta/4R, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty)$. Therefore, we derive the desired upper bound on T_2:

$$
T_2 \leq \frac{104R^2}{3n}\log\mathcal{N}(\delta/4R, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty) + \left(4 + \frac{1}{2R}\right)\delta.
$$

□

B.1.3 Proof of Lemma 4.3

Proof. To construct a covering for $\mathcal{F}(R, \kappa, L, p, K)$, we discretize each weight parameter by a uniform grid with grid size h. Recall we write $f \in \mathcal{F}(R, \kappa, L, p, K)$ as $f = W_L \cdot \mathrm{ReLU}(W_{L-1} \cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1}) + \mathbf{b}_L$. Let $f, f' \in \mathcal{F}$ with all the weight parameters at most h from each other. Denoting the weight matrices in f, f' as $W_L, \dots, W_1, \mathbf{b}_L, \dots, \mathbf{b}_1$ and $W_L', \dots, W_1', \mathbf{b}_L', \dots, \mathbf{b}_1'$, respectively, we bound the ℓ_∞ difference $\|f - f'\|_\infty$ as

$$\begin{aligned}
&\|f - f'\|_\infty \\
&= \big\| W_L \cdot \mathrm{ReLU}(W_{L-1} \cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1}) + \mathbf{b}_L \\
&\quad - (W_L' \cdot \mathrm{ReLU}(W_{L-1}' \cdots \mathrm{ReLU}(W_1' \mathbf{x} + \mathbf{b}_1') \cdots + \mathbf{b}_{L-1}') - \mathbf{b}_L') \big\|_\infty \\
&\leq \|\mathbf{b}_L - \mathbf{b}_L'\|_\infty + \|W_L - W_L'\|_1 \|W_{L-1} \cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1}\|_\infty \\
&\quad + \|W_L\|_1 \big\| W_{L-1} \cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1} - (W_{L-1}' \cdots \mathrm{ReLU}(W_1' \mathbf{x} + \mathbf{b}_1') \cdots + \mathbf{b}_{L-1}') \big\|_\infty \\
&\leq h + hp \|W_{L-1} \cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1}\|_\infty \\
&\quad + \kappa p \big\| W_{L-1} \cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1} - (W_{L-1}' \cdots \mathrm{ReLU}(W_1' \mathbf{x} + \mathbf{b}_1') \cdots + \mathbf{b}_{L-1}') \big\|_\infty .
\end{aligned}$$

We derive the following bound on $\|W_{L-1} \cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1}\|_\infty$:

$$\begin{aligned}
&\|W_{L-1} \cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1}\|_\infty \\
&\leq \|W_{L-1}(\cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots)\|_\infty + \|\mathbf{b}_{L-1}\|_\infty \\
&\leq \|W_{L-1}\|_1 \|W_{L-2}(\cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots) + \mathbf{b}_{L-2}\|_\infty + \kappa \\
&\leq \kappa p \|W_{L-2}(\cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots) + \mathbf{b}_{L-2}\|_\infty + \kappa \\
&\overset{(i)}{\leq} (\kappa p)^{L-1} B + \kappa \sum_{i=0}^{L-3} (\kappa p)^i \\
&\leq (\kappa p)^{L-1} B + \kappa (\kappa p)^{L-2},
\end{aligned}$$

where (i) is obtained by induction and $\|\mathbf{x}\|_\infty \leq B$. The last inequality holds, since $\kappa p > 1$. Substituting back into the bound for $\|f - f'\|_\infty$, we have

$$
\begin{aligned}
& \|f - f'\|_\infty \\
\leq\ & \kappa p \left\| W_{L-1} \cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1} - (W'_{L-1} \cdots \mathrm{ReLU}(W'_1 \mathbf{x} + \mathbf{b}'_1) \cdots + \mathbf{b}'_{L-1}) \right\|_\infty \\
& + h + hp \left[(\kappa p)^{L-1} B + \kappa (\kappa p)^{L-2} \right] \\
\leq\ & \kappa p \left\| W_{L-1} \cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1} - (W'_{L-1} \cdots \mathrm{ReLU}(W'_1 \mathbf{x} + \mathbf{b}'_1) \cdots + \mathbf{b}'_{L-1}) \right\|_\infty \\
& + h(pB + 2)(\kappa p)^{L-1} \\
\overset{(i)}{\leq}\ & (\kappa p)^{L-1} \left\| W_1 \mathbf{x} + \mathbf{b}_1 - W'_1 \mathbf{x} - \mathbf{b}'_1 \right\|_\infty + h(L-1)(pB+2)(\kappa p)^{L-1} \\
\leq\ & hL(pB+2)(\kappa p)^{L-1},
\end{aligned}
$$

where (i) is obtained by induction. We choose h satisfying $hL(pB + 2)(\kappa p)^{L-1} = \delta$. Then discretizing each parameter uniformly into $2\kappa/h$ grid points yields a δ-covering on $\mathcal{F}$. Note that there are $\binom{Lp^2}{K} \leq (Lp^2)^K$ different choices of K non-zero entries out of Lp^2 total weight parameters. Therefore, the covering number is upper bounded by

$$
\mathcal{N}(\delta, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty) \leq (Lp^2)^K \left(\frac{2\kappa}{h} \right)^K \leq \left(\frac{2L^2(pB+2)\kappa^L p^{L+1}}{\delta} \right)^K .
$$

□

B.1.4 Proof of Bias-variance Trade-off

Proof of Theorem 4.1. We recall the bias and variance decomposition

$$
\begin{aligned}
& \mathbb{E}\left[\int_{\mathcal{M}} \left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x}) \right)^2 d\mathcal{D}_x(\mathbf{x}) \right] \\
= & \underbrace{\mathbb{E}\left[\frac{2}{n} \sum_{i=1}^{n} (\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2 \right]}_{T_1}
\end{aligned}
$$

$$\underbrace{+\,\mathbb{E}\left[\int_{\mathcal{M}}\left(\widehat{f}_n(\mathbf{x})-f_0(\mathbf{x})\right)^2 d\mathcal{D}_x(\mathbf{x})\right]-\mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i)-f_0(\mathbf{x}_i))^2\right]}_{T_2}.$$

Combining the upper bounds on T_1 and T_2 in Lemma 4.1 and 4.2, we can derive

$$\begin{aligned}
\mathbb{E}\left[\int_{\mathcal{M}}\left(\widehat{f}_n(\mathbf{x})-f_0(\mathbf{x})\right)^2 d\mathcal{D}_x(\mathbf{x})\right] \leq\ & 4\inf_{f\in\mathcal{F}(R,\kappa,L,p,K)}\int_{\mathcal{M}}(f(\mathbf{x})-f_0(\mathbf{x}))^2 d\mathcal{D}_x(\mathbf{x})\\
&+48\sigma^2\frac{\log\mathcal{N}(\delta,\mathcal{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)+2}{n}\\
&+8\sqrt{6}\sqrt{\frac{\log\mathcal{N}(\delta,\mathcal{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)+2}{n}}\sigma\delta\\
&+\frac{104R^2}{3n}\log\mathcal{N}(\delta/4R,\mathcal{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)\\
&+\left(4+\frac{1}{2R}+8\sigma\right)\delta.
\end{aligned}$$

By our choice of $\mathcal{F}(R,\kappa,L,p,K)$, there exists a network class which can yield a function f satisfying $\|f-f_0\|_\infty\leq\epsilon$ for $\epsilon\in(0,1)$. We will choose ϵ later for the bias-variance tradeoff. Such a network consists of $L=\widetilde{O}\left(\log\frac{1}{\epsilon}\right)$ layers and $K=\widetilde{O}\left(\left(\epsilon^{-\frac{d}{s+\alpha}}+D\right)\log\frac{1}{\epsilon}\right)$ weight parameters. Invoking the upper bound of the covering number in Lemma 4.3, we derive

$$\begin{aligned}
&\mathbb{E}\left[\int_{\mathcal{M}}\left(\widehat{f}_n(\mathbf{x})-f_0(\mathbf{x})\right)^2 d\mathcal{D}_x(\mathbf{x})\right]\\
&\leq 4\epsilon^2+\frac{48\sigma^2}{n}\left(K\log\left(2R^2L^2(pB+2)\kappa^Lp^{L+1}/\delta\right)+2\right)\\
&\quad+8\sqrt{6}\sqrt{\frac{K\log\left(2RL^2(pB+2)\kappa^Lp^{L+1}/\delta\right)}{n}}\sigma\delta\\
&\quad+\frac{104R^2}{3n}K\log\left(8R^2L^2(pB+2)\kappa^Lp^{L+1}/\delta\right)\\
&\quad+\left(4+\frac{1}{2R}+8\sigma\right)\delta\\
&=\widetilde{O}\Bigg(\epsilon^2+\frac{R^2+\sigma^2}{n}\left(\epsilon^{-\frac{d}{s+\alpha}}+D\right)\log\frac{1}{\epsilon}\log\frac{L^2(\kappa p)^{L+1}}{\delta}\\
&\qquad+\sigma\delta\sqrt{\frac{\left(\epsilon^{-\frac{d}{s+\alpha}}+D\right)\log\frac{1}{\epsilon}\log\frac{L^2(\kappa p)^{L+1}}{\delta}}{n}}+\sigma\delta+\frac{\sigma^2}{n}\Bigg).
\end{aligned}\tag{B.10}$$

Now we choose ϵ to satisfy $\epsilon^2 = \frac{1}{n}\epsilon^{-\frac{d}{s+\alpha}}$, which gives $\epsilon = n^{-\frac{s+\alpha}{d+2(s+\alpha)}}$. It suffices to pick $\delta = \frac{1}{n}$. Substitute both ϵ and δ into (Eq. B.10), we deduce the desired estimation error bound

$$\begin{aligned}
&\mathbb{E}\left[\int_{\mathcal{M}} \left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 d\mathcal{D}_x(\mathbf{x})\right] \\
&= \widetilde{O}\Bigg(\epsilon^2 + \frac{R^2+\sigma^2}{n}\left(\epsilon^{-\frac{d}{s+\alpha}} + D\right)\log\frac{1}{\epsilon}\log\frac{L^2(\kappa p)^{L+1}}{\delta} \\
&\qquad + \sigma\delta\sqrt{\frac{\left(\epsilon^{-\frac{d}{s+\alpha}} + D\right)\log\frac{1}{\epsilon}\log\frac{L^2(\kappa p)^{L+1}}{\delta}}{n}} + \sigma\delta + \frac{\sigma^2}{n}\Bigg) \\
&\le c(R^2+\sigma^2)\left(n^{-\frac{2(s+\alpha)}{d+2(s+\alpha)}} + \frac{D}{n}\right)\log^3 n,
\end{aligned}$$

where constant c depends on depending on $\log D$, d, s, τ, B, the surface area of $\mathcal{M}$, and the upper bounds of derivatives of the coordinate systems ϕ_i's and partition of unity ρ_i's, up to order s. □

APPENDIX C
OMITTED PROOFS IN CHAPTER 5

C.1 Detailed Proofs in Euclidean Space

In this section, we no longer use bold-faced letters to represent vectors; instead, we use normal font lower-case letters. We also slightly alter the dimension notation D to lower case letter d.

C.1.1 Proof of Lemma 5.3

Proof. We introduce the empirical data distribution as an intermediate term for bounding $d_{\mathcal{H}^\beta}((g_\theta^*)_\sharp\rho, \mu)$. Using the triangle inequality, we derive

$$\begin{aligned}
& d_{\mathcal{H}^\beta}((g_\theta^*)_\sharp\rho, \mu) \\
\leq\ & d_{\mathcal{H}^\beta}((g_\theta^*)_\sharp\rho, \widehat{\mu}_n) + d_{\mathcal{H}^\beta}(\widehat{\mu}_n, \mu) \\
=\ & d_{\mathcal{F}_{\rm NN}}((g_\theta^*)_\sharp\rho, \widehat{\mu}_n) + d_{\mathcal{H}^\beta}((g_\theta^*)_\sharp\rho, \widehat{\mu}_n) - d_{\mathcal{F}_{\rm NN}}((g_\theta^*)_\sharp\rho, \widehat{\mu}_n) \\
& + d_{\mathcal{H}^\beta}(\widehat{\mu}_n, \mu) \\
\overset{(i)}{\leq}\ & d_{\mathcal{F}_{\rm NN}}((g_\theta^*)_\sharp\rho, \widehat{\mu}_n) + 2 \sup_{f\in\mathcal{H}^\beta} \inf_{f_\omega\in\mathcal{F}_{\rm NN}} \|f - f_\omega\|_\infty + d_{\mathcal{H}^\beta}(\widehat{\mu}_n, \mu),
\end{aligned} \tag{C.1}$$

where step (i) is obtained by rewriting $d_{\mathcal{H}^\beta}((g_\theta^*)_\sharp\rho, \widehat{\mu}_n) - d_{\mathcal{F}_{\rm NN}}((g_\theta^*)_\sharp\rho, \widehat{\mu}_n)$ as

$$\begin{aligned}
& d_{\mathcal{H}^\beta}((g_\theta^*)_\sharp\rho, \widehat{\mu}_n) - d_{\mathcal{F}_{\rm NN}}((g_\theta^*)_\sharp\rho, \widehat{\mu}_n) \\
=\ & \sup_{f\in\mathcal{H}^\beta} \left[\mathbb{E}_{x\sim(g_\theta^*)_\sharp\rho}[f(x)] - \mathbb{E}_{x\sim\widehat{\mu}_n}[f(x)]\right] \\
& - \sup_{f_\omega\in\mathcal{F}_{\rm NN}} \left[\mathbb{E}_{x\sim(g_\theta^*)_\sharp\rho}[f_\omega(x)] - \mathbb{E}_{x\sim\widehat{\mu}_n}[f_\omega(x)]\right] \\
=\ & \sup_{f\in\mathcal{H}^\beta} \inf_{f_\omega\in\mathcal{F}_{\rm NN}} \left[\mathbb{E}_{x\sim(g_\theta^*)_\sharp\rho}[f(x)] - \mathbb{E}_{x\sim\widehat{\mu}_n}[f(x)]\right]
\end{aligned}$$

$$
\begin{aligned}
&\quad - \left[\mathbb{E}_{x\sim(g_\theta^*)_\sharp\rho}[f_\omega(x)] - \mathbb{E}_{x\sim\widehat{\mu}_n}[f_\omega(x)]\right] \\
&= \sup_{f\in\mathcal{H}^\beta} \inf_{f_\omega\in\mathcal{F}_{\text{NN}}} \mathbb{E}_{x\sim(g_\theta^*)_\sharp\rho}[f(x) - f_\omega(x)] - \mathbb{E}_{x\sim\widehat{\mu}_n}[f(x) - f_\omega(x)] \\
&\leq \sup_{f\in\mathcal{H}^\beta} \inf_{f_\omega\in\mathcal{F}_{\text{NN}}} \mathbb{E}_{x\sim(g_\theta^*)_\sharp\rho}[|f(x) - f_\omega(x)|] + \mathbb{E}_{x\sim\widehat{\mu}_n}[|f(x) - f_\omega(x)|] \\
&\leq 2 \sup_{f\in\mathcal{H}^\beta} \inf_{f_\omega\in\mathcal{F}_{\text{NN}}} \|f - f_\omega\|_\infty .
\end{aligned}
$$

Now we bound $d_{\mathcal{F}_{\text{NN}}}((g_\theta^*)_\sharp\rho, \widehat{\mu}_n)$ using a similar triangle inequality trick:

$$
\begin{aligned}
d_{\mathcal{F}_{\text{NN}}}((g_\theta^*)_\sharp\rho, \widehat{\mu}_n) &= \inf_{g_\theta\in\mathcal{G}_{\text{NN}}} d_{\mathcal{F}_{\text{NN}}}((g_\theta)_\sharp\rho, \widehat{\mu}_n) \\
&\leq \inf_{g_\theta\in\mathcal{G}_{\text{NN}}} d_{\mathcal{F}_{\text{NN}}}((g_\theta)_\sharp\rho, \mu) + d_{\mathcal{F}_{\text{NN}}}(\mu, \widehat{\mu}_n) \\
&= \inf_{g_\theta\in\mathcal{G}_{\text{NN}}} d_{\mathcal{F}_{\text{NN}}}((g_\theta)_\sharp\rho, \mu) - d_{\mathcal{H}^\beta_\infty}((g_\theta)_\sharp\rho, \mu) \\
&\quad + d_{\mathcal{H}^\beta_\infty}((g_\theta)_\sharp\rho, \mu) + d_{\mathcal{F}_{\text{NN}}}(\mu, \widehat{\mu}_n) \\
&\leq 2 \sup_{f\in\mathcal{H}^\beta} \inf_{f_\omega\in\mathcal{F}_{\text{NN}}} \|f - f_\omega\|_\infty + \inf_{g_\theta\in\mathcal{G}_{\text{NN}}} d_{\mathcal{H}^\beta_\infty}((g_\theta)_\sharp\rho, \mu) \\
&\quad + d_{\mathcal{F}_{\text{NN}}}(\mu, \widehat{\mu}_n),
\end{aligned}
$$

where the last inequality holds by the identity $\mathcal{H}^\beta_\infty \subset \mathcal{H}^\beta$. Substituting the above ingredients into (Eq. C.1), we have

$$
\begin{aligned}
d_{\mathcal{H}^\beta}((g_\theta^*)_\sharp\rho, \mu) \leq &\underbrace{\inf_{g_\theta\in\mathcal{G}_{\text{NN}}} d_{\mathcal{H}^\beta_\infty}((g_\theta)_\sharp\rho, \mu)}_{\mathcal{E}_1:\ \text{generator approximation error}} \\
&+ 4 \underbrace{\sup_{f\in\mathcal{H}^\beta} \inf_{f_\omega\in\mathcal{F}_{\text{NN}}} \|f - f_\omega\|_\infty}_{\mathcal{E}_2:\ \text{discriminator approximation error}} \\
&+ \underbrace{d_{\mathcal{H}^\beta}(\widehat{\mu}_n, \mu) + d_{\mathcal{F}_{\text{NN}}}(\mu, \widehat{\mu}_n)}_{\mathcal{E}_3:\ \text{statistical error}} .
\end{aligned}
$$

The proof is complete. □

C.1.2 Proof of Lemma 5.4

Proof. Without loss of generality, we assume $\mathcal{Z} = \mathcal{X} = [0, 1]^d$. Otherwise, we can rescale the domain to be a subset of $[0, 1]^d$. By Monge map (Lemma 5.1), there exists a mapping $T = [T_1, \ldots, T_d] : \mathcal{Z} \mapsto \mathcal{X}$ such that $T_\sharp \nu = \mu$. Such a mapping is Hölder continuous, i.e., each coordinate mapping T_i for $i = 1, \ldots, d$ belongs to $\mathcal{H}^{\alpha+1}$. We approximate each function T_i using the network architecture identified in Theorem 3.1. Specifically, given approximation error $\delta \in (0, 1)$. There exists a network architecture with no more than $c(\log \frac{1}{\delta} + 1)$ layers and $c'\delta^{-\frac{d}{\alpha+1}}(\log \frac{1}{\delta} + 1)$ neurons and weight parameters, such that with properly chosen weight parameters, yields an approximation $\widehat{T}_i$ of T_i satisfying $\|\widehat{T}_i - T_i\|_\infty \leq \delta$. Applying this argument d times, we form an approximation $g_\theta = [\widehat{T}_1, \ldots, \widehat{T}_d]$ of T. We show $(g_\theta)_\sharp \rho$ satisfies the following IPM bound

$$
\begin{aligned}
& d_{\mathcal{H}^\beta_\infty}((g_\theta)_\sharp \rho, \mu) \\
&= d_{\mathcal{H}^\beta}((g_\theta)_\sharp \rho, T_\sharp \rho) \\
&= \sup_{f \in \mathcal{H}^\beta} \mathbb{E}_{x \sim (g_\theta)_\sharp \rho}[f(x)] - \mathbb{E}_{y \sim T_\sharp \rho}[f(y)] \\
&= \sup_{f \in \mathcal{H}^\beta} \mathbb{E}_{z \sim \rho}[f(g_\theta(z))] - \mathbb{E}_{z \sim \rho}[f(T(z))] \\
&\leq \mathbb{E}_{z \sim \rho}\left[\|g_\theta(z) - T(z)\|_\infty\right] \\
&= \mathbb{E}_{z \sim \rho}\left[\left\|[\widehat{T}_1(z) - T_1(z), \ldots, \widehat{T}_d(z) - T_d(z)]^\top\right\|_\infty\right] \\
&\leq \delta.
\end{aligned}
$$

Therefore, choosing $\delta = \epsilon_1$ gives rise to $d_{\mathcal{H}^\beta_\infty}((g_\theta)_\sharp \rho, \mu) \leq \epsilon_1$. □

C.1.3 Proof of Lemma 5.5

Proof. Using Theorem 3.1 immediately yields a network architecture for uniformly approximating functions in $\mathcal{H}^\beta(\mathcal{X})$. Specifically, let the approximation error be $\epsilon_2 > 0$.

We choose the network architecture $\mathcal{F}_{\rm NN}$ consisting of $\bar{L} = O\big(\log(1/\epsilon_2)\big)$ layers and $\bar{K} = O\big(\epsilon_2^{-d/\beta}\log(1/\epsilon_2)\big)$ total number of neurons and weight parameters. The maximum width is $\bar{p} = O\big(\epsilon_2^{-d/\beta}\big)$. Meanwhile, for any function $f \in \mathcal{H}^\beta(\mathcal{X})$, we have $\|f\|_{\mathcal{H}^\beta} \leq C$. Threfore, it is enough to choose $\bar{R} = C$ and $\bar{\kappa} = C$. Accordingly, for any $f \in \mathcal{H}^\beta(\mathcal{X})$, there exists a function $\widehat{f}_\omega$ given by the network architecture $\mathcal{F}_{\rm NN}(\bar{R}, \bar{\kappa}, \bar{L}, \bar{p}, \bar{K})$, such that $\|f - \widehat{f}_\omega\|_\infty \leq \epsilon_2$. To this end, we can establish that for any $f \in \mathcal{H}^\beta(\mathcal{X})$, inequality $\inf_{f_\omega \in \mathcal{F}_{\rm NN}} \|f - f_\omega\|_\infty \leq \epsilon_2$ holds. □

C.1.4 Proof of Lemma 5.6

Proof. The proof utilizes the symmetrization technique and Dudley's entropy integral, which can be found in empirical process theory [221, 96]. We prove here for completeness. Let $y_1, \ldots, y_n$ be i.i.d. samples from μ, independent of x_i's. By symmetrization, we derive

$$\begin{aligned}
\mathbb{E}[d_{\mathcal{F}}(\widehat{\mu}_n, \mu)] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}_{y\sim\mu}[f(y)]\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}_{\substack{y_i\sim\mu,\\ i=1,\ldots,n}} \frac{1}{n}\sum_{i=1}^n f(y_i)\right] \\
&\leq \mathbb{E}_x\mathbb{E}_y\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (f(x_i) - f(y_i))\right] \\
&= \mathbb{E}_x\mathbb{E}_y\mathbb{E}_\xi\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \xi_i(f(x_i) - f(y_i))\right] \\
&= 2\mathbb{E}_{x,\xi}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \xi_i f(x_i)\right],
\end{aligned}$$

where ξ_i's are i.i.d. Rademacher random variables, i.e., $\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = \frac{1}{2}$. The next step is to discretize the function space $\mathcal{F}$. Let $\{\delta_i\}_{i=1}^k$ be a decreasing series of real numbers with $\delta_{i+1} < \delta_i$. We construct a collection of coverings on $\mathcal{F}$ under the function ℓ_∞ norm with accuracy δ_i. Denote the δ_i-covering number as $\mathcal{N}(\delta_i, \mathcal{F}, \|\cdot\|_\infty)$. For a given f, denote the closest element (in the ℓ_∞ sense) to f in the δ_i covering as $f^{(i)}$ for $i = 1, \ldots, k$.

We expand $\mathbb{E}_{x,\xi}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i f(x_i)\right]$ as a telescoping sum as

$$\begin{aligned}\mathbb{E}_{x,\xi}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i f(x_i)\right] \leq \mathbb{E}&\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i(f(x_i)-f^k(x_i))\right]\\ &+\sum_{j=1}^{k-1}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i(f^{(j+1)}(x_i)-f^{(j)}(x_i))\right]\\ &+\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i f^{(1)}(x_i)\right].\end{aligned}$$

We choose $\delta_1 = \mathrm{diam}(\mathcal{F})$, i.e., the diameter of the class $\mathcal{F}$. Then $f^{(1)}$ can be arbitrarily picked from $\mathcal{F}$. Therefore, the last term $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i f^{(1)}(x_i)\right] = 0$ since ξ_i's are symmetric. The first term $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i(f(x_i)-f^k(x_i))\right]$ can be bounded by Cauchy-Schwarz inequality:

$$\begin{aligned}&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i(f(x_i)-f^k(x_i))\right]\\ &\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sqrt{(\sum_{i=1}^{n}\xi_i^2)(\sum_{i=1}^{n}(f(x_i)-f^{(k)}(x_i))^2)}\right]\\ &\leq \delta_k.\end{aligned}$$

We now bound each term in the telescoping sum

$$\sum_{j=1}^{k-1}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i(f^{(j+1)}(x_i)-f^{(j)}(x_i))\right].$$

Observe

$$\begin{aligned}\left\|f^{(j+1)}-f^{(j)}\right\|_\infty &= \left\|f^{(j+1)}-f+f-f^{(j)}\right\|_\infty\\ &\leq \left\|f^{(j+1)}-f\right\|_\infty+\left\|f-f^{(j)}\right\|_\infty\\ &\leq \delta_{j+1}+\delta_j.\end{aligned}$$

By Massart's lemma [222], we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i(f^{(j+1)}(x_i)-f^{(j)}(x_i))\right]\\
&\leq \frac{(\delta_{j+1}+\delta_j)\sqrt{2\log(\mathcal{N}(\delta_j,\mathcal{F},\|\cdot\|_\infty)\mathcal{N}(\delta_{j+1},\mathcal{F},\|\cdot\|_\infty))}}{\sqrt{n}}\\
&\leq \frac{2(\delta_{j+1}+\delta_j)\sqrt{\log\mathcal{N}(\delta_{j+1},\mathcal{F},\|\cdot\|_\infty)}}{\sqrt{n}}.
\end{aligned}$$

Summing up all the terms indexed by j, we establish

$$\mathbb{E}_{x,\xi}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i f(x_i)\right]\leq \delta_k+2\sum_{j=1}^{k-1}\frac{(\delta_{j+1}+\delta_j)\sqrt{\log\mathcal{N}(\delta_{j+1},\mathcal{F},\|\cdot\|)}}{\sqrt{n}}.$$

It suffices to set $\delta_{j+1}=\frac{1}{2}\delta_j$. Invoking the identity $\delta_{j+1}+\delta_j=6(\delta_{j+1}-\delta_{j+2})$, we derive

$$\begin{aligned}
\mathbb{E}_{x,\xi}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\xi_i f(x_i)\right] &\leq \delta_k+12\sum_{j=1}^{k-1}\frac{(\delta_{j+1}-\delta_{j+2})\sqrt{\log\mathcal{N}(\delta_{j+1},\mathcal{F},\|\cdot\|_\infty)}}{\sqrt{n}}\\
&\leq \delta_k+\frac{12}{\sqrt{n}}\int_{\delta_{k+1}}^{\delta_2}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_\infty)}d\epsilon\\
&\leq \inf_{\delta}\ 2\delta+\frac{12}{\sqrt{n}}\int_{\delta}^{\delta_1}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_\infty)}d\epsilon.
\end{aligned}$$

By the assumption, we pick $\delta_1=M$ and set the δ_1-covering with only one element $f=0$. This yields the desired result

$$\mathbb{E}\left[d_{\mathcal{F}}(\widehat{\mu}_n,\mu)\right]\leq 2\inf_{0<\delta<M}\left(2\delta+\frac{12}{\sqrt{n}}\int_{\delta}^{M}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_\infty)}d\epsilon\right).$$

□

C.2 Detailed proofs for Low-dimensional Linear Subspace

C.2.1 Proof of Lemma 5.7

Proof. We replicate the error decomposition in (Eq. C.1) by taking $\beta = 1$,

$$\begin{aligned} W_1((U^* \circ g_\theta^*)_\sharp \rho, \mu) &\le W_1((U^* \circ g_\theta^*)_\sharp \rho, \widehat{\mu}_n) + W_1(\widehat{\mu}_n, \mu) \\ &= d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}((U^* \circ g_\theta^*)_\sharp \rho, \widehat{\mu}_n) \\ &\quad + W_1((U^* \circ g_\theta^*)_\sharp \rho, \widehat{\mu}_n) - d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}((U^* \circ g_\theta^*)_\sharp \rho, \widehat{\mu}_n) \\ &\quad + W_1(\widehat{\mu}_n, \mu). \end{aligned} \tag{C.2}$$

Using the optimality of (U^*, g_θ^*), we further bound $d_{\mathcal{F}_{\mathrm{NN}}}((U^* \circ g_\theta^*)_\sharp \rho, \widehat{\mu}_n)$ in the last display as

$$\begin{aligned} & d_{\mathcal{F}_{\mathrm{NN}}}((U^* \circ g_\theta^*)_\sharp \rho, \widehat{\mu}_n) \\ \le\ & d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}((U^* \circ g_\theta^*)_\sharp \rho, \mu) + d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}(\mu, \widehat{\mu}_n) \\ =\ & \inf_{U \circ g_\theta \in \mathcal{G}_{\mathrm{NN}}^{\mathrm{ld}}} d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}((U \circ g_\theta)_\sharp \rho, \mu) + d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}(\mu, \widehat{\mu}_n) \\ \overset{(i)}{=}\ & \inf_{U \circ g_\theta \in \mathcal{G}_{\mathrm{NN}}^{\mathrm{ld}}} d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}((U \circ g_\theta)_\sharp \rho, \mu) - d_{\mathcal{H}_\infty^1}((U \circ g_\theta)_\sharp \rho, \mu) \\ & \qquad + d_{\mathcal{H}_\infty^1}((U \circ g_\theta)_\sharp \rho, \mu) + d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}(\mu, \widehat{\mu}_n), \end{aligned} \tag{C.3}$$

where in (i), discriminative class $\mathcal{H}_\infty^1$ follows the same definition in Lemma 5.3 with $\beta = 1$.

By Assumption 5.5 and the optimal transport theory in Lemma 5.1, we rewrite the data distribution μ as a pushforward distribution $\mu = (A \circ T^{\mathrm{ld}})_\sharp \rho$, where $T^{\mathrm{ld}} : \mathbb{R}^q \mapsto \mathbb{R}^q$ is an $(\alpha + 1)$-Hölder continuous transport plan. Accordingly, we rewrite the empirical data distribution $\widehat{\mu}_n$ as $\widehat{\mu}_n = (A \circ T^{\mathrm{ld}})_\sharp \widehat{\rho}_n$, with $\widehat{\rho}_n$ an empirical version of ρ. Applying Lemma Theorem 3.1 and using the same argument in Theorem 5.1 for approximating $A \circ T^{\mathrm{ld}}$, we obtain $A \circ \widetilde{g}_\theta \in \mathcal{G}_{\mathrm{NN}}^{\mathrm{ld}}$ as a proper approximation. Note that we have chosen $U = A$ in representing $A \circ T^{\mathrm{ld}}$ for simplicity. Substituting these notations into (Eq. C.3) gives rise to

$$\begin{aligned}
& d_{\mathcal{F}_{\mathrm{NN}}}((U^* \circ g_\theta^*)_\sharp \rho, \widehat{\mu}_n) \\
\overset{(i)}{\leq} & \, d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \left((A \circ \widetilde{g}_\theta)_\sharp \rho, (A \circ T^{\mathrm{ld}})_\sharp \rho\right) - d_{\mathcal{H}_\infty^1} \left((A \circ \widetilde{g}_\theta)_\sharp \rho, (A \circ T^{\mathrm{ld}})_\sharp \rho\right) \\
& + d_{\mathcal{H}_\infty^1} \left((A \circ \widetilde{g}_\theta)_\sharp \rho, (A \circ T^{\mathrm{ld}})_\sharp \rho\right) + d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}(\mu, \widehat{\mu}_n) \\
\overset{(ii)}{\leq} & \, d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \left((A \circ \widetilde{g}_\theta)_\sharp \rho, (A \circ T^{\mathrm{ld}})_\sharp \rho\right) - d_{\mathcal{H}_\infty^1} \left((A \circ \widetilde{g}_\theta)_\sharp \rho, (A \circ T^{\mathrm{ld}})_\sharp \rho\right) \\
& + \left\|A \circ \widetilde{g}_\theta - A \circ T^{\mathrm{ld}}\right\|_\infty + d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}(\mu, \widehat{\mu}_n),
\end{aligned} \tag{C.4}$$

where inequality (i) holds by instantiating the infimum in (Eq. C.3) to $A \circ \widetilde{g}_\theta$, and inequality (ii) follows by the definition of IPM over $\mathcal{H}_\infty^1$ class. We substitute (Eq. C.4) into (Eq. C.2), which leads to

$$\begin{aligned}
& W_1((U^* \circ g_\theta^*)_\sharp \rho, \mu) \\
\leq & \underbrace{\left\|A \circ \widetilde{g}_\theta - A \circ T^{\mathrm{ld}}\right\|_\infty}_{\text{generator approximation error}} + \underbrace{W_1(\widehat{\mu}_n, \mu) + d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}(\widehat{\mu}_n, \mu)}_{\text{statistical error}} \\
& + \underbrace{W_1\left((U^* \circ g_\theta^*)_\sharp \rho, (A \circ T^{\mathrm{ld}})_\sharp \widehat{\rho}_n)\right) - d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}\left((U^* \circ g_\theta^*)_\sharp \rho, (A \circ T^{\mathrm{ld}})_\sharp \widehat{\rho}_n\right)}_{\text{discriminator approximation error (HARD)}} \\
& + \underbrace{d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \left((A \circ \widetilde{g}_\theta)_\sharp \rho, (A \circ T^{\mathrm{ld}})_\sharp \rho\right) - d_{\mathcal{H}_\infty^1} \left((A \circ \widetilde{g}_\theta)_\sharp \rho, (A \circ T^{\mathrm{ld}})_\sharp \rho\right)}_{\text{discriminator approximation error (EASY)}}.
\end{aligned} \tag{C.5}$$

Two discriminator approximation error terms share a similar formulation, and we can further provide a simplified upper bound on them. Denote $\|f\|_{\mathrm{Lip}}$ as the lipschitz constant of function f, and consider the (HARD) term for example.

$$\begin{aligned}
& W_1\left((U^* \circ g_\theta^*)_\sharp \rho, (A \circ T^{\mathrm{ld}})_\sharp \widehat{\rho}_n)\right) - d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}\left((U^* \circ g_\theta^*)_\sharp \rho, (A \circ T^{\mathrm{ld}})_\sharp \widehat{\rho}_n\right) \\
= & \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \mathbb{E}_{z \sim \rho}\left[f \circ U^* \circ g_\theta^*(z)\right] - \mathbb{E}_{z \sim \widehat{\rho}_n}\left[f \circ A \circ T^{\mathrm{ld}}(z)\right] \\
& - \sup_{f_\omega \circ V^\top \in \mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \mathbb{E}_{z \sim \rho}\left[f_\omega \circ V^\top \circ U^* \circ g_\theta^*(z)\right] - \mathbb{E}_{z \sim \widehat{\rho}_n}\left[f_\omega \circ V^\top \circ A \circ T^{\mathrm{ld}}(z)\right]
\end{aligned}$$

$$\begin{aligned}
&\leq \sup_{\|f\|_{\mathrm{Lip}}\leq 1} \inf_{f_\omega \circ V^\top \in \mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \Big\{ \left| \mathbb{E}_{z\sim\rho}\left[(f\circ U^* - f_\omega \circ V^\top \circ U^*) \circ g_\theta^*(z) \right] \right| \\
&\qquad\qquad + \left| \mathbb{E}_{z\sim\widehat{\rho}_n}\left[(f\circ A - f_\omega \circ V^\top \circ A) \circ T^{\mathrm{ld}}(z) \right] \right| \Big\} \\
&\leq \sup_{\|f\|_{\mathrm{Lip}}\leq 1} \inf_{f_\omega \circ V^\top \in \mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \left\| f\circ U^* - f_\omega \circ V^\top U^* \right\|_\infty + \left\| f\circ A - f_\omega \circ V^\top A \right\|_\infty . \qquad \text{(C.6)}
\end{aligned}$$

Applying the same argement to the (EASY) error term yields

$$\begin{aligned}
&d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}\left((A\circ \widetilde{g}_\theta)_\sharp \rho, (A\circ T^{\mathrm{ld}})_\sharp \rho \right) - d_{\mathcal{H}_\infty^1}\left((A\circ \widetilde{g}_\theta)_\sharp \rho, (A\circ T^{\mathrm{ld}})_\sharp \rho \right) \\
&\leq 2 \sup_{\|f\|_{\mathrm{Lip}}\leq 1} \inf_{f_\omega \circ V^\top \in \mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \left\| f\circ A - f_\omega \circ V^\top A \right\|_\infty . \qquad \text{(C.7)}
\end{aligned}$$

Note that we already used the fact that $\mathcal{H}_\infty^1$ is a subset of $\mathcal{H}^1$. Plugging (Eq. C.6) and (Eq. C.7) into (Eq. C.5) and taking infimum over $\widetilde{g}_\theta$, we obtain the desired oracle inequality,

$$\begin{aligned}
&W_1((U^*\circ g_\theta^*)_\sharp \rho, \mu) \\
&\leq \underbrace{\inf_{g: A\circ g \in \mathcal{G}_{\mathrm{NN}}^{\mathrm{ld}}} \left\| A\circ g - A\circ T^{\mathrm{ld}} \right\|_\infty}_{\text{generator approximation error}} + \underbrace{W_1(\widehat{\mu}_n, \mu) + d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}(\widehat{\mu}_n, \mu)}_{\text{statistical error}} \\
&\quad + \underbrace{\sup_{\|f\|_{\mathrm{Lip}}\leq 1} \inf_{f_\omega \circ V^\top \in \mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \left\| f\circ U^* - f_\omega \circ V^\top U^* \right\|_\infty + \left\| f\circ A - f_\omega \circ V^\top A \right\|_\infty}_{\text{discriminator approximation error (HARD)}} \\
&\quad + \underbrace{2 \sup_{\|f\|_{\mathrm{Lip}}\leq 1} \inf_{f_\omega \circ V^\top \in \mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}} \left\| f\circ A - f_\omega \circ V^\top A \right\|_\infty}_{\text{discriminator approximation error (EASY)}} .
\end{aligned}$$

The proof is complete. □

C.2.2 Proof of Lemma 5.8

Proof. The proof consists of two steps: 1) construction of a piecewise linear function for approximating 1-Lipschitz functions, which can be implemented by a ReLU neural network; 2) establishing the global Lipschitz continuity of the neural network, in addition to the L^∞ approximation error guarantee.

Step 1). Given a positive integer $N > 0$, we evenly choose $(N+1)^q$ points in the hypercube $[0, 1]^q$, denoted as m/N with $m = [m_1, \dots, m_q]^\top \in \{0, \dots, N\}^q$. We define a univariate trapezoid function (see graphical illustration in Figure C.1)

$$\phi(a) = \begin{cases} 1, & |a| < 1 \\ 2 - |a|, & |a| \in [1, 2] \\ 0, & |a| > 2 \end{cases} .$$

Then for any $x \in [0, 1]^q$, we define a partition of unity based on a product of trapezoid functions indexed by m,

$$\xi_m(x) = \prod_{k=1}^{q} \phi\left(3N\left(x_k - \frac{m_k}{N}\right)\right).$$

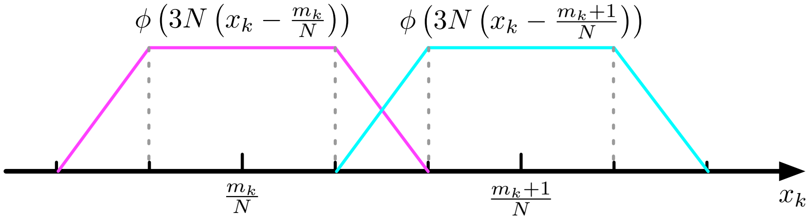

$\phi\left(3N\left(x_k - \frac{m_k}{N}\right)\right)$
$\phi\left(3N\left(x_k - \frac{m_k+1}{N}\right)\right)$
$\frac{m_k}{N}$
$\frac{m_k+1}{N}$
x_k

Figure C.1: Trapezoid function in one dimension.

For any target 1-Lipschitz function f, it is more convenient to write its Lipschitz continuity with respect to the ℓ_∞ norm, i.e.,

$$|f(x) - f(y)| \leq \|x - y\|_2 \leq \sqrt{q}\,\|x - y\|_\infty . \tag{C.8}$$

We now define a collection of piecewise constant functions

$$P_m(x) = f(m) \quad \text{for} \quad m \in \{0, \dots, N\}^q.$$

We claim that $\widetilde{f}(x) = \sum_m \xi_m(x) P_m(x)$ is an approximation of f, with an approximation error evaluated as

$$\begin{aligned}
\sup_{x \in [0,1]^q} \left| \widetilde{f}(x) - f(x) \right| &= \sup_{x \in [0,1]^q} \left| \sum_m \xi_m(x) \left(P_m(x) - f(x) \right) \right| \\
&\leq \sup_{x \in [0,1]^q} \sum_{m : |x_k - m_k/N| \leq \frac{2}{3N}} |P_m(x) - f(x)| \\
&= \sup_{x \in [0,1]^q} \sum_{m : |x_k - m_k/N| \leq \frac{2}{3N}} |f(m) - f(x)| \\
&\leq \sqrt{q} 2^{q+1} \frac{1}{3N},
\end{aligned}$$

where the last inequality follows from the Lipschitz continuity in (Eq. C.8) and the fact that there are at most 2^q terms in the summation.

We use a ReLU network to implement $\widetilde{f}$. It turns out that we only need to implement the multiplication operation in ξ_m. For scalars $a, b \in [0,1]$, we rewrite ab as $\left(\frac{a+b}{2}\right)^2 - \left(\frac{|a-b|}{2}\right)^2$. We know neural networks can approximate a univariate quadratic function on $[0,1]$ as

$$a^2 \approx \widehat{h}_K(a) = a - \sum_{k=1}^{K} \frac{1}{2^{2k}} g_k(a), \quad \text{with} \quad g_k = \underbrace{g \circ \cdots \circ g}_{k \text{ compositions}}, \tag{C.9}$$

where $g(a) = 2\mathrm{ReLU}(a) - 4\mathrm{ReLU}(a - 0.5) + 2\mathrm{ReLU}(a-1)$. The L_∞ approximation error of $\widehat{h}_K$ is $2^{-(2K+2)}$ (A proof can be found in [40, Proposition 2] or [214, Lemma 1]). We approximate ξ_m recursively using univariate quadratic functions. Specifically, we construct

$$\xi_m(x) \approx \widehat{\xi}_m(x) = \widehat{\times}\left(\phi(3N(x_q - m_q/N)), \widehat{\times}\left(\phi(3N(x_{q-1} - m_{q-1}/N)), \dots \right) \right), \tag{C.10}$$

where $\widehat{\times}(a,b) = \widehat{h}_K((a+b)/2) - \widehat{h}_K(|a-b|/2)$ for $a, b \in [0,1]$. Then the network for approximating f is obtained as

$$f(x) \approx \widehat{f}(x) = \sum_m \widehat{\xi}_m(x) f(m). \tag{C.11}$$

We bound L_∞ approximation error of $\widehat{f}$ as

$$\begin{aligned}\left\|\widehat{f}-f\right\|_\infty &\le \left\|\widehat{f}-\widetilde{f}\right\|_\infty + \left\|\widetilde{f}-f\right\|_\infty \\ &\le \sup_{x\in[0,1]^q}\left|\sum_m \left(\widehat{\xi}_m(x)-\xi_m(x)\right)P_m(x)\right| + \sqrt{q}2^{q+1}\frac{1}{3N} \\ &\le \|f\|_\infty \sup_{x\in[0,1]^q}\left|\sum_m \left(\widehat{\xi}_m(x)-\xi_m(x)\right)\right| + \sqrt{q}2^{q+1}\frac{1}{3N} \\ &\le 2^q \|f\|_\infty \left\|\widehat{\xi}_m-\xi_m\right\|_\infty + \sqrt{q}2^{q+1}\frac{1}{3N} \\ &\le q2^q \|f\|_\infty 2^{-2K-1} + \sqrt{q}2^{q+1}\frac{1}{3N},\end{aligned}$$

where the last inequality follows from recursively decomposing $\left\|\widehat{\xi}_m-\xi_m\right\|_\infty$ into q terms as

$$\begin{aligned}\left\|\widehat{\xi}_m-\xi_m\right\|_\infty \le &\left\|\widehat{\times}\left(\phi(3N(x_q-m_q/N)), \widehat{\times}\left(\phi(3N(x_{q-1}-m_{q-1}/N)),\dots\right)\right)\right. \\ &\left. -\phi(3N(x_q-m_q/N))\cdot\widehat{\times}\left(\phi(3N(x_{q-1}-m_{q-1}/N)),\dots\right)\right\|_\infty \\ &+\dots \\ &+\phi(3N(x_q-m_q/N))\cdots\phi(3N(x_3-m_3/N)) \\ &\cdot\left\|\widehat{\times}\left(\phi(3N(x_2-m_2/N)),\phi(3N(x_1-m_1/N))\right)\right. \\ &\quad\left. -\phi(3N(x_2-m_2/N))\phi(3N(x_1-m_1/N)\right\|_\infty\end{aligned}$$

and observing

$$\begin{aligned}|\widehat{\times}(a,b)-ab| &\le \left|\widehat{h}_K((a+b)/2)-(a+b)^2/4\right| + \left|\widehat{h}_K(|a-b|/2)-(a-b)^2/4\right| \\ &\le 2\cdot 2^{-2K-2} = 2^{-2K-1}\end{aligned}$$

for any $a,b\in[0,1]$.

Step 2). The following lemma establishes the Lipschitz continuity of $\widehat{f}$ with respect to the

ℓ_∞ norm.

Lemma C.1. *Let* $\widehat{f}$ *be defined in* (Eq. C.11). *Then for any* $x, y \in [0,1]^q$, *it holds*

$$\left|\widehat{f}(x) - \widehat{f}(y)\right| \leq 3q \left(3 + 2(N \left\|f\right\|_\infty + 1) \cdot q 2^{-K+q-1} \frac{1 - \left(q2^{-K}\right)^q}{1 - q2^{-K}}\right) \left\|x - y\right\|_\infty .$$

The proof is deferred to Appendix C.3. Given Lemma C.1, we choose $N = \left\lceil \frac{\sqrt{q}2^{q+1}}{\epsilon_2} \right\rceil$ and K satisfying

$$2(N \left\|f\right\|_\infty + 1) \cdot q 2^{-K+q-1} \frac{1 - \left(q2^{-K}\right)^q}{1 - q2^{-K}} \leq \frac{1}{3},$$

which implies $K = \left\lceil \log \frac{12 q^{3/2}(\|f\|_\infty + 1)}{\epsilon_2} + 2q \right\rceil$. As a result, we check the L_∞ approximation error of $\widehat{f}$ as

$$\begin{aligned}\left\|\widehat{f} - f\right\|_\infty &\leq q 2^q \left\|f\right\|_\infty 2^{-2K-1} + \sqrt{q} 2^{q+1} \frac{1}{3N} \\ &\leq \frac{1}{9q^2 2^{3q+5}(\left\|f\right\|_\infty + 1)} \epsilon_2^2 + \frac{1}{3}\epsilon_2 \\ &\leq \epsilon_2 .\end{aligned}$$

Meanwhile, with the choice of K and N, Lemma C.1 implies that for any $x, y \in [0,1]^q$, it holds

$$\begin{aligned}\left|\widehat{f}(x) - \widehat{f}(y)\right| &\leq 3q \left(3 + 2(N \left\|f\right\|_\infty + 1) \cdot q 2^{-K+q-1} \frac{1 - \left(q2^{-K}\right)^q}{1 - q2^{-K}}\right) \left\|x - y\right\|_\infty \\ &\leq 10q \left\|x - y\right\|_\infty .\end{aligned}$$

The remaining step is to characterize the size of the ReLU network for implementing $\widehat{f}$. Construction (Eq. C.11) suggests that the network consists of $(N+1)^q$ parallel subnetworks. In each subnetwork, we need to implement $\widehat{\xi}_m$ defined in (Eq. C.10), where the subnetwork architecture consists of K layers and the width is bounded by a constant (since

$\widehat{h}_K$ is realizable by a width-3 network). Putting together all the parallel subnetworks, we conclude that the whole network architecture consists of K layers and the width is bounded by $O((N+1)^q)$. Substituting our choice of N and K into the network size, we obtain $L = O\left(\log \frac{1}{\epsilon_2} + q\right)$ and $p = O(\epsilon_2^{-q})$. The total number of neurons and nonzero weight parameters in the network is $J = O(Lp)$.

The last step is to ensure that each weight parameter in $\widehat{f}$ is bounded by a constant. The only caveat stems from the trapezoid function in ξ_m, which is rescaled by $3N$ (see equation (Eq. C.10)). We use a deep network to implement $\phi(3N(x_k - \frac{m_k}{N}))$. Consider a basic step function $s(x) = 2\mathrm{ReLU}(x) - 2\mathrm{ReLU}(x-1)$, whose j-th order composition is

$$s_j = s \circ \cdots \circ s = \begin{cases} 0, & x < 0 \\ 2^j x, & x \in [0, 1/2^{j-1}] \\ 2, & x > 1/2^j \end{cases} .$$

Setting $j = \lceil \log(3N) \rceil + 1$, we observe that s_j has a slope of at least $6N$. We use $s_j/2$ to realize the left linear segments in $\phi(3N(x_k - \frac{m_k}{N}))$. For the right linear segments, we can use $1 - s_j/2$ instead. In this way, we increment the network architecture for implementing $\widehat{f}$ by a depth of $\lceil \log(3N) \rceil + 1 = O\left(\log 1/\epsilon_2 + q\right)$ and a width of 4, while each weight parameter in the network is bounded by a constant. To summarize the network architecture, we have

$$L = O\left(\log 1/\epsilon_2 + q\right), \quad p = O\left(\epsilon_2^{-q}\right), \quad J = O\left(\epsilon_2^{-q}(\log 1/\epsilon_2 + q)\right),$$
$$\kappa = O(1), \quad R = \sqrt{q}.$$

The bound on R is obtained by combining Lipschitz continuity (Eq. C.8) with $f(0) = 0$.

□

C.2.3 Proof of Lemma 5.9

Proof. Given the choice of generator and discriminator network classes, we show that at a global optimizer (U^*, g_θ^*), it holds

$$W_1\left((U^* \circ g_\theta^*)_\sharp \rho, \mu\right) \leq \left(1 + 4\sqrt{q}\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-1} \mathbb{E}_{z\sim\rho}\left[\|g_\theta^*(z)\|_2\right]\right) \cdot (\bar{\gamma}\epsilon_1 + 3\epsilon_2). \tag{C.12}$$

Suppose for the purpose of contradiction, we have

$$W_1\left((U^* \circ g_\theta^*)_\sharp \rho, \mu\right) > \left(1 + 4\sqrt{q}\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-1} \mathbb{E}_{z\sim\rho}\left[\|g_\theta^*(z)\|_2\right]\right) \cdot (\bar{\gamma}\epsilon_1 + 3\epsilon_2). \tag{C.13}$$

We will prove that there exists (V, f_ω) such that

$$\begin{aligned} d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}\left((U^* \circ g_\theta^*)_\sharp \rho, \mu\right) &\geq \mathbb{E}_{z\sim\rho}\left[f_\omega(V^\top U^* g_\theta^*(z))\right] - \mathbb{E}_{x\sim\mu}[f_\omega(V^\top x)] \\ &> \bar{\gamma}\epsilon_1. \end{aligned} \tag{C.14}$$

On the other hand, by choosing $U^* = A$ and g_θ with $\left\|T^{\mathrm{ld}} - g_\theta\right\|_\infty \leq \epsilon_1/q$, we have

$$d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}\left((U^* \circ g_\theta)_\sharp \rho, \mu\right) \leq \bar{\gamma}\epsilon_1, \tag{C.15}$$

since discriminator is $\bar{\gamma}$-Lipschitz with respect to the L_∞ norm. Putting (Eq. C.14) and (Eq. C.15) together, we conclude that (U^*, g_θ^*) cannot be a global optimizer. Therefore, (Eq. C.12) holds true. It remains to establish (Eq. C.14). Since the discriminator network can approximate any 1-Lipschitz function by Lemma 5.8, it is convenient to show the following sufficient condition for (Eq. C.14),

$$\sup_{V} W_1\left((V^\top U^* \circ g_\theta^*)_\sharp \rho, V_\sharp^\top \mu\right) > \bar{\gamma}\epsilon_1 + 3\epsilon_2. \tag{C.16}$$

In fact, (Eq. C.16) implies that for any $\delta \in (0, \epsilon_2)$, there exists a discriminative function f_0 and matrix V_0 such that $\mathbb{E}_{z\sim\rho}[f_0(V_0^\top U^* g_\theta^*(z))] - \mathbb{E}_{x\sim\mu}[f_0(V_0^\top x)] > \bar{\gamma}\epsilon_1 + 3\epsilon_2 + 2d_{\mathcal{F}_{\text{NN}}^{\text{ld}}}(\widehat{\mu}_n, \mu) - \delta$. By choosing f_ω as an ϵ_2-approximation of f_0 and $V = V_0$, we obtain

$$\begin{aligned}
&\mathbb{E}_{z\sim\rho}\left[f_\omega(V_0^\top U^* g_\theta^*(z))\right] - \mathbb{E}_{x\sim\mu}[f_\omega(V_0^\top x)] \\
&= \mathbb{E}_{z\sim\rho}\left[f_\omega(V_0^\top U^* g_\theta^*(z))\right] \\
&\quad - \mathbb{E}_{x\sim\mu}[f_\omega(V_0^\top x)] - \mathbb{E}_{z\sim\rho}[f_0(V_0^\top U^* g_\theta^*(z))] - \mathbb{E}_{x\sim\mu}[f_0(V_0^\top x)] \\
&\quad + \mathbb{E}_{z\sim\rho}[f_0(V_0^\top U^* g_\theta^*(z))] - \mathbb{E}_{x\sim\mu}[f_0(V_0^\top x)] \\
&> \bar{\gamma}\epsilon_1 + 3\epsilon_2 - \delta - 2\left\|f_\omega - f_0\right\|_\infty \\
&> \bar{\gamma}\epsilon_1,
\end{aligned}$$

which establishes (Eq. C.14).

To ease the presentation, we recall that $\epsilon = \bar{\gamma}\epsilon_1 + 3\epsilon_2$. We now consider two complementary cases for establishing (Eq. C.16),

- **(Case 1)** $\frac{1}{q}\left|\operatorname{tr}\left(A^\top U^*\right)\right| < 1 - 2\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\text{ld}}(z)\right]\right)^{-2}\epsilon^2$;
- **(Case 2)** $\frac{1}{q}\left|\operatorname{tr}\left(A^\top U^*\right)\right| \geq 1 - 2\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\text{ld}}(z)\right]\right)^{-2}\epsilon^2$,

where T_i^{ld} denotes the i-th coordinate mapping. Note that **(Case 2)** says that the column spaces of $A, U^\star$ are nearly identical. We tackle the two cases separately. To further ease the analysis, we assume without loss of generality that $a_i^\top u_i^\star \geq 0$ for $i = 1, \ldots, q$, where a_i and $u_i^\star$ are column vectors of A and $U^\star$, respectively. Otherwise we can replace a_i with $-a_i$ and T_i^{ld} with $-T_i^{\text{ld}}$ simultaneously. As a result, we may remove the absolute values in **(Case 1)** and **(Case 2)** for simplicity.

• **(Case 1)** We show that there exists an index I such that the corresponding column vectors a_I and u_I^* are sufficiently mis-aligned in direction. Specifically, given $\frac{1}{q}\operatorname{tr}\left(A^\top U^\star\right) < 1 - 2\mathbb{E}_{z\sim\rho}^{-2}\left[\min_i T_i^{\mathrm{ld}}(z)\right]\epsilon^2$, we expand the expression as

$$\frac{1}{q}\operatorname{tr}\left(A^\top U^\star\right) = \frac{1}{q}\sum_{i=1}^{q} a_i^\top u_i^\star < 1 - 2\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-2}\epsilon^2.$$

Since $a_i^\top u_i^\star \in [0,1]$ for $i = 1,\dots,q$, by the Pigeonhole principle, we deduce that there exists an index I with

$$a_I^\top u_I^\star < 1 - 2\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-2}\epsilon^2. \tag{C.17}$$

Now we prove that the mis-alignment of a_I and u_I^* already results in a sufficient separation between the generated distribution and data distribution, in terms of projected Wasserstein distance. By definition, we have

$$\begin{aligned}
&W_1\left(\left(V^\top U^* g_\theta^*\right)_\sharp \rho, V_\sharp^\top \mu\right)\\
&= \sup_{f\in \mathrm{Lip}_1(\mathbb{R}^q)} \mathbb{E}_{z\sim\rho}\left[f\left(V^\top U^* g_\theta^*(z)\right)\right] - \mathbb{E}_{z\sim\rho}\left[f\left(V^\top A T^{\mathrm{ld}}(z)\right)\right]\\
&= \sup_{f\in \mathrm{Lip}_1(\mathbb{R}^q)} \mathbb{E}_{z\sim\rho}\left[f\left(V^\top \sum_{i=1}^{q} u_i^*(g_\theta^*)_i(z)\right)\right] - \mathbb{E}_{z\sim\rho}\left[f\left(V^\top \sum_{i=1}^{q} a_i T_i^{\mathrm{ld}}(z)\right)\right]. \qquad (C.18)
\end{aligned}$$

We choose the projection matrix V to be a rank-1 matrix with only the I-th column nonzero, i.e.,

$$V = \left[\mathbf{0}_{d\times(I-1)}, \quad \frac{a_I - u_I^\star}{\left\|a_I - u_I^\star\right\|_2}, \quad \mathbf{0}_{d\times(q-I)}\right].$$

We further choose a specific testing function f to derive a lower bound on (Eq. C.18). Let $f(x) = w^\top x$ be linear with $w_I = 1$ and $w_i = 0$ for $i \neq I$. Substituting our choice of V and

f into (Eq. C.18), we obtain

$$\begin{aligned}
& W_1\left(\left(V^\top U^* g^*_\theta\right)_\sharp \rho, V_\sharp^\top \mu\right) \\
& \geq \mathbb{E}_{z\sim\rho}\left[w^\top V^\top \sum_{i=1}^{q} u^*_i (g^*_\theta)_i(z)\right] - \mathbb{E}_{z\sim\rho}\left[w^\top V^\top \sum_{i=1}^{q} a_i T^{\mathrm{ld}}_i(z)\right] \\
& = \frac{1 - a_I^\top u^*_I}{\|a_I - u^*_I\|_2}\, \mathbb{E}_{z\sim\rho}\left[(g^*_\theta)_I(z) + T^{\mathrm{ld}}_I(z)\right] \\
& = \frac{1}{2}\, \|a_I - u^*_I\|_2\, \mathbb{E}_{z\sim\rho}\left[(g^*_\theta)_I(z) + T^{\mathrm{ld}}_I(z)\right]. \qquad \text{(C.19)}
\end{aligned}$$

Using (Eq. C.17), we lower bound

$$\|a_I - u^*_I\|_2 = \sqrt{2 - 2a_I^\top u^*_I} > 2\left(\min_i \mathbb{E}_{z\sim\rho}\left[T^{\mathrm{ld}}_i(z)\right]\right)^{-1} \epsilon.$$

Substituting into (Eq. C.19), we conclude

$$\begin{aligned}
W_1\left(\left(V^\top U^* g^*_\theta\right)_\sharp \rho, V_\sharp^\top \mu\right) &> \epsilon \cdot \left(\min_i \mathbb{E}_{z\sim\rho}\left[T^{\mathrm{ld}}_i(z)\right]\right)^{-1} \mathbb{E}_{z\sim\rho}[(g^*_\theta)_I(z) + T^{\mathrm{ld}}_I(z)] \\
&> \epsilon.
\end{aligned}$$

• **(Case 2)** The assertion of **(Case 2)** translates to several useful spectral norm bounds. We first observe

$$\begin{aligned}
\|A - U^*\|_2^2 \leq \|A - U^*\|_{\mathrm{F}}^2 &= \operatorname{tr}\left((A - U^*)^\top (A - U^*)\right) \\
&= \operatorname{tr}\left(2I - A^\top U^* - (U^*)^\top A\right) \\
&\leq 4q\left(\min_i \mathbb{E}_{z\sim\rho}\left[T^{\mathrm{ld}}_i(z)\right]\right)^{-2} \epsilon^2. \qquad \text{(C.20)}
\end{aligned}$$

Taking square root on both sides of (Eq. C.20), we have $\|A - U^*\|_2 \leq 2\sqrt{q}\left(\min_i \mathbb{E}_{z\sim\rho}\left[T^{\mathrm{ld}}_i(z)\right]\right)^{-1} \epsilon$. In addition, since A has orthonormal columns, we have

$$\left\|I - A_0^\top A_\star\right\|_2 = \left\|A_0^\top (A_0 - A_\star)\right\|_2 \leq \|A_0\|_2 \|A_0 - A_\star\|_2$$

$$\leq 2\sqrt{q}\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-1}\epsilon. \tag{C.21}$$

We use a similar proof strategy as in **(Case 1)** by choosing a specific projection matrix $V = A$, and evaluate the Wasserstein distance

$$\begin{aligned}
&W_1\left(\left(V^\top U^* g_\theta^*\right)_\sharp \rho, V_\sharp^\top \mu\right) \\
&= \sup_{f\in\mathrm{Lip}_1(\mathbb{R}^q)} \mathbb{E}_{z\sim\rho}\left[f(T^{\mathrm{ld}}(z))\right] - \mathbb{E}_{z\sim\rho}\left[f(A^\top U^* g_\theta^*(z))\right] \\
&= \sup_{f\in\mathrm{Lip}_1(\mathbb{R}^q)} \mathbb{E}_{z\sim\rho}\left[f(T^{\mathrm{ld}}(z))\right] - \mathbb{E}_{z\sim\rho}\left[f(g_\theta^*(z))\right] \\
&\quad + \mathbb{E}_{z\sim\rho}\left[f(g_\theta^*(z))\right] - \mathbb{E}_{z\sim\rho}\left[f(A^\top U^* g_\theta^*(z))\right] \\
&\geq \sup_{f\in\mathrm{Lip}_1(\mathbb{R}^q)} \mathbb{E}_{z\sim\rho}\left[f(T^{\mathrm{ld}}(z))\right] - \mathbb{E}_{z\sim\rho}\left[f(g_\theta^*(z))\right] - \mathbb{E}_{z\sim\rho}\left[\left\|(I - A^\top U^*)g_\theta^*(z)\right\|_2\right] \\
&\geq \sup_{f\in\mathrm{Lip}_1(\mathbb{R}^q)} \mathbb{E}_{z\sim\rho}\left[f(T^{\mathrm{ld}}(z))\right] - \mathbb{E}_{z\sim\rho}\left[f(g_\theta^*(z))\right] - \left\|I - A^\top U^*\right\|_2 \mathbb{E}_{z\sim\rho}\left[\|g_\theta^*(z)\|_2\right] \\
&= \underbrace{W_1(T_\sharp^{\mathrm{ld}}\rho, (g_\theta^*)_\sharp\rho)}_{(\spadesuit)} - \underbrace{\left\|I - A^\top U^*\right\|_2 \mathbb{E}_{z\sim\rho}\left[\|g_\theta^*(z)\|_2\right]}_{(\clubsuit)}.
\end{aligned} \tag{C.22}$$

Invoking inequality (Eq. C.21), ($\clubsuit$) assumes the upper bound

$$(\clubsuit) \leq 2\sqrt{q}\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-1} \mathbb{E}\left[\|g_\theta^*(z)\|_2\right]\epsilon.$$

To lower bound ($\spadesuit$), we prove a lower bound on $W_1\left((AT^{\mathrm{ld}})_\sharp\rho, (U^* g_\theta^*)_\sharp\rho\right)$. The triangle inequality implies

$$W_1\left((AT^{\mathrm{ld}})_\sharp\rho, (U^* g_\theta^*)_\sharp\rho\right) \leq W_1\left((AT^{\mathrm{ld}})_\sharp\rho, (Ag_\theta^*)_\sharp\rho\right) + W_1\left((Ag_\theta^*)_\sharp\rho, (U^* g_\theta^*)_\sharp\rho\right).$$

We bound the second term in the right-hand side above as

$$W_1\left((Ag_\theta^*)_\sharp\rho, (U^* g_\theta^*)_\sharp\rho\right) = \sup_{f\in\mathrm{Lip}_1(\mathbb{R}^d)} \mathbb{E}_{z\sim\rho}\left[f\left(Ag_\theta^*(z)\right)\right] - \mathbb{E}_{z\sim\rho}[f(U^* g_\theta^*(z))]$$

$$\begin{aligned}
&\overset{(i)}{\leq} \mathbb{E}_{z\sim\rho}\left[\|Ag_\theta^*(z) - U^* g_\theta^*(z)\|_2\right] \\
&\leq \|A - U^*\|_2 \, \mathbb{E}_{z\sim\rho}\left[\|g_\theta^*(z)\|_2\right] \\
&\overset{(ii)}{\leq} 2\sqrt{q}\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-1} \mathbb{E}_{z\sim\rho}\left[\|g_\theta^*(z)\|_2\right]\epsilon,
\end{aligned}$$

where inequality (i) invokes the Lipschitz continuity of f and inequality (ii) invokes (Eq. C.20). Recall that in (Eq. C.13), we assume

$$W_1\left((AT^{\mathrm{ld}})_\sharp\rho, (U^* g_\theta^*)_\sharp\rho\right) > \left(1 + 4\sqrt{q}\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-1} \mathbb{E}_{z\sim\rho}\left[\|g_\theta^*(z)\|_2\right]\right)\epsilon.$$

Thus, we have

$$W_1\left((U^*T^{\mathrm{ld}})_\sharp\rho, (U^* g_\theta^*)_\sharp\rho\right) > \left(1 + 2\sqrt{q}\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-1} \mathbb{E}_{z\sim\rho}\left[\|g_\theta^*(z)\|_2\right]\right)\epsilon,$$

which implies

$$(\spadesuit) = W_1(T_\sharp^{\mathrm{ld}}\rho, (g_\theta^*)_\sharp\rho) > \left(1 + 2\sqrt{q}\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-1} \mathbb{E}_{z\sim\rho}\left[\|g_\theta^*(z)\|_2\right]\right)\epsilon.$$

Combining the bounds of $(\spadesuit)$ and $(\clubsuit)$ and substituting into (Eq. C.22), we obtain

$$W_1\left(\left(V^\top U^* g_\theta^*\right)_\sharp \rho, V_\sharp^\top \mu\right) \geq (\spadesuit) - (\clubsuit) > \epsilon,$$

which checks **(Case 2)**. Putting **(Case 1)** and **(Case 2)** together, we establish inequality (Eq. C.16). Consequently, (Eq. C.14) holds true and therefore, (Eq. C.12) is valid for a global optimizer (U^*, g_θ^*).

Next, we show given (Eq. C.12), the column space of A and U^* are approximately equal. In particular, we show the following bound

$$\frac{1}{2q}\|A - U^*\|_{\mathrm{F}}^2 \leq 2 \cdot \left(1 + 4\sqrt{q}\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-1} \mathbb{E}_{z\sim\rho}\left[\|g_\theta^*(z)\|_2\right]\right)^2$$

$$\cdot \left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-2} \epsilon^2.$$

Suppose not. We expand the squared Frobenius norm $\|A - U^*\|_{\mathrm{F}}^2$ as

$$\begin{aligned}
\frac{1}{2q}\|A - U^*\|_{\mathrm{F}}^2 &= \frac{1}{2q}\operatorname{tr}\left((A-U^*)^\top (A-U^*)\right) \\
&= \frac{1}{2q}\operatorname{tr}\left(2I - A^\top U^* - (U^*)^\top A\right) \\
&= 1 - \frac{1}{2q}\operatorname{tr}\left(A^\top U^* + (U^*)^\top A\right) \\
&= 1 - \frac{1}{q}\operatorname{tr}\left(A^\top U^*\right).
\end{aligned}$$

From the last display above, we deduce

$$\begin{aligned}
\frac{1}{q}\operatorname{tr}\left(A^\top U^*\right) < 1 - 2\cdot\left(1 + 4\sqrt{q}\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-1}\mathbb{E}_{z\sim\rho}\left[\|g_\theta^*(z)\|_2\right]\right)^2 \\
\cdot \left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-2}\epsilon^2.
\end{aligned}$$

We consider distinguish $(U^*\circ g_\theta^*)_\sharp\rho$ and μ by a linear testing function $f(x) = \frac{(a_I - u_I^*)^\top}{\left\|a_I - u_I^*\right\|_2}x$, where the index I verifies

$$\begin{aligned}
a_I^\top u_I^* < 1 - 2\cdot\left(1 + 4\sqrt{q}\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-1}\mathbb{E}_{z\sim\rho}\left[\|g_\theta^*(z)\|_2\right]\right)^2 \\
\cdot \left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-2}\epsilon^2.
\end{aligned}$$

Repeating the same argument in **(Case 1)**, we deduce

$$W_1\left((U^*\circ g_\theta^*)_\sharp\rho, \mu\right) > \left(1 + 4\sqrt{q}\left(\min_i \mathbb{E}_{z\sim\rho}\left[T_i^{\mathrm{ld}}(z)\right]\right)^{-1}\mathbb{E}_{z\sim\rho}\left[\|g_\theta^*(z)\|_2\right]\right)\epsilon,$$

which contradicts (Eq. C.15). The proof is complete. □

C.2.4 Proof of Lemma 5.10

Proof. We bound $W_1(\widehat{\mu}_n, \mu)$ first. Denote $\nu = A_\sharp^\top \mu$ and $\widehat{\nu}_n = A_\sharp^\top \widehat{\mu}_n$. By Assumption 5.4, we write $W_1(\widehat{\mu}_n, \mu)$ as

$$\begin{aligned}
W_1(\widehat{\mu}_n, \mu) &= W_1(A_\sharp \widehat{\nu}_n, A_\sharp \nu) \\
&= \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \mathbb{E}_{x \sim A_\sharp \widehat{\nu}_n}[f(x)] - \mathbb{E}_{y \sim A_\sharp \nu}[f(y)] \\
&= \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \mathbb{E}_{x \sim \widehat{\nu}_n}[f(Ax)] - \mathbb{E}_{y \sim \nu}[f(Ay)] \\
&\overset{(i)}{=} \sup_{g = f \circ A} \mathbb{E}_{x \sim \widehat{\nu}_n}[g(x)] - \mathbb{E}_{y \sim \nu}[g(y)] \\
&\leq W_1(\widehat{\nu}_n, \nu). \qquad \text{(C.23)}
\end{aligned}$$

where in (i), the composite function $g = f \circ A : \mathbb{R}^q \mapsto \mathbb{R}$ is Lipschitz continuous, whose Lipschitz constant is bounded by 1. Applying Lemma 5.6, with the function class being 1-Lipschitz functions on $[0, 1]^q$, we derive

$$\begin{aligned}
W_1(\widehat{\nu}_n, \nu) &\leq 4 \inf_{\delta \in (0, \sqrt{q})} \left(\delta + \frac{6}{\sqrt{n}} \int_\delta^{\sqrt{q}} \sqrt{\log \mathcal{N}(\tau, \mathcal{H}^1([0,1]^q), \|\cdot\|_\infty)} d\tau \right) \\
&\overset{(i)}{\leq} 4 \inf_\delta \left(\delta + \frac{6}{\sqrt{n}} \int_\delta^{\sqrt{q}} \tau^{-q/2} d\tau \right) \\
&\overset{(ii)}{\leq} O\left(n^{-1/q} \log n\right), \qquad \text{(C.24)}
\end{aligned}$$

where in (i), we substitute a covering number bound $\log \mathcal{N}(\tau, \mathcal{H}^1([0,1]^q), \|\cdot\|_\infty) = O\left((1/\tau)^q\right)$, and in (ii), we take $\delta = n^{-1/q}$ and distinguish two cases depending on q:

- ($q = 2$). Inequality (i) can be simplified as

$$\begin{aligned}
W_1(\widehat{\nu}_n, \nu) &\leq \frac{4}{\sqrt{n}} + \frac{24}{\sqrt{n}} \log(\sqrt{qn}) \\
&= O\left(n^{-1/q} \log n\right).
\end{aligned}$$

- $(q > 2)$. Inequality (i) can be computed as

$$\begin{aligned}W_1(\widehat{\nu}_n, \nu) &\le 4n^{-1/q} + \frac{24}{\sqrt{n}}\frac{1}{1-q/2}\left((\sqrt{q})^{-q/2+1} - \left(n^{-1/q}\right)^{-q/2+1}\right) \\ &= O\left(n^{-1/q} + n^{-1/2}\right).\end{aligned}$$

Applying Lemma 5.6 again, with the function class being $\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}$, we further have

$$\begin{aligned}d_{\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}}(\widehat{\mu}_n, \mu) &\le 4 \inf_{\delta\in(0,\sqrt{q})}\left(\delta + \frac{6}{\sqrt{n}}\int_{\delta}^{\sqrt{q}}\sqrt{\log \mathcal{N}(\tau, \mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}, \|\cdot\|_\infty)}d\tau\right) \\ &\overset{(i)}{\le} 4\inf_{\delta}\left(\delta + \frac{6}{\sqrt{n}}\int_{\delta}^{\sqrt{q}}\sqrt{\bar{K}\log\frac{2\bar{L}^2(\bar{p}+2)(\bar{\kappa}\bar{p})^{\bar{L}+1}}{\tau}}d\tau\right) \\ &\overset{(ii)}{=} O\left(\frac{1}{n} + \frac{1}{\sqrt{n}}\sqrt{\bar{K}\bar{L}\log(\bar{L}\bar{p}n)}\right), \qquad \text{(C.25)}\end{aligned}$$

where in (i), we invoke Lemma 4.3 instantiated to $\mathcal{F}_{\mathrm{NN}}^{\mathrm{ld}}$, and in (ii), we set $\delta = \frac{1}{n}$.

□

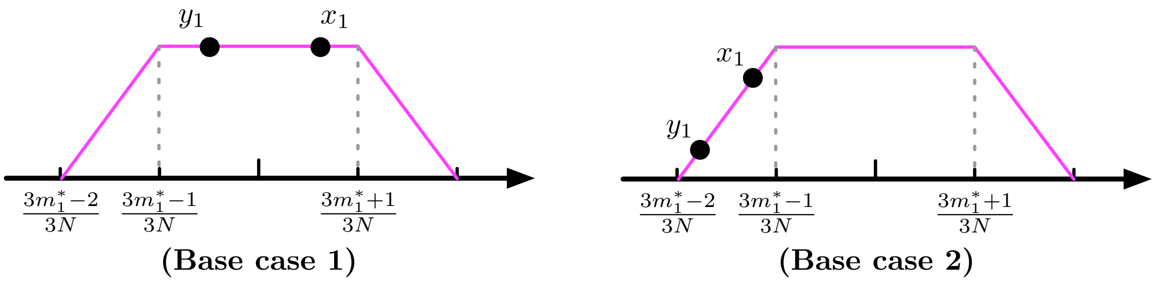
C.3 Proof of Lemma C.1

Proof. We begin by considering two points $x, y \in [0,1]^q$ differing in only one coordinate. Without loss of generality, we assume $x_1 - y_1 \ge 0$, while $x_j - y_j = 0$ for $j = 2, \dots, q$. We have two base cases:

- **(Base case 1)** there exists $m_1^* \in \{0, \dots, N\}$ such that $x_1, y_1 \in \left[\frac{3m_1^*-1}{3N}, \frac{3m_1^*+1}{3N}\right]$;
- **(Base case 2)** there exists $m_1^* \in \{0, \dots, N\}$ such that $x_1, y_1 \in \left[\frac{3m_1^*-2}{3N}, \frac{3m_1^*-1}{3N}\right]$.

In both base cases, x_1 and y_1 are close enough within distance $2/3N$. Later, we will reduce general positions of $x_1, y_1 \in [0,1]$ to a collection of base bases. A graphical illustration of base cases are given in Figure C.2.

In **(Base case 1)**, we have $\phi(3N(x_1 - m_1^*/N)) = \phi(3N(y_1 - m_1^*/N)) = 1$ and $\phi(3N(x_1 - m_1/N)) = \phi(3N(y_1 - m_1/N)) = 0$ for any $m_1 \neq m_1^*$. Therefore, the



(Base case 1)

(Base case 2)

Figure C.2: Illustration of **(Base case 1)** and **(Base case 2)**.

equality $\widehat{\xi}_m(x) = \widehat{\xi}_m(y)$ holds true for any $m \in \{0, \ldots, N\}^q$. Consequently, we deduce $\widehat{f}(x) - \widehat{f}(y) = 0$.

In **(Base case 2)**, the analysis is more complicated. We first observe that $\phi(3N(x_1 - m_1^*/N)) = 3Nx_1 - 3m_1^* + 2$ and $\phi(3N(x_1 - (m_1^* - 1)/N)) = -3Nx_1 + 3m_1^* - 1$ are both nonzero, while $\phi(3N(x_1 - m_1/N)) = 0$ for any $m_1 \notin \{m_1^* - 1, m_1^*\}$ (the same holds for y_1). Denote $m_{\backslash 1} = [m_2, \ldots, m_q]^\top$ as all the entries in m except the first entry m_1. We rewrite $\widehat{f}(x)$ as

$$\begin{aligned}\widehat{f}(x) &= \sum_m \widehat{\xi}_m(x) P_m(x) \\ &= \sum_{m=[m_1^*, m_{\backslash 1}^\top]^\top} \widehat{\xi}_m(x) P_m(x) + \sum_{m=[m_1^*-1, m_{\backslash 1}^\top]^\top} \widehat{\xi}_m(x) P_m(x).\end{aligned}$$

The second equality above holds, since $\widehat{\xi}_m(x) = 0$ whenever $m_1 \notin \{m_1^* - 1, m_1^*\}$. Furthermore, we have

$$\begin{aligned}\left|\widehat{f}(x) - \widehat{f}(y)\right| = \Bigg| &\sum_{m_{\backslash 1}: m=[m_1^*, m_{\backslash 1}^\top]^\top} \left(\widehat{\xi}_m(x) - \widehat{\xi}_m(y)\right) f(m) \\ &+ \sum_{m_{\backslash 1}: m=[m_1^*-1, m_{\backslash 1}^\top]^\top} \left(\widehat{\xi}_m(x) - \widehat{\xi}_m(y)\right) f(m)\Bigg|. \qquad \text{(C.26)}\end{aligned}$$

In order to bound the right-hand side of (Eq. C.26), we establish several regularity properties of $\widehat{\xi}_m$ based on Lemma C.4. The first result proves the monotonicity of $\widehat{\xi}_m$.

Lemma C.2. *Let* $\widehat{\xi}_m$ *be defined in* (Eq. C.10). *Consider two points* $x = [x_1, \dots, x_i, \dots, x_q]^\top$ *and* $x' = [x_1, \dots, x_i', \dots, x_q]^\top$ *only differing in the* i*-th coordinate. Denote* m_i^* *satisfying* $x_i, x_i' \in \left[\frac{3m_i^*-2}{3N}, \frac{3m_i^*-1}{3N}\right]$. *Then it holds*

$$\left(\widehat{\xi}_m(x) - \widehat{\xi}_m(x')\right)(x_i - x_i') \geq 0 \quad \textit{for} \quad m = [m_1, \dots, m_i^*, \dots, m_q]^\top \quad \textit{and}$$
$$\left(\widehat{\xi}_m(x) - \widehat{\xi}_m(x')\right)(x_i - x_i') \leq 0 \quad \textit{for} \quad m = [m_1, \dots, m_i^* - 1, \dots, m_q]^\top.$$

The proof is deferred to Appendix C.3.1. Next, we show first-order continuity of $\widehat{\xi}_m$.

Lemma C.3. *Let* $\widehat{\xi}_m$ *be defined in* (Eq. C.10). *Consider two points* $x = [x_1, \dots, x_i, \dots, x_q]^\top$ *and* $x' = [x_1, \dots, x_i', \dots, x_q]^\top$ *only differing in the* i*-th coordinate. Then for any* m*, it holds*

$$\begin{aligned} 3N \prod_{j \neq i} \max\left\{\phi(3N(x_j - m_j/N)) - \frac{1}{2^K}, 0\right\} |x_i - x_i'| &\leq \left|\widehat{\xi}_m(x) - \widehat{\xi}_m(x')\right| \\ &\leq 3N \prod_{j \neq i} \left(\phi(3N(x_j - m_j/N)) + \frac{1}{2^K}\right) |x_i - x_i'|. \end{aligned}$$

The proof is deferred to Appendix subsubsection C.3.1. Using Lemma C.2 and C.3, we are able to bound the right-hand side of (Eq. C.26). We partition all the values of $m_{\setminus 1}$ into two complementary disjoint sets:

$$\mathcal{A}_{\leq 0} = \left\{m_{\setminus 1} : f\left(\frac{[m_1^*, m_{\setminus 1}^\top]^\top}{N}\right) f\left(\frac{[m_1^* - 1, m_{\setminus 1}^\top]^\top}{N}\right) \leq 0\right\} \quad \text{and}$$
$$\mathcal{A}_{>0} = \left\{m_{\setminus 1} : f\left(\frac{[m_1^*, m_{\setminus 1}^\top]^\top}{N}\right) f\left(\frac{[m_1^* - 1, m_{\setminus 1}^\top]^\top}{N}\right) > 0\right\}.$$

In $\mathcal{A}_{\leq 0}$, by the Lipschitz continuity of f, we have

$$\left| f\left(\frac{[m_1^*, m_{\setminus 1}^\top]^\top}{N}\right) - f\left(\frac{[m_1^* - 1, m_{\setminus 1}^\top]^\top}{N}\right)\right| \leq 1/N.$$

If either $\left|f\left([m_1^*, m_{\setminus 1}^\top]^\top/N\right)\right| > \frac{1}{N}$ or $\left|f\left([m_1^* - 1, m_{\setminus 1}^\top]^\top/N\right)\right| > \frac{1}{N}$, then $f\left([m_1^*, m_{\setminus 1}^\top]^\top/N\right)$

and $f\big([m_1^*, m_{\backslash 1}^\top]^\top/N\big)$ should be both positive or negative. Their product must be positive. As a result, we deduce that in $\mathcal{A}_{\leq 0}$,

$$\left|f\left(\frac{[m_1^*, m_{\backslash 1}^\top]^\top}{N}\right)\right| \leq \frac{1}{N} \quad \text{and} \quad \left|f\left(\frac{[m_1^* - 1, m_{\backslash 1}^\top]^\top}{N}\right)\right| \leq \frac{1}{N}$$

hold simultaneously.

In $\mathcal{A}_{>0}$, $f\left([m_1^*, m_{\backslash 1}^\top]^\top/N\right)$ and $f\left([m_1^*, m_{\backslash 1}^\top]^\top/N\right)$ are both positive or negative. We rewrite (Eq. C.26) according to the partition of $\mathcal{A}_{\leq 0}$ and $\mathcal{A}_{>0}$ on $m_{\backslash 1}$:

$$\left|\widehat{f}(x) - \widehat{f}(y)\right| = (\spadesuit) + (\clubsuit), \tag{C.27}$$

where

$$\begin{aligned}
(\spadesuit) = \Bigg| & \sum_{m_{\backslash 1} \in \mathcal{A}_{\leq 0} : m = [m_1^*, m_{\backslash 1}^\top]^\top} \left(\widehat{\xi}_m(x) - \widehat{\xi}_m(y)\right) f(m/N) \\
& + \sum_{m_{\backslash 1} \in \mathcal{A}_{\leq 0} : m = [m_1^* - 1, m_{\backslash 1}^\top]^\top} \left(\widehat{\xi}_m(x) - \widehat{\xi}_m(y)\right) f(m/N)\Bigg|, \\
(\clubsuit) = \Bigg| & \sum_{m_{\backslash 1} \in \mathcal{A}_{>0} : m = [m_1^*, m_{\backslash 1}^\top]^\top} \left(\widehat{\xi}_m(x) - \widehat{\xi}_m(y)\right) f(m/N) \\
& + \sum_{m_{\backslash 1} \in \mathcal{A}_{>0} : m = [m_1^* - 1, m_{\backslash 1}^\top]^\top} \left(\widehat{\xi}_m(x) - \widehat{\xi}_m(y)\right) f(m/N)\Bigg|.
\end{aligned}$$

For term $(\spadesuit)$, we bound it by

$$\begin{aligned}
& (\spadesuit) \\
& \leq \left|\sum_{m_{\backslash 1} \in \mathcal{A}_{\leq 0} : m = [m_1^*, m_{\backslash 1}^\top]^\top} \left(\widehat{\xi}_m(x) - \widehat{\xi}_m(y)\right) \frac{1}{N}\right| \\
& \quad + \left|\sum_{m_{\backslash 1} \in \mathcal{A}_{\leq 0} : m = [m_1^* - 1, m_{\backslash 1}^\top]^\top} \left(\widehat{\xi}_m(x) - \widehat{\xi}_m(y)\right) \frac{1}{N}\right|
\end{aligned}$$

$$
\begin{aligned}
&\overset{(i)}{\leq} \frac{1}{N} \sum_{m_{\setminus 1}:\widehat{\xi}_m(x)\neq 0, \widehat{\xi}_m(y)\neq 0} 6N|x_1 - y_1| \prod_{k\geq 2} \min\left\{\phi(3N(x_k - m_k/N)) + \frac{1}{2^K}, 1\right\} \\
&\leq 6|x_1 - y_1| \sum_{m_{\setminus 1}:\widehat{\xi}_m(x)\neq 0, \widehat{\xi}_m(y)\neq 0} \prod_{k\geq 2} \left(\phi(3N(x_k - m_k/N)) + \frac{1}{2^K}\right) \\
&\overset{(ii)}{\leq} 6|x_1 - y_1| \sum_{m_{\setminus 1}:\widehat{\xi}_m(x)\neq 0, \widehat{\xi}_m(y)\neq 0} \left[\prod_{k\geq 2} \phi(3N(x_k - m_k/N)) + \sum_{j=1}^{q} 2^{-jK}\binom{q}{j}\right] \\
&\overset{(iii)}{\leq} 6\left(1 + q2^{-K+q-1}\frac{1-\left(q2^{-K}\right)^q}{1-q2^{-K}}\right)|x_1 - y_1|, \qquad \text{(C.28)}
\end{aligned}
$$

where inequality (i) invokes Lemma C.3 and neglects terms involving $\widehat{\xi}_m(x) = \widehat{\xi}_m(y) = 0$ and inequality (ii) expands the product $\prod_{k\geq 2}\left(\phi(3N(x_k - m_k/N)) + \frac{1}{2^K}\right)$ by noting $\phi(3N(x_k - m_k/N)) \leq 1$. To see inequality (iii), we first observe that there are at most 2^{q-1} terms in the summation, due to the definition of ϕ. Then we bound $\sum_{m_{\setminus 1}:\widehat{\xi}_m(x)\neq 0, \widehat{\xi}_m(y)\neq 0} \sum_{j=1}^{q} 2^{-jK}\binom{q}{j}$ as

$$
\begin{aligned}
\sum_{m_{\setminus 1}:\widehat{\xi}_m(x)\neq 0, \widehat{\xi}_m(y)\neq 0} \sum_{j=1}^{q} 2^{-jK}\binom{q}{j} &\leq \sum_{m_{\setminus 1}:\widehat{\xi}_m(x)\neq 0, \widehat{\xi}_m(y)\neq 0} \sum_{j=1}^{q} 2^{-jK} q^j \\
&\leq 2^{q-1} q 2^{-K} \frac{1-\left(q2^{-K}\right)^q}{1-q2^{-K}} \\
&= q2^{-K+q-1}\frac{1-\left(q2^{-K}\right)^q}{1-q2^{-K}}.
\end{aligned}
$$

Meanwhile, $\prod_{k\geq 2}\phi(3N(x_k - m_k/N))$ is indeed a partition of unity on a $(d-1)$-dimensional unit cude. Therefore, we have

$$
\sum_{m_{\setminus 1}:\widehat{\xi}_m(x)\neq 0, \widehat{\xi}_m(y)\neq 0} \prod_{k\geq 2} \phi(3N(x_k - m_k/N)) = 1.
$$

For term (♣), we leverage the cancellation in the two summations. We assume without loss of generality, $f\left([m_1^*, m_{\setminus 1}^\top]^\top/N\right) > 0$ and $f\left([m_1^* - 1, m_{\setminus 1}^\top]^\top/N\right) > 0$ for $m_{\setminus 1} \in \mathcal{A}_{>0}$.

Otherwise, replacing f by $-f$ won't change term (♣). Therefore, we derive

$$
\begin{aligned}
&(\clubsuit) \\
&\overset{(i)}{\leq} \Bigg| \sum_{m_{\setminus 1} \in \mathcal{A}_{>0} : m = [m_1^*, m_{\setminus 1}^\top]^\top} \left(\widehat{\xi}_m(x) - \widehat{\xi}_m(y) \right) f(m/N) \\
&\quad - \sum_{m_{\setminus 1} \in \mathcal{A}_{>0} : m = [m_1^* - 1, m_{\setminus 1}^\top]^\top} \left| \widehat{\xi}_m(x) - \widehat{\xi}_m(y) \right| f(m/N) \Bigg| \\
&\overset{(ii)}{\leq} 3N|x_1 - y_1| \\
&\quad \cdot \sum_{m_{\setminus 1} : \widehat{\xi}_m(x) \neq 0, \widehat{\xi}_m(y) \neq 0} \Bigg| \prod_{k \geq 2} \left(\phi(3N(x_k - m_k/N)) + \frac{1}{2^K} \right) f\left([m_1^*, m_{\setminus 1}^\top]^\top / N \right) \\
&\qquad\qquad - \prod_{k \geq 2} \max \left\{ \phi(3N(x_k - m_k/N)) - \frac{1}{2^K}, 0 \right\} f\left([m_1^* - 1, m_{\setminus 1}^\top]^\top / N \right) \Bigg| \\
&\overset{(iii)}{\leq} 3N|x_1 - y_1| \sum_{m_{\setminus 1} : \widehat{\xi}_m(x) \neq 0, \widehat{\xi}_m(y) \neq 0} \Bigg(\prod_{k \geq 2} \phi(3N(x_k - m_k/N)) \\
&\qquad\qquad \cdot \left| f([m_1^*, m_{\setminus 1}^\top]^\top / N) - f([m_1^* - 1, m_{\setminus 1}^\top]^\top / N) \right| \Bigg) \\
&\quad + 6N|x_1 - y_1| \, \|f\|_\infty \cdot q 2^{-K+q-1} \frac{1 - \left(q 2^{-K} \right)^q}{1 - q 2^{-K}} \\
&\leq 3 \left(1 + 2N \, \|f\|_\infty \cdot q 2^{-K+q-1} \frac{1 - \left(q 2^{-K} \right)^q}{1 - q 2^{-K}} \right) |x_1 - y_1|,
\end{aligned}
\tag{C.29}
$$

where inequality (i) uses the monotonicity of $\widehat{\xi}_m$ in Lemma C.2, inequality (ii) invokes Lemma C.3, and inequality (iii) follows from the same argument of (iii) in (Eq. C.28). Combining (Eq. C.28), (Eq. C.29) and substituting into (Eq. C.27), we obtain

$$
\left| \widehat{f}(x) - \widehat{f}(y) \right| \leq 3 \left(3 + 2(N \, \|f\|_\infty + 1) \cdot q 2^{-K+q-1} \frac{1 - \left(q 2^{-K} \right)^q}{1 - q 2^{-K}} \right) |x_1 - y_1|. \tag{C.30}
$$

Given two base cases, we proceed to show Lipschitz continuity of $\widehat{f}$. We first partition $[0, 1]$ into two types of sub-intervals,

$$\textit{(Type 1)}\ \left[\frac{3k-1}{3N}, \frac{3k+1}{3N}\right] \bigcap [0,1] \quad \text{and} \quad \textit{(Type 2)}\ \left[\frac{3k+1}{3N}, \frac{3k+2}{3N}\right] \bigcap [0,1],$$

where $k \leq N$ is an integer. We observe that on a *(Type 1)* sub-interval, **(Base case 1)** applies; while on a *(Type 2)* sub-interval, **(Base case 2)** applies. Depending on the location of x_1 and y_1, we discuss four situations.

(Situation 1): x_1 belongs to a *(Type 1)* sub-interval and y_1 belongs to a *(Type 1)* sub-interval. If the two sub-intervals coincide, we obtain **(Base case 1)**. There is nothing to show. Otherwise, we denote integer k_x such that $x_1 \in \left[\frac{3k_x-1}{3N}, \frac{3k_x+1}{3N}\right] \bigcap [0,1]$ and integer $k_y < k_x$ such that $y_1 \in \left[\frac{3k_y-1}{3N}, \frac{3k_y+1}{3N}\right] \bigcap [0,1]$. See Figure C.3 for an illustration.

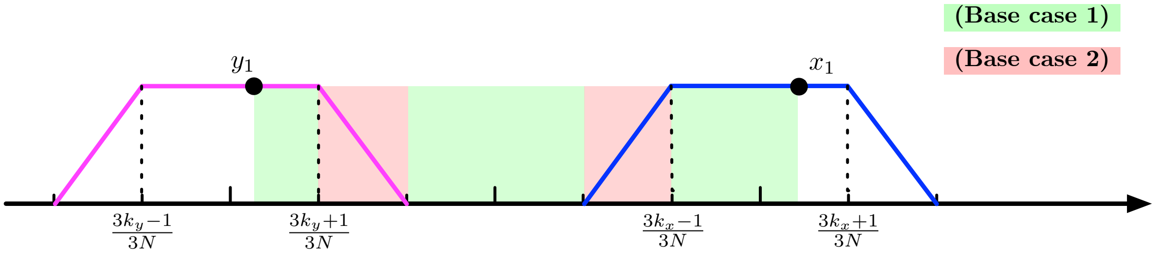


Figure C.3: Demonstration of **(Situation 1)** with $k_x = k_y + 2$. We can decompose such an situation into a serial of alternating **(Base case 1)** (green) and **(Base case 2)** (red). The function value difference can be obtained by aggregating differences in each base case.

We can derive

$$\begin{aligned}
\left|\widehat{f}(x) - \widehat{f}(y)\right| \leq & \left|\widehat{f}(x) - \widehat{f}\left(\left[\frac{3k_x-1}{3N}, x_{\setminus 1}^\top\right]^\top\right)\right| \\
& + \left|\widehat{f}\left(\left[\frac{3k_x-1}{3N}, x_{\setminus 1}^\top\right]^\top\right) - \widehat{f}\left(\left[\frac{3k_y+1}{3N}, y_{\setminus 1}^\top\right]^\top\right)\right| \\
& + \left|\widehat{f}\left(\left[\frac{3k_y+1}{3N}, y_{\setminus 1}^\top\right]^\top\right) - \widehat{f}(y)\right| \\
\overset{(i)}{=} & \left|\widehat{f}\left(\left[\frac{3k_x-1}{3N}, x_{\setminus 1}^\top\right]^\top\right) - \widehat{f}\left(\left[\frac{3k_y+1}{3N}, y_{\setminus 1}^\top\right]^\top\right)\right|,
\end{aligned}$$

where inequality (i) follows from **(Base case 1)**. If $k_x = k_y + 1$, then we can apply **(Base

case 2) to show

$$\begin{aligned}\left|\widehat{f}(x)-\widehat{f}(y)\right| &\leq \left|\widehat{f}\left(\left[\frac{3k_x-1}{3N}, x_{\backslash 1}^\top\right]^\top\right)-\widehat{f}\left(\left[\frac{3k_y+1}{3N}, y_{\backslash 1}^\top\right]^\top\right)\right| \\ &\leq 3\left(3+2(N\|f\|_\infty+1)\cdot q2^{-K+q-1}\frac{1-\left(q2^{-K}\right)^q}{1-q2^{-K}}\right)\frac{1}{3N} \\ &\leq 3\left(3+2(N\|f\|_\infty+1)\cdot q2^{-K+q-1}\frac{1-\left(q2^{-K}\right)^q}{1-q2^{-K}}\right)|x_1-y_1|.\end{aligned}$$

Otherwise, we have

$$\begin{aligned}&\left|\widehat{f}(x)-\widehat{f}(y)\right| \\ &\leq \left|\widehat{f}\left(\left[\frac{3k_x-1}{3N}, x_{\backslash 1}^\top\right]^\top\right)-\widehat{f}\left(\left[\frac{3k_y+1}{3N}, y_{\backslash 1}^\top\right]^\top\right)\right| \\ &\leq \left|\widehat{f}\left(\left[\frac{3k_x-1}{3N}, x_{\backslash 1}^\top\right]^\top\right)-\widehat{f}\left(\left[\frac{3(k_x-1)+1}{3N}, x_{\backslash 1}^\top\right]^\top\right)\right| \\ &\quad+\left|\widehat{f}\left(\left[\frac{3(k_x-1)+1}{3N}, x_{\backslash 1}^\top\right]^\top\right)-\widehat{f}\left(\left[\frac{3(k_y+1)-1}{3N}, y_{\backslash 1}^\top\right]^\top\right)\right| \\ &\quad+\left|\widehat{f}\left(\left[\frac{3(k_y+1)-1}{3N}, y_{\backslash 1}^\top\right]^\top\right)-\widehat{f}\left(\left[\frac{3k_y+1}{3N}, y_{\backslash 1}^\top\right]^\top\right)\right| \\ &\overset{(i)}{\leq} 3\left(3+2(N\|f\|_\infty+1)\cdot q2^{-K+q-1}\frac{1-\left(q2^{-K}\right)^q}{1-q2^{-K}}\right)\frac{2}{3N} \\ &\quad+\left|\widehat{f}\left(\left[\frac{3(k_x-1)+1}{3N}, x_{\backslash 1}^\top\right]^\top\right)-\widehat{f}\left(\left[\frac{3(k_y+1)-1}{3N}, y_{\backslash 1}^\top\right]^\top\right)\right|,\end{aligned}$$

where inequality (i) is obtained by applying **(Base case 2)** twice. To complete the argument, we can replace $k_x = k_x-1$ and $k_y = k_y+1$ and repeat the derivation to accumulate all the differences yielded on a *(Type 2)* sub-interval, until $k_x - i = k_y + i + 1$ or $k_x - i = k_y + i$ for some integer i. Consequently, noting that the total length of *(Type 2)* interval between

x_1 and y_1 is always smaller than $|x_1 - y_1|$, we deduce

$$\left|\widehat{f}(x) - \widehat{f}(y)\right| \leq 3\left(3 + 2(N\left\|f\right\|_\infty + 1) \cdot q2^{-K+q-1}\frac{1 - \left(q2^{-K}\right)^q}{1 - q2^{-K}}\right)|x_1 - y_1|.$$

(Situation 2): x_1 belongs to a *(Type 1)* sub-interval and y_1 belongs to a *(Type 2)* sub-interval. We aim to reduce this situation to **(Situation 1)**. Following the same notation, we denote $x_1 \in \left[\frac{3k_x - 1}{3N}, \frac{3k_x + 1}{3N}\right] \bigcap [0, 1]$ for some integer k_x and $y_1 \in \left[\frac{3k_y + 1}{3N}, \frac{3k_y + 2}{3N}\right] \bigcap [0, 1]$ for $k_y < k_x$. Triangle inequality yields

$$\begin{aligned}
&\left|\widehat{f}(x) - \widehat{f}(y)\right| \\
\leq &\left|\widehat{f}(x) - \widehat{f}\left(\left[\frac{3(k_y + 1) - 1}{3N}, y_{\setminus 1}^\top\right]^\top\right)\right| + \left|\widehat{f}\left(\left[\frac{3(k_y + 1) - 1}{3N}, x_{\setminus 1}^\top\right]^\top\right) - \widehat{f}(y)\right| \\
\leq &\left|\widehat{f}(x) - \widehat{f}\left(\left[\frac{3(k_y + 1) - 1}{3N}, y_{\setminus 1}^\top\right]^\top\right)\right| \\
&+ 3\left(3 + 2(N\left\|f\right\|_\infty + 1) \cdot q2^{-K+q-1}\frac{1 - \left(q2^{-K}\right)^q}{1 - q2^{-K}}\right)\left|\frac{3(k_y + 1) - 1}{3N} - y_1\right|.
\end{aligned}$$

We now observe that term $\left|\widehat{f}(x) - \widehat{f}\left(\left[\frac{3(k_y+1)-1}{3N}, y_{\setminus 1}^\top\right]^\top\right)\right|$ falls into **(Situation 1)**. A straightforward adaptation of the argument in **(Situation 1)** gives rise to

$$\left|\widehat{f}(x) - \widehat{f}(y)\right| \leq 3\left(3 + 2(N\left\|f\right\|_\infty + 1) \cdot q2^{-K+q-1}\frac{1 - \left(q2^{-K}\right)^q}{1 - q2^{-K}}\right)|x_1 - y_1|.$$

(Situation 3): x_1 belongs to a *(Type 2)* sub-interval and y_1 belongs to a *(Type 1)* sub-interval. The analysis is analogous to **(Situation 2)** by switching x_1 and y_1. Denoting $x_1 \in \left[\frac{3k_x + 1}{3N}, \frac{3k_x + 2}{3N}\right] \bigcap [0, 1]$ for some integer k_x and $y_1 \in \left[\frac{3k_y - 1}{3N}, \frac{3k_y + 1}{3N}\right] \bigcap [0, 1]$ for $k_y \leq k_x$, we derive

$$\left|\widehat{f}(x) - \widehat{f}(y)\right|$$

$$
\begin{aligned}
&\leq \left|\widehat{f}(x) - \widehat{f}\left(\left[\frac{3k_x+1}{3N}, x_{\backslash 1}^\top\right]^\top\right)\right| + \left|\widehat{f}\left(\left[\frac{3k_x+1}{3N}, x_{\backslash 1}^\top\right]^\top\right) - \widehat{f}(y)\right| \\
&\leq \left|\widehat{f}\left(\left[\frac{3k_x+1}{3N}, x_{\backslash 1}^\top\right]^\top\right) - \widehat{f}(y)\right| \\
&\quad + 3\left(3 + 2(N\|f\|_\infty + 1) \cdot q2^{-K+q-1}\frac{1-\left(q2^{-K}\right)^q}{1-q2^{-K}}\right)\left|x_1 - \frac{3k_x+1}{3N}\right|.
\end{aligned}
$$

Note that $\left|\widehat{f}\left(\left[\frac{3k_x+1}{3N}, x_{\backslash 1}^\top\right]^\top\right) - \widehat{f}(y)\right|$ falls into **(Situation 1)**. Therefore, the desired Lipschitz continuity (Eq. C.30) holds.

(Situation 4): x_1 belongs to a *(Type 2)* sub-interval and y_1 belongs to a *(Type 2)* sub-interval. If the two sub-intervals coincide, this recovers **(Base case 2)**, and there is nothing to show. Otherwise, applying the analysis in **(Situation 2)** and **(Situation 3)** consecutively to move x_1 first into a *(Type 1)* sub-interval and then y_1, we reduce this situation to **(Situation 1)** again. Therefore, Lipschitz continuity in (Eq. C.30) still holds true.

Combining all four situations, for any x, y only differing in the first coordinate, it holds

$$
\left|\widehat{f}(x) - \widehat{f}(y)\right| \leq 3\left(3 + 2(N\|f\|_\infty + 1) \cdot q2^{-K+q-1}\frac{1-\left(q2^{-K}\right)^q}{1-q2^{-K}}\right)|x_1 - y_1|.
$$

The proof is complete for general x, y by aggregating coordinate-wise differences and the fact $\sum_{i=1}^q |x_i - y_i| = \|x-y\|_1 \leq q\|x-y\|_\infty$. □

C.3.1 Proofs of supporting results for Lemma C.1

Before we present omitted proofs in Lemma C.1, we study the regularity of the approximated square function $\widehat{h}_K$, which will be frequently used in proving Lemma C.2 and Lemma C.3.

Regularity of $\widehat{h}_K$

Given a function $g : [0,1] \mapsto \mathbb{R}$, for any $x \in (0,1)$, we define upper and lower slopes at x, denoted by $\overline{\mathsf{slope}}_g$ and $\underline{\mathsf{slope}}_g$, respectively, as

$$\overline{\mathsf{slope}}_g(x) = \limsup_{\Delta \to 0} \frac{g(x+\Delta) - g(x)}{\Delta},$$
$$\underline{\mathsf{slope}}_g(x) = \liminf_{\Delta \to 0} \frac{g(x+\Delta) - g(x)}{\Delta}.$$

The definition above coincides with upper and lower derivatives of a univariate function. We use "slope" instead of derivatives as we will instantiate the definition to the piecewise linear function $\widehat{h}_K$. We show several useful properties.

Lemma C.4. *For a given positive integer K, let $\widehat{h}_K$ be defined on $[0,1]$ as in* (Eq. C.9). *Then the following identities hold.*

1. *For any $x \in (0,1)$, we have*

$$\overline{\mathsf{slope}}_{\widehat{h}_K}(x) = \lim_{\Delta \to 0^+} \frac{\widehat{h}_K(x+\Delta) - \widehat{h}_K(x)}{\Delta} \quad \textit{and} \tag{C.31}$$
$$\underline{\mathsf{slope}}_{\widehat{h}_K}(x) = \lim_{\Delta \to 0^+} \frac{\widehat{h}_K(x) - \widehat{h}_K(x-\Delta)}{\Delta}. \tag{C.32}$$

2. *Given an integer i, we denote $\mathcal{B}_K(i) = [b_1, \ldots, b_K]^\top \in \{0,1\}^K$ as the K-bit binary encoding of i, that is, $i = \sum_{k=1}^K b_k 2^{k-1}$. Then for any $x \in (0,1)$, we have*

$$\overline{\mathsf{slope}}_{\widehat{h}_K}(x) = 1 + \sum_{k=1}^K \Big(2\left[\mathcal{B}_K\left(\lfloor 2^K \cdot x \rfloor\right)\right]_{K-k+1} - 1\Big) 2^{-k} \quad \textit{and} \tag{C.33}$$
$$\underline{\mathsf{slope}}_{\widehat{h}_K}(x) = 1 + \sum_{k=1}^K \Big(2\left[\mathcal{B}_K\left(\lceil 2^K \cdot x \rceil - 1\right)\right]_{K-k+1} - 1\Big) 2^{-k}. \tag{C.34}$$

Proof of Lemma C.4. By construction, g_k is a piecewise linear function. Each of its linear segment is supported on a sub-interval $[i/2^k, (i+1)/2^k]$ for $i = 1, \ldots, 2^k - 1$. Therefore, it can be checked that $\widehat{h}_K$ is also a piecewise linear function, since it is a linear combination of g_k's. Furthermore, the i-th linear segment in g_k has a slope $(-1)^i 2^k$, i.e., $g_k' = (-1)^i 2^k$ on open interval $\big(i/2^k, (i+1)/2^k\big)$. As a result, $\widehat{h}_K$ is differentiable on $\big(i/2^K, (i+1)/2^K\big)$ and its derivative satisfies

$$\begin{aligned}\widehat{h}_K'(x) &= 1 - \sum_{k=1}^{K} \frac{1}{2^{2k}} g_k'(x) \\ &= 1 - \sum_{k=1}^{K} (-1)^{\left\lfloor i/2^{K-k} \right\rfloor} \frac{1}{2^k} \quad \text{for any } x \in \big(i/2^K, (i+1)/2^K\big)\,. \end{aligned} \tag{C.35}$$

We observe $i = \lfloor x \cdot 2^K \rfloor$ for $x \in \big(i/2^K, (i+1)/2^K\big)$, which implies $x \cdot 2^K - 1 < i \leq x \cdot 2^K$. For any $k = 1, \ldots, K$, we have

$$\frac{x \cdot 2^K - 1}{2^{K-k}} < i/2^{K-k} \leq \frac{x \cdot 2^K}{2^{K-k}} \quad \Longrightarrow \quad x \cdot 2^k - 1 < i/2^{K-k} \leq x \cdot 2^k.$$

Thus, we deduce $\left\lfloor i/2^{K-k} \right\rfloor = \left\lfloor x \cdot 2^k \right\rfloor$ and (Eq. C.35) can be simplified as

$$\widehat{h}_K'(x) = 1 - \sum_{k=1}^{K} (-1)^{\left\lfloor x \cdot 2^k \right\rfloor} \frac{1}{2^k}. \tag{C.36}$$

We claim

$$(-1)^{\left\lfloor x \cdot 2^k \right\rfloor} = -2b_{K-k+1} + 1 \quad \text{for} \quad k = 1, \ldots, K, \tag{C.37}$$

where b_j is the j-th entry of the K-bit binary encoding $\mathcal{B}_K\left(\left\lfloor x \cdot 2^K \right\rfloor\right)$. In other words, the parity of $\left\lfloor x \cdot 2^k \right\rfloor$ is encoded by b_{K-k+1}. In particular, if $\left\lfloor x \cdot 2^k \right\rfloor$ is odd (resp. even), $b_{K-k+1} = 1$ (resp. $b_{K-k+1} = 0$).

To show the claim in (Eq. C.37), we first prove

$$\left\lfloor x \cdot 2^k \right\rfloor = \sum_{j=1}^{k} 2^{k-j} b_{K-j+1}$$

for any $k = 1, \ldots, K$. Indeed, we can show the following sandwich inequality

$$\left\lfloor \frac{\lfloor x \cdot 2^K \rfloor}{2^{K-k}} \right\rfloor \overset{(i)}{\leq} \lfloor x \cdot 2^k \rfloor \overset{(ii)}{\leq} \frac{\lfloor x \cdot 2^K \rfloor}{2^{K-k}}. \tag{C.38}$$

Inequality (i) holds, since $\left\lfloor \frac{\lfloor x \cdot 2^K \rfloor}{2^{K-k}} \right\rfloor \leq \left\lfloor \frac{x \cdot 2^K}{2^{K-k}} \right\rfloor = \lfloor x \cdot 2^k \rfloor$; inequality (ii) holds, since $2^{K-k} \lfloor x \cdot 2^k \rfloor = \lfloor 2^{K-k} \lfloor x \cdot 2^k \rfloor \rfloor \leq \lfloor x \cdot 2^K \rfloor$. Substituting $\lfloor x \cdot 2^K \rfloor = \sum_{j=1}^{K} b_j 2^{j-1}$ into (Eq. C.38), we derive

$$\begin{aligned} \left\lfloor \sum_{j > K-k} b_j 2^{j-1-K+k} + \sum_{j \leq K-k} b_j 2^{j-1-K+k} \right\rfloor &\leq \lfloor x \cdot 2^k \rfloor \\ &\leq \sum_{j > K-k} b_j 2^{j-1-K+k} + \underbrace{\sum_{j \leq K-k} b_j 2^{j-1-K+k}}_{(\spadesuit)<1}. \end{aligned}$$

Due to $(\spadesuit) < 1$, we conclude $\lfloor x \cdot 2^k \rfloor = \sum_{j=1}^{k} 2^{k-j} b_{K-j+1}$. Consequently, we deduce $\lfloor x \cdot 2^k \rfloor \equiv b_{K-k+1} \pmod 2$, which verifies the claim by noting $(-1)^{\lfloor x \cdot 2^k \rfloor} = (-1)^{b_{K-k+1}} = -2b_{K-k+1} + 1$.

Substituting (Eq. C.37) into (Eq. C.36), for $x \in \big(i/2^K, (i+1)/2^K\big)$, we obtain

$$\widehat{h}'_K(x) = 1 + \sum_{k=1}^{K} (2b_{K-k+1} - 1) 2^{-k} \quad \text{with} \quad \mathcal{B}_K\left(\lfloor x \cdot 2^K \rfloor\right) = [b_1, \ldots, b_K]^\top. \tag{C.39}$$

To establish the first assertion in Lemma C.4, we only need to consider end points of each linear segment of $\widehat{h}_K$. Otherwise, when $x \in \big(i/2^K, (i+1)/2^K\big)$ for some $i = 0, \ldots, 2^K - 1$, (Eq. C.39) shows $\widehat{h}_K$ is differentiable at x, and therefore, (Eq. C.31) holds true. Consider an end point $x = i/2^K$ for some $i = 1, \ldots, 2^K - 1$. We evaluate left and right derivatives of $\widehat{h}_K$ at x. We denote left and right derivatives as $\partial^- \widehat{h}_K$ and $\partial^+ \widehat{h}_K$, respectively. Using (Eq. C.39) again, we derive

$$\partial^- \widehat{h}_K(x) = \lim_{\Delta \to 0^+} \frac{\widehat{h}_K(x - \Delta) - \widehat{h}_K(x)}{\Delta}$$

$$
\begin{aligned}
&= \lim_{y \to x^-} \widehat{h}'_K(y) \\
&= 1 + \sum_{k=1}^{K} \left(2[\mathcal{B}_K(i-1)]_{K-k+1} - 1\right) 2^{-k}, \\
\partial^+ \widehat{h}_K(x) &= \lim_{\Delta \to 0^+} \frac{\widehat{h}_K(x+\Delta) - \widehat{h}_K(x)}{\Delta} \\
&= \lim_{y \to x^+} \widehat{h}'_K(y) \\
&= 1 + \sum_{k=1}^{K} \left(2[\mathcal{B}_K(i)]_{K-k+1} - 1\right) 2^{-k}.
\end{aligned}
$$

We note $\partial^+ \widehat{h}_K(x) \geq \partial^- \widehat{h}_K(x)$, and therefore, for $x = i/2^K$, we obtain

$$
\begin{aligned}
\overline{\mathsf{slope}}_{\widehat{h}_K}(x) &= \limsup_{\Delta \to 0} \frac{\widehat{h}_K(x+\Delta) - \widehat{h}_K(x)}{\Delta} = \partial^+ \widehat{h}_K(x), \\
\underline{\mathsf{slope}}_{\widehat{h}_K}(x) &= \liminf_{\Delta \to 0} \frac{\widehat{h}_K(x+\Delta) - \widehat{h}_K(x)}{\Delta} = \partial^- \widehat{h}_K(x).
\end{aligned}
\tag{C.40}
$$

This establishes the first assertion in Lemma C.4.

To show the second assertion, we also tackle separately when x is an end point of a linear segment or inside a linear segment of $\widehat{h}_K$. Suppose $x \in \left(i/2^K, (i+1)/2^K\right)$ for some $i = 0, \ldots, 2^K - 1$. We check that $\lfloor x \cdot 2^K \rfloor = \lceil x \cdot 2^K \rceil - 1 = i$. It implies (Eq. C.33) and (Eq. C.34) are both equal to (Eq. C.39). On the other hand, suppose $x = i/2^K$ for some $i = 1, \ldots, 2^K - 1$, we check $\lfloor x \cdot 2^K \rfloor = \lceil x \cdot 2^K \rceil = i$. Therefore, (Eq. C.33) and (Eq. C.34) coincide with $\partial^+ \widehat{h}_K(x)$ and $\partial^- \widehat{h}_K(x)$, respectively. In combination with (Eq. C.40), we verify that (Eq. C.33) and (Eq. C.34) hold for any $x \in (0,1)$. The proof is complete. □

For later convenience, we define slopes at end points $x = 0$ and $x = 1$ as

$$
\begin{aligned}
&\overline{\mathsf{slope}}_{\widehat{h}_K}(1) = 2, && \underline{\mathsf{slope}}_{\widehat{h}_K}(1) = \lim_{x \to 1^-} \widehat{h}'_K(x) = 2 - 2^{-K}, \\
&\underline{\mathsf{slope}}_{\widehat{h}_K}(0) = 0, && \overline{\mathsf{slope}}_{\widehat{h}_K}(0) = \lim_{x \to 0^+} \widehat{h}'_K(x) = 2^{-K}.
\end{aligned}
$$

Proof of Lemma C.2

Proof. We first show $\widehat{\times}(x, a)$ is monotone in x for any fixed a. Let $x_1 \leq x_2 \in [0, 1]$. (We slightly abuse the notation here. Note that x_1, x_2 are scalars.) By the construction of $\widehat{\times}$, we have

$$\begin{aligned}\widehat{\times}(x_2, a) - \widehat{\times}(x_1, a) = &\underbrace{\widehat{h}_K\left(\frac{x_2 + a}{2}\right) - \widehat{h}_K\left(\frac{x_1 + a}{2}\right)}_{(A)} \\ &- \underbrace{\left(\widehat{h}_K\left(\frac{|x_2 - a|}{2}\right) - \widehat{h}_K\left(\frac{|x_1 - a|}{2}\right)\right)}_{(B)}.\end{aligned}$$

By the triangle inequality, we observe

$$\left|\frac{|x_1 - a|}{2} - \frac{|x_2 - a|}{2}\right| \leq \left|\frac{|x_1 - a - x_2 + a|}{2}\right| = \left|\frac{x_1 + a}{2} - \frac{x_2 + a}{2}\right|,$$

and $\frac{x_2+a}{2} \geq \max\left\{\frac{|x_1-a|}{2}, \frac{|x_2-a|}{2}, \frac{x_1+a}{2}\right\}$.

We need to compare the differences in term (A) and (B) in the following two cases.

- If $\frac{x_1+a}{2} \geq \max\left\{\frac{|x_2-a|}{2}, \frac{|x_1-a|}{2}\right\}$, we have

$$\begin{aligned}(A) - (B) \geq &\overline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x_1 + a}{2}\right)\left|\frac{x_2 + a}{2} - \frac{x_1 + a}{2}\right| \\ &- \underline{\mathsf{slope}}_{\widehat{h}_K}\left(\max\left\{\frac{|x_2 - a|}{2}, \frac{|x_1 - a|}{2}\right\}\right)\left|\frac{|x_2 - a|}{2} - \frac{|x_1 - a|}{2}\right|.\end{aligned}$$

By the triangle inequality, we observe

$$\left|\frac{|x_1 - a|}{2} - \frac{|x_2 - a|}{2}\right| \leq \left|\frac{|x_1 - a - x_2 + a|}{2}\right| = \left|\frac{x_1 + a}{2} - \frac{x_2 + a}{2}\right|.$$

Meanwhile, by Lemma C.4, $\overline{\mathsf{slope}}_{\widehat{h}_K}(z_1) \geq \underline{\mathsf{slope}}_{\widehat{h}_K}(z_2)$ whenever $z_1 \geq z_2$. Therefore, we verify $(A)-(B)\geq 0$.

• If on the contrary, $\frac{x_1+a}{2} < \max\left\{\frac{|x_2-a|}{2}, \frac{|x_1-a|}{2}\right\}$, by removing overlapping pieces, we have

$$\begin{aligned}
&(A)-(B)\\
&=\widehat{h}_K\left(\frac{x_2+a}{2}\right)-\widehat{h}_K\left(\max\left\{\frac{|x_2-a|}{2},\frac{|x_1-a|}{2}\right\}\right)\\
&\quad-\left(\widehat{h}_K\left(\frac{x_1+a}{2}\right)-\widehat{h}_K\left(\min\left\{\frac{|x_2-a|}{2},\frac{|x_1-a|}{2}\right\}\right)\right)\\
&\geq\overline{\mathsf{slope}}_{\widehat{h}_K}\left(\max\left\{\frac{|x_2-a|}{2},\frac{|x_1-a|}{2}\right\}\right)\left|\frac{x_2+a}{2}-\max\left\{\frac{|x_2-a|}{2},\frac{|x_1-a|}{2}\right\}\right|\\
&\quad-\underline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x_1+a}{2}\right)\left|\frac{x_1+a}{2}-\min\left\{\frac{|x_2-a|}{2},\frac{|x_1-a|}{2}\right\}\right|\\
&\overset{(i)}{\geq}0,
\end{aligned}$$

where inequality (i) holds, since

$$\begin{aligned}
&\left|\frac{x_2+a}{2}-\max\left\{\frac{|x_2-a|}{2},\frac{|x_1-a|}{2}\right\}\right|\\
=&\left|\frac{x_1+a}{2}-\frac{x_2+a}{2}\right|-\left|\max\left\{\frac{|x_2-a|}{2},\frac{|x_1-a|}{2}\right\}-\frac{x_1+a}{2}\right|\\
\geq&\left|\frac{|x_1-a|}{2}-\frac{|x_2-a|}{2}\right|-\left|\max\left\{\frac{|x_2-a|}{2},\frac{|x_1-a|}{2}\right\}-\frac{x_1+a}{2}\right|\\
=&\left|\frac{x_1+a}{2}-\min\left\{\frac{|x_2-a|}{2},\frac{|x_1-a|}{2}\right\}\right|
\end{aligned}$$

and $\overline{\mathsf{slope}}_{\widehat{h}_K}\left(\max\left\{\frac{|x_2-a|}{2},\frac{|x_1-a|}{2}\right\}\right)\geq\underline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x_1+a}{2}\right)$.

Combining the two cases above, we deduce $(A)-(B)\geq 0$, and $\widehat{\times}(x,a)$ is monotone in x for any fixed a. By symmetry, $\widehat{\times}(a,x)$ is also monotone. When $m=[m_1,\ldots,m_i^*,\ldots,m_q]^\top$, $\phi(3N(x_i-m_i^*/N))=3Nx_i-3m_i^*+2$, which is increasing in x_i. By construction of $\widehat{\xi}_m$ in (Eq. C.10) and the monotonicity of composite functions, we deduce the monotonicity of $\widehat{\xi}_m$. Similarly, when $m=[m_1,\ldots,m_i^*-1,\ldots,m_q]^\top$, we have $\phi(3N(x_i-(m_i^*-1)/N))=-3Nx_i+3m_i^*-1$ — decreasing in x_i. Therefore, $\widehat{\xi}_m$ is decreasing with respect to the i-th coordinate in x. The proof is complete. $\square$

Proof of Lemma C.3

Proof. We first analyze the Lipschitz continuity of $\widehat{\times}$. Let's fix $a \in [0,1]$ and recall $\widehat{\times}(x,a) = \widehat{h}_K\left(\frac{x+a}{2}\right) - \widehat{h}_K\left(\frac{|x-a|}{2}\right)$. We observe that $\widehat{\times}(x,a)$ is a piecewise linear function in x, due to $\widehat{h}_K$ being piecewise linear. Therefore, to characterize the Lipschitz continuity of $\widehat{\times}$, it suffices to evaluate the steepest and flattest slopes of $\widehat{\times}(x,a)$ as x varies in $[0,1]$. Specifically, we define

$$\mathsf{SteepSlope}\left(\widehat{\times}(\cdot,a)\right) = \sup_{x\in(0,1)} \operatorname*{limsup}_{\Delta\to 0} \frac{\widehat{\times}(x+\Delta,a) - \widehat{\times}(x,a)}{\Delta}, \tag{C.41}$$

$$\mathsf{FlatSlope}\left(\widehat{\times}(\cdot,a)\right) = \inf_{x\in(0,1)} \operatorname*{liminf}_{\Delta\to 0} \frac{\widehat{\times}(x+\Delta,a) - \widehat{\times}(x,a)}{\Delta}. \tag{C.42}$$

• **Steepest slope**. We consider two cases depending on the value of x, namely, $0 < x \leq a$ and $a < x < 1$.

$\star$ *(Case 1)* When $a < x < 1$, we rewrite $\widehat{\times}(x,a)$ as $\widehat{\times}(x,a) = \widehat{h}_K\left(\frac{x+a}{2}\right) - \widehat{h}_K\left(\frac{x-a}{2}\right)$. Substituting into (Eq. C.41), we obtain

$$\begin{aligned}
&\operatorname*{limsup}_{\Delta\to 0} \frac{\widehat{\times}(x+\Delta,a) - \widehat{\times}(x,a)}{\Delta} \\
&= \operatorname*{limsup}_{\Delta\to 0} \frac{\widehat{h}_K\left(\frac{x+\Delta+a}{2}\right) - \widehat{h}_K\left(\frac{x+\Delta-a}{2}\right) - \widehat{h}_K\left(\frac{x+a}{2}\right) + \widehat{h}_K\left(\frac{x-a}{2}\right)}{\Delta} \\
&= \operatorname*{limsup}_{\Delta\to 0} \frac{\widehat{h}_K\left(\frac{x+\Delta+a}{2}\right) - \widehat{h}_K\left(\frac{x+a}{2}\right) - \left[\widehat{h}_K\left(\frac{x+\Delta-a}{2}\right) - \widehat{h}_K\left(\frac{x-a}{2}\right)\right]}{\Delta}.
\end{aligned}$$

Lemma C.4 implies that $\widehat{h}_K$ is strictly monotone increasing. Hence, for any Δ, $\widehat{h}_K\left(\frac{x+\Delta+a}{2}\right) - \widehat{h}_K\left(\frac{x+a}{2}\right)$ and $\widehat{h}_K\left(\frac{x+\Delta-a}{2}\right) - \widehat{h}_K\left(\frac{x-a}{2}\right)$ are both positive or negative depending on the sign of Δ. Moreover, Lemma C.4 shows that $\overline{\mathsf{slope}}_{\widehat{h}_K}$ and $\underline{\mathsf{slope}}_{\widehat{h}_K}$ are monotone increasing. As a result, we have

$$
\begin{aligned}
&\limsup_{\Delta\to 0}\frac{\widehat{\times}(x+\Delta,a)-\widehat{\times}(x,a)}{\Delta}\\
&\le \limsup_{\Delta\to 0}\frac{\widehat{h}_K\left(\frac{x+\Delta+a}{2}\right)-\widehat{h}_K\left(\frac{x+a}{2}\right)}{\Delta}-\liminf_{\Delta\to 0}\frac{\widehat{h}_K\left(\frac{x+\Delta-a}{2}\right)-\widehat{h}_K\left(\frac{x-a}{2}\right)}{\Delta}\\
&=\frac{1}{2}\overline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x+a}{2}\right)-\frac{1}{2}\underline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x-a}{2}\right).
\end{aligned}
$$

Using Lemma C.4, we can upper bound $\overline{\mathsf{slope}}_{\widehat{h}_K}(z)$ and lower bound $\underline{\mathsf{slope}}_{\widehat{h}_K}(z)$ for any $z\in(0,1)$ as

$$
\frac{1}{2}\overline{\mathsf{slope}}_{\widehat{h}_K}(z)\le\min\left\{z+\frac{1}{2^{K+1}},1\right\}, \tag{C.43}
$$

$$
\frac{1}{2}\underline{\mathsf{slope}}_{\widehat{h}_K}(z)\ge\max\left\{z-\frac{1}{2^{K+1}},0\right\}. \tag{C.44}
$$

The upper bound (Eq. C.43) is a consequence of (Eq. C.33). Specifically, for any $a\in(0,1)$, it holds

$$
\begin{aligned}
\overline{\mathsf{slope}}_{\widehat{h}_K}(z)&=1+\sum_{k=1}^{K}\Big(2\left[\mathcal{B}_K\left(\left\lfloor z\cdot 2^K\right\rfloor\right)\right]_{K-k+1}-1\Big)2^{-k}\\
&=2^{-K}+2\sum_{k=1}^{K}\frac{\left[\mathcal{B}_K\left(\left\lfloor z\cdot 2^K\right\rfloor\right)\right]_{K-k+1}2^{K-k}}{2^K}\\
&=2^{-K}+2\frac{\left\lfloor z\cdot 2^K\right\rfloor}{2^K}\\
&\le 2z+2^{-K}.
\end{aligned}
$$

In combination with 2 being a natural upper bound of $\overline{\mathsf{slope}}_{\widehat{h}_K}$ and rescaling by $1/2$, (Eq. C.43) holds true. The lower bound (Eq. C.44) is a consequence of (Eq. C.34). We have

$$
\underline{\mathsf{slope}}_{\widehat{h}_K}(z)=1+\sum_{k=1}^{K}\Big(2\left[\mathcal{B}_K\left(\left\lceil z\cdot 2^K\right\rceil-1\right)\right]_{K-k+1}-1\Big)2^{-k}
$$

$$
\begin{aligned}
&= 2^{-K} + 2\sum_{k=1}^{K} \frac{\left[\mathcal{B}_K\left(\lceil z\cdot 2^K\rceil - 1\right)\right]_{K-k+1} 2^{K-k}}{2^K} \\
&= 2^{-K} + 2\frac{\lceil z\cdot 2^K\rceil - 1}{2^K} \\
&\geq 2z - 2^{-K}.
\end{aligned}
$$

Combining with 0 being a natural lower bound of $\underline{\mathsf{slope}}_{\widehat{h}_K}$, we establish (Eq. C.44). To this end, (Eq. C.43) and (Eq. C.44) together yield

$$
\begin{aligned}
&\sup_{a<x<1} \limsup_{\Delta\to 0} \frac{\widehat{\times}(x+\Delta,a) - \widehat{\times}(x,a)}{\Delta} \\
\leq& \sup_{a<x<1} \frac{1}{2}\overline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x+a}{2}\right) - \frac{1}{2}\underline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x-a}{2}\right) \\
\leq& \sup_{a<x<1} \min\left\{\frac{x+a}{2} + \frac{1}{2^{K+1}}, 1\right\} - \max\left\{\frac{x-a}{2} - \frac{1}{2^{K+1}}, 0\right\} \\
=& \min\left\{a + \frac{1}{2^K}, 1\right\}. \qquad \text{(C.45)}
\end{aligned}
$$

⋆ *(Case 2)* When $0 < x \leq a$, the analysis is similar. We have $\widehat{\times}(x,a) = \widehat{h}_K\left(\frac{x+a}{2}\right) - \widehat{h}_K\left(\frac{a-x}{2}\right)$, and derive

$$
\begin{aligned}
&\limsup_{\Delta\to 0} \frac{\widehat{\times}(x+\Delta,a) - \widehat{\times}(x,a)}{\Delta} \\
=& \limsup_{\Delta\to 0} \frac{\widehat{h}_K\left(\frac{x+\Delta+a}{2}\right) - \widehat{h}_K\left(\frac{a-x-\Delta}{2}\right) - \widehat{h}_K\left(\frac{x+a}{2}\right) + \widehat{h}_K\left(\frac{a-x}{2}\right)}{\Delta} \\
=& \limsup_{\Delta\to 0} \frac{\widehat{h}_K\left(\frac{x+\Delta+a}{2}\right) - \widehat{h}_K\left(\frac{x+a}{2}\right) + \left[\widehat{h}_K\left(\frac{a-x}{2}\right) - \widehat{h}_K\left(\frac{a-x-\Delta}{2}\right)\right]}{\Delta} \\
=& \limsup_{\Delta\to 0} \frac{\widehat{h}_K\left(\frac{x+\Delta+a}{2}\right) - \widehat{h}_K\left(\frac{x+a}{2}\right)}{\Delta} + \frac{\widehat{h}_K\left(\frac{a-x}{2}\right) - \widehat{h}_K\left(\frac{a-x-\Delta}{2}\right)}{\Delta}.
\end{aligned}
$$

We also notice that $\widehat{h}_K\left(\frac{x+\Delta+a}{2}\right) - \widehat{h}_K\left(\frac{x+a}{2}\right)$ and $\widehat{h}_K\left(\frac{a-x}{2}\right) - \widehat{h}_K\left(\frac{a-x-\Delta}{2}\right)$ have the same sign depending on Δ. In the case of $\Delta > 0$, we have

$$
\lim_{\Delta\to 0^+} \frac{\widehat{\times}(x+\Delta,a) - \widehat{\times}(x,a)}{\Delta} = \frac{1}{2}\overline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x+a}{2}\right) + \frac{1}{2}\underline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{a-x}{2}\right).
$$

Using (Eq. C.33) and (Eq. C.34), we derive

$$\begin{aligned}
&\overline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x+a}{2}\right)+\underline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{a-x}{2}\right)\\
&=2^{-K}+2\frac{\left\lfloor (x+a)\cdot 2^{K-1}\right\rfloor}{2^K}+2^{-K}+2\frac{\left\lceil (a-x)\cdot 2^{K-1}\right\rceil-1}{2^K}\\
&=\frac{\left\lfloor (x+a)\cdot 2^{K-1}\right\rfloor+\left\lceil (a-x)\cdot 2^{K-1}\right\rceil}{2^{K-1}}\\
&\leq\frac{(x+a)\cdot 2^{K-1}+(a-x)\cdot 2^{K-1}+1}{2^{K-1}}\\
&=2a+2^{-K+1},
\end{aligned}$$

which implies $\lim_{\Delta\to 0^+}\frac{\widehat{\times}(x+\Delta,a)-\widehat{\times}(x,a)}{\Delta}\leq a+2^{-K}$ for any $0<x\leq a$. In the case of $\Delta<0$, we have

$$\begin{aligned}
\lim_{\Delta\to 0^-}\frac{\widehat{\times}(x+\Delta,a)-\widehat{\times}(x,a)}{\Delta}&=\frac{1}{2}\underline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x+a}{2}\right)+\frac{1}{2}\overline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{a-x}{2}\right)\\
&=\frac{\left\lceil (a+x)\cdot 2^{K-1}\right\rceil+\left\lfloor (a-x)\cdot 2^{K-1}\right\rfloor}{2^{K-1}}\\
&\leq\frac{(a+x)\cdot 2^{K-1}+1+(a-x)\cdot 2^{K-1}}{2^{K-1}}\\
&=2a+2^{-K+1},
\end{aligned}$$

which implies $\lim_{\Delta\to 0^-}\frac{\widehat{\times}(x+\Delta,a)-\widehat{\times}(x,a)}{\Delta}\leq a+2^{-K}$ for any $0<x\leq a$. Combining both $\Delta>0$ and $\Delta<0$ cases, we conclude

$$\sup_{0<x\leq a}\ \operatorname*{limsup}_{\Delta\to 0}\frac{\widehat{\times}(x+\Delta,a)-\widehat{\times}(x,a)}{\Delta}\leq a+2^{-K}. \tag{C.46}$$

Putting (Eq. C.45) and (Eq. C.46) together, we deduce

$$\mathsf{SteepSlope}\left(\widehat{\times}(\cdot,a)\right)\leq a+\frac{1}{2^K}.$$

• **Flattest slope**. We also discuss two cases, i.e., $0<x\leq a$, $a<x<1$.

⋆ *(Case 1)* When $a < x < 1$, following the same computation for the steepest slope, we have

$$\begin{aligned}
&\liminf_{\Delta \to 0} \frac{\widehat{\times}(x+\Delta, a) - \widehat{\times}(x, a)}{\Delta} \\
&\geq \liminf_{\Delta \to 0} \frac{\widehat{h}_K\left(\frac{x+\Delta+a}{2}\right) - \widehat{h}_K\left(\frac{x+a}{2}\right)}{\Delta} - \limsup_{\Delta \to 0} \frac{\widehat{h}_K\left(\frac{x+\Delta-a}{2}\right) - \widehat{h}_K\left(\frac{x-a}{2}\right)}{\Delta} \\
&= \frac{1}{2}\underline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x+a}{2}\right) - \frac{1}{2}\overline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x-a}{2}\right) \\
&\overset{(i)}{\geq} \max\left\{\frac{x+a}{2} - \frac{1}{2^{K+1}}, 0\right\} - \min\left\{\frac{x-a}{2} + \frac{1}{2^{K+1}}, 1\right\} \\
&\overset{(ii)}{\geq} \max\left\{a - \frac{1}{2^K}, 0\right\},
\end{aligned} \tag{C.47}$$

where inequality (i) invokes (Eq. C.43) and (Eq. C.44), and inequality (ii) uses the natural lower bound $\underline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x+a}{2}\right) - \overline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x-a}{2}\right) \geq 0$ since $x + a > x - a$.

⋆ *(Case 2)* When $0 < x \leq a$, we derive

$$\begin{aligned}
&\liminf_{\Delta \to 0} \frac{\widehat{\times}(x+\Delta, a) - \widehat{\times}(x, a)}{\Delta} \\
&= \liminf_{\Delta \to 0} \frac{\widehat{h}_K\left(\frac{x+\Delta+a}{2}\right) - \widehat{h}_K\left(\frac{x+a}{2}\right) + \left[\widehat{h}_K\left(\frac{a-x}{2}\right) - \widehat{h}_K\left(\frac{a-x-\Delta}{2}\right)\right]}{\Delta} \\
&= \liminf_{\Delta \to 0} \frac{\widehat{h}_K\left(\frac{x+\Delta+a}{2}\right) - \widehat{h}_K\left(\frac{x+a}{2}\right)}{\Delta} + \frac{\widehat{h}_K\left(\frac{a-x}{2}\right) - \widehat{h}_K\left(\frac{a-x-\Delta}{2}\right)}{\Delta}.
\end{aligned}$$

We distinguish the limit depending on Δ being positive or negative. If $\Delta > 0$, we have

$$\begin{aligned}
&\lim_{\Delta \to 0^+} \frac{\widehat{\times}(x+\Delta, a) - \widehat{\times}(x, a)}{\Delta} \\
&= \frac{1}{2}\overline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x+a}{2}\right) + \frac{1}{2}\underline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{a-x}{2}\right) \\
&= \frac{1}{2}\left(2^{-K} + 2\frac{\lfloor (x+a) \cdot 2^{K-1} \rfloor}{2^K} + 2^{-K} + 2\frac{\lceil (a-x) \cdot 2^{K-1} \rceil - 1}{2^K}\right) \\
&= \frac{\lfloor (x+a) \cdot 2^{K-1} \rfloor + \lceil (a-x) \cdot 2^{K-1} \rceil}{2^K} \\
&\overset{(i)}{\geq} \max\left\{\frac{(a+x) \cdot 2^{K-1} - 1 + (a-x) \cdot 2^{K-1}}{2^K}, 0\right\}
\end{aligned}$$

$$= \max\left\{a - \frac{1}{2^K}, 0\right\},$$

where inequality (i) uses 0 being a natural lower bound of $\overline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x+a}{2}\right)$ and $\underline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{a-x}{2}\right)$. If $\Delta < 0$, we have

$$\begin{aligned}
&\lim_{\Delta \to 0^-} \frac{\widehat{\times}(x+\Delta, a) - \widehat{\times}(x, a)}{\Delta} \\
&= \frac{1}{2}\underline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{x+a}{2}\right) + \frac{1}{2}\overline{\mathsf{slope}}_{\widehat{h}_K}\left(\frac{a-x}{2}\right) \\
&= \frac{\left\lceil (a+x)\cdot 2^{K-1} \right\rceil + \left\lfloor (a-x)\cdot 2^{K-1} \right\rfloor}{2^K} \\
&\geq \max\left\{\frac{(a+x)\cdot 2^{K-1} + (a-x)\cdot 2^{K-1} - 1}{2^K}, 0\right\} \\
&= \max\left\{a - 2^{-K}, 0\right\}.
\end{aligned}$$

Combining both $\Delta > 0$ and $\Delta < 0$, we deduce

$$\sup_{0 < x \leq a} \operatorname*{liminf}_{\Delta \to 0} \frac{\widehat{\times}(x+\Delta, a) - \widehat{\times}(x, a)}{\Delta} \geq \max\{a - 2^{-K}, 0\}. \tag{C.48}$$

Putting (Eq. C.47) and (Eq. C.48) together, we deduce

$$\mathsf{FlatSlope}\left(\widehat{\times}(\cdot, a)\right) \geq \max\left\{a - \frac{1}{2^K}, 0\right\}.$$

To complete the proof, we observe that $\widehat{\xi}_m$ is a composition of q approximate product operations $\widehat{\times}$. For $x = [x_1, \dots, x_i, \dots, x_q]^\top$ and $x' = [x_1, \dots, x_i', \dots, x_q]^\top$ only differing in the i-th coordinate, recursively applyling (Eq. C.41) and (Eq. C.42), we derive

$$\begin{aligned}
3N \prod_{j \neq i} \max\left\{\phi(3N(x_j - m_j/N)) - \frac{1}{2^K}, 0\right\} |x_i - x_i'| &\leq \left|\widehat{\xi}_m(x) - \widehat{\xi}_m(x')\right| \\
&\leq 3N \prod_{j \neq i} \left(\phi(3N(x_j - m_j/N)) + \frac{1}{2^K}\right) |x_i - x_i'|.
\end{aligned}$$

The proof is complete. □

C.4 Proof of Lemma 5.2

Proof. The compactness of $\mathcal{Y}$ follows from $\mathcal{X}$ being compact and A being a continuous transformation. To see $\mathcal{Y}$ is also convex, we consider $y_1, y_2 \in \mathcal{Y}$ and any $\lambda \in (0,1)$. Since $\mathcal{Y} = A^\top \mathcal{X}$, we have $x_1, x_2 \in \mathcal{X}$ such that $A^\top x_1 = y_1$ and $A^\top x_2 = y_2$. Then we have

$$\lambda y_1 + (1-\lambda) y_2 = A^\top(\lambda x_1 + (1-\lambda)x_2) \in A^\top \mathcal{X} = \mathcal{Y}.$$

Therefore, $\mathcal{Y}$ is convex.

To check $A\mathcal{Y} = \mathcal{X}$, we show $A\mathcal{Y} \subset \mathcal{X}$ and $\mathcal{X} \subset A\mathcal{Y}$ hold true simultaneously. Let $y \in \mathcal{Y}$, then there exists $x \in \mathcal{X}$ such that $y = A^\top x$. Due to Assumption 5.4, we write $x \in \mathcal{X}$ as $x = Az$. As a result, we derive $Ay = AA^\top x = AA^\top Az = Az = x \in \mathcal{X}$. Thus, $A\mathcal{Y} \subset \mathcal{X}$. On the other hand, any $x \in \mathcal{X}$ can be written as $x = Az$, which implies $A^\top x = z \in \mathcal{Y}$, since A has orthonormal columns. Therefore, $\mathcal{X} \subset A\mathcal{Y}$. Combining two arguments together, we deduce $\mathcal{X} = A\mathcal{Y}$. □

APPENDIX D
OMITTED PROOFS IN CHAPTER 6

D.1 Doubly Robust Policy Learning using Neural Networks without Uniformly Bounded Weight Parameters

We prove statistical guarantees of doubly robust policy learning using neural networks without the uniform boundedness condition on weight parameters. To ease the presentation, we focus on the discrete-action setting. The network architecture can now be determined by three parameters, width, depth, and output range:

$$\widetilde{\mathcal{F}}(L, p, R) = \{f \mid f \text{ has the form of (Eq. 2.2) with } L \text{ layers and width bounded by } p, \|f\|_\infty \leq R\}.$$

We assume the expected reward and propensity score functions are $\mathcal{C}^s(\mathcal{M})$ (s-order continuously differentiable) for some positive integer s, which can be embedded in Hölder $\mathcal{H}^{s+1}(\mathcal{M})$. A formal statement is given in the following assumption.

Assumption C.4. *For a given integer $s > 0$, we assume $\mu_{A_j}(\mathbf{x}) \in \mathcal{C}^s(\mathcal{M})$ and $e_{A_j}(\mathbf{x}) \in \mathcal{C}^s(\mathcal{M})$ for $j = 1, \dots, |\mathcal{A}|$. Moreover, for a fixed $\mathcal{C}^\infty$ atlas of $\mathcal{M}$, there exists $M_2 > 0$ such that*

$$\max_j \left\|\mu_{A_j}\right\|_{\mathcal{C}^s} \leq M_2 \quad \textit{and} \quad \max_j \|\log e_{A_j}\|_{\mathcal{C}^s} \leq M_2.$$

Assumption C.4 can be viewed as a special case of Assumption 6.A.3, as $\mathcal{C}^s(\mathcal{M})$ is a subspace of $\mathcal{H}^{s_1}(\mathcal{M})$ for any real number s_1 with $s_1 > s$.

We reload the network architectures for estimating μ_{A_j}'s and e_{A_j}'s as

$$\mathcal{F}_{\text{NN}} = \widetilde{\mathcal{F}}(L_1, p_1, R_1) \quad \text{and} \quad \mathcal{G}_{\text{NN}} = \widetilde{\mathcal{F}}(L_2, p_2, R_2), \text{ respectively.} \tag{D.1}$$

Using the doubly robust method, in **Stage 1**, we can show an analogy of Lemma 6.1, when $\mathcal{F}_{\mathrm{NN}}$ and $\mathcal{G}_{\mathrm{NN}}$ are properly chosen.

Lemma D.1. *Suppose Assumption 3.1, 3.2, 6.1, 6.A.2, and C.4 hold. We choose*

$$L_1 = O(\widetilde{L}_1 \log \widetilde{L}_1), \quad p_1 = O(\widetilde{p}_1 \log \widetilde{p}_1), \quad R_1 = M, \quad \textit{with} \quad \widetilde{L}_1 \widetilde{p}_1 = O\left((\eta n_1)^{\frac{d}{2(d+2s)}}\right) \tag{D.2}$$

for $\mathcal{F}_{\mathrm{NN}}$ in (Eq. D.1) *and*

$$L_2 = O(\widetilde{L}_2 \log \widetilde{L}_2), \quad p_2 = O(|\mathcal{A}| \widetilde{p}_2 \log \widetilde{p}_2), \quad R_2 = M, \quad \textit{with} \quad \widetilde{L}_2 \widetilde{p}_2 = O\left(|\mathcal{A}|^{\frac{3s}{2s+d}} n_1^{\frac{d}{2(d+2s)}}\right) \tag{D.3}$$

for $\mathcal{G}_{\mathrm{NN}}$ in (Eq. D.1)*. Then for any $j = 1, \ldots, |\mathcal{A}|$, we have*

$$\mathbb{E}_{\mathcal{S}_1}\left[\left\|\widehat{\mu}_{A_j} - \mu_{A_j}\right\|_{L^2}^2\right] \leq C_1 (M^2 + \sigma^2)(\eta n_1)^{-\frac{2s}{2s+d}} \log^6(\eta n_1),$$
$$\mathbb{E}_{\mathcal{S}_1}\left[\left\|\widehat{e}_{A_j} - e_{A_j}\right\|_{L^2}^2\right] \leq C_2 M^2 |\mathcal{A}|^{\frac{8s+d}{2s+d}} n_1^{-\frac{2s}{2s+d}} \log^6 n_1,$$

where C_1, C_2 depend on $\log D$, B, τ and the surface area of $\mathcal{M}$.

Compared to Lemma 6.1, Lemma D.1 attains the same rate of convergence (if α is an integer in Lemma 6.1) up to some logarithmic factor dependent on sample size n_1. More interestingly, (Eq. D.2) and (Eq. D.3) suggest that Lemma D.1 holds for arbitrarily chosen width and depth for $\mathcal{F}_{\mathrm{NN}}$ and $\mathcal{G}_{\mathrm{NN}}$, as long as the product of width and depth satisfies certain requirement. In contrast, Lemma 6.1 requires a fixed ratio between network width and depth.

Lemma D.1 can be proved using the same analytical framework of Lemma 6.1, in combination of a new approximation guarantee of weight unbounded networks for approximating $\mathcal{C}^s(\mathcal{M})$ functions. Besides, we need a new analysis on the statistical complexity of weight unbounded networks.

Proof of Lemma D.1. We successively present the three steps in proving Lemma D.1.

• **Step 1: $\mathcal{C}^s(\mathcal{M})$ function approximation using $\widetilde{\mathcal{F}}$.** We begin with a universal function approximation theory of $\widetilde{\mathcal{F}}$.

Lemma D.2. *Suppose Assumption 3.1 and 3.2 hold. For any integers $\widetilde{L}, \widetilde{p} > 0$ and $f \in \mathcal{C}^s(\mathcal{M})$ with $\|f\|_{\mathcal{C}^s} \leq M$, there exists a network architecture $\widetilde{\mathcal{F}}(L, p, R)$ with*

$$L = O(\widetilde{L}\log\widetilde{L}), \quad p = O(\widetilde{p}\log\widetilde{p}), \quad \textit{and} \quad R = M,$$

giving rise to a network $\widetilde{f}$ satisfying

$$\|\widetilde{f} - f\|_\infty \leq C\left(\widetilde{L}\widetilde{p}\right)^{-\frac{2s}{d}}, \tag{D.4}$$

where C is a constant depending on s, d, M, B, τ, and the surface area of $\mathcal{M}$.

We emphasize that Lemma D.2 allows arbitrary choice of width L and depth p, and the approximation error is purely dependent on the product of width and depth.

Lemma D.2 is a generalization of Lemma 17 in [223]. To prove Lemma D.2, we repeat the argument in the proof of Lemma 17 in [223]. In particular, we only need to invoke Theorem 1.1 in [224] in replacement of Lemma 8 in the proof of Lemma 17 in [223].

• **Step 2: Statistical estimation guarantees of μ_{A_j}'s.** Given the approximation guarantee of $\widetilde{\mathcal{F}}$ for implementing $\mathcal{C}^s(\mathcal{M})$ functions, we prove statistical estimation error bound of estimating expected reward function μ_{A_j}'s.

Lemma D.3. *Suppose Assumption 3.1, 3.2, 6.1, 6.A.2, and C.4 hold. There exists a network architecture $\mathcal{F}_{\mathrm{NN}} = \widetilde{\mathcal{F}}(L, p, R)$ satisfying*

$$L = O(\widetilde{L}\log\widetilde{L}), \quad p = O(\widetilde{p}\log\widetilde{p}), \quad R = M \quad \textit{with} \quad \widetilde{L}\widetilde{p} = O\left((n_{A_j})^{\frac{d}{2(d+2s)}}\right),$$

such that for each $j = 1, \dots, |\mathcal{A}|$, the empirical risk minimizer $\widehat{\mu}_{A_j}$ in (Eq. 6.4) *satisfies*

$$\mathbb{E}\left[\left\|\widehat{\mu}_{A_j} - \mu_{A_j}\right\|_{L^2}^2\right] \leq C_1(M^2 + \sigma^2) n_1^{-\frac{2s}{2s+d}} \log^6 n_1,$$

where C_1 depends on $\log D$*, B, τ and the surface area of $\mathcal{M}$.*

Lemma D.3 is obtained by a bias-variance tradeoff. Specifically, Lemma D.2 already characterizes the bias term. The remaining task is to bound the variance term. A difficulty arises as $\widetilde{\mathcal{F}}$ can have unbounded weight parameters. However, we observe that the ReLU activation is positive homogeneous. Therefore, we can rescale layers in $\widetilde{\mathcal{F}}$ while maintaining the output unchanged. Combining with the output range bound, we can still bound the complexity of $\widetilde{\mathcal{F}}$ (A precise argument can be found in [223, Lemma 11–12]). The full proof can be found in the proof of Theorem 2 in [223], except we replace Lemma 8 in the proof of [223, Theorem 2] by Lemma D.2.

- **Step 3: Statistical estimation guarantees of e_{A_j}'s**. The estimation error of propensity scores can be obtained very similar to **Step 2**. By repeating the argument in the proof of Lemma 6.1, with a replacement of Theorem 2 in [214] by Lemma D.2. □

With estimation guarantees on μ_{A_j} and e_{A_j} for $j = 1, \dots, |\mathcal{A}|$, we proceed to **Stage 2** of the doubly robust method, where we choose proper policy network class for policy learning. Suppose we are competing with $\mathcal{C}^\ell$ oracle policies. That is, we denote π_ℓ^* as

$$\pi_\ell^* = \underset{\pi \in \Pi_{\mathcal{C}^\ell}}{\operatorname{argmax}} \, \mathbb{E}[Q(\pi(\mathbf{x}))],$$

where $\Pi_{\mathcal{C}^\ell}$ consists of policies

$$\Pi_{\mathcal{C}^\ell} = \left\{ \text{Softmax}[\nu_1(\mathbf{x}), \dots, \nu_{|\mathcal{A}|}(\mathbf{x})]^\top : \nu_j \in \mathcal{C}^\ell(\mathcal{M}) \text{ and } \|\nu_j\|_{\mathcal{C}^\ell(\mathcal{M})} \leq M \text{ for } j = 1, \dots, |\mathcal{A}| \right\}.$$

With a slight abuse of notation, we reload the policy network $\Pi_{\text{NN}}^{|\mathcal{A}|}$ by

$$\Pi_{\text{NN}}^{|\mathcal{A}|} = \left\{ \text{Softmax}(f) \text{ with } f = [f_1, \dots, f_{|\mathcal{A}|}]^\top \text{ such that } f_k : \mathcal{M} \to \mathbb{R} \in \widetilde{\mathcal{F}}(L_\Pi, p_\Pi, R_\Pi) \right\}. \tag{D.5}$$

Recall $\widehat{\pi}_{\text{DR}}$ is the optimal policy learned from $\Pi_{\text{NN}}^{|\mathcal{A}|}$ using the doubly robust method. We establish the following Theorem Theorem 6.1 bounding its regret.

Theorem D.1. *Suppose Assumption 3.1, 3.2, 6.1, 6.A.2, and C.4 hold. Under the setup in Lemma D.1, if the network parameters of* $\Pi_{\text{NN}}^{|\mathcal{A}|}$ *are chosen with*

$$L_{\widetilde{\Pi}} = O(\widetilde{L} \log \widetilde{L}), \quad p_{\widetilde{\Pi}} = O\big(|\mathcal{A}| \widetilde{p} \log \widetilde{p}\big), \quad R_\Pi = M \quad \textit{for} \quad \widetilde{L}\widetilde{p} = O\left(n^{\frac{d}{2(d+2\ell)}}\right),$$

then with probability no less than $1 - C_1 |\mathcal{A}| n^{-\frac{\ell}{2\ell+d}}$ *over the randomness of data* $\mathcal{S}_1$ *and* $\mathcal{S}_2$*, the following bound holds*

$$\begin{aligned} R(\pi_\ell^*, \widehat{\pi}_{\text{DR}}) \leq{}& C |\mathcal{A}|^2 n^{-\frac{\ell}{2\ell+d}} \log^3 n \\ & + \eta^{-1} |\mathcal{A}| \sqrt{\frac{1}{n_2} \sum_{i=n_1+1}^{n} \left(\widehat{\mu}_{A_j}(\mathbf{x}_i) - \mu_{A_j}(\mathbf{x}_i)\right)^2} \sqrt{\frac{1}{n_2} \sum_{i=n_1+1}^{n} \left(\widehat{e}_{A_j}(\mathbf{x}_i) - e_{A_j}(\mathbf{x}_i)\right)^2}, \end{aligned}$$

where $C_1 > 0$ *is an absolute constant and* C *depends on* $\log D$*,* d*,* B*,* M*,* τ*,* η*,* ζ*, and the surface area of* $\mathcal{M}$*.*

Combining Theorem D.1 and Lemma D.1, we obtain the following corollary providing a concrete convergence of $\widehat{\pi}_{\text{DR}}$ to π_ℓ^*.

Corollary D.1. *Suppose Assumption 3.1, 3.2, 6.1, 6.A.2, and C.4 hold. If the network structures are chosen as in Lemma D.1 and Theorem D.1, the following regret bound holds with probability no less than* $1 - C_1 n^{-\frac{s \wedge \ell}{2(s \wedge \ell)+d}} \log^6 n$*,*

$$R(\pi_\ell^*, \widehat{\pi}_{\text{DR}}) \leq C |\mathcal{A}|^{\frac{20s+7d}{2(2s+d)}} n^{-\frac{s \wedge \ell}{2(s \wedge \ell)+d}} \log^3 n$$

where C_1 is an absolute constant, and C depends on $\log D$*,* d*,* B*,* M*,* σ*,* τ*,* η*,* s*,* ζ*, and the surface area of* $\mathcal{M}$*.*

In comparison with Corollary 6.1, we conclude that removing the uniform boundedness condition of weight parameters in neural networks does not deteriorate the performance of doubly robust policy learning method. In particular, the learned policy attains the same rate of convergence (up to a log factor) compared to the optimal policy in a constrained oracle class ($\mathcal{C}^\ell$ policies).

We would like to point out that extensions to learning unconstrained policies as well as continuous-action settings are both plausible, using networks with unbounded weights. The analysis is almost identical, expect we invoke the approximation theory Lemma D.2 and statistical estimation theory Lemma D.3. On the other hand, we suspect the analysis can be extended to general Hölder regularity, i.e., $\mathcal{H}^\alpha(\mathcal{M})$ for any $\alpha \geq 1$. Specifically, we can show Lemma D.2 and D.3 hold for any $f \in \mathcal{H}^\alpha(\mathcal{M})$ with $\alpha \in (0, 1]$ [225, Theorem 1.1]. Then for any Hölder class $\mathcal{H}^\alpha(\mathcal{M})$, we decompose it into $\mathcal{C}^{\lfloor \alpha \rfloor}(\mathcal{M})$ and $\mathcal{H}^{\alpha - \lfloor \alpha \rfloor}(\mathcal{M})$, where we can apply Lemma D.2 and D.3 separately. Integrating together yields a result for general $\mathcal{H}^\alpha(\mathcal{M})$ regularity. However, detailed analysis can be tedious and a bit involved. We omit here for simplicity.

Proof of Theorem D.1. We rely on the analytical framework in the proof of Theorem 6.1. We first follow (Eq. D.16) to decompose the regret into (I_1) and (II_2). We then discuss necessary modifications in the proof of Theorem 6.1 in order to prove Theorem D.1.

• **Bounding** (I_1)**.** To bound (I_1), instead of using [207], we use Lemma D.2 and deduce that for any $\epsilon \in (0, 1)$ and integers $\widetilde{L}, \widetilde{p}$ satisfying $\widetilde{L}\widetilde{p} = \epsilon^{-\frac{d}{2\ell}}$, there exists a network architecture $\widetilde{\mathcal{F}}(L, p, R)$ with

$$L = O\left(\widetilde{L} \log \widetilde{L}\right), \ p = O\left(\widetilde{p} \log \widetilde{p}\right), \ R = M, \tag{D.6}$$

such that for each $\mu^*_{A_j} \in \mathcal{H}^\ell(\mathcal{M})$, there exists $\widetilde{\mu}_{A_j} \in \widetilde{\mathcal{F}}(L,p,R)$ satisfying

$$\|\widetilde{\mu}_{A_j} - \mu^*_{A_j}\|_\infty \le \epsilon. \tag{D.7}$$

We denote $\widetilde{\pi} = \mathrm{Softmax}([\widetilde{\mu}_{A_1}, \ldots, \widetilde{\mu}_{A_{|\mathcal{A}|}}]^\top) \in \mathbb{R}^{|\mathcal{A}|}$, which implies $\widetilde{\pi} \in \Pi^{|\mathcal{A}|}_{\mathrm{NN}}$ with

$$L_\Pi = L, \quad p_\Pi = |\mathcal{A}|p, \quad R_\Pi = R \tag{D.8}$$

for L, p, R defined in (Eq. D.6). We can then deduce

$$\|\widetilde{\pi} - \pi^*_\beta\|_\infty \le \epsilon,$$

where $\|\widetilde{\pi} - \pi^*_\beta\|_\infty = \sup_{\mathbf{x}\in\mathcal{M}} \max_j |[\widetilde{\pi}(\mathbf{x}) - \pi^*_\beta(\mathbf{x})]_j|$, and

$$\begin{aligned}(\mathrm{I}_1) = Q(\pi^*_\beta) - Q(\widehat{\pi}^*) \le Q(\pi^*_\beta) - Q(\widetilde{\pi}) &= \mathbb{E}\left[\left\langle [\mu_{A_1}(\mathbf{x}), \ldots, \mu_{A_{|\mathcal{A}|}}(\mathbf{x})]^\top, \pi^*(\mathbf{x}) - \widetilde{\pi}(\mathbf{x})\right\rangle\right] \\ &\le M|\mathcal{A}|\epsilon. \end{aligned} \tag{D.9}$$

• **Bounding** (II_1)**.** The bound of (II_2) can be derived by exactly following the derivation in the proof of Theorem 6.1, except $\Pi^{|\mathcal{A}|}_{\mathrm{NN}}$ is defined as in (Eq. D.5).

• **Putting** (I_1) **and** (II_1) **together.** Putting the upper bound of (I_1) in (Eq. D.9) and the upper bound of (II_1) in (Eq. D.40) together, we can get the same bound as in (Eq. D.41). We next derive an upper bound of the covering number $\mathcal{N}(\theta, \Pi^{|\mathcal{A}|}_{\mathrm{NN}}, \|\cdot\|_\Gamma)$ using the concept of uniform covering number. Define a cover with respect to samples as

Definition D.1 (Cover with respect to samples)**.** *Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}$. Given a set of samples $X = \{\mathbf{x}_k\}_{k=1}^m \subset \mathbb{R}^{d_1}$, for any $\delta > 0$, a function set $\mathcal{S}_f(X)$ is a δ-cover of F with respect to X if for any $f \in \mathcal{F}$, there exists $f^* \in \mathcal{S}_f(X)$ such that*

$$\|f(\mathbf{x}_k) - f^*(\mathbf{x}_k)\|_\infty \le \delta, \quad \forall 1 \le k \le m.$$

The uniform covering is defined as

Definition D.2 (Uniform covering number, Section 10.2 of [226])**.** *Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d$ to $\mathbb{R}$. For any set of samples $X = \{\mathbf{x}_k\}_{k=1}^m \subset \mathbb{R}^d$, denote*

$$\mathcal{F}|_X = \{(f(\mathbf{x}_1), ..., f(\mathbf{x}_m)) : f \in \mathcal{F}\}.$$

For any $\delta > 0$, the uniform covering number of $\mathcal{F}$ with m samples is defined as

$$\mathcal{N}(\delta, \mathcal{F}, m) = \max_{X\subset\mathbb{R}^d, |X|=m} \min_{\mathcal{S}_f(X)} \{|\mathcal{S}_f(X)| : \mathcal{S}_f(X) \textit{ is a } \delta\textit{-cover of } \mathcal{F} \textit{ with respect to } X\}. \tag{D.10}$$

Note that the metric $\|\cdot\|_\Gamma$ depends on the set $\{\mathbf{x}_i\}_{i=n_1+1}^n$. We can follow the proof of Lemma D.5 to show that

$$\mathcal{N}(\theta, \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, \|\cdot\|_\Gamma) \leq \mathcal{N}\left(\theta/(|\mathcal{A}|M + 2M/\eta), \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, n_2\right). \tag{D.11}$$

In the proof, we modify the construction of $\pi^{(1)}, \pi^{(2)}$ so that $|\mu_{A_j}^{(1)}(\mathbf{x}_i) - \mu_{A_j}^{(2)}(\mathbf{x}_i)| \leq \theta$ for any $j = 1, ..., |\mathcal{A}|$ and $i = n_1 + 1, ..., n$.

According to [223, Lemma 11–12], we have

$$\log\mathcal{N}(\theta, \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, \|\cdot\|_\Gamma) \leq C|\mathcal{A}|p_\Pi^2 L_\Pi^2 \log(p_\Pi^2 L_\Pi)(\log R_\Pi + \log\theta^{-1} + \log n), \tag{D.12}$$

and thus

$$\log\mathcal{N}(\theta, \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, \|\cdot\|_\Gamma) \leq C|\mathcal{A}|p_\Pi^2 L_\Pi^2 \log(p_\Pi^2 L_\Pi)(\log R_\Pi + \log\theta^{-1} + \log n). \tag{D.13}$$

Substituting the choice of the network architecture (Eq. D.8) into (Eq. D.13) gives rise to

$$\log\mathcal{N}(\theta, \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, \|\cdot\|_\Gamma) \leq C|\mathcal{A}|\epsilon^{-\frac{d}{\ell}} \log^5 \frac{1}{\epsilon}\left(\log\frac{1}{\theta} + \log n\right). \tag{D.14}$$

Following the derivation of (Eq. D.45), we can prove Theorem D.1 by substituting (Eq. D.14) and (Eq. D.43) into (Eq. D.41), and setting $\epsilon = \delta = \lambda = n_2^{-\frac{\ell}{2\ell+d}}$. □

D.2 Proof of Regret Bound

For readability, we present a proof sketch of Theorem 6.1, Theorem 6.2, Theorem 6.3 first and leave all the technical proofs to Appendix D.3.

D.2.1 Proof of Learning Hölder Policy

Proof of Theorem 6.1. We denote

$$\widehat{\pi}^* = \underset{\pi \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}}{\operatorname{argmax}} \; Q(\pi), \tag{D.15}$$

which is the optimal policy given by the neural network class $\Pi_{\mathrm{NN}}^{|\mathcal{A}|}$ defined in (Eq. 6.19). The regret can be decomposed as

$$R(\pi_\beta^*, \widehat{\pi}_{\mathrm{DR}}) = \underbrace{Q(\pi_\beta^*) - Q(\widehat{\pi}^*)}_{(\mathrm{I}_1)} + \underbrace{Q(\widehat{\pi}^*) - Q(\widehat{\pi}_{\mathrm{DR}})}_{(\mathrm{II}_1)}. \tag{D.16}$$

In (Eq. D.16), (I_1) is the approximation error (bias) of the optimal Hölder policy π_β^* by the neural network class $\Pi_{\mathrm{NN}}^{|\mathcal{A}|}$, and (II_1) represents the variance of the estimated policy in $\Pi_{\mathrm{NN}}^{|\mathcal{A}|}$. We next derive the bounds for both terms.

• **Bounding** (I_1)**.** Recall that π_β^* is the Hölder continuous optimal policy in $\Pi_{\mathcal{H}^\beta}$. By defnintion, we can write $\pi_\beta^* = \mathrm{Softmax}([\mu_{A_1}^*, \ldots, \mu_{A_{|\mathcal{A}|}}^*]^\top) \in \mathbb{R}^{|\mathcal{A}|}$ where $\mu_{A_j}^* \in \mathcal{H}^\beta(\mathcal{M})$, for $j = 1, \ldots, |\mathcal{A}|$. According to [214], Hölder functions can be uniformly approximated by a neural network class, if the network parameters are properly chosen. For any $\epsilon \in (0, 1)$

there exists a network architecture $\mathcal{F}(L, p, K, \kappa, R)$ with

$$L = O\left(\log \frac{1}{\epsilon}\right),\ p = O\left(\epsilon^{-\frac{d}{\beta}}\right),\ K = O\left(\epsilon^{-\frac{d}{\beta}} \log \frac{1}{\epsilon}\right),\ \kappa = \max\{B, M, \sqrt{d}, \tau^2\},\ R = M, \tag{D.17}$$

such that for each $\mu^*_{A_j} \in \mathcal{H}^\beta(\mathcal{M})$, there exists $\widetilde{\mu}_{A_j} \in \mathcal{F}(L, p, K, \kappa, R)$ satisfying

$$\|\widetilde{\mu}_{A_j} - \mu^*_{A_j}\|_\infty \leq \epsilon. \tag{D.18}$$

The constants hidden in $O(\cdot)$ depend on $\log D$, d, B, M, τ, β, and the surface area of $\mathcal{M}$. We denote $\widetilde{\pi} = \mathrm{Softmax}([\widetilde{\mu}_{A_1}, \dots, \widetilde{\mu}_{A_{|\mathcal{A}|}}]^\top) \in \mathbb{R}^{|\mathcal{A}|}$, which implies $\widetilde{\pi} \in \Pi^{|\mathcal{A}|}_{\mathrm{NN}}$ with

$$L_\Pi = L, \quad p_\Pi = |\mathcal{A}|p, \quad K_\Pi = |\mathcal{A}|K, \quad \kappa_\Pi = \kappa, \quad R_\Pi = R \tag{D.19}$$

for L, p, K, κ, R defined in (Eq. D.17). Based on (Eq. D.18) and the Lipschitz continuity of the Softmax function, we have

$$\|\widetilde{\pi} - \pi^*_\beta\|_\infty \leq \epsilon,$$

where $\|\widetilde{\pi} - \pi^*_\beta\|_\infty = \sup_{\mathbf{x}\in\mathcal{M}} \max_j |[\widetilde{\pi}(\mathbf{x}) - \pi^*_\beta(\mathbf{x})]_j|$ with $[\widetilde{\pi}(\mathbf{x}) - \pi^*_\beta(\mathbf{x})]_j$ denoting the j-th element of $\widetilde{\pi}(\mathbf{x}) - \pi^*_\beta(\mathbf{x})$. Therefore we bound I_1 as

$$\begin{aligned}(\mathrm{I}_1) &= Q(\pi^*_\beta) - Q(\widehat{\pi}^*) \leq Q(\pi^*_\beta) - Q(\widetilde{\pi}) \\ &= \mathbb{E}\left[\left\langle [\mu_{A_1}(\mathbf{x}), \dots, \mu_{A_{|\mathcal{A}|}}(\mathbf{x})]^\top, \pi^*(\mathbf{x}) - \widetilde{\pi}(\mathbf{x})\right\rangle\right] \leq M|\mathcal{A}|\epsilon. \end{aligned} \tag{D.20}$$

- **Bounding** (II_1)**.** We introduce an intermediate reward function $\widetilde{Q}$ to decompose the variance term (II_1). Define

$$
\begin{aligned}
&\widetilde{Q}(\pi) = \frac{1}{n_2} \sum_{i=n_1+1}^{n} \left\langle \widetilde{\Gamma}_i, \pi(\mathbf{x}_i) \right\rangle \\
&\text{with } \widetilde{\Gamma}_i = \frac{y_i - \mu_{\mathbf{a}_i}(\mathbf{x}_i)}{e_{\mathbf{a}_i}(\mathbf{x}_i)} \cdot \mathbf{a}_i + [\mu_{A_1}(\mathbf{x}_i), \dots, \mu_{A_{|\mathcal{A}|}}(\mathbf{x}_i)]^\top \in \mathbb{R}^{|\mathcal{A}|}.
\end{aligned} \tag{D.21}
$$

Note that $\widetilde{Q}$ has the same form as $\widehat{Q}$ while the estimated propensity score $\widehat{e}_{\mathbf{a}}$ and expected reward $\widehat{\mu}_{\mathbf{a}}$ are replaced by their ground truth $e_{\mathbf{a}}$ and $\mu_{\mathbf{a}}$, respectively.

We decompose (II_1) as

$$
\begin{aligned}
(\mathrm{II}_1) &= Q(\widehat{\pi}^*) - Q(\widehat{\pi}_{\mathrm{DR}}) \\
&= \widehat{Q}(\widehat{\pi}^*) - \widehat{Q}(\widehat{\pi}_{\mathrm{DR}}) + Q(\widehat{\pi}^*) - \widehat{Q}(\widehat{\pi}^*) + \widehat{Q}(\widehat{\pi}_{\mathrm{DR}}) - Q(\widehat{\pi}_{\mathrm{DR}}) \\
&\le Q(\widehat{\pi}^*) - Q(\widehat{\pi}_{\mathrm{DR}}) + \widehat{Q}(\widehat{\pi}_{\mathrm{DR}}) - \widehat{Q}(\widehat{\pi}^*) \\
&\le \sup_{\pi_1, \pi_2 \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}} Q(\pi_1) - Q(\pi_2) - \left(\widetilde{Q}(\pi_1) - \widetilde{Q}(\pi_2)\right) \\
&\quad + \sup_{\pi_1, \pi_2 \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}} \widetilde{Q}(\pi_1) - \widetilde{Q}(\pi_2) - \left(\widehat{Q}(\pi_1) - \widehat{Q}(\pi_2)\right) \\
&\le \underbrace{\sup_{\pi_1, \pi_2 \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}} \Delta(\pi_1, \pi_2) - \widetilde{\Delta}(\pi_1, \pi_2)}_{\mathcal{E}_1} + \underbrace{\sup_{\pi_1, \pi_2 \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}} \widetilde{\Delta}(\pi_1, \pi_2) - \widehat{\Delta}(\pi_1, \pi_2)}_{\mathcal{E}_2},
\end{aligned} \tag{D.22}
$$

where $\Delta(\pi_1, \pi_2) = Q(\pi_1) - Q(\pi_2)$, $\widetilde{\Delta}(\pi_1, \pi_2) = \widetilde{Q}(\pi_1) - \widetilde{Q}(\pi_2)$ and $\widehat{\Delta}(\pi_1, \pi_2) = \widehat{Q}(\pi_1) - \widehat{Q}(\pi_2)$. The first inequality in (Eq. D.22) come from (Eq. D.15) which implies $\widehat{Q}(\widehat{\pi}_{\mathrm{DR}}) \le \widehat{Q}(\widehat{\pi}^*)$. In this decomposition, $\mathcal{E}_1$ corresponds to the difference between Q and $\widetilde{Q}$ which can be bounded using the metric entropy argument, since $\widetilde{Q}$ is unbiased, i.e. $\mathbb{E}[\widetilde{Q}(\pi)] = Q(\pi)$. The second term $\mathcal{E}_2$ corresponds to the error between $\widetilde{Q}$ and $\widehat{Q}$, which can be bounded in terms of the estimation errors of the e_{A_j}'s and the μ_{A_j}'s.

Bounding $\mathcal{E}_1$. We first show that $\mathbb{E}\left[\widetilde{Q}(\pi)\right] = Q(\pi)$:

$$
\mathbb{E}\left[\widetilde{Q}(\pi)\right] = \mathbb{E}\left[\left\langle \mathbb{E}\left[\frac{y - \mu_{\mathbf{a}}(\mathbf{x})}{e_{\mathbf{a}}(\mathbf{x})}\mathbf{a}\middle|\mathbf{x}\right], \pi(\mathbf{x})\right\rangle + \left\langle [\mu_{A_1}(\mathbf{x}), \dots, \mu_{A_{|\mathcal{A}|}}(\mathbf{x})]^\top, \pi(\mathbf{x})\right\rangle\right]
$$

$$= \mathbb{E}\left[\left\langle [\mu_{A_1}(\mathbf{x}), \ldots, \mu_{A_{|\mathcal{A}|}}(\mathbf{x})]^\top, \pi(\mathbf{x})\right\rangle\right] = \mathbb{E}[Y(\pi(\mathbf{x}))] = Q(\pi), \tag{D.23}$$

which further implies $\mathbb{E}\left[\widetilde{\Delta}(\pi_1, \pi_2)\right] = \Delta(\pi_1, \pi_2)$. In (Eq. D.23), the second equality holds since

$$\mathbb{E}\left[\frac{y - \mu_{\mathbf{a}}(\mathbf{x})}{e_{\mathbf{a}}(\mathbf{x})}\mathbf{a}\middle|\mathbf{x}\right] = \frac{\mathbb{E}[y|\mathbf{x}] - \mu_{\mathbf{a}}(\mathbf{x})}{e_{\mathbf{a}}(\mathbf{x})}\mathbb{E}[\mathbf{a}|\mathbf{x}] = 0$$

by Assumption 6.1. Therefore we can write

$$\mathcal{E}_1 = \sup_{\pi_1, \pi_2 \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}} \widetilde{\Delta}(\pi_1, \pi_2) - \mathbb{E}[\widetilde{\Delta}(\pi_1, \pi_2)] \tag{D.24}$$

with

$$\widetilde{\Delta} = \frac{1}{n_2}\sum_{i=n_1+1}^{n}\left\langle \widetilde{\Gamma}_i, \pi_1(\mathbf{x}_i)\right\rangle - \frac{1}{n_2}\sum_{i=n_1+1}^{n}\left\langle \widetilde{\Gamma}_i, \pi_2(\mathbf{x}_i)\right\rangle.$$

We derive a bound of $\mathcal{E}_1$ using the following lemma which is be proved by symmetrization and Dudley's entropy integral [227, 221] in Appendix D.3.3:

Lemma D.4. *Let $\Pi : \mathcal{M} \to \mathbb{R}^{|\mathcal{A}|}$ be a policy space on $|\mathcal{A}|$ actions such that any $\pi \in \Pi$ maps a covariate $\mathbf{x} \in \mathcal{M}$ to $\pi(\mathbf{x})$ in the simplex of $\mathbb{R}^{|\mathcal{A}|}$, and $\mathcal{S}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a set of i.i.d. samples, where $\mathbf{x}_i$ is sampled from a probability distribution $\mathbb{P}$ supported on $\mathcal{M}$ and $y_i \in \mathbb{R}$. For any $(\mathbf{x}, y)$, we define $\mathring{\Gamma}(\mathbf{x}, y) \in \mathbb{R}^{|\mathcal{A}|}$ as a function of the sample $(\mathbf{x}, y)$. Assume that there exists a constant $J \geq 0$, such that*

$$\sup_{(\mathbf{x}, y) \in \mathcal{M} \times \mathbb{R}} |\mathring{\Gamma}(\mathbf{x}, y)| \leq J. \tag{D.25}$$

For any policies $\pi_1, \pi_2 \in \Pi$, define

$$\mathring{\Delta}(\pi_1, \pi_2) = \frac{1}{n}\sum_{i=1}^{n}\left\langle \mathring{\Gamma}_i, \pi_1(\mathbf{x}_i)\right\rangle - \frac{1}{n}\sum_{i=1}^{n}\left\langle \mathring{\Gamma}_i, \pi_2(\mathbf{x}_i)\right\rangle \quad \textit{and} \tag{D.26}$$

$$\mathcal{D}(\Pi) = \sup_{\pi_1,\pi_2\in\Pi} \mathring{\Delta}(\pi_1,\pi_2) - \mathbb{E}[\mathring{\Delta}(\pi_1,\pi_2)] \tag{D.27}$$

with the shorthand $\mathring{\Gamma}_i = \mathring{\Gamma}(\mathbf{x}_i, y_i)$. *Then the following bound holds*

$$\mathcal{D}(\Pi) \leq \inf_{\lambda}\ 4\lambda + \frac{96}{\sqrt{n}} \int_{\lambda}^{\max_{\pi\in\Pi}\|\pi\|_\Gamma} \sqrt{\log \mathcal{N}(\theta, \Pi, \|\cdot\|_\Gamma)}d\theta + 12J\sqrt{\frac{\log 1/\delta}{2n}} \tag{D.28}$$

with probability no less than $1-2\delta$ *over* $\mathcal{S}_n$, *where* $\|\pi\|_\Gamma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\langle\mathring{\Gamma}_i, \pi(\mathbf{x}_i)\rangle^2}$.

A key observation is that when taking $\mathring{\Gamma} = \widetilde{\Gamma}$ defined in (Eq. D.21) and $\Pi = \Pi_{\mathrm{NN}}^{|\mathcal{A}|}$, we have $\mathcal{E}_1 = \mathcal{D}(\Pi)$ in (Eq. D.24). To apply Lemma D.4 for bounding $\mathcal{E}_1$, we only need to verify the assertion (Eq. D.25). In fact, due to Assumption 6.A.2, we see that y, $\mu_{A_j}(\mathbf{x})$, and $e_{A_j}(\mathbf{x})$ are all bounded. A simple calculation yields $\sup_{(\mathbf{x},y)\in\mathcal{M}\times\mathbb{R}} |\mathring{\Gamma}(\mathbf{x},y)| \leq J = 2M/\eta + M$. Therefore, we bound $\mathcal{E}_1$ as

$$\mathcal{E}_1 \leq \inf_{\lambda}\ 4\lambda + \frac{96}{\sqrt{n_2}} \int_{\lambda}^{\max_{\pi\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}}\|\pi\|_\Gamma} \sqrt{\log \mathcal{N}(\theta, \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, \|\cdot\|_\Gamma)}d\theta + (24M/\eta + 12M)\sqrt{\frac{\log 1/\delta}{2n_2}} \tag{D.29}$$

with probability no less than $1-2\delta$.

Bounding $\mathcal{E}_2$. The $\mathcal{E}_2$ term depends on the difference between $\widetilde{\Gamma}_i$ and $\widehat{\Gamma}_i$, where $\widetilde{\Gamma}_i$ and $\widehat{\Gamma}_i$ are defined in (Eq. D.21) and (Eq. 6.7), respectively. In $\mathcal{E}_2$, we have

$$\begin{aligned}
\widetilde{\Delta}(\pi_1,\pi_2) - \widehat{\Delta}(\pi_1,\pi_2) &= \frac{1}{n_2}\sum_{i=n_1+1}^{n} \left\langle \pi_1(\mathbf{x}_i) - \pi_2(\mathbf{x}_i), \widetilde{\Gamma}_i - \widehat{\Gamma}_i \right\rangle \\
&= \sum_{j=1}^{|\mathcal{A}|} \frac{1}{n_2} \sum_{i=n_1+1}^{n} (\pi_{1,j}(\mathbf{x}_i) - \pi_{2,j}(\mathbf{x}_i)) \left(\widetilde{\Gamma}_{i,j} - \widehat{\Gamma}_{i,j}\right),
\end{aligned}$$

where $\pi_{k,j}$ and $\widetilde{\Gamma}_{i,j}$ denote the j-th element of π_k and $\widetilde{\Gamma}_i$, respectively. Define

$$\Lambda_j(\pi_1,\pi_2) = \frac{1}{n_2}\sum_{i=n_1+1}^{n} (\pi_{1,j}(\mathbf{x}_i) - \pi_{2,j}(\mathbf{x}_i)) \left(\widetilde{\Gamma}_{i,j} - \widehat{\Gamma}_{i,j}\right) \in \mathbb{R}, \quad \text{for } j = 1,\ldots,|\mathcal{A}|.$$

Then we can write $\widetilde{\Delta}(\pi_1,\pi_2)-\widehat{\Delta}(\pi_1,\pi_2)$ as

$$\widetilde{\Delta}(\pi_1,\pi_2)-\widehat{\Delta}(\pi_1,\pi_2)=\sum_{j=1}^{|\mathcal{A}|}\Lambda_j(\pi_1,\pi_2). \tag{D.30}$$

The error term $\widetilde{\Gamma}_{i,j}-\widehat{\Gamma}_{i,j}$ in $\Lambda_j(\pi_1,\pi_2)$ depends on the estimation error of $\widehat{\mu}_{A_j}$ and $\widehat{e}_{A_j}$. Based on the source of the error, we decompose each $\Lambda_j(\pi_1,\pi_2)$ into three terms:

$$\begin{aligned}
\Lambda_j(\pi_1,\pi_2) &= \frac{1}{n_2}\sum_{i=n_1+1}^{n}(\pi_{1,j}(\mathbf{x}_i)-\pi_{2,j}(\mathbf{x}_i))\Bigg[\left(\frac{y_i-\mu_{A_j}(\mathbf{x}_i)}{e_{A_j}(\mathbf{x}_i)}\mathbb{1}_{\{\mathbf{a}_i=A_j\}}+\mu_{A_j}(\mathbf{x}_i)\right)\\
&\quad-\left(\frac{y_i-\widehat{\mu}_{A_j}(\mathbf{x}_i)}{\widehat{e}_{A_j}(\mathbf{x}_i)}\mathbb{1}_{\{\mathbf{a}_i=A_j\}}+\widehat{\mu}_{A_j}(\mathbf{x}_i)\right)\Bigg]\\
&=\frac{1}{n_2}\sum_{i=n_1+1}^{n}(\pi_{1,j}(\mathbf{x}_i)-\pi_{2,j}(\mathbf{x}_i))\left(\mu_{A_j}(\mathbf{x}_i)-\widehat{\mu}_{A_j}(\mathbf{x}_i)\right)\\
&\quad+\frac{1}{n_2}\sum_{i=n_1+1}^{n}(\pi_{1,j}(\mathbf{x}_i)-\pi_{2,j}(\mathbf{x}_i))\mathbb{1}_{\{\mathbf{a}_i=A_j\}}\left(\frac{y_i-\mu_{A_j}(\mathbf{x}_i)}{e_{A_j}(\mathbf{x}_i)}-\frac{y_i-\widehat{\mu}_{A_j}(\mathbf{x}_i)}{\widehat{e}_{A_j}(\mathbf{x}_i)}\right)\\
&=S_j^{(1)}(\pi_{1,j},\pi_{2,j})+S_j^{(2)}(\pi_{1,j},\pi_{2,j})+S_j^{(3)}(\pi_{1,j},\pi_{2,j}),
\end{aligned}\tag{D.31}$$

where

$$\begin{aligned}
S_j^{(1)}(\pi_{1,j},\pi_{2,j})&=\frac{1}{n_2}\sum_{i=n_1+1}^{n}(\pi_{1,j}(\mathbf{x}_i)-\pi_{2,j}(\mathbf{x}_i))\left(\mu_{A_j}(\mathbf{x}_i)-\widehat{\mu}_{A_j}(\mathbf{x}_i)\right)\left(1-\frac{\mathbb{1}_{\{\mathbf{a}_i=A_j\}}}{e_{A_j}(\mathbf{x}_i)}\right),\\
S_j^{(2)}(\pi_{1,j},\pi_{2,j})&=\frac{1}{n_2}\sum_{\{n_1+1\le i\le n|\mathbf{a}_i=A_j\}}(\pi_{1,j}(\mathbf{x}_i)-\pi_{2,j}(\mathbf{x}_i))\left(y_i-\mu_{A_j}(\mathbf{x}_i)\right)\left(\frac{1}{e_{A_j}(\mathbf{x}_i)}-\frac{1}{\widehat{e}_{A_j}(\mathbf{x}_i)}\right),\\
S_j^{(3)}(\pi_{1,j},\pi_{2,j})&=\frac{1}{n_2}\sum_{\{n_1+1\le i\le n|\mathbf{a}_i=A_j\}}(\pi_{1,j}(\mathbf{x}_i)-\pi_{2,j}(\mathbf{x}_i))\left(\widehat{\mu}_{A_j}(\mathbf{x}_i)-\mu_{A_j}(\mathbf{x}_i)\right)\left(\frac{1}{\widehat{e}_{A_j}(\mathbf{x}_i)}-\frac{1}{e_{A_j}(\mathbf{x}_i)}\right).
\end{aligned}$$

Here $S_j^{(1)}(\pi_{1,j},\pi_{2,j})$ and $S_j^{(2)}(\pi_{1,j},\pi_{2,j})$ can be bounded using Lemma D.4. $S_j^{(3)}(\pi_{1,j},\pi_{2,j})$ contains the product of the estimation error of $\widehat{\mu}_{A_j}$ and $\widehat{e}_{A_j}$, which gives the doubly robust property. According to (Eq. D.30) and (Eq. D.31),

$$\begin{aligned}\mathcal{E}_2 &= \sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} \left[\sum_{j=1}^{|\mathcal{A}|} S_j^{(1)}(\pi_{1,j},\pi_{2,j}) + S_j^{(2)}(\pi_{1,j},\pi_{2,j}) + S_j^{(3)}(\pi_{1,j},\pi_{2,j})\right] \\ &\leq \sum_{j=1}^{|\mathcal{A}|} \sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(1)}(\pi_{1,j},\pi_{2,j}) + \sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(2)}(\pi_{1,j},\pi_{2,j}) + \sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(3)}(\pi_{1,j},\pi_{2,j}).\end{aligned} \tag{D.32}$$

In the rest of the proof, when there is no ambiguity, we omit the dependency on $(\pi_{1,j},\pi_{2,j})$ and use the notations $S_j^{(1)}$, $S_j^{(2)}$ and $S_j^{(3)}$. We next derive the bounds for the $S_j^{(1)}$, $S_j^{(2)}$ and $S_j^{(3)}$ terms in the right hand side of (Eq. D.32) respectively.

Bounding $\sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(1)}$: For $S_j^{(1)}$, one can show that $\mathbb{E}[S_j^{(1)}]=0$:

$$\begin{aligned}&\mathbb{E}[S_j^{(1)}] \\ &= \mathbb{E}\left[\frac{1}{n_2}\sum_{i=n_1+1}^{n} \mathbb{E}\left[(\pi_{1,j}(\mathbf{x}_i)-\pi_{2,j}(\mathbf{x}_i))\left(\mu_{A_j}(\mathbf{x}_i)-\widehat{\mu}_{A_j}(\mathbf{x}_i)\right)\left(1-\frac{\mathbb{1}_{\{\mathbf{a}_i=A_j\}}}{e_{A_j}(\mathbf{x}_i)}\right)\middle|\mathbf{x}_i\right]\right] \\ &= \mathbb{E}\left[\frac{1}{n_2}\sum_{i=n_1+1}^{n} (\pi_{1,j}(\mathbf{x}_i)-\pi_{2,j}(\mathbf{x}_i))\left(\mu_{A_j}(\mathbf{x}_i)-\widehat{\mu}_{A_j}(\mathbf{x}_i)\right)\mathbb{E}\left[1-\frac{\mathbb{1}_{\{\mathbf{a}_i=A_j\}}}{e_{A_j}(\mathbf{x}_i)}\middle|\mathbf{x}_i\right]\right]=0.\end{aligned}$$

Denote

$$\bar{\Gamma}^{(1,j)}(\mathbf{x}_i) = \left(\mu_{A_j}(\mathbf{x}_i)-\widehat{\mu}_{A_j}(\mathbf{x}_i)\right)\left(1-\frac{\mathbb{1}_{\{\mathbf{a}_i=A_j\}}}{e_{A_j}(\mathbf{x}_i)}\right)\in\mathbb{R},$$

then we have

$$\begin{aligned}\sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(1)} &= \sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(1)} - \mathbb{E}\left[S_j^{(1)}\right] \\ &= \sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} \frac{1}{n_2}\sum_{i=n_1+1}^{n}(\pi_{1,j}(\mathbf{x}_i)-\pi_{2,j}(\mathbf{x}_i))\,\bar{\Gamma}^{(1,j)}(\mathbf{x}_i) \\ &\quad - \mathbb{E}\left[\frac{1}{n_2}\sum_{i=n_1+1}^{n}(\pi_{1,j}(\mathbf{x}_i)-\pi_{2,j}(\mathbf{x}_i))\,\bar{\Gamma}^{(1,j)}(\mathbf{x}_i)\right].\end{aligned} \tag{D.33}$$

The expression in (Eq. D.33) resembles the same form as $\mathcal{D}$ in (Eq. D.27) with $\mathring{\Gamma} = \bar{\Gamma}^{(1,j)}(\mathbf{x})$ and $\Pi = \Pi_{\mathrm{NN}}^{|\mathcal{A}|}$. Therefore, we can estimate $\sup_{\pi_1,\pi_2 \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(1)}$ using Lemma D.4. Due to Assumption 6.A.2, for any $\mathbf{x} \in \mathcal{M}$, we have $|\bar{\Gamma}^{(1,j)}(\mathbf{x})| \leq 2M/\eta$. After substituting $J = 2M/\eta$ in Lemma D.4, we have

$$\sup_{\pi_1,\pi_2 \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(1)} \leq \inf_{\lambda} \; 4\lambda + \frac{96}{\sqrt{n_2}} \int_{\lambda}^{\max_{\pi \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}} \|\pi\|_{\Gamma}} \sqrt{\log \mathcal{N}(\theta, \Pi, \|\cdot\|_{\Gamma})} d\theta \\ + (24M/\eta) \sqrt{\frac{\log 1/\delta}{2n_2}} \tag{D.34}$$

with probability no less than $1 - 2\delta$.

Bounding $\sup_{\pi_1,\pi_2 \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(2)}$**:** Similarly, one can show $\mathbb{E}\left[S_j^{(2)}\right] = 0$. Denote

$$\bar{\Gamma}^{(2,j)}(\mathbf{x}_i, y_i) = \big(y_i - \mu_{A_j}(\mathbf{x}_i)\big) \left(\frac{1}{e_{A_j}(\mathbf{x}_i)} - \frac{1}{\widehat{e}_{A_j}(\mathbf{x}_i)} \right) \mathbb{1}_{\{\mathbf{a}_i = A_j\}} \in \mathbb{R}.$$

We follow the same calculation in (Eq. D.33) to express $\sup_{\pi_1,\pi_2 \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(2)}$ in the same form as $\mathcal{D}$ in (Eq. D.27) with $\mathring{\Gamma} = \bar{\Gamma}^{(2,j)}$ and $\Pi = \Pi_{\mathrm{NN}}^{|\mathcal{A}|}$.

An upper bound of $\sup_{(\mathbf{x},y)} |\bar{\Gamma}^{(2,j)}(\mathbf{x})|$ can be derived as follows. With $\mathcal{G}_{\mathrm{NN}}$ chosen in (Eq. 6.15), its output is bounded by M, which implies $\widehat{e}_{A_j} \geq (|\mathcal{A}| e^{2M})^{-1}$. Thus

$$\left| \frac{1}{e_{A_j}(\mathbf{x}_i)} - \frac{1}{\widehat{e}_{A_j}(\mathbf{x}_i)} \right| \leq |\mathcal{A}| e^{2M},$$

since $M \geq -\log \eta$ by (Eq. 6.12). By Assumption 6.A.2 and (Eq. 6.12), we have $\sup_{\mathbf{x},y} |y - \mu_{A_j}(\mathbf{x})| \leq 2M$ hold for any $j = 1, \ldots, |\mathcal{A}|$. Therefore, we have

$$\sup_{(\mathbf{x},y) \in \mathcal{M} \times \mathbb{R}} |\Gamma^{(2,j)}(\mathbf{x})| \leq 2|\mathcal{A}| e^{2M} M. \tag{D.35}$$

Using Lemma D.4 and substituting $J = 2|\mathcal{A}|e^{2M}M$ give rise to

$$\begin{aligned}\sup_{\pi_1,\pi_2\in\Pi_{\text{NN}}^{|\mathcal{A}|}} S_j^{(2)} \leq \inf_{\lambda}\ 4\lambda + \frac{96}{\sqrt{n_2}}\int_{\lambda}^{\max_{\pi\in\Pi_{\text{NN}}^{|\mathcal{A}|}}\|\pi\|_\Gamma}\sqrt{\log\mathcal{N}(\theta,\Pi,\|\cdot\|_\Gamma)}d\theta \\ + (24|\mathcal{A}|e^{2M}M)\sqrt{\frac{\log 1/\delta}{2n_2}} \qquad \text{(D.36)}\end{aligned}$$

with probability no less than $1-2\delta$.

Bounding $\sup_{\pi_1,\pi_2\in\Pi_{\text{NN}}^{|\mathcal{A}|}} S_j^{(3)}$**:** We next derive an upper bound of $\sup_{\pi_1,\pi_2\in\Pi_{\text{NN}}^{|\mathcal{A}|}} S_j^{(3)}$ as the product of the estimation errors of the $\widehat{\mu}_{A_j}$'s and the $\widehat{e}_{A_j}$'s:

$$\begin{aligned}&\sup_{\pi_1,\pi_2\in\Pi_{\text{NN}}^{|\mathcal{A}|}} S_j^{(3)} \\ &= \frac{1}{n_2}\sup_{\pi_1,\pi_2\in\Pi_{\text{NN}}^{|\mathcal{A}|}}\sum_{\{n_1+1\leq i\leq n|\mathbf{a}_i=A_j\}}(\pi_{1,j}(\mathbf{x}_i)-\pi_{2,j}(\mathbf{x}_i))\left(\widehat{\mu}_{A_j}(\mathbf{x}_i)-\mu_{A_j}(\mathbf{x}_i)\right)\left(\frac{1}{\widehat{e}_{A_j}(\mathbf{x}_i)}-\frac{1}{e_{A_j}(\mathbf{x}_i)}\right) \\ &\leq \frac{1}{n_2}\sum_{i=n_1+1}^{n}\left|\widehat{\mu}_{A_j}(\mathbf{x}_i)-\mu_{A_j}(\mathbf{x}_i)\right|\left|\frac{1}{\widehat{e}_{A_j}(\mathbf{x}_i)}-\frac{1}{e_{A_j}(\mathbf{x}_i)}\right| \\ &\qquad\qquad (\text{since } |\pi_{1,j}(\mathbf{x}_i)-\pi_{2,j}(\mathbf{x}_i)|\leq 1) \\ &\leq \sqrt{\frac{1}{n_2}\sum_{i=n_1+1}^{n}\left(\widehat{\mu}_{A_j}(\mathbf{x}_i)-\mu_{A_j}(\mathbf{x}_i)\right)^2}\sqrt{\frac{1}{n_2}\sum_{i=n_1+1}^{n}\left(\frac{1}{\widehat{e}_{A_j}(\mathbf{x}_i)}-\frac{1}{e_{A_j}(\mathbf{x}_i)}\right)^2} \\ &\qquad\qquad (\text{by Cauchy-Schwarz}) \\ &\leq \eta^{-1}|\mathcal{A}|e^{2M}\sqrt{\frac{1}{n_2}\sum_{i=n_1+1}^{n}\left(\widehat{\mu}_{A_j}(\mathbf{x}_i)-\mu_{A_j}(\mathbf{x}_i)\right)^2}\sqrt{\frac{1}{n_2}\sum_{i=n_1+1}^{n}\left(\widehat{e}_{A_j}(\mathbf{x}_i)-e_{A_j}(\mathbf{x}_i)\right)^2}, \\ &\qquad\qquad \text{(D.37)}\end{aligned}$$

where the last inequality holds since $e_{A_j} \geq \eta$ by Assumption 6.A.2 and $\widehat{e}_{A_j} \geq$

$(|\mathcal{A}|e^{2M})^{-1}$. We denote

$$\omega_j = \eta^{-1}|\mathcal{A}|e^{2M}\sqrt{\frac{1}{n_2}\sum_{i=n_1+1}^{n}\big(\widehat{\mu}_{A_j}(\mathbf{x}_i) - \mu_{A_j}(\mathbf{x}_i)\big)^2}\sqrt{\frac{1}{n_2}\sum_{i=n_1+1}^{n}\big(\widehat{e}_{A_j}(\mathbf{x}_i) - e_{A_j}(\mathbf{x}_i)\big)^2}, \tag{D.38}$$

and write $\sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(3)} \le \omega_j$.

Putting the $S_j^{(1)}$, $S_j^{(2)}$ and $S_j^{(3)}$ terms together: Combining (Eq. D.34), (Eq. D.36), (Eq. D.37) gives rise to

$$\begin{aligned}\sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} \Lambda_j(\pi_1,\pi_2) &\le \sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(1)} + \sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(2)} + \sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} S_j^{(3)} \\ &\le \inf_{\lambda}\ 8\lambda + \frac{192}{\sqrt{n_2}}\int_{\lambda}^{\max_{\pi\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}}\|\pi\|_\Gamma}\sqrt{\log\mathcal{N}(\theta,\Pi_{\mathrm{NN}}^{|\mathcal{A}|},\|\cdot\|_\Gamma)d\theta} + 48|\mathcal{A}|e^{2M}M\sqrt{\frac{\log 1/\delta}{2n_2}} + \omega_j\end{aligned}$$

with probability no less than $1-4\delta$ where we used $e^{2M} \ge \eta^{-1}$ according to (Eq. 6.12).

According to (Eq. D.32), we can apply the union probability bound for $j = 1,\dots,|\mathcal{A}|$ and obtain

$$\begin{aligned}\mathcal{E}_2 &= \sup_{\pi_1,\pi_2\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}} \widetilde{\Delta}(\pi_1,\pi_2) - \widehat{\Delta}(\pi_1,\pi_2) \\ &\le \inf_{\lambda}\ 8|\mathcal{A}|\lambda + \frac{192|\mathcal{A}|}{\sqrt{n_2}}\int_{\lambda}^{\max_{\pi\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}}\|\pi\|_\Gamma}\sqrt{\log\mathcal{N}(\theta,\Pi_{\mathrm{NN}}^{|\mathcal{A}|},\|\cdot\|_\Gamma)d\theta} \\ &\quad + 48|\mathcal{A}|^2e^{2M}M\sqrt{\frac{\log 1/\delta}{2n_2}} + \sum_{j=1}^{|\mathcal{A}|}\omega_j\end{aligned} \tag{D.39}$$

with probability no less than $1-4|\mathcal{A}|\delta$.

Combining (Eq. D.29) and (Eq. D.39), we have

$$(\mathrm{II}_1) \le \inf_{\lambda}\ (8|\mathcal{A}|+4)\lambda + \frac{192|\mathcal{A}|+96}{\sqrt{n_2}}\int_{\lambda}^{\max_{\pi\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}}\|\pi\|_\Gamma}\sqrt{\log\mathcal{N}(\theta,\Pi_{\mathrm{NN}}^{|\mathcal{A}|},\|\cdot\|_\Gamma)d\theta}$$

$$+\sum_{j=1}^{|\mathcal{A}|}\omega_j + \left(72|\mathcal{A}|^2 e^{2M} M + 12M\right)\sqrt{\frac{\log 1/\delta}{2n_2}} \tag{D.40}$$

with probability at least $1-6|\mathcal{A}|\delta$.

- **Putting** (I_1), (II_1) **together.** Putting our estimates of (I_1) in (Eq. D.20) and (II_1) in (Eq. D.40) together, we get

$$R(\pi^*_\beta, \widehat{\pi}_{\mathrm{DR}}) \leq |\mathcal{A}|M\epsilon + \inf_\lambda \ (8|\mathcal{A}|+4)\lambda + \frac{192|\mathcal{A}|+96}{\sqrt{n_2}}\int_\lambda^{\max_{\pi\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}}\|\pi\|_\Gamma} \sqrt{\log\mathcal{N}(\theta, \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, \|\cdot\|_\Gamma)d\theta}$$
$$+\sum_{j=1}^{|\mathcal{A}|}\omega_j + 84|\mathcal{A}|^2 e^{2M} M\sqrt{\frac{\log 1/\delta}{2n_2}} \tag{D.41}$$

with probability at least $1-6|\mathcal{A}|\delta$. The upper bound in (Eq. D.41) depends on the covering number $\mathcal{N}(\theta, \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, \|\cdot\|_\Gamma)$ and the integral upper limit $\max_{\pi\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}}\|\pi\|_\Gamma$ which can be estimated by the following lemmas (see the proofs in Appendix D.3.4 and Appendix D.3.5 respectively):

Lemma D.5. *Suppose Assumption 6.A.2 and 6.A.3 hold and define* $\Pi_{\mathrm{NN}}^{|\mathcal{A}|}$ *according to (Eq. 6.19). Then*

$$\mathcal{N}(\theta, \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, \|\cdot\|_\Gamma) \leq \left(\frac{2(|\mathcal{A}|M + 2M/\eta)L_\Pi^2(p_\Pi R/|\mathcal{A}|+2)\kappa_\Pi^L(p_\Pi/|\mathcal{A}|)^{L_\Pi+1}}{\theta}\right)^{K_\Pi}. \tag{D.42}$$

Lemma D.6. *Suppose Assumptions 6.A.2 and 6.A.3 hold. For any* $\pi\in\Pi_{\mathrm{NN}}^{|\mathcal{A}|}$*, the following holds*

$$\|\pi\|_\Gamma^2 \leq (2M/\eta + |\mathcal{A}|M)^2. \tag{D.43}$$

Setting the network parameter as in (Eq. D.19) and using (Eq. D.42), we have

$$\log\mathcal{N}(\theta, \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, \|\cdot\|_\Gamma) \leq C_1|\mathcal{A}|\epsilon^{-\frac{d}{\beta}}\log\frac{1}{\epsilon}\left(\log^2\frac{1}{\epsilon} + \log\frac{1}{\theta}\right) \tag{D.44}$$

with C_1 depending on $\log D$, d, B, τ, η, β, and the surface area of $\mathcal{M}$.

Substituting (Eq. D.44) and (Eq. D.43) into (Eq. D.41) gives

$$\begin{aligned}
R(\pi_\beta^*, \widehat{\pi}_{\mathrm{DR}}) \leq{} & |\mathcal{A}|M\epsilon + \sum_{j=1}^{|\mathcal{A}|}\omega_j + 84|\mathcal{A}|^2 e^{2M} M\sqrt{\frac{\log 1/\delta}{2n_2}} \\
& + \inf_\lambda \; 12|\mathcal{A}|\lambda + \frac{288|\mathcal{A}|}{\sqrt{n_2}}\int_\lambda^{|\mathcal{A}|M+2M/\eta}\sqrt{C_1|\mathcal{A}|\epsilon^{-\frac{d}{\beta}}\log\frac{1}{\epsilon}\left(\log^2\frac{1}{\epsilon}+\log\frac{1}{\theta}\right)}d\theta \\
\leq{} & |\mathcal{A}|M\epsilon + \sum_{j=1}^{|\mathcal{A}|}\omega_j + 84|\mathcal{A}|^2 e^{2M} M\sqrt{\frac{\log 1/\delta}{2n_2}} + \inf_\lambda \; 12|\mathcal{A}|\lambda \\
& + C_2\frac{288|\mathcal{A}|^{3/2}}{\sqrt{n_2}}M\eta^{-1}\epsilon^{-\frac{d}{2\beta}}\sqrt{\log\frac{1}{\epsilon}\left(\log^2\frac{1}{\epsilon}+\log\frac{1}{\lambda}\right)} \qquad \text{(D.45)}
\end{aligned}$$

with probability no less than $1 - 6|\mathcal{A}|\delta$ and C_2 depending on $\log D$, d, B, τ, η, β, and the surface area of $\mathcal{M}$. Setting $\epsilon = n_2^{-\frac{\beta}{2\beta+d}}, \delta = n_2^{-\frac{\beta}{2\beta+d}}, \lambda = n_2^{-\frac{\beta}{2\beta+d}}$ implies (Eq. 6.20) and (Eq. 6.21) in Theorem 6.1.

□

D.2.2 Proof of Corollary 6.1

Proof. Corollary 6.1 is proved based on Theorem 6.1 and Lemma 6.1. We first derive an upper bound of the ω_j's using Lemma 6.1. Taking an expectation on the both sides of (Eq. D.38) gives rise to

$$\begin{aligned}
\mathbb{E}[\omega_j] &\leq \eta^{-1}|\mathcal{A}|e^{2M}\mathbb{E}\left[\sqrt{\frac{1}{n_2}\sum_{i=n_1+1}^{n}\left(\widehat{\mu}_{A_j}(\mathbf{x}_i)-\mu_{A_j}(\mathbf{x}_i)\right)^2}\sqrt{\frac{1}{n_2}\sum_{i=n_1+1}^{n}\left(\widehat{e}_{A_j}(\mathbf{x}_i)-e_{A_j}(\mathbf{x}_i)\right)^2}\right] \\
&\leq \eta^{-1}|\mathcal{A}|e^{2M}\sqrt{\mathbb{E}\left[\|\widehat{\mu}_{A_j}-\mu_{A_j}\|_{L^2}^2\right]}\sqrt{\mathbb{E}\left[\|e_{A_j}-\widehat{e}_{A_j}\|_{L^2}^2\right]} \\
&\leq C_1 e^{2M}(M+\sigma)\eta^{-\frac{3\alpha+d}{2\alpha+d}}|\mathcal{A}|^{\frac{8\alpha+3d}{2(2\alpha+d)}}n_1^{-\frac{2\alpha}{2\alpha+d}}\log^3 n_1,
\end{aligned}$$

where the second inequality is due to Jensen's inequality and the unconfoundedness condition in Assumption 6.1, the last inequality is due to Lemma 6.1, and C_1 is a constant depending on $\log D, d, B, \tau, \alpha$ and the surface area of $\mathcal{M}$.

By Markov's inequality, for any $\delta > 0$,

$$\mathbb{P}\left(\omega_j > \delta\right) \leq \frac{\mathbb{E}\left[\omega_j\right]}{\delta} \leq \frac{1}{\delta} C_1 G_1 n_1^{-\frac{2\alpha}{2\alpha+d}} \log^3 n_1, \tag{D.46}$$

where $G_1 = e^{2M}(M+\sigma)\eta^{-\frac{3\alpha+d}{2\alpha+d}}|\mathcal{A}|^{\frac{8\alpha+3d}{2(2\alpha+d)}}$. Applying a union probability bound gives rise to

$$\mathbb{P}\left(\sum_{j=1}^{|\mathcal{A}|} \omega_j > |\mathcal{A}|\delta\right) \leq \frac{C_1}{\delta}|\mathcal{A}| G_1 n_1^{-\frac{2\alpha}{2\alpha+d}} \log^3 n_1. \tag{D.47}$$

Substituting (Eq. D.47) into (Eq. 6.21) and setting $\delta = C_1|\mathcal{A}|G_1 n_1^{-\frac{\alpha}{2\alpha+d}}$, we get

$$\begin{aligned} R(\pi^*_\beta, \widehat{\pi}_{\mathrm{DR}}) &\leq C_2 e^{2M}|\mathcal{A}|^2(M+\sigma) n_2^{-\frac{\beta}{2\beta+d}} \log^{3/2} n_2 + C_1|\mathcal{A}|^2 G_1 n_1^{-\frac{\alpha}{2\alpha+d}} \\ &\leq C_3 e^{2M}|\mathcal{A}|^{\frac{16\alpha+7d}{2(2\alpha+d)}}(M+\sigma) n^{-\frac{\alpha\wedge\beta}{2(\alpha\wedge\beta)+d}} \log^{3/2} n \end{aligned}$$

with probability no less than $1 - C_4 n^{-\frac{\alpha\wedge\beta}{2(\alpha\wedge\beta)+d}} \log^3 n$, where C_4 is an absolute constant, C_2, C_3 are constants depending on $\log D, d, B, \tau, \alpha, \beta, \eta$, and the surface area of $\mathcal{M}$.

□

D.2.3 Proof of Learning Unconstrained Policy

Proof. Proof of Theorem Theorem 6.2.In Theorem 6.2, π^* is the unconstrained optimal policy. We prove Theorem 6.2 in a similar manner as we prove Theorem 6.1. We first decompose the regret using an oracle inequality:

$$R(\pi^*, \widehat{\pi}_{\mathrm{DR}}) = \underbrace{Q(\pi^*) - Q(\widehat{\pi}^*)}_{(\mathrm{I}_2)} + \underbrace{Q(\widehat{\pi}^*) - Q(\widehat{\pi}_{\mathrm{DR}})}_{(\mathrm{II}_2)}, \tag{D.48}$$

where $\widehat{\pi}^*$ is the same as in (Eq. D.15). In (Eq. D.48), (I_2) is the bias of approximating π^* by the policy class $\Pi^{|\mathcal{A}|}_{\mathrm{NN}(H)}$, and (II_2) is the same as (II_1) in (Eq. D.16) which can be bounded similarly.

Following the proof of Theorem Theorem 6.1 and Corollary 6.1, we can derive that

$$(\mathrm{II}_2) \leq C_1 e^{2M} |\mathcal{A}|^{\frac{16\alpha+7d}{2(2\alpha+d)}} (M+\sigma) n^{-\frac{\alpha}{2\alpha+d}} \log^{3/2} n \log^{1/2}(1/H)$$

with probability no less than $1 - C_2 n^{-\frac{\alpha}{2\alpha+d}} \log^3 n$ where C_2 is an absolute constant and C_1 depends on $\log D, d, B, \tau, \alpha$, and the surface area of $\mathcal{M}$. In addition, $\widehat{\pi}_{\mathrm{DR}} \in \Pi_{\mathrm{NN}(H)}^{|\mathcal{A}|}$ with $L_\Pi, p_\Pi, K_\Pi, \kappa_\Pi$ and R_Π given in (Eq. 6.20). It remains to show $(\mathrm{I}_2) \leq 2cMt^q + M|\mathcal{A}|^2 \exp\left[\left(-Mt + 2n^{-\frac{2\alpha}{2\alpha+d}}\right)/H\right]$ for any $t \in (0,1)$.

Bounding (I_2). We estimate $Q(\pi^*) - Q(\widehat{\pi}^*)$ on two regions. The first region is, for any given $t \in (0,1)$,

$$\chi_t = \left\{ \mathbf{x} \mid \mathbf{x} \in \mathcal{M}, \mu_{A_{j^*(\mathbf{x})}}(\mathbf{x}) - \max_{j \neq j^*(\mathbf{x})} \mu_{A_j}(\mathbf{x}) \leq Mt \right\}$$

with $j^*(\mathbf{x}) = \operatorname{argmax}_j \mu_{A_j}(\mathbf{x})$. On χ_t, the gap between $\mu_{A_{j^*(\mathbf{x})}}(\mathbf{x})$, the reward of the optimal action, and the reward of the second optimal action is smaller than Mt. Assumption 6.A.4 yields $\mathbb{P}(\chi_t) \leq ct^q$. The second region is

$$\chi_t^{\complement} = \left\{ \mathbf{x} \mid \mathbf{x} \in \mathcal{M}, \mu_{A_{j^*(\mathbf{x})}}(\mathbf{x}) - \max_{j \neq j^*(\mathbf{x})} \mu_{A_j}(\mathbf{x}) > Mt \right\}$$

on which the gap between $\mu_{A_{j^*}}(\mathbf{x})$ and the reward of any other action is larger than Mt.

For any policy π, we have

$$Q(\pi) = \mathbb{E}[Y(\pi(\mathbf{x}))] = \int_{\mathcal{M}} \langle \boldsymbol{\mu}(\mathbf{x}), \pi(\mathbf{x}) \rangle \, d\mathbb{P}(\mathbf{x}),$$

where $\boldsymbol{\mu}(\mathbf{x}) = [\mu_{A_1}(\mathbf{x}), \ldots, \mu_{A_{|\mathcal{A}|}}(\mathbf{x})]^\top$. According to [214, Theorem 2], for any $\epsilon \in (0,1)$, there is a neural network architecture $\mathcal{F}(L, p, K, \kappa, R)$ with

$$L = O(\log 1/\epsilon), \; p = O\left(\epsilon^{-\frac{d}{\alpha}}\right), \; K = O\left(\epsilon^{-\frac{d}{\alpha}} \log 1/\epsilon\right), \; \kappa = \max\{B, M, \sqrt{d}, \tau^2\}, \; R = M,$$

such that for each μ_{A_j}, there exists $\widetilde{\mu}_{A_j} \in \mathcal{F}(L, p, K, \kappa, R)$ and $\|\widetilde{\mu}_{A_j} - \mu_{A_j}\|_\infty \leq \epsilon$.

Define $\widetilde{\pi} = \mathrm{Softmax}_H(\widetilde{\mu}_{A_1}, \ldots, \widetilde{\mu}_{A_{|\mathcal{A}|}})$. Since $\widehat{\pi}^* = \arg\max_{\pi \in \Pi_{\mathrm{NN}(H)}^{|\mathcal{A}|}} Q(\pi)$, $\widetilde{\pi} \in \Pi_{\mathrm{NN}(H)}^{|\mathcal{A}|}$, and $Q(\widehat{\pi}^*) \geq Q(\widetilde{\pi})$, we have

$$\begin{aligned} Q(\pi^*) - Q(\widehat{\pi}^*) \leq Q(\pi^*) - Q(\widetilde{\pi}) &= \int_{\mathcal{M}} \langle \boldsymbol{\mu}(\mathbf{x}), \pi^*(\mathbf{x}) - \widetilde{\pi}(\mathbf{x}) \rangle \, d\mathbb{P}(\mathbf{x}) \\ &= \int_{\chi_t} \langle \boldsymbol{\mu}(\mathbf{x}), \pi^*(\mathbf{x}) - \widetilde{\pi}(\mathbf{x}) \rangle \, d\mathbb{P}(\mathbf{x}) + \int_{\chi_t^{\complement}} \langle \boldsymbol{\mu}(\mathbf{x}), \pi^*(\mathbf{x}) - \widetilde{\pi}(\mathbf{x}) \rangle \, d\mathbb{P}(\mathbf{x}). \end{aligned} \tag{D.49}$$

The first integral in (Eq. D.49) can be bounded as

$$\begin{aligned} \int_{\chi_t} \langle \boldsymbol{\mu}(\mathbf{x}), \pi^*(\mathbf{x}) - \widetilde{\pi}(\mathbf{x}) \rangle \, d\mathbb{P}(\mathbf{x}) &\leq \int_{\chi_t} \|\boldsymbol{\mu}(\mathbf{x})\|_\infty \|\pi^*(\mathbf{x}) - \widetilde{\pi}(\mathbf{x})\|_1 d\mathbb{P}(\mathbf{x}) \\ &\leq 2M \int_{\chi_t} 1 d\mathbb{P}(\mathbf{x}) \leq 2cMt^q, \end{aligned} \tag{D.50}$$

where $\|\boldsymbol{\mu}(\mathbf{x})\|_\infty = \max_j |\mu_{A_j}(\mathbf{x})|$ and $\|\pi(\mathbf{x})\|_1 = \sum_{j=1}^{|\mathcal{A}|} |[\pi(\mathbf{x})]_j|$. For the second integral, we first derive an upper bound of $\|\pi^*(\mathbf{x}) - \widetilde{\pi}(\mathbf{x})\|_\infty$. Since π^* is the unconstrained optimal policy, represented by a one-hot vector $\pi^*(\mathbf{x}) = A_{j^*(\mathbf{x})}$, we deduce

$$[\pi^*(\mathbf{x}) - \widetilde{\pi}(\mathbf{x})]_j = \begin{cases} -\frac{\exp(\widetilde{\mu}_{A_j}(\mathbf{x})/H)}{\sum_k \exp(\widetilde{\mu}_{A_k}(\mathbf{x})/H)} & \text{if } \pi^*(\mathbf{x}) \neq A_j, \\ \frac{\sum_{k \neq j} \exp(\widetilde{\mu}_{A_k}(\mathbf{x})/H)}{\sum_k \exp(\widetilde{\mu}_{A_k}(\mathbf{x})/H)} & \text{if } \pi^*(\mathbf{x}) = A_j, \end{cases}$$

$$|[\pi^*(\mathbf{x}) - \widetilde{\pi}(\mathbf{x})]_j| \leq \begin{cases} \frac{\max_{k \neq j^*} \exp((\mu_{A_k}(\mathbf{x}) + \epsilon)/H)}{\exp((\mu_{A_{j^*}}(\mathbf{x}) - \epsilon)/H)} & \text{if } j \neq j^*(\mathbf{x}), \\ \frac{|\mathcal{A}| \max_{k \neq j^*} \exp((\mu_{A_k}(\mathbf{x}) + \epsilon)/H)}{\exp((\mu_{A_{j^*}}(\mathbf{x}) - \epsilon)/H)} & \text{if } j = j^*(\mathbf{x}). \end{cases}$$

Therefore $\|\pi^*(\mathbf{x}) - \widetilde{\pi}(\mathbf{x})\|_\infty \leq |\mathcal{A}| \exp\left(\max_{k \neq j^*(\mathbf{x})} \left(\mu_{A_k}(\mathbf{x}) - \mu_{A_{j^*}}(\mathbf{x}) + 2\epsilon\right)/H\right)$. Thus

$$\int_{\chi_t^{\complement}} \langle \boldsymbol{\mu}(\mathbf{x}), \pi^*(\mathbf{x}) - \widetilde{\pi}(\mathbf{x}) \rangle \, d\mathbb{P}(\mathbf{x}) \leq \int_{\chi_t^{\complement}} \|\boldsymbol{\mu}(\mathbf{x})\|_1 \|(\pi^*(\mathbf{x}) - \widetilde{\pi}(\mathbf{x}))\|_\infty d\mathbb{P}(\mathbf{x})$$

$$\leq \int_{\mathcal{X}_t^{\complement}} M|\mathcal{A}|^2 \exp\left((-Mt+2\epsilon)/H\right) d\mathbb{P}(\mathbf{x}) \leq M|\mathcal{A}|^2 \exp\left((-Mt+2\epsilon)/H\right). \qquad \text{(D.51)}$$

Combining (Eq. D.50) and (Eq. D.51), and setting $\epsilon = n^{-\frac{\alpha}{2\alpha+d}}$ give rise to $Q(\pi^*) - Q(\widehat{\pi}^*) \leq 2cMt^q + M|\mathcal{A}|^2 \exp\left((-Mt + 2n^{-\frac{\alpha}{2\alpha+d}})/H\right)$ which completes the proof.

□

D.2.4 Proof of Policy Learning with Continuous Actions

Proof of Theorem Theorem 6.3. We denote

$$\widehat{\pi}^* = \operatorname*{argmax}_{\pi \in \Pi^V_{\mathrm{NN}(H)}} Q^{(\mathrm{D})}(\pi),$$

where $Q^{(\mathrm{D})}(\pi)$ is defined in (Eq. 6.28). The regret can be decomposed as

$$R(\pi^*_{\mathrm{C}}, \widehat{\pi}_{\text{C-DR}}) = \underbrace{Q(\pi^*_{\mathrm{C}}) - Q^{(\mathrm{D})}(\widehat{\pi}^*)}_{(\mathrm{I}_3)} + \underbrace{Q^{(\mathrm{D})}(\widehat{\pi}^*) - Q^{(\mathrm{D})}(\widehat{\pi}_{\mathrm{C-DR}})}_{(\mathrm{II}_3)} + \underbrace{Q^{(\mathrm{D})}(\widehat{\pi}_{\mathrm{C-DR}}) - Q(\widehat{\pi}_{\mathrm{C-DR}})}_{(\mathrm{III}_3)}. \qquad \text{(D.52)}$$

In (Eq. D.52), (I_3) is the bias of approximating the optimal policy π^*_{C} using the neural network policy class $\Pi^V_{\mathrm{NN}(H)}$ in the discretized setting. (II_3) is the variance of the estimated policy in $\Pi^V_{\mathrm{NN}(H)}$. (III_3) characterizes the difference between the discretized policy reward and the continuous policy reward of $\widehat{\pi}_{\mathrm{C-DR}}$. We next derive the bounds for each part.

Bounding (I_3)**.** By Assumption 6.B.3, $\mu_{I_j} \in \mathcal{H}^\alpha(\mathcal{M})$. According to [214], Hölder functions can be uniformly approximated by a neural network class if the network parameters are properly chosen. For any $\epsilon \in (0,1)$ there exists a network architecture $\mathcal{F}(L, p, K, \kappa, R)$ with

$$L = O(\log 1/\epsilon), p = O\left(\epsilon^{-\frac{d}{\alpha}}\right), K = O\left(\epsilon^{-\frac{d}{\alpha}} \log 1/\epsilon\right), \kappa = \max\{B, M, \sqrt{d}, \tau^2\}, R = M, \qquad \text{(D.53)}$$

such that if the weight parameters are properly chosen, we have $\widetilde{\mu}_{I_j} \in \mathcal{F}(L, p, K, \kappa, R)$ satisfying

$$\|\widetilde{\mu}_{I_j} - \mu_{I_j}\|_\infty \leq \epsilon.$$

We then define an intermediate policy

$$\widetilde{\pi} = \mathrm{Softmax}_H(\widetilde{\mu}_{I_1}, \ldots, \widetilde{\mu}_{I_V}).$$

Let $A^*(\mathbf{x}) = \mathrm{argmax}_{A \in [0,1]} \mu(\mathbf{x}, A)$. Then $\pi^*_{\mathrm{C}}(\mathbf{x}) = A^*(\mathbf{x})$. After defining $\boldsymbol{\mu}(\mathbf{x}) = [\mu_{I_1}(\mathbf{x}), \ldots, \mu_{I_V}(\mathbf{x})]^\top \in \mathbb{R}^V$, we can bound (I_3) as

$$\begin{aligned}
(\mathrm{I}_3) &= Q(\pi^*_{\mathrm{C}}) - Q^{(\mathrm{D})}(\widehat{\pi}^*) \leq Q(\pi^*_{\mathrm{C}}) - Q^{(\mathrm{D})}(\widetilde{\pi}) \\
&= \int_{\mathcal{M}} \mu(\mathbf{x}, A^*(\mathbf{x})) d\mathbb{P}(\mathbf{x}) - \int_{\mathcal{M}} \langle \boldsymbol{\mu}(\mathbf{x}), \widetilde{\pi}(\mathbf{x}) \rangle d\mathbb{P}(\mathbf{x}) \\
&= \underbrace{\int_{\mathcal{M}} \mu(\mathbf{x}, A^*(\mathbf{x})) d\mathbb{P}(\mathbf{x}) - \int_{\mathcal{M}} \langle \boldsymbol{\mu}(\mathbf{x}), [\mathbb{1}_{\{A^*(\mathbf{x}) \in I_1\}}, \ldots, \mathbb{1}_{\{A^*(\mathbf{x}) \in I_V\}}]^\top \rangle d\mathbb{P}(\mathbf{x})}_{T_1} \\
&\quad + \underbrace{\int_{\mathcal{M}} \left\langle \boldsymbol{\mu}(\mathbf{x}), [\mathbb{1}_{\{A^*(\mathbf{x}) \in I_1\}}, \ldots, \mathbb{1}_{\{A^*(\mathbf{x}) \in I_V\}}]^\top \right\rangle d\mathbb{P}(\mathbf{x}) - \int_{\mathcal{M}} \langle \boldsymbol{\mu}(\mathbf{x}), \widetilde{\pi}(\mathbf{x}) \rangle d\mathbb{P}(\mathbf{x})}_{T_2}.
\end{aligned} \tag{D.54}$$

If $A^*(\mathbf{x}) \in I_j$, we denote $j^*(\mathbf{x}) = j$ and $I_*(\mathbf{x}) = I_j$. According to Assumption 6.B.3 and (Eq. 6.34), M is a Lipschitz constant of the function $\mu(\mathbf{x}, \cdot)$ for any fixed $\mathbf{x} \in \mathcal{M}$. Since $A^*(\mathbf{x}) \in I_*(\mathbf{x})$, $|\mu(\mathbf{x}, A^*(\mathbf{x})) - \mu_{I_*(\mathbf{x})}(\mathbf{x})| \leq M/V$ for any $\mathbf{x} \in \mathcal{M}$. Hence T_1 can be bounded as

$$T_1 = \int_{\mathcal{M}} \mu(\mathbf{x}, A^*(\mathbf{x})) - \mu_{I_*(\mathbf{x})}(\mathbf{x}) d\mathbb{P}(\mathbf{x}) \leq M/V. \tag{D.55}$$

We then derive the bound for T_2 on two regions. The first region is

$$\chi_{t,\gamma} = \{\mathbf{x} | \mu(\mathbf{x}, A^*(\mathbf{x})) - \mu(\mathbf{x}, A) \leq Mt \text{ given } |A - A^*(\mathbf{x})| \geq \gamma\}$$

and the second region is $\chi_{t,\gamma}^{\complement}$. According to Assumption 6.B.4, $\mathbb{P}(\chi_{t,\gamma}) \leq ct^q(1-\gamma)$.

T_2 is decomposed as

$$\begin{aligned} T_2 = &\int_{\chi_{t,\gamma}} \left\langle \boldsymbol{\mu}(\mathbf{x}), [\mathbb{1}_{\{A^*(\mathbf{x})\in I_1\}}, \dots, \mathbb{1}_{\{A^*(\mathbf{x})\in I_V\}}]^\top - \widetilde{\pi}(\mathbf{x}) \right\rangle d\mathbb{P}(\mathbf{x}) \\ &+ \int_{\chi_{t,\gamma}^{\complement}} \left\langle \boldsymbol{\mu}(\mathbf{x}), [\mathbb{1}_{\{A^*(\mathbf{x})\in I_1\}}, \dots, \mathbb{1}_{\{A^*(\mathbf{x})\in I_V\}}]^\top - \widetilde{\pi}(\mathbf{x}) \right\rangle d\mathbb{P}(\mathbf{x}). \end{aligned} \tag{D.56}$$

The first integral in (Eq. D.56) is bounded as

$$\int_{\chi_{t,\gamma}} \left\langle \boldsymbol{\mu}(\mathbf{x}), [\mathbb{1}_{\{A^*(\mathbf{x})\in I_1\}}, \dots, \mathbb{1}_{\{A^*(\mathbf{x})\in I_V\}}]^\top - \widetilde{\pi}(\mathbf{x}) \right\rangle d\mathbb{P}(\mathbf{x}) \leq 2cMt^q(1-\gamma). \tag{D.57}$$

We then derive an upper bound of the second integral in (Eq. D.56) in a way similar to the derivation of (Eq. D.51). Denote

$$\Xi(\mathbf{x}) = \left[-\frac{\exp(\widetilde{\mu}_{I_1}(\mathbf{x})/H)}{\sum_{j=1}^V \exp(\widetilde{\mu}_{I_j}(\mathbf{x})/H)}, \dots, \frac{\sum_{j\neq j^*(\mathbf{x})} \exp(\widetilde{\mu}_{I_j}(\mathbf{x})/H)}{\sum_{j=1}^V \exp(\widetilde{\mu}_{I_j}(\mathbf{x})/H)}, \dots, -\frac{\exp(\widetilde{\mu}_{I_V}(\mathbf{x})/H)}{\sum_{j=1}^V \exp(\widetilde{\mu}_{I_j}(\mathbf{x})/H)} \right]^\top.$$

Similar to (Eq. D.51), we have

$$\begin{aligned} &\int_{\chi_{t,\gamma}^{\complement}} \left\langle \boldsymbol{\mu}(\mathbf{x}), [\mathbb{1}_{\{A^*(\mathbf{x})\in I_1\}}, \dots, \mathbb{1}_{\{A^*(\mathbf{x})\in I_V\}}]^\top - \widetilde{\pi}(\mathbf{x}) \right\rangle d\mathbb{P}(\mathbf{x}) = \int_{\chi_{t,\gamma}^{\complement}} \langle \boldsymbol{\mu}(\mathbf{x}), \Xi(\mathbf{x}) \rangle \, d\mathbb{P}(\mathbf{x}) \\ \leq &\int_{\chi_{t,\gamma}^{\complement}} \|\boldsymbol{\mu}(\mathbf{x})\|_1 \|\Xi(\mathbf{x})\|_\infty d\mathbb{P}(\mathbf{x}) \leq VM \int_{\chi_{t,\gamma}^{\complement}} \|\Xi(\mathbf{x})\|_\infty d\mathbb{P}(\mathbf{x}). \end{aligned} \tag{D.58}$$

To derive an upper bound of $\|\Xi\|_\infty$, we need a lower bound of $\mu_{I_*(\mathbf{x})}(\mathbf{x}) - \mu_{I_j}(\mathbf{x})$ for any $1 \leq j \leq V$ and $j \neq j^*(\mathbf{x})$. By Assumption 6.B.3, For any j and $\widetilde{A} \in I_j$, one has

$$|\mu(\mathbf{x}, \widetilde{A}) - \mu(\mathbf{x}, A_j)| \leq M/V,$$

and

$$|\mu_{I_j}(\mathbf{x}) - \mu(\mathbf{x}, A_j)| \leq \frac{1}{|I_j|} \int_{I_j} |\mu(\mathbf{x}, A) - \mu(\mathbf{x}, A_j)| dA \leq M/V$$

where $|I_j| = 1/V$ represents the length of I_j.

As a result, on $\chi_{t,\gamma}^{\complement}$, for any $j \neq j^*(\mathbf{x})$, we have

$$\begin{aligned}&\mu_{I_*(\mathbf{x})}(\mathbf{x}) - \mu_{I_j}(\mathbf{x}) \geq \mu(\mathbf{x}, A_{j^*(\mathbf{x})}) - \mu(\mathbf{x}, A_j) - 2M/V \\ &\geq \mu(\mathbf{x}, A^*(\mathbf{x})) - \mu(\mathbf{x}, A_j) - |\mu(\mathbf{x}, A^*(\mathbf{x})) - \mu(\mathbf{x}, A_{j^*(\mathbf{x})})| - 2M/V \geq Mt - 3M/V,\end{aligned}$$

where the last inequality holds for two reasons: (1) $A^*(\mathbf{x}) \in I_*(\mathbf{x})$ and $A_{j^*(\mathbf{x})} \in I_*(\mathbf{x})$; (2) We set $V < 1/(2\gamma)$, and then $j \neq j^*(\mathbf{x})$ implies $|A_j - A^*(\mathbf{x})| \geq 1/(2V) \geq \gamma$. We then deduce

$$\|\Xi\|_\infty \leq (V-1)\exp(-(Mt - 3M/V - 2\epsilon)/H). \tag{D.59}$$

Plugging (Eq. D.59) into (Eq. D.58), we have

$$\begin{aligned}&\int_{\chi_{t,\gamma}^{\complement}} \left\langle \boldsymbol{\mu}(\mathbf{x}), [\mathbb{1}_{\{A^*(\mathbf{x}) \in I_1\}}, \ldots, \mathbb{1}_{\{A^*(\mathbf{x}) \in I_V\}}]^\top - \widetilde{\pi}(\mathbf{x}) \right\rangle d\mathbb{P}(\mathbf{x}) \\ \leq& VM \cdot (V-1)\exp(-(Mt - 3M/V - 2\epsilon)/H) \leq MV^2 \exp(-(Mt - 3M/V - 2\epsilon)/H).\end{aligned} \tag{D.60}$$

Substituting (Eq. D.55), (Eq. D.57) and (Eq. D.60) into (Eq. D.54), if $V < 1/(2\gamma)$, we have

$$(\mathrm{I}_3) \leq \frac{M}{V} + 2cMt^q(1-\gamma) + MV^2 \exp\left(-(Mt - 3M/V - 2\epsilon)/H\right). \tag{D.61}$$

Bounding (II_3)**.** (II_3) has the same form as (II_1) in (Eq. D.16). We derive the upper bound by following the same procedure while $|\mathcal{A}|$ is replaced V. Besides, we need to express the estimation error of $\widehat{\mu}_{I_j}$'s and $\widehat{e}_{I_j}$'s in terms of V. Note that $e_{I_j} \geq \eta/V$. By Lemma 6.1, we can find $\widehat{\mu}_{I_j} \in \mathcal{F}(L_1, p_1, K_1, \kappa_1, R_1)$ with

$$L_1 = O(\log(\eta n_1/V)), \quad p_1 = O\left((\eta n_1/V)^{\frac{d}{2\alpha+d}}\right), \quad K_1 = O\left((\eta n_1/V)^{\frac{d}{2\alpha+d}} \log(\eta n_1/V)\right),$$
$$\kappa_1 = \max\{B, M, \sqrt{d}, \tau^2\}, \quad R_1 = M,$$

such that

$$\mathbb{E}\left[\|\widehat{\mu}_{I_j} - \mu_{I_j}\|_{L^2}^2\right] \le C_1(M^2 + \sigma^2)(\eta n_1/V)^{-\frac{2\alpha}{2\alpha+d}} \log^3(\eta n_1/V) \tag{D.62}$$

with C_1 being a constant depending on $\log D, B, \tau, \alpha$ and the surface area of $\mathcal{M}$. Similarly, we can find $\widehat{g} \in \mathcal{F}(L_2, p_2, K_2, \kappa_2, R_2)$ with

$$L_2 = O(\log(n_1/V)),\ p_2 = O\left(V^{\frac{2\alpha}{2\alpha+d}} n_1^{\frac{d}{2\alpha+d}}\right),\ K_2 = O\left(V^{\frac{2\alpha}{2\alpha+d}} n_1^{\frac{d}{2\alpha+d}} \log(n_1/V)\right),$$
$$\kappa_2 = \max\{B, M, \sqrt{d}, \tau^2\},\ R_2 = M,$$

such that

$$\mathbb{E}[\|\widehat{e}_{I_j} - e_{I_j}\|_{L^2}^2] \le C_2 M^2 V^{\frac{4\alpha+d}{2\alpha+d}} n_1^{-\frac{2\alpha}{2\alpha+d}} \log^3 n_1 \tag{D.63}$$

with C_2 depending on $\log D, B, M, \alpha$ and the surface area of $\mathcal{M}$.

Following the proof of Corollary 6.1 and using (Eq. D.62) and (Eq. D.63), we rewrite (Eq. D.47) as

$$\mathbb{P}\left(\sum_{j=1}^{V} \omega_j \ge V\delta_1\right) \le \frac{C_3 G_2}{\delta_1} V^{\frac{14\alpha+5d}{2(2\alpha+d)}} n_1^{-\frac{2\alpha}{2\alpha+d}} \log^3 n_1 \tag{D.64}$$

with $G_2 = e^{2M}(M+\sigma)\eta^{-\frac{3\alpha+d}{2\alpha+d}}$ and C_3 being an constant depending on $\log D, B, \tau, \eta, \alpha$ and the surface area of $\mathcal{M}$.

By replacing $|\mathcal{A}|$ by V and η by η/V in $\mathcal{E}_1$ and $\mathcal{E}_2$ in the proof of Theorem 6.1, and substituting (Eq. D.64), one derives

$$
\begin{aligned}
Q^{(\mathrm{D})}(\widehat{\pi}^*) - Q^{(\mathrm{D})}(\widehat{\pi}_{\text{C-DR}}) \leq V\delta_1 + 84e^{2M}V^2M\sqrt{\frac{\log 1/\delta}{2n_2}} + \inf_{\lambda}\ 12V\lambda \\
+ \frac{288V}{\sqrt{n_2}}\int_{\lambda}^{VM+2VM/\eta}\Big[K_\Pi \log\big(\theta^{-1}(VM+2VM/\eta)\times \\
L_\Pi^2(p_\Pi B/V+2)\max(\kappa_\Pi, 1/H)^{L_\Pi}(p_\Pi/V)^{L_\Pi+1}\big)\Big]^{1/2} d\theta
\end{aligned}
$$

with probability no less than

$$
1 - 6V\delta - \frac{C_3G_2}{\delta_1}V^{\frac{14\alpha+5d}{2(2\alpha+d)}}n_1^{-\frac{2\alpha}{2\alpha+d}}\log^3 n_1.
$$

Here $\widehat{\pi}_{\text{C-DR}} \in \Pi^V_{\mathrm{NN}(H)}$ where the network class $\Pi^V_{\mathrm{NN}(H)}$ has the parameters

$$
L_\Pi = L,\ p_\Pi = O(Vp),\ K_\Pi = O(VK),\ \kappa_\Pi = \kappa,\ R_\Pi = R
$$

with L, p, K, κ and R defined in (Eq. D.53).

Setting $\epsilon = n^{-\frac{\alpha}{2\alpha+d}}, \delta = n^{-\frac{\alpha}{2\alpha+d}}, \delta_1 = C_3G_2V^{\frac{3}{2}}n_1^{-\frac{\alpha}{2\alpha+d}}$ and $\lambda = V^{\frac{3}{2}}n_2^{-\frac{\alpha}{2\alpha+d}}$ gives rise to

$$
\begin{aligned}
L_\Pi = O(\log n), p_\Pi = O\left(Vn^{\frac{d}{2\alpha+d}}\right), K_\Pi = O\left(Vn^{\frac{d}{2\alpha+d}}\log n\right), \\
\kappa_\Pi = \max\{B, M, \sqrt{d}, \tau^2\}, R_\Pi = M,
\end{aligned}
$$

and

$$
(\mathrm{II}_3) \leq C_4e^{2M}(M+\sigma)V^{\frac{5}{2}}n^{-\frac{\alpha}{2\alpha+d}}\log^{3/2} n \log^{1/2}(1/H) \qquad \text{(D.65)}
$$

with probability no less than $1 - C_5V^{\frac{4\alpha+d}{2\alpha+d}}n^{-\frac{\alpha}{2\alpha+d}}\log^3 n$, where C_4 is a constant depending on $\log D, B, \tau, \eta, \alpha$, and the surface area of $\mathcal{M}$, C_5 is an absolute constant.

Bounding (III_3)**.** According to Lemma 6.2,

$$(\mathrm{III}_3) = Q^{(\mathrm{D})}(\widehat{\pi}_{\text{C-DR}}) - Q(\widehat{\pi}_{\text{C-DR}}) \leq M/V. \tag{D.66}$$

Putting all ingredients together. Putting (Eq. D.61), (Eq. D.65) and (Eq. D.66) together and using $\epsilon = n^{-\frac{\alpha}{2\alpha+d}}$ give rise to

$$\begin{aligned} R(\pi_{\mathrm{C}}^*, \widehat{\pi}_{\text{C-DR}}) \leq{} & \frac{2M}{V} + C_4 e^{2M}(M+\sigma)V^{\frac{5}{2}} n^{-\frac{\alpha}{2\alpha+d}} \log^{3/2} n \log^{1/2} 1/H \\ & + 2cMt^q(1-\gamma) + MV^2 \exp\left(-\left(Mt - 3M/V - 2n^{-\frac{\alpha}{2\alpha+d}}\right)/H\right) \end{aligned}$$

with probability no less than $1 - C_5 V^{\frac{4\alpha+d}{2\alpha+d}} n^{-\frac{\alpha}{2\alpha+d}} \log^3 n$ for any t and $\gamma < 1/4V$.

Setting $V = n^{\frac{2\alpha}{7(2\alpha+d)}}, \gamma = \frac{1}{4V}$ and $t > \frac{2}{V} + 2\epsilon/M$, we get

$$\begin{aligned} R(\pi_{\mathrm{C}}^*, \widehat{\pi}_{\text{C-DR}}) \leq{} & C_4 e^{2M}(M+\sigma) n^{-\frac{2\alpha}{7(2\alpha+d)}} \log^{3/2} n \log^{1/2} 1/H \\ & + 2cMt^q + Mn^{\frac{4\alpha}{7(2\alpha+d)}} \exp\left(-\left(Mt - 4Mn^{-\frac{2\alpha}{7(2\alpha+d)}}\right)/H\right) \end{aligned}$$

for any $t \in (2(1+1/M)n^{-\frac{2\alpha}{7(2\alpha+d)}}, 1)$ with probability no less than $1 - C_6 n^{-\frac{6\alpha^2+5\alpha d}{7(2\alpha+d)^2}} \log^3 n$, where C_6 is an absolute constant. In addition, $\widehat{\pi}_{\text{C-DR}} \in \Pi_{\mathrm{NN}(H)}^V$ with $L_\Pi, p_\Pi, K_\Pi, \kappa_\Pi, R_\Pi$ defined in (Eq. 6.37), $\widehat{\mu}_{I_j} \in \mathcal{F}(L_1, p_1, K_1, \kappa_1, R_1)$ for $j = 1, \ldots, V$ with the parameters defined in (Eq. 6.35), $\widehat{g} \in \mathcal{F}(L_2, p_2, K_2, \kappa_2, R_2)$ with the parameters defined in (Eq. 6.36).

□

D.3 Technical Proofs

D.3.1 Proof of Lemma 6.1

Proof. We first derive the error bound $\|\widehat{\mu}_{A_j} - \mu_{A_j}\|_{L^2}$ for any j. Note that $\widehat{\mu}_{A_j} \in \mathcal{F}(L_1, p_1, K_1, \kappa_1, R_1)$ is the minimizer of (Eq. 6.4). If we choose

$$\begin{aligned} L_1 = O(\log n_{A_j}),\ p_1 = O\left(n_{A_j}^{\frac{d}{2\alpha+d}}\right),\ K_1 = O\left(n_{A_j}^{\frac{d}{2\alpha+d}} \log n_{A_j}\right), \\ \kappa_1 = \max\{B, M, \sqrt{d}, \tau^2\},\ R_1 = M, \end{aligned} \tag{D.67}$$

then according to [214, Theorem 1], for each j, we have

$$\mathbb{E}\left[\|\widehat{\mu}_{A_j} - \mu_{A_j}\|_{L^2}^2\right] \leq C_1(M^2 + \sigma^2) n_{A_j}^{-\frac{2\alpha}{2\alpha+d}} \log^3 n_{A_j}, \tag{D.68}$$

where $n_{A_j} = \sum_{i=1}^{n_1} \mathbb{1}_{\{\mathbf{a}_i = A_j\}}$ and C_1 is a constant only depending on $\log D, B, \tau$ and the surface area of $\mathcal{M}$. In (Eq. D.68) the expectation is taken with respect to the randomness of samples.

Next, we derive a high probability lower bound of n_{A_j} for all j's in terms of n_1. By Assumption 6.A.2(ii), $\mathbb{E}(n_{A_j}/n_1) \geq \eta$. By [228, Lemma 29], we have

$$\mathbb{P}\left(\left|\frac{n_{A_j}}{n_1} - \mathbb{E}\left(\frac{n_{A_j}}{n_1}\right)\right| \geq \frac{1}{2}\mathbb{E}\left(\frac{n_{A_j}}{n_1}\right)\right) \leq 2\exp\left(-\frac{3}{28} n_1 \mathbb{E}\left(\frac{n_{A_j}}{n_1}\right)\right).$$

Thus $n_{A_j} \geq \eta n_1/2$ holds with probability at least $1 - 2\exp\left(-\frac{3}{28}\eta n_1\right)$. Denote the event $E_1 = \{n_{A_j} \geq \eta n_1/2\}$ and its complement by $E_1^{\complement}$. When n_1 (so as n) is large enough, we have

$$\begin{aligned} \mathbb{E}\left[\|\widehat{\mu}_{A_j} - \mu_{A_j}\|_{L^2}^2\right] &= \mathbb{E}\left[\|\widehat{\mu}_{A_j} - \mu_{A_j}\|_{L^2}^2 | E_1\right] \mathbb{P}(E_1) + \mathbb{E}\left[\|\widehat{\mu}_{A_j} - \mu_{A_j}\|_{L^2}^2 | E_1^{\complement}\right] \mathbb{P}\left(E_1^{\complement}\right) \\ &\leq C_1(M^2 + \sigma^2)(\eta n_1)^{-\frac{2\alpha}{2\alpha+d}} \log^3(\eta n_1) + 2C_1(M^2 + \sigma^2)\exp\left(-\frac{3}{28}\eta n_1\right) \end{aligned}$$

$$\leq C_2(M^2+\sigma^2)(\eta n_1)^{-\frac{2\alpha}{2\alpha+d}} \log^3(\eta n_1),$$

where C_2 is a constant depending on $\log D, B, \tau$ and the surface area of $\mathcal{M}$. Substituting $n_{A_j} = \eta n_1$ into (Eq. D.67) gives rise to $\widehat{\mu}_{A_j} \in \mathcal{F}(L_1, p_1, K_1, \kappa_1, R_1)$ with $L_1, p_1, K_1, \kappa_1, R_1$ in (Eq. 6.14).

To estimate $\mathbb{E}\left[\|\widehat{e}_{A_j} - e_{A_j}\|_{L^2}^2\right]$, we use $\mathcal{H}^{\alpha}_{|\mathcal{A}|-1}(\mathcal{M})$ to denote the space of the $|\mathcal{A}|-1$ dimensional vectors whose elements are in $\mathcal{H}^{\alpha}(\mathcal{M})$. We denote $g^* = \left[g_{A_1}, \ldots, g_{A_{|\mathcal{A}|-1}}\right]^\top$ with $g_{A_j} = \log \frac{e_{A_j}}{e_{A_{|\mathcal{A}|}}}$. According to Assumption 6.A.3, $g_{A_j} = \log e_{A_j} - \log e_{A_{|\mathcal{A}|}} \in \mathcal{H}^{\alpha}$, $\|g_{A_j}\|_{\mathcal{H}^{\alpha}} \leq M$ and $g^* \in \mathcal{H}^{\alpha}_{|\mathcal{A}|-1}(\mathcal{M})$. Let $\widehat{g}$ be the minimizer of (Eq. 6.5). From [229, Corollary 4] and the proof of [206, Theorem 2], setting $\mathcal{G}_{\text{NN}} = \mathcal{F}(L, p, K, \kappa, R)$ gives rise to

$$\begin{aligned} &\|\widehat{g} - g^*\|_{L^2}^2 \\ &\leq C_3 M^2 \left(\frac{|\mathcal{A}| LK \log K}{n_1} \log n_1 + \frac{|\mathcal{A}|(\log\log n_1 + \gamma)}{n_1} + |\mathcal{A}| \sup_{g' \in \mathcal{H}^{\alpha}_{|\mathcal{A}|-1}(\mathcal{M})} \inf_{g \in \mathcal{G}_{NN}} \|g^* - g'\|_\infty^2 \right) \end{aligned}$$

with probability at least $1 - \exp(-\gamma)$, where C_3 is an absolute constant.

According to [214, Theorem 2], for any $\epsilon_2 \in (0,1)$, there exists a neural network architecture $\mathcal{F}(L, p, K, \kappa, R)$ with

$$L = O\left(\log \frac{1}{\epsilon_2}\right), p = O\left(|\mathcal{A}| \epsilon_2^{-\frac{d}{\alpha}}\right), K = O\left(|\mathcal{A}| \epsilon_2^{-\frac{d}{\alpha}} \log \frac{1}{\epsilon_2}\right),$$
$$\kappa = \max\{B, M, \sqrt{d}, \tau^2\}, R = M$$

such that for any $g \in \mathcal{H}^{\alpha}_{|\mathcal{A}|-1}(\mathcal{M})$, there exists $\widetilde{g} \in \mathcal{F}(L, p, K, \kappa, R)$ with $\|\widetilde{g} - g\|_\infty \leq \epsilon_2$, where $\|g\|_\infty = \sup_{\mathbf{x} \in \mathcal{M}} \max_j |g_j(\mathbf{x})|$. Setting $\epsilon_2 = |\mathcal{A}|^{\frac{\alpha}{2\alpha+d}} n_1^{-\frac{\alpha}{2\alpha+d}}, \gamma = |\mathcal{A}|^{-\frac{d}{2\alpha+d}} n_1^{\frac{d}{2\alpha+d}}$ gives rise to $\mathcal{G}_{\text{NN}} = \mathcal{F}(L_2, p_2, K_2, \kappa_2, R_2)$ with

$$L_2 = O(\log(n_1/|\mathcal{A}|)),\ p_2 = O\left(|\mathcal{A}|^{\frac{2\alpha}{2\alpha+d}} n_1^{\frac{d}{2\alpha+d}}\right),\ K_2 = O\left(|\mathcal{A}|^{\frac{2\alpha}{2\alpha+d}} n_1^{\frac{d}{2\alpha+d}} \log(n_1/|\mathcal{A}|)\right),$$

$$\kappa_2 = \max\{B, M, \sqrt{d}, \tau^2\},\ R_2 = M \tag{D.69}$$

which implies (Eq. 6.15).Then with probability no less than $1 - \exp\left(-|\mathcal{A}|^{-\frac{d}{2\alpha+d}} n_1^{\frac{d}{2\alpha+d}}\right)$, we deduce

$$\|\widehat{g} - g^*\|_{L^2}^2 \leq C_4 M^2 |\mathcal{A}|^{\frac{4\alpha+d}{2\alpha+d}} n_1^{-\frac{2\alpha}{2\alpha+d}} \log^3 n_1$$

with C_4 depending on $\log D, B, \tau$ and the surface area of $\mathcal{M}$. Denote the event

$$E_2 = \left\{\|\widehat{g} - g^*\|_{L^2}^2 \leq C_4 M^2 |\mathcal{A}|^{\frac{4\alpha+d}{2\alpha+d}} n_1^{-\frac{2\alpha}{2\alpha+d}} \log^3 n_1\right\}.$$

When n_1 (so as n) is large enough, we obtain

$$\begin{aligned}
\mathbb{E}[\|\widehat{g} - g^*\|_{L^2}^2] &= \mathbb{E}[\|\widehat{g} - g^*\|_{L^2}^2 | E_2]\mathbb{P}(E_2) + \mathbb{E}[\|\widehat{g} - g^*\|_{L^2}^2 | E_2^{\complement}]\mathbb{P}(E_2^{\complement}) \\
&\leq C_4 M^2 |\mathcal{A}|^{\frac{4\alpha+d}{2\alpha+d}} n_1^{-\frac{2\alpha}{2\alpha+d}} \log^3 n_1 + 4(M^2 + \sigma^2) \exp\left(-|\mathcal{A}|^{-\frac{d}{2\alpha+d}} n_1^{\frac{d}{2\alpha+d}}\right) \\
&\leq C_5 M^2 |\mathcal{A}|^{\frac{4\alpha+d}{2\alpha+d}} n_1^{-\frac{2\alpha}{2\alpha+d}} \log^3 n_1
\end{aligned}$$

with C_5 depending on $\log D, B, \tau$ and the surface area of $\mathcal{M}$.

Define $r_j(g) = \frac{\exp([g]_j)}{1+\sum_{k=1}^{|\mathcal{A}|-1} \exp([g]_k)}$ for $j = 1, \ldots, |\mathcal{A}| - 1$. Since $\|\nabla r_j\|_\infty \leq 1$ for any j, we have

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{e}_{A_j} - e_{A_j}\|_{L^2}^2\right] &= \mathbb{E}\left[\|r_j(\widehat{g}) - r_j(g^*)\|_{L^2}^2\right] \\
&\leq \mathbb{E}\left[(\|\nabla r_j\|_\infty \|\widehat{g} - g^*\|_{L^2})^2\right] \leq C_5 M^2 |\mathcal{A}|^{\frac{4\alpha+d}{2\alpha+d}} n_1^{-\frac{2\alpha}{2\alpha+d}} \log^3 n_1.
\end{aligned}$$

Similarly, one can show $\mathbb{E}\left[\|\widehat{e}_{A_{|\mathcal{A}|}} - e_{A_{|\mathcal{A}|}}\|_{L^2}^2\right] \leq C_5 M^2 |\mathcal{A}|^{\frac{4\alpha+d}{2\alpha+d}} n_1^{-\frac{2\alpha}{2\alpha+d}} \log^3 n_1$. □

D.3.2 Proof of Lemma 6.2

Proof. Recall that

$$
\begin{aligned}
Q^{(\mathrm{D})}(\pi) &= \int_{\mathcal{M}} \left\langle [\mu_{I_1}(\mathbf{x}), \ldots, \mu_{I_V}(\mathbf{x})]^\top, \pi(\mathbf{x}) \right\rangle d\mathbb{P}(\mathbf{x}), \\
Q(\pi) &= \int_{\mathcal{M}} \left\langle [\mu_{A_1}(\mathbf{x}), \ldots, \mu_{A_V}(\mathbf{x})]^\top, \pi(\mathbf{x}) \right\rangle d\mathbb{P}(\mathbf{x}).
\end{aligned}
$$

Since L_μ is a uniform Lipschitz constant of $\mu(\mathbf{x}, \cdot)$ for any $\mathbf{x} \in \mathcal{M}$, we derive

$$
Q^{(\mathrm{D})}(\pi) - Q(\pi) = \int_{\mathcal{M}} \left\langle [\mu_{I_1}(\mathbf{x}) - \mu(\mathbf{x}, A_1), \ldots, \mu_{I_V}(\mathbf{x}) - \mu(\mathbf{x}, A_V)]^\top, \pi(\mathbf{x}) \right\rangle d\mathbb{P}(\mathbf{x}) \leq L_\mu / V.
$$

□

D.3.3 Proof of Lemma D.4

Proof. We first use McDiarmid's inequality (Lemma D.11) to show $\mathcal{D}(\Pi)$ concentrates around $\mathbb{E}[\mathcal{D}(\Pi)]$ and then derive a bound of $\mathbb{E}[\mathcal{D}(\Pi)]$. To simplify the notation, we omit the domain Π in $\mathcal{D}$.

We denote $\{\mathring{\Gamma}'_i\}_{i=1}^n$ as the counterpart of $\{\mathring{\Gamma}\}_{i=1}^n$ when one sample $(\mathbf{x}_k, \Gamma_k)$ is replaced by $(\mathbf{x}_k, \Gamma'_k)$ for any k with $1 \leq k \leq n$. $\mathring{\Delta}'(\pi_1, \pi_2)$ and $\mathcal{D}'$ are defined analogously. We have

$$
\begin{aligned}
|\mathcal{D} - \mathcal{D}'| &\leq \sup_{\pi_1, \pi_2 \in \Pi} \mathring{\Delta}(\pi_1, \pi_2) - \mathring{\Delta}'(\pi_1, \pi_2) \leq \sup_{\pi_1, \pi_2 \in \Pi} \frac{1}{n} \left\langle \mathring{\Gamma}_k - \mathring{\Gamma}'_k, \pi_1(\mathbf{x}_k) - \pi_2(\mathbf{x}_k) \right\rangle \\
&\leq \frac{1}{n} \left\| \mathring{\Gamma}_k - \mathring{\Gamma}'_k \right\|_\infty \|\pi_1(\mathbf{x}_k) - \pi_2(\mathbf{x}_k)\|_1 \leq \frac{2}{n} \left\| \mathring{\Gamma}_k - \mathring{\Gamma}'_k \right\|_\infty \leq \frac{4}{n} J, \qquad \text{(D.70)}
\end{aligned}
$$

where $\|\cdot\|_\infty$ and $\|\cdot\|_1$ stand for the ℓ^∞ and ℓ^1 norm for vectors. Applying Lemma D.11 with $f = \mathcal{D}$, we have

$$
\mathbb{P}\left(\mathcal{D} - \mathbb{E}[\mathcal{D}] \geq t\right) \leq \exp\left(-2nt^2 / \left(16J^2\right)\right). \qquad \text{(D.71)}
$$

Setting $t = 4J\sqrt{\frac{\log 1/\delta}{2n}}$ gives rise to

$$\mathcal{D} \leq \mathbb{E}[\mathcal{D}] + 4J\sqrt{\frac{\log 1/\delta}{2n}} \tag{D.72}$$

with probability no less than $1 - \delta$.

We next derive a bound of $\mathbb{E}[\mathcal{D}]$ by symmetrization:

$$\begin{aligned}\mathbb{E}[\mathcal{D}] &= \mathbb{E}\left[\sup_{\pi_1,\pi_2\in\Pi} \mathring{\Delta}(\pi_1,\pi_2) - \mathbb{E}\left[\mathring{\Delta}(\pi_1,\pi_2)\right]\right] \leq \mathbb{E}\left[\sup_{\pi_1,\pi_2\in\Pi} \mathring{\Delta}(\pi_1,\pi_2) - \mathring{\Delta}_{\text{copy}}(\pi_1,\pi_2)\right] \\ &= \mathbb{E}\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi} \boldsymbol{\xi} \odot \left(\mathring{\Delta}(\pi_1,\pi_2) - \mathring{\Delta}_{\text{copy}}(\pi_1,\pi_2)\right)\right] = 2\mathbb{E}\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi} \boldsymbol{\xi} \odot \mathring{\Delta}(\pi_1,\pi_2)\right],\end{aligned}$$

where $\mathring{\Delta}_{\text{copy}}$ denotes $\mathring{\Delta}$ using independent copies of samples and $\boldsymbol{\xi} = [\xi_1,\ldots,\xi_n]^\top$ with ξ_i's being i.i.d. Rademacher variables which take value 1 or -1 with the same probability. Here $\boldsymbol{\xi} \odot \mathring{\Delta}$ denotes the entry-wise product of ξ and $\mathring{\Delta}$, i.e.,

$$\boldsymbol{\xi} \odot \mathring{\Delta}(\pi_1,\pi_2) := \frac{1}{n}\sum_{i=1}^{n} \xi_i \left\langle \mathring{\Gamma}_i, \pi_1(\mathbf{x}_i) - \pi_2(\mathbf{x}_i)\right\rangle.$$

We next apply Lemma D.11 with $f = \mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi} \boldsymbol{\xi} \odot \mathring{\Delta}(\pi_1,\pi_2)\right]$. Again, we denote $\{\mathring{\Gamma}'_i\}_{i=1}^n$ as the counterpart of $\{\mathring{\Gamma}\}_{i=1}^n$ when one sample $(\mathbf{x}_k, \mathring{\Gamma}_k)$ is replaced by $(\mathbf{x}'_k, \mathring{\Gamma}'_k)$ for any k with $1 \leq k \leq n$. $\mathring{\Delta}'(\pi_1,\pi_2)$ is defined analogously. We get

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi} \boldsymbol{\xi} \odot \mathring{\Delta}(\pi_1,\pi_2)\right] - \mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi} \boldsymbol{\xi} \odot \mathring{\Delta}'(\pi_1,\pi_2)\right] \\ \leq& \mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi} \frac{1}{n}\xi_k \left\langle \mathring{\Gamma}_k - \mathring{\Gamma}'_k, \pi_1(\mathbf{x}_k) - \pi_2(\mathbf{x}_k)\right\rangle\right] \\ \leq& \frac{1}{n}\left\|\mathring{\Gamma}_k - \mathring{\Gamma}'_k\right\|_\infty \left\|\pi_1(\mathbf{x}_k) - \pi_2(\mathbf{x}_k)\right\|_1 \leq \frac{4}{n}J.\end{aligned} \tag{D.73}$$

Applying Lemma D.11 with $f = \mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi} \boldsymbol{\xi} \odot \widetilde{\Delta}(\pi_1,\pi_2)\right]$ gives rise to

$$\mathbb{P}\left(\mathbb{E}\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi} \boldsymbol{\xi} \odot \mathring{\Delta}(\pi_1,\pi_2)\right] - \mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi} \boldsymbol{\xi} \odot \mathring{\Delta}(\pi_1,\pi_2)\right] \geq t\right) \leq \exp\left(-2nt^2/\left(16J^2\right)\right). \tag{D.74}$$

Setting $t = 4J\sqrt{\frac{\log 1/\delta}{2n_2}}$ gives rise to

$$\mathbb{P}\left(\mathbb{E}\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi} \boldsymbol{\xi} \odot \mathring{\Delta}(\pi_1,\pi_2)\right] - \mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi} \boldsymbol{\xi} \odot \mathring{\Delta}(\pi_1,\pi_2)\right] \geq 4J\sqrt{\frac{\log 1/\delta}{2n}}\right) \leq \delta. \tag{D.75}$$

The following lemma provides an upper bound of $\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi} \boldsymbol{\xi} \odot \mathring{\Delta}(\pi_1,\pi_2)\right]$ (see a proof in Appendix D.3.6):

Lemma D.7. *Let $\boldsymbol{\xi}$ be a set of Rademacher random variable and $\mathring{\Delta}(\pi_1,\pi_2)$ defined in (Eq. D.26). Then the following bound holds*

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi} \boldsymbol{\xi} \odot \mathring{\Delta}(\pi_1,\pi_2)\right] \leq \inf_{\lambda}\ 2\lambda + \frac{48}{\sqrt{n}}\int_{\lambda}^{\max_{\pi\in\Pi}\|\pi\|_\Gamma} \sqrt{\log\mathcal{N}(\theta,\Pi,\|\cdot\|_\Gamma)}d\theta, \tag{D.76}$$

where $\mathcal{N}(\theta,\Pi_{\mathrm{NN}}^{|\mathcal{A}|},\|\cdot\|_\Gamma)$ is the θ-covering number (see Definition D.3) of Π with respect to the measure $\|\pi\|_\Gamma = \sqrt{\frac{1}{n}\sum_{i=1}^n \langle \mathring{\Gamma}_i, \pi(\mathbf{x}_i)\rangle^2}$.

Substituting (Eq. D.76) into (Eq. D.75) yields

$$\mathbb{E}[\mathcal{D}] \leq \inf_{\lambda}\ 4\lambda + \frac{96}{\sqrt{n}}\int_{\lambda}^{\max_{\pi\in\Pi}\|\pi\|_\Gamma} \sqrt{\log\mathcal{N}(\theta,\Pi,\|\cdot\|_\Gamma)}d\theta + 8J\sqrt{\frac{\log 1/\delta}{2n}} \tag{D.77}$$

with probability no less than $1-\delta$.

Substituting (Eq. D.77) into (Eq. D.72) give rise to

$$\mathcal{D} \leq \inf_{\lambda}\ 4\lambda + \frac{96}{\sqrt{n}}\int_{\lambda}^{\max_{\pi\in\Pi}\|\pi\|_\Gamma} \sqrt{\log\mathcal{N}(\theta,\Pi,\|\cdot\|_\Gamma)}d\theta + 12J\sqrt{\frac{\log 1/\delta}{2n}} \tag{D.78}$$

with probability no less than $1 - 2\delta$. □

D.3.4 Proof of Lemma D.5

Proof. We derive the bound of the covering number $\mathcal{N}(\theta, \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, \|\cdot\|_{\Gamma})$ using the covering number of the neural network class $\mathcal{N}(\theta, \mathcal{F}(L, p, K, \kappa, R), \|\cdot\|_{\infty})$. Let $\pi^{(1)} = \mathrm{Softmax}(\mu_{A_1}^{(1)}, \ldots, \mu_{A_{|\mathcal{A}|}}^{(1)})$ and $\pi^{(2)} = \mathrm{Softmax}(\mu_{A_1}^{(2)}, \ldots, \mu_{A_{|\mathcal{A}|}}^{(2)})$ be two policies in $\Pi_{\mathrm{NN}}^{|\mathcal{A}|}(L_{\Pi}, p_{\Pi}, K_{\Pi}, \kappa_{\Pi}, R_{\Pi})$ such that for each j, $\|\mu_{A_j}^{(1)} - \mu_{A_j}^{(2)}\|_{\infty} \leq \theta$. By Assumption 6.A.2 and (Eq. 6.12), $\|\widetilde{\Gamma}_i\|_1 \leq 2M/\eta + |\mathcal{A}|M$ for any $n_1 \leq i \leq n$. Therefore we have

$$\begin{aligned}
\left\|\pi^{(1)} - \pi^{(2)}\right\|_{\Gamma}^2 &= \frac{1}{n_2} \sum_{i=n_1+1}^{n} \left\langle \widetilde{\Gamma}_i, (\pi^{(1)} - \pi^{(2)})(\mathbf{x}_i) \right\rangle^2 \\
\leq& \frac{1}{n_2} \sum_{i=n_1+1}^{n} \left\|\widetilde{\Gamma}_i\right\|_1^2 \left\|(\pi^{(1)} - \pi^{(2)})(\mathbf{x}_i)\right\|_{\infty}^2 \leq (|\mathcal{A}|M + 2M/\eta))^2 \theta^2.
\end{aligned}$$

Thus we obtain

$$\mathcal{N}(\theta, \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, \|\cdot\|_{\Gamma}) \leq \mathcal{N}\left(\theta/(|\mathcal{A}|M + 2M/\eta), \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, \|\cdot\|_{\infty}\right). \tag{D.79}$$

Since for every $\pi \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}$ with $L_{\Pi} = L, p_{\Pi} = |\mathcal{A}|p, K_{\Pi} = |\mathcal{A}|K, \kappa_{\Pi} = \kappa, R_{\Pi} = R$, it contains $|\mathcal{A}|$ parallel ReLU networks in $\mathcal{F}(L, p, K, \kappa, R)$ with an additional softmax layer, we have $\mathcal{N}(\theta, \Pi_{\mathrm{NN}}^{|\mathcal{A}|}, \|\cdot\|_{\infty}) \leq \mathcal{N}(\theta, \mathcal{F}(L, p, K, \kappa, R), \|\cdot\|_{\infty})^{|\mathcal{A}|}$. From [214, Proof of Theorem 3.1], we have

$$\mathcal{N}(\theta, \mathcal{F}(L, p, K, \kappa, R), \|\cdot\|_{\infty}) \leq \left(\frac{2L^2(pR+2)\kappa^L p^{L+1}}{\theta}\right)^K.$$

We get

$$\mathcal{N}(\theta, \Pi_{\mathrm{NN}}, \|\cdot\|_{\infty}) \leq \left(\frac{2L^2(pR+2)\kappa^L p^{L+1}}{\theta}\right)^{|\mathcal{A}|K}. \tag{D.80}$$

Combining (Eq. D.79) and (Eq. D.80) proves Lemma D.5. □

D.3.5 Proof of Lemma D.6

Proof. For any $\pi \in \Pi_{\mathrm{NN}}^{|\mathcal{A}|}$,

$$\|\pi\|_\Gamma^2 \leq \frac{1}{n_2} \sum_{i=n_1+1}^{n} \langle \widetilde{\Gamma}_i, \pi(\mathbf{x}_i) \rangle^2 \leq \frac{1}{n_2} \sum_{i=n_1+1}^{n} \|\widetilde{\Gamma}_i\|_1^2 \|\pi(\mathbf{x}_i)\|_\infty^2 \leq \frac{1}{n_2} \sum_{i=n_1+1}^{n} \|\widetilde{\Gamma}_i\|_1^2.$$

By Assumption 6.A.2 and (Eq. 6.12), $\|\widetilde{\Gamma}_i\|_1 \leq 2M/\eta + |\mathcal{A}|M$ for any $n_1 \leq i \leq n$. Therefore we obtain

$$\|\pi\|_\Gamma^2 \leq (2M/\eta + |\mathcal{A}|M)^2.$$

□

D.3.6 Proof of Lemma D.7

We first define the covering number of a set.

Definition D.3. *Let $\mathcal{F}$ be a set equipped with metric ρ. For any $\delta > 0$, a δ-covering of $\mathcal{F}$ is a set $\{f_1, \dots, f_N\} \subset \mathcal{F}$ such that for any $f \in \mathcal{F}$, there exists f_k for $1 \leq k \leq N$ with $\rho(f_k, f) \leq \theta$. The δ-covering number of $\mathcal{F}$ is defined as*

$$\mathcal{N}(\delta, \mathcal{F}, \rho) = \inf\{N : \textit{ there exists } \{f_1, \dots, f_N\} \textit{ which is a } \theta\textit{-covering of } \mathcal{F}\}. \quad \text{(D.81)}$$

Proof of Lemma D.7. To bound $\mathbb{E}_{\boldsymbol{\xi}} \left[\sup_{\pi_1, \pi_2 \in \Pi} \boldsymbol{\xi} \odot \mathring{\Delta}(\pi_1, \pi_2) \right]$ with respect to the measure $\|\cdot\|_\Gamma$, we construct a series of I coverings of Π with resolutions $\{\delta_i\}_{i=1}^{I}$ satisfying $\delta_{i+1} = \frac{1}{2}\delta_i$. The elements in the (i)-th covering are denoted as $\{\pi_i^{(i)}\}_{i=1}^{N^{(i)}}$, where the $N^{(i)}$'s are to be determined later. Thus for any $\pi \in \Pi$, there exists $\pi^{(i)}$ in the (i)-th covering such that

$$\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left\langle \mathring{\Gamma}_i, \pi(\mathbf{x}_i) - \pi^{(i)}(\mathbf{x}_i) \right\rangle^2} \leq \delta_i.$$

Let $\pi_1^{(i)}$ denote the closest element of π_1 in the (i)-th covering, and $\pi_2^{(i)}$ is defined analogously. We now expand $\pi_1 - \pi_2$ using a telescoping sum:

$$\pi_1 - \pi_2 = \left(\pi_1 - \pi_1^{(I)} + \sum_{i=1}^{I-1} \pi_1^{(i+1)} - \pi_1^{(i)} + \pi_1^{(1)} \right) - \left(\pi_2 - \pi_2^{(I)} + \sum_{i=1}^{I-1} \pi_2^{(i+1)} - \pi_2^{(i)} + \pi_2^{(1)} \right). \tag{D.82}$$

Substituting (Eq. D.82) into $\mathbb{E}_{\boldsymbol{\xi}} \left[\sup_{\pi_1, \pi_2 \in \Pi} \boldsymbol{\xi} \otimes \mathring{\Delta}(\pi_1, \pi_2) \right]$, due to the bi-linearity of $\mathring{\Delta}$, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}} \left[\sup_{\pi_1, \pi_2 \in \Pi} \boldsymbol{\xi} \otimes \mathring{\Delta}(\pi_1, \pi_2) \right] \\
\leq &\mathbb{E}_{\boldsymbol{\xi}} \left[\sup_{\pi_1 \in \Pi} \frac{1}{n} \sum_{i=1}^{n} \xi_i \left\langle \mathring{\Gamma}_i, \left(\pi_1 - \pi_1^{(I)} + \sum_{i=1}^{I-1} \pi_1^{(i+1)} - \pi_1^{(i)} + \pi_1^{(1)} \right) (\mathbf{x}_i) \right\rangle \right] \\
&+ \mathbb{E}_{\boldsymbol{\xi}} \left[\sup_{\pi_2 \in \Pi} \frac{1}{n} \sum_{i=1}^{n} \xi_i \left\langle \widetilde{\Gamma}_i, \left(\pi_2 - \pi_2^{(I)} + \sum_{i=1}^{I-1} \pi_2^{(i+1)} - \pi_2^{(i)} + \pi_2^{(1)} \right) (\mathbf{x}_i) \right\rangle \right].
\end{aligned} \tag{D.83}$$

By the construction of the coverings, we immediately have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}} \left[\sup_{\pi_1 \in \Pi} \frac{1}{n} \sum_{i=1}^{n} \xi_i \left\langle \mathring{\Gamma}_i, \left(\pi_1 - \pi_1^{(I)} \right) (\mathbf{x}_i) \right\rangle \right] \\
\leq &\mathbb{E}_{\boldsymbol{\xi}} \left[\sup_{\pi_1 \in \Pi} \frac{1}{n} \|\boldsymbol{\xi}\|_2 \sqrt{\sum_{i=1}^{n} \left\langle \mathring{\Gamma}_i, \left(\pi_1 - \pi_1^{(I)} \right) (\mathbf{x}_i) \right\rangle^2} \right] \leq \delta_I.
\end{aligned} \tag{D.84}$$

We can also check

$$\begin{aligned}
&\sqrt{\sum_{i=1}^{n} \left\langle \mathring{\Gamma}_i, \pi^{(i+1)}(\mathbf{x}_i) - \pi^{(i)}(\mathbf{x}_i) \right\rangle^2} = \sqrt{\sum_{i=1}^{n} \left\langle \mathring{\Gamma}_i, \pi^{(i+1)}(\mathbf{x}_i) - \pi(\mathbf{x}_i) + \pi(\mathbf{x}_i) - \pi^{(i)}(\mathbf{x}_i) \right\rangle^2} \\
\leq &\sqrt{2 \sum_{i=1}^{n} \left\langle \mathring{\Gamma}_i, \pi^{(i+1)}(\mathbf{x}_i) - \pi(\mathbf{x}_i) \right\rangle^2 + 2 \sum_{i=1}^{n} \left\langle \mathring{\Gamma}_i, \pi(\mathbf{x}_i) - \pi^{(i)}(\mathbf{x}_i) \right\rangle^2} \\
\leq &\sqrt{2n(\delta_{i+1}^2 + \delta_i^2)} \leq \sqrt{2n}(\delta_{i+1} + \delta_i).
\end{aligned} \tag{D.85}$$

Using Lemma D.12, we have

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1\in\Pi}\frac{1}{n}\sum_{i=1}^{n}\xi_i\left\langle\mathring{\Gamma}_i,\pi_1^{(i+1)}(\mathbf{x}_i)-\pi_1^{(i)}(\mathbf{x}_i)\right\rangle\right]$$
$$\le\frac{2(\delta_{i+1}+\delta_i)\sqrt{\log(\mathcal{N}(\delta_i,\Pi,\|\cdot\|_\Gamma)\mathcal{N}(\delta_{i+1},\Pi,\|\cdot\|_\Gamma))}}{\sqrt{n}}\le\frac{4(\delta_{i+1}+\delta_i)\sqrt{\log\mathcal{N}(\delta_{i+1},\Pi,\|\cdot\|_\Gamma)}}{\sqrt{n}}, \tag{D.86}$$

where the metric in the covering is $\|\pi\|_\Gamma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\langle\mathring{\Gamma}_i,\pi(\mathbf{x}_i)\right\rangle^2}$. Substituting (Eq. D.84), (Eq. D.86) into (Eq. D.83), and invoking the identity $\delta_{i+1}+\delta_i=6(\delta_{i+1}-\delta_{i+2})$ yield

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi}\boldsymbol{\xi}\otimes\mathring{\Delta}(\pi_1,\pi_2)\right]\le2\delta_I+\sum_{i=1}^{I-1}\frac{8(\delta_{i+1}+\delta_i)\sqrt{\log\mathcal{N}(\delta_{i+1},\Pi,\|\cdot\|_\Gamma)}}{\sqrt{n}}$$
$$\le2\delta_I+\frac{48(\delta_{i+1}-\delta_{i+2})\sqrt{\log\mathcal{N}(\delta_{i+1},\Pi,\|\cdot\|_\Gamma)}}{\sqrt{n}}\le2\delta_I+\frac{48}{\sqrt{n}}\int_{\delta_I}^{\delta_1}\sqrt{\log\mathcal{N}(\tau,\Pi,\|\cdot\|_\Gamma)}d\tau.$$

Choosing $\delta_1=\max_{\pi\in\Pi}\|\pi\|_\Gamma$ so that the first covering only consists of one element, we derive

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\pi_1,\pi_2\in\Pi}\boldsymbol{\xi}\odot\mathring{\Delta}(\pi_1,\pi_2)\right]\le\inf_{\lambda}\ 2\lambda+\frac{48}{\sqrt{n}}\int_{\lambda}^{\max_{\pi\in\Pi}\|\pi\|_\Gamma}\sqrt{\log\mathcal{N}(\theta,\Pi,\|\cdot\|_\Gamma)}d\theta.$$

□

D.4 Helper Lemmas

Lemma D.8. *Let $f(\mathbf{x},A)$ be any function defined on $\mathcal{M}\times[0,1]$. Assume there exists $M>0$ such that*

$$\sup_{A\in[0,1]}\|f(\cdot,A)\|_{\mathcal{H}^\alpha(\mathcal{M})}\le M \textit{ and } \sup_{\mathbf{x}\in\mathcal{M}}|f(\mathbf{x},A)-f(\mathbf{x},\widetilde{A})|\le M|A-\widetilde{A}|,\ \forall A,\widetilde{A}\in[0,1]. \tag{D.87}$$

Then $F^{(I)} = \int_I f(\mathbf{x}, A)dA \in \mathcal{H}^\alpha(\mathcal{M})$ *satisfies* $\|F^{(I)}\|_{\mathcal{H}^\alpha(\mathcal{M})} \leq M|I|$ *for any interval* $I \subset [0,1]$ *where* $|I|$ *is the length of* I*.*

Proof of Lemma D.8. To show $F^{(I)} \in \mathcal{H}^\alpha(\mathcal{M})$, it is sufficient to show $\|F^{(I)}\|_{\mathcal{H}^\alpha(U)} < \infty$ for any chart (U, ϕ) of $\mathcal{M}$. For simplicity, we denote

$$F_\phi^{(I)}(\mathbf{z}) = F^{(I)} \circ \phi^{-1}(\mathbf{z}),\ f_\phi(\mathbf{z}, A) = f(\phi^{-1}(\mathbf{z}), A)$$

for $\mathbf{z} \in \phi(U)$. Then $F_\phi^{(I)}(\mathbf{z}) = \int_I f_\phi(\mathbf{z}, A)dA$.

We first consider $0 < \alpha \leq 1$. In this case, we have

$$\begin{aligned}\|F_\phi^{(I)}\|_{\mathcal{H}^\alpha(\phi(U))} &= \sup_{\mathbf{z}\neq\mathbf{y}\in\phi(U)} \frac{|F_\phi^{(I)}(\mathbf{z}) - F_\phi^{(I)}(\mathbf{y})|}{\|\mathbf{z}-\mathbf{y}\|_2^\alpha} \\ &\leq \sup_{\mathbf{z}\neq\mathbf{y}\in\phi(U)} \int_I \frac{|f_\phi(\mathbf{z}, A) - f_\phi(\mathbf{y}, A)|}{\|\mathbf{z}-\mathbf{y}\|_2^\alpha} dA \leq M|I| < \infty, \qquad \text{(D.88)}\end{aligned}$$

which implies $F^{(I)} \in \mathcal{H}^\alpha(\mathcal{M})$.

Next we consider $\alpha > 1$. We first show that $\partial^{\mathbf{s}} F_\phi^{(I)}(\mathbf{z}) = \int_I \partial^{\widetilde{\mathbf{s}}} f_\phi(\mathbf{z}, A)dA$ for any $|\mathbf{s}| \leq \lceil \alpha - 1 \rceil$ where $\widetilde{\mathbf{s}} = [\mathbf{s}^\top, 0]^\top$. Let $\{h_n\}_{n=1}^\infty$ be any sequence converging to 0. When $|\mathbf{s}| = 1$, by definition, we have

$$\partial^{\mathbf{s}} F_\phi^{(I)}(\mathbf{z}) = \lim_{n\to\infty} \frac{F_\phi^{(I)}(\mathbf{z} + h_n\mathbf{s}) - F_\phi^{(I)}(\mathbf{z})}{h_n} = \lim_{n\to\infty} \int_I \frac{f_\phi(\mathbf{z} + h_n\mathbf{s}, A) - f_\phi(\mathbf{z}, A)}{h_n} dA.$$

Since $\|f_\phi(\mathbf{z}, A)\|_{\mathcal{H}^\alpha(\phi(U))} \leq M$ for any fixed $A \in [0,1]$, by the mean value theorem,

$$\left| \frac{f_\phi(\mathbf{z} + h_n\mathbf{s}, A) - f_\phi(\mathbf{z}, A)}{h_n} \right| \leq \max_{\widetilde{\mathbf{z}}\in\phi(U)} |\partial^{\widetilde{\mathbf{s}}} f_\phi(\widetilde{\mathbf{z}}, A)| \leq M.$$

Since

$$\lim_{n\to\infty} \frac{f_\phi(\mathbf{z} + h_n\mathbf{s}, A) - f_\phi(\mathbf{z}, A)}{h_n} = \partial^{\widetilde{\mathbf{s}}} f_\phi(\mathbf{z}, A) \qquad \text{(D.89)}$$

and by the dominated convergence theorem, we obtain

$$\begin{aligned}\partial^{\mathbf{s}} F_{\phi}^{(I)}(\mathbf{z}) &= \lim_{n\to\infty} \frac{F_{\phi}^{(I)}(\mathbf{z}+h_n\mathbf{s}) - F_{\phi}^{(I)}(\mathbf{z})}{h_n} \\ &= \int_I \lim_{n\to\infty} \frac{f_{\phi}(\mathbf{z}+h_n\mathbf{s}, A) - f_{\phi}(\mathbf{z}, A)}{h_n} dA = \int_I \partial^{\tilde{\mathbf{s}}} f_{\phi}(\mathbf{z}, A) dA.\end{aligned}$$

Similarly, for any $|\mathbf{s}| \leq \lceil \alpha - 1 \rceil$, $\partial^{\tilde{\mathbf{s}}} f(\mathbf{x}, A)$ can be expressed in the form similar to (Eq. D.89) using the Taylor series. Following the same procedure, one can show

$$\partial^{\mathbf{s}} F_{\phi}^{(I)}(\mathbf{z}) = \int_I \partial^{\tilde{\mathbf{s}}} f_{\phi}(\mathbf{z}, A) dA$$

for any $|\mathbf{s}| \leq \lceil \alpha - 1 \rceil$. Therefore we have

$$\max_{|\mathbf{s}| \leq \lceil \alpha - 1 \rceil} \sup_{\mathbf{z} \in \phi(U)} |\partial^{\mathbf{s}} F_{\phi}^{(I)}| \leq M|I| < \infty, \tag{D.90}$$

where $|I|$ represents the length of I.

On the other hand,

$$\begin{aligned}&\max_{|\mathbf{s}| = \lceil \alpha - 1 \rceil} \sup_{\mathbf{z} \neq \mathbf{y} \in \phi(U)} \frac{|\partial^{\mathbf{s}} F_{\phi}^{(I)}(\mathbf{z}) - \partial^{\mathbf{s}} F_{\phi}^{(I)}(\mathbf{y})|}{\|\mathbf{z} - \mathbf{y}\|_2^{\alpha - \lceil \alpha - 1 \rceil}} \\ \leq &\max_{|\mathbf{s}| = \lceil \alpha - 1 \rceil} \sup_{\mathbf{z} \neq \mathbf{y} \in \phi(U)} \int_I \frac{|\partial^{\tilde{\mathbf{s}}} f_{\phi}(\mathbf{z}, A) - \partial^{\tilde{\mathbf{s}}} f_{\phi}(\mathbf{y}, A)|}{\|\mathbf{z} - \mathbf{y}\|_2^{\alpha - \lceil \alpha - 1 \rceil}} dA \leq M|I| < \infty.\end{aligned} \tag{D.91}$$

Combining (Eq. D.90) and (Eq. D.91) gives $\|F^{(I)}\|_{\mathcal{H}^{\alpha}(U)} < \infty$ for any chart (U, ϕ) which implies $F^{(I)} \in \mathcal{H}^{\alpha}(\mathcal{M})$. □

Lemma D.9. *Assume Assumption 3.1 and 3.2 hold. Let $f, g \in \mathcal{H}^{\alpha}(\mathcal{M})$ with $\inf_{\mathbf{x} \in \mathcal{M}} g(\mathbf{x}) \geq \eta > 0$. Let $M > 0$ be a constant such that $\|f\|_{\mathcal{H}^{\alpha}(\mathcal{M})} \leq M$ and $\|g\|_{\mathcal{H}^{\alpha}(\mathcal{M})} \leq M$. Then we have $f/g \in \mathcal{H}^{\alpha}(\mathcal{M})$ with $\|f/g\|_{\mathcal{H}^{\alpha}(\mathcal{M})} \leq 2^{\frac{5+\lceil \alpha - 1 \rceil}{2} \lceil \alpha - 1 \rceil} (M/\eta)^{2^{\lceil \alpha \rceil}} (2B + 1)$.*

Proof of Lemma D.9. It is sufficiently to show $\|f/g\|_{\mathcal{H}^{\alpha}(U)} < \infty$ for any chart (U, ϕ) of

$\mathcal{M}$. For simplicity, denote

$$f_\phi(\mathbf{z}) = f \circ \phi^{-1}(\mathbf{z}),\ g_\phi(\mathbf{z}) = g \circ \phi^{-1}(\mathbf{z})$$

for any $\mathbf{z} \in \phi(U)$.

We first consider $0 < \alpha \leq 1$. In this case,

$$\begin{aligned}
\|f/g\|_{\mathcal{H}^\alpha(U)} = &\sup_{\mathbf{z}\neq\mathbf{y}\in\phi(U)} \frac{|(f_\phi(\mathbf{z})/g_\phi(\mathbf{z})) - (f_\phi(\mathbf{y})/g_\phi(\mathbf{y}))|}{\|\mathbf{z}-\mathbf{y}\|_2^\alpha} \\
\leq &\frac{|f_\phi(\mathbf{z})g_\phi(\mathbf{y}) - f_\phi(\mathbf{z})g_\phi(\mathbf{z}) + f_\phi(\mathbf{z})g_\phi(\mathbf{z}) - f_\phi(\mathbf{y})g_\phi(\mathbf{z})|}{g_\phi(\mathbf{y})g_\phi(\mathbf{z})\|\mathbf{z}-\mathbf{y}\|_2^\alpha} \\
\leq &\frac{1}{\eta^2}\left(M\frac{|g_\phi(\mathbf{y}) - g_\phi(\mathbf{z})|}{\|\mathbf{z}-\mathbf{y}\|_2^\alpha} + M\frac{|f_\phi(\mathbf{z}) - f_\phi(\mathbf{y})|}{\|\mathbf{z}-\mathbf{y}\|_2^\alpha}\right) \leq 2M^2/\eta^2 < \infty
\end{aligned} \tag{D.92}$$

which implies $f/g \in \mathcal{H}^\alpha(\mathcal{M})$.

We next consider the case $\alpha > 1$. We first show $|\partial^{\mathrm{s}}(f_\phi(\mathbf{z})/g_\phi(\mathbf{z}))| < \infty$ for $|\mathrm{s}| \leq \lceil \alpha - 1 \rceil$. When $|\mathrm{s}| = 1$, we have

$$\left|\partial^{\mathrm{s}}\left(\frac{f_\phi(\mathbf{z})}{g_\phi(\mathbf{z})}\right)\right| = \left|\frac{\partial^{\mathrm{s}} f_\phi(\mathbf{z})g_\phi(\mathbf{z}) - f_\phi(\mathbf{z})\partial^{\mathrm{s}} g_\phi(\mathbf{z})}{g_\phi^2(\mathbf{z})}\right| \leq 2M^2/\eta^2.$$

For any $|\mathrm{s}| \leq \lceil \alpha - 1 \rceil$, following this process, one can show

$$\left|\partial^{\mathrm{s}}\left(\frac{f_\phi(\mathbf{z})}{g_\phi(\mathbf{z})}\right)\right| = \frac{\sum_{i=1}^{2^{\frac{1+|\mathrm{s}|}{2}|\mathrm{s}|}} G_i}{g_\phi^{2^{|\mathrm{s}|}}} \leq 2^{\frac{1+|\mathrm{s}|}{2}|\mathrm{s}|}(M/\eta)^{2^{|\mathrm{s}|}} < \infty, \tag{D.93}$$

where each G_i is the product of $2^{|\mathrm{s}|}$ terms from $\{\partial^{\bar{\mathrm{s}}} f_\phi, \partial^{\bar{\mathrm{s}}} f_\phi \| \bar{\mathrm{s}}| \leq |\mathrm{s}|\}$.

On the other hand, note that for any s with $|\mathrm{s}| = 1$, we have

$$\begin{aligned}
&\left|\partial^{\mathrm{s}}\left(\frac{f_\phi(\mathbf{z})}{g_\phi(\mathbf{z})}\right) - \partial^{\mathrm{s}}\left(\frac{f_\phi(\mathbf{y})}{g_\phi(\mathbf{y})}\right)\right| \\
= &\left|\frac{\partial^{\mathrm{s}} f_\phi(\mathbf{z})g_\phi(\mathbf{z}) - f_\phi(\mathbf{z})\partial^{\mathrm{s}} g_\phi(\mathbf{z})}{g_\phi^2(\mathbf{z})} - \frac{\partial^{\mathrm{s}} f_\phi(\mathbf{y})g_\phi(\mathbf{y}) - f_\phi(\mathbf{y})\partial^{\mathrm{s}} g_\phi(\mathbf{y})}{g_\phi^2(\mathbf{y})}\right|
\end{aligned}$$

$$
\begin{aligned}
&=\left|\frac{g_\phi^2(\mathbf{y})\partial^{\mathbf{s}} f_\phi(\mathbf{z})g_\phi(\mathbf{z})-g_\phi^2(\mathbf{z})\partial^{\mathbf{s}} f_\phi(\mathbf{y})g_\phi(\mathbf{y})-\left(g_\phi^2(\mathbf{y})f_\phi(\mathbf{z})\partial^{\mathbf{s}} g_\phi(\mathbf{z})-g_\phi^2(\mathbf{z})f(\mathbf{y})\partial^{\mathbf{s}} g_\phi(\mathbf{y})\right)}{g_\phi^2(\mathbf{z})g_\phi^2(\mathbf{y})}\right| \\
&\leq\frac{1}{\eta^4}\Big|g_\phi(\mathbf{y})g_\phi(\mathbf{z})\left[\partial^{\mathbf{s}} f_\phi(\mathbf{z})g_\phi(\mathbf{y})-\partial^{\mathbf{s}} f_\phi(\mathbf{z})g_\phi(\mathbf{z})+\partial^{\mathbf{s}} f_\phi(\mathbf{z})g_\phi(\mathbf{z})-\partial^{\mathbf{s}} f_\phi(\mathbf{y})g_\phi(\mathbf{z})\right] \\
&\quad+g_\phi^2(\mathbf{y})f_\phi(\mathbf{z})\partial^{\mathbf{s}} g_\phi(\mathbf{z})-g_\phi^2(\mathbf{y})f_\phi(\mathbf{z})\partial^{\mathbf{s}} g_\phi(\mathbf{y})+g_\phi^2(\mathbf{y})f_\phi(\mathbf{z})\partial^{\mathbf{s}} g_\phi(\mathbf{y})-g_\phi^2(\mathbf{y})f_\phi(\mathbf{y})\partial^{\mathbf{s}} g_\phi(\mathbf{y}) \\
&\quad+g_\phi^2(\mathbf{y})f_\phi(\mathbf{y})\partial^{\mathbf{s}} g_\phi(\mathbf{y})-g_\phi(\mathbf{y})g_\phi(\mathbf{z})f_\phi(\mathbf{y})\partial^{\mathbf{s}} g_\phi(\mathbf{y}) \\
&\quad+g_\phi(\mathbf{y})g_\phi(\mathbf{z})f_\phi(\mathbf{y})\partial^{\mathbf{s}} g_\phi(\mathbf{y})-g_\phi^2(\mathbf{z})f_\phi(\mathbf{y})\partial^{\mathbf{s}} g_\phi(\mathbf{y})\Big| \\
&\leq\frac{M^3}{\eta^4}\left[3|g_\phi(\mathbf{z})-g_\phi(\mathbf{y})|+|f_\phi(\mathbf{z})-f_\phi(\mathbf{y})|+|\partial^{\mathbf{s}} g_\phi(\mathbf{z})-\partial^{\mathbf{s}} g_\phi(\mathbf{y})|+|\partial^{\mathbf{s}} f_\phi(\mathbf{z})-\partial^{\mathbf{s}} f_\phi(\mathbf{y})|\right] \\
&\leq\frac{M^3}{\eta^4}\left[4M\|\mathbf{z}-\mathbf{y}\|+|\partial^{\mathbf{s}} g_\phi(\mathbf{z})-\partial^{\mathbf{s}} g_\phi(\mathbf{y})|+|\partial^{\mathbf{s}} f_\phi(\mathbf{z})-\partial^{\mathbf{s}} f_\phi(\mathbf{y})|\right].
\end{aligned}
$$

Analogously, for any $|\mathbf{s}| \leq \lceil \alpha-1 \rceil$, one can show

$$
\begin{aligned}
&\left|\partial^{\mathbf{s}}\left(\frac{f_\phi(\mathbf{z})}{g_\phi(\mathbf{z})}\right)-\partial^{\mathbf{s}}\left(\frac{f_\phi(\mathbf{y})}{g_\phi(\mathbf{y})}\right)\right| \\
\leq&(M/\eta)^{2^{|\mathbf{s}|+1}-1}\left(C_1M\|\mathbf{z}-\mathbf{y}\|+C_2|\partial^{\mathbf{s}} g_\phi(\mathbf{z})-\partial^{\mathbf{s}} g_\phi(\mathbf{y})|+C_3|\partial^{\mathbf{s}} f_\phi(\mathbf{z})-\partial^{\mathbf{s}} f_\phi(\mathbf{y})|\right)
\end{aligned}
$$

for some absolute constants C_1, C_2, C_3 such that $C_1+C_2+C_3=2^{\frac{5+|\mathbf{s}|}{2}|\mathbf{s}|}$. Thus we deduce

$$
\begin{aligned}
&\max_{|\mathbf{s}|\leq\lceil\alpha-1\rceil}\sup_{\mathbf{z}\neq\mathbf{y}\in\phi(U)}\frac{|\partial^{\mathbf{s}}\left(f_\phi(\mathbf{z})/g_\phi(\mathbf{z})\right)-\partial^{\mathbf{s}}\left(f_\phi(\mathbf{y})/g_\phi(\mathbf{y})\right)|}{\|\mathbf{z}-\mathbf{y}\|_2^{\alpha-\lceil\alpha-1\rceil}} \\
\leq&(M/\eta)^{2^{\lceil\alpha-1\rceil+1}-1}(2C_1MB+(C_2+C_3)M)<2^{\frac{5+\lceil\alpha-1\rceil}{2}\lceil\alpha-1\rceil}(M/\eta)^{2^{\lceil\alpha\rceil}}(2B+1)<\infty.
\end{aligned}
\tag{D.94}
$$

Combining (Eq. D.93) and (Eq. D.94) yields

$$
\|f/g\|_{\mathcal{H}^\alpha(U)}<2^{\frac{5+\lceil\alpha-1\rceil}{2}\lceil\alpha-1\rceil}(M/\eta)^{2^{\lceil\alpha\rceil}}(2B+1)<\infty \tag{D.95}
$$

for any chart (U,ϕ) of $\mathcal{M}$ which implies $f/g\in\mathcal{H}^\alpha(\mathcal{M})$.

□

Lemma D.10. *Assume Assumption 3.1 and 3.2. Let* $f\in\mathcal{H}^\alpha(\mathcal{M})$ *with* $\alpha>1$ *and* $f(\mathbf{x})\geq$

$\eta > 0$. *Let* $M > 0$ *be a constant such that* $\|f\|_{\mathcal{H}^\alpha(\mathcal{M})} \leq M$. *Then we have* $\log f \in \mathcal{H}^\alpha(\mathcal{M})$ *with* $\|\log f\|_{\mathcal{H}^\alpha(\mathcal{M})} \leq 2^{\frac{5+\lceil \alpha-2\rceil}{2}\lceil \alpha-2\rceil}(M/\eta)^{2^{\lceil \alpha-1\rceil}}(2B+1)$.

Proof of Lemma D.10. It is sufficiently to show $\|\log f\|_{\mathcal{H}^\alpha(U)} < \infty$ for any chart (U, ϕ) of $\mathcal{M}$. For simplicity, denote $f_\phi(\mathbf{z}) = f \circ \phi^{-1}(\mathbf{z})$ for any $\mathbf{z} \in \phi(U)$. We further denote $M > 0$ such that $\|f\|_{\mathcal{H}^\alpha(U)} \leq M$.

For any $|\mathbf{s}| = 1$, $\partial^{\mathbf{s}} \log f_\phi(\mathbf{z}) = \frac{\partial^{\mathbf{s}} f_\phi(\mathbf{z})}{f_\phi(\mathbf{z})}$. Note that $\partial^{\mathbf{s}} f_\phi(\mathbf{z}) \in \mathcal{H}^{\alpha-1}(\phi(U))$. According to Lemma D.9,

$$\frac{\partial^{\mathbf{s}} f_\phi}{f_\phi} \in \mathcal{H}^{\alpha-1}(\phi(U)) \tag{D.96}$$

for any $|\mathbf{s}| = 1$. Combining (Eq. D.96) and

$$\max_{|\mathbf{s}|=1, \mathbf{z}\in\phi(U)} \left|\frac{\partial^{\mathbf{s}} f_\phi(\mathbf{z})}{f_\phi(\mathbf{z})}\right| \leq M/\eta,$$

we have $\|\log f\|_{\mathcal{H}^\alpha(U)} < \infty$ with $\|\log f\|_{\mathcal{H}^\alpha(U)} \leq 2^{\frac{5+\lceil \alpha-2\rceil}{2}\lceil \alpha-2\rceil}(M/\eta)^{2^{\lceil \alpha-1\rceil}}(2B+1)$ which proves Lemma D.10.

□

The following two lemmas are extensively used in the previous proofs.

Lemma D.11 (McDiarmid's inequality ([230]))**.** *Let* $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ *be independent random variables and* $f : \mathcal{X}^n \to \mathbb{R}$ *be a map. If for any* i *and* $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_i' \in \mathcal{X}$, *the following holds*

$$|f(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n) - f(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_i', \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n)| \leq c_i,$$

then for any $t > 0$,

$$\mathbb{P}(|f(\mathbf{x}_1, \ldots, \mathbf{x}_n) - \mathbb{E}[f] \geq t|) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Lemma D.12 (Massart's lemma ([231, Lemma 5.2]))**.** *Let $\mathcal{X}$ be some finite set in $\mathbb{R}^m$ and $\epsilon_1, \ldots, \epsilon_m$ be independent Rademacher random variables. Then*

$$\mathbb{E}\left[\sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{m}\sum_{i=1}^{m} \epsilon_i x_i\right] \leq \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\| \frac{\sqrt{2\log|\mathcal{X}|}}{m}.$$

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[3] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015.

[4] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2013, pp. 6645–6649.

[5] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[6] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent trends in deep learning based natural language processing,” *IEEE Press Ser. Comput. Intell.*, vol. 13, no. 3, pp. 55–75, 2018.

[7] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: Review, opportunities and challenges,” *Brief Bioinform.*, vol. 19, no. 6, pp. 1236–1246, 2017.

[8] F. Jiang *et al.*, “Artificial intelligence in healthcare: Past, present and future,” *Stroke Vasc. Neurol.*, vol. 2, no. 4, pp. 230–243, 2017.

[9] S. Gu, E. Holly, T. Lillicrap, and S. Levine, “Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates,” in *2017 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2017, pp. 3389–3396.

[10] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[11] D. Amodei *et al.*, “Deep speech 2: End-to-end speech recognition in english and mandarin,” in *International conference on machine learning*, PMLR, 2016, pp. 173–182.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[13] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[14] A. B. Tsybakov, *Introduction to nonparametric estimation*, 1st. Springer Publishing Company, Incorporated, 2008.

[15] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[16] S. Osher, Z. Shi, and W. Zhu, “Low dimensional manifold model for image processing,” *SIAM J. Imaging Sci.*, vol. 10, no. 4, pp. 1669–1690, 2017.

[17] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, “Hate speech detection with comment embeddings,” in *Proceedings of the 24th international conference on world wide web*, ACM, 2015, pp. 29–30.

[18] J. B. Tenenbaum, V. De Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[19] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[20] R. R. Coifman *et al.*, “Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps,” *Proc. Natl. Acad. Sci.*, vol. 102, no. 21, pp. 7426–7431, 2005.

[21] W. K. Allard, G. Chen, and M. Maggioni, “Multi-scale geometric methods for data sets ii: Geometric multi-resolution analysis,” *Appl. Comput. Harmon. Anal.*, vol. 32, no. 3, pp. 435–462, 2012.

[22] L. Tu, *An introduction to manifolds*, ser. Universitext. Springer New York, 2010, ISBN: 9781441973993.

[23] J. M. Lee, *Riemannian manifolds: an introduction to curvature*. Springer Science & Business Media, 2006, vol. 176.

[24] ——, *Introduction to Riemannian manifolds*. Springer, 2018.

[25] H. Federer, “Curvature measures,” *Trans. Amer. Math. Soc.*, vol. 93, no. 3, pp. 418–491, 1959.

[26] E. Aamari, J. Kim, F. Chazal, B. Michel, A. Rinaldo, and L. Wasserman, "Estimating the reach of a manifold," *Electron. J. Stat.*, vol. 13, no. 1, pp. 1359–1399, 2019.

[27] L. Slobodeckij, "Generalized sobolev spaces and their applications to boundary value problems of partial differential equations, leningrad," *Gos. Ped. Inst. Ucep. Zap*, vol. 197, pp. 54–112, 1958.

[28] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*. Springer Science & Business Media, 1993, vol. 303.

[29] H. Triebel, *Theory of function spaces II*, ser. Monographs in Mathematics. Birkhäuser Basel, 1992.

[30] B. Irie and S. Miyake, "Capabilities of three-layered perceptrons," in *IEEE International Conference on Neural Networks*, vol. 1, 1988, p. 218.

[31] K.-I. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural networks*, vol. 2, no. 3, pp. 183–192, 1989.

[32] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control Signals Systems*, vol. 2, no. 4, pp. 303–314, 1989.

[33] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.

[34] C. K. Chui and X. Li, "Approximation by ridge functions and neural networks with one hidden layer," *J. Approx. Theory*, vol. 70, no. 2, pp. 131–141, 1992.

[35] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," *Neural Networks*, vol. 6, no. 6, pp. 861–867, 1993.

[36] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[37] H. N. Mhaskar, "Neural networks for optimal approximation of smooth and analytic functions," *Neural Comput.*, vol. 8, no. 1, pp. 164–177, 1996.

[38] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," in *Advances in Neural Information Processing Systems*, 2017, pp. 6231–6239.

[39] B. Hanin, "Universal function approximation by deep neural nets with bounded width and relu activations," *arXiv preprint arXiv:1708.02691*, 2017.

[40] D. Yarotsky, “Error bounds for approximations with deep relu networks,” *Neural Networks*, vol. 94, pp. 103–114, 2017.

[41] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[42] P. J. Bickel, B. Li, *et al.*, “Local polynomial regression on unknown manifolds,” in *Complex datasets and inverse problems*, Institute of Mathematical Statistics, 2007, pp. 177–186.

[43] R. R. Coifman and M. Maggioni, “Diffusion wavelets,” *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 53–94, 2006.

[44] C. K. Chui and H. N. Mhaskar, “Deep nets for local manifold learning,” *arXiv preprint arXiv:1607.07110*, 2016.

[45] U. Shaham, A. Cloninger, and R. R. Coifman, “Provable approximation properties for deep neural networks,” *Appl. Comput. Harmon. Anal.*, vol. 44, no. 3, pp. 537–557, 2018.

[46] R. A. DeVore, R. Howard, and C. Micchelli, “Optimal nonlinear approximation,” *Manuscripta Math.*, vol. 63, no. 4, pp. 469–478, 1989.

[47] P. Niyogi, S. Smale, and S. Weinberger, “Finding the homology of submanifolds with high confidence from random samples,” *Discrete Comput. Geom.*, vol. 39, no. 1-3, pp. 419–441, 2008.

[48] J. H. Conway, N. J. A. Sloane, and E. Bannai, *Sphere-packings, lattices, and groups*. Berlin, Heidelberg: Springer-Verlag, 1987, ISBN: 0-387-96617-X.

[49] H. Liu, M. Chen, T. Zhao, and W. Liao, “Besov function approximation and binary classification on low-dimensional manifolds using convolutional residual networks,” in *International Conference on Machine Learning*, PMLR, 2021, pp. 6770–6780.

[50] T. 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*, 2019.

[51] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[52] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[53] T. B. Brown *et al.*, “Language models are few-shot learners,” *arXiv preprint arXiv:2005.14165*, 2020.

[54] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017.

[55] S. Bubeck and M. Sellke, “A universal law of robustness via isoperimetry,” *arXiv preprint arXiv:2105.12806*, 2021.

[56] K. Hornik, M. Stinchcombe, and H. White, “Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks,” *Neural networks*, vol. 3, no. 5, pp. 551–560, 1990.

[57] P. Cardaliaguet and G. Euvrard, “Approximation of a function and its derivative with a neural network,” *Neural networks*, vol. 5, no. 2, pp. 207–220, 1992.

[58] I. Gühring, G. Kutyniok, and P. Petersen, “Error bounds for approximations with deep relu neural networks in w s, p norms,” *Analysis and Applications*, vol. 18, no. 05, pp. 803–859, 2020.

[59] S. Hon and H. Yang, “Simultaneous neural network approximations in sobolev spaces,” *arXiv preprint arXiv:2109.00161*, 2021.

[60] S. Gu and L. Rigazio, “Towards deep neural network architectures robust to adversarial examples,” *arXiv preprint arXiv:1412.5068*, 2014.

[61] M. Hein and M. Andriushchenko, “Formal guarantees on the robustness of a classifier against adversarial manipulation,” *arXiv preprint arXiv:1705.08475*, 2017.

[62] T.-W. Weng *et al.*, “Evaluating the robustness of neural networks: An extreme value theory approach,” *arXiv preprint arXiv:1801.10578*, 2018.

[63] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, “Virtual adversarial training: A regularization method for supervised and semi-supervised learning,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[64] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[65] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, “Theoretically principled trade-off between robustness and accuracy,” in *International Conference on Machine Learning*, 2019.

[66] S. Zagoruyko and N. Komodakis, “Wide residual networks,” *arXiv preprint arXiv:1605.07146*, 2016.

[67] B. Wu, J. Chen, D. Cai, X. He, and Q. Gu, “Do wider neural networks really help adversarial robustness?” *arXiv preprint arXiv:2010.01279*, 2020.

[68] H. Liu, M. Chen, S. Er, W. Liao, T. Zhang, and T. Zhao, “Benefits of overparameterized convolutional residual networks: Function approximation under smoothness constraint,” *arXiv preprint arXiv:2206.04569*, 2022.

[69] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[70] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 315–323.

[71] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, “Visualizing the loss landscape of neural nets,” in *Advances in Neural Information Processing Systems*, 2018, pp. 6389–6399.

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

[73] S. Arora, N. Cohen, and E. Hazan, “On the optimization of deep networks: Implicit acceleration by overparameterization,” *arXiv preprint arXiv:1802.06509*, 2018.

[74] L. Wasserman, *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[75] G. Wahba, *Spline models for observational data*. Siam, 1990, vol. 59.

[76] N. S. Altman, “An introduction to kernel and nearest-neighbor nonparametric regression,” *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, 1992.

[77] J. Fan and I. Gijbels, *Local polynomial modelling and its applications*, ser. Monographs on statistics and applied probability series. Chapman & Hall, 1996.

[78] A. R. Barron, “Complexity regularization with application to artificial neural networks,” in *Nonparametric functional estimation and related topics*, Springer, 1991, pp. 561–576.

[79] D. F. McCaffrey and A. R. Gallant, “Convergence rates for single hidden layer feedforward networks,” *Neural Networks*, vol. 7, no. 1, pp. 147–158, 1994.

[80] M. Hamers and M. Kohler, “Nonasymptotic bounds on the L_2 error of neural network regression estimates,” *Ann. Inst. Statist. Math.*, vol. 58, no. 1, pp. 131–151, 2006.

[81] M. Kohler and A. Krzyżak, “Adaptive regression estimation with multilayer feedforward neural networks,” *J. Nonparametr. Stat.*, vol. 17, no. 8, pp. 891–913, 2005.

[82] ——, “Nonparametric regression based on hierarchical interaction models,” *IEEE Trans. Inform. Theory*, vol. 63, no. 3, pp. 1620–1630, 2016.

[83] M. Kohler and J. Mehnert, “Analysis of the rate of convergence of least squares neural network regression estimates in case of measurement errors,” *Neural Networks*, vol. 24, no. 3, pp. 273–279, 2011.

[84] J. Schmidt-Hieber, “Nonparametric regression using deep neural networks with relu activation function,” *arXiv preprint arXiv:1708.06633*, 2017.

[85] Y. Kim, I. Ohn, and D. Kim, “Fast convergence rates of deep neural networks for classification,” *arXiv preprint arXiv:1812.03599*, 2018.

[86] I. Ohn and Y. Kim, “Smooth function approximation by deep neural networks with general activation functions,” *Entropy*, vol. 21, no. 7, p. 627, 2019.

[87] P. J. Bickel and B. Li, “Local polynomial regression on unknown manifolds,” *Lecture Notes-Monograph Series*, vol. 54, pp. 177–186, 2007.

[88] M.-Y. Cheng and H.-t. Wu, “Local linear regression on manifolds and its geometric interpretation,” *J. Amer. Statist. Assoc.*, vol. 108, no. 504, pp. 1421–1434, 2013.

[89] W. Liao, M. Maggioni, and S. Vigogna, “Multiscale regression on unknown manifolds,” *arXiv preprint arXiv:2101.05119*, 2021.

[90] S. Kpotufe, “k-NN regression adapts to local intrinsic dimension,” in *Advances in Neural Information Processing Systems*, 2011, pp. 729–737.

[91] S. Kpotufe and V. K. Garg, “Adaptivity to local smoothness and dimension in kernel regression,” in *Advances in Neural Information Processing Systems*, 2013, pp. 3075–3083.

[92] Y. Yang, S. T. Tokdar, *et al.*, “Minimax-optimal nonparametric regression in high dimensions,” *Ann. Statist.*, vol. 43, no. 2, pp. 652–674, 2015.

[93] J. Schmidt-Hieber, “Deep relu network approximation of functions on a manifold,” *arXiv preprint arXiv:1908.00695*, 2019.

[94] R. Nakada and M. Imaizumi, “Adaptive approximation and generalization of deep neural network with intrinsic dimensionality,” *J. Mach. Learn. Res.*, vol. 21, no. 174, pp. 1–38, 2020.

[95] A. Cloninger and T. Klock, “Relu nets adapt to intrinsic dimensionality beyond the target domain,” *arXiv preprint arXiv:2008.02545*, 2020.

[96] A. van der Vaart and J. Wellner, *Weak convergence and empirical processes: with applications to statistics*, ser. Springer Series in Statistics. Springer, 1996.

[97] T. Hu, Z. Shang, and G. Cheng, “Sharp rate of convergence for deep neural network classifiers under the teacher-student setting,” *arXiv preprint arXiv:2001.06892*, 2020.

[98] M. Kohler, A. Krzyzak, and B. Walter, “On the rate of convergence of image classifiers based on convolutional neural networks,” *arXiv preprint arXiv:2003.01526*, 2020.

[99] M. Kohler and S. Langer, “Statistical theory for image classification using deep convolutional neural networks with cross-entropy loss,” *arXiv preprint arXiv:2011.13602*, 2020.

[100] A. Nitanda and T. Suzuki, “Functional gradient boosting based on residual network perception,” in *International Conference on Machine Learning*, 2018, pp. 3819–3828.

[101] F. Huang, J. Ash, J. Langford, and R. Schapire, “Learning deep resnet blocks sequentially using boosting theory,” in *International Conference on Machine Learning*, 2018, pp. 2058–2067.

[102] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” *arXiv preprint arXiv:1605.05396*, 2016.

[103] C. Ledig *et al.*, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[104] K. Schawinski, C. Zhang, H. Zhang, L. Fowler, and G. K. Santhanam, “Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit,” *Monthly Notices of the Royal Astronomical Society: Letters*, vol. 467, no. 1, pp. L110–L114, 2017.

[105] A. Brock, J. Donahue, and K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” *arXiv preprint arXiv:1809.11096*, 2018.

[106] V. Volz, J. Schrum, J. Liu, S. M. Lucas, A. Smith, and S. Risi, “Evolving mario levels in the latent space of a deep convolutional generative adversarial network,” in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2018, pp. 221–228.

[107] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.

[108] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” in *Advances in neural information processing systems*, 2016, pp. 2234–2242.

[109] A. Müller, “Integral probability metrics and their generating classes of functions,” *Adv. Appl. Prob.*, vol. 29, no. 2, pp. 429–443, 1997.

[110] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang, “Generalization and equilibrium in generative adversarial nets (gans),” *arXiv preprint arXiv:1703.00573*, 2017.

[111] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *International conference on machine learning*, 2010, pp. 807–814.

[112] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier nonlinearities improve neural network acoustic models,” in *Proc. icml*, vol. 30, 2013, p. 3.

[113] Y. Bai, T. Ma, and A. Risteski, “Approximability of discriminators implies diversity in gans,” *arXiv preprint arXiv:1806.10586*, 2018.

[114] T. Liang, “On how well generative adversarial networks learn densities: Nonparametric and parametric results,” *arXiv preprint arXiv:1811.03179*, 2018.

[115] N. Schreuder, V.-E. Brunel, and A. Dalalyan, “Statistical guarantees for generative models without domination,” in *Algorithmic Learning Theory*, PMLR, 2021, pp. 1051–1071.

[116] A. Block, Z. Jia, Y. Polyanskiy, and A. Rakhlin, “Intrinsic dimension estimation,” *arXiv preprint arXiv:2106.04018*, 2021.

[117] C. Villani, *Optimal transport: old and new*. New York, NY, USA: Springer Science & Business Media, 2008, vol. 338.

[118] F. Santambrogio, “Models and applications of optimal transport in economics, traffic and urban planning,” *arXiv preprint arXiv:1009.3857*, 2010.

[119] A. Galichon, *A survey of some recent applications of optimal transport methods to econometrics*, 2017.

[120] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

[121] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 9, pp. 1853–1865, 2016.

[122] G. Monge, *Mémoire sur le calcul intégral des équations aux différences partielles*. Paris, France: Imprimerie royale, 1784.

[123] L. A. Caffarelli, "The regularity of mappings with a convex potential," *Journal of the American Mathematical Society*, vol. 5, no. 1, pp. 99–104, 1992.

[124] ——, "Boundary regularity of maps with convex potentials," *Communications on pure and applied mathematics*, vol. 45, no. 9, pp. 1141–1151, 1992.

[125] ——, "Boundary regularity of maps with convex potentials–ii," *Annals of mathematics*, pp. 453–496, 1996.

[126] J. I. Urbas, "Regularity of generalized solutions of monge-ampere equations," *Mathematische Zeitschrift*, vol. 197, no. 3, pp. 365–393, 1988.

[127] J. Urbas, "On the second boundary value problem for equations of monge-ampere type," *Journal fur die Reine und Angewandte Mathematik*, vol. 487, pp. 115–124, 1997.

[128] J. Moser, "On the volume elements on a manifold," *Transactions of the American Mathematical Society*, vol. 120, no. 2, pp. 286–294, 1965.

[129] R. Tang and Y. Yang, "Minimax rate of distribution estimation on unknown submanifold under adversarial losses," *arXiv preprint arXiv:2202.09030*, 2022.

[130] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein, "The intrinsic dimension of images and its impact on learning," *arXiv preprint arXiv:2104.08894*, 2021.

[131] A. Virmaux and K. Scaman, "Lipschitz regularity of deep neural networks: Analysis and efficient estimation," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[132] P. Pauli, A. Koch, J. Berberich, P. Kohler, and F. Allgöwer, "Training robust neural networks using lipschitz bounds," *IEEE Control Systems Letters*, vol. 6, pp. 121–126, 2021.

[133] H. Gouk, E. Frank, B. Pfahringer, and M. J. Cree, "Regularisation of neural networks by enforcing lipschitz continuity," *Machine Learning*, vol. 110, no. 2, pp. 393–416, 2021.

[134] J. Wang, R. Gao, and Y. Xie, "Two-sample test using projected wasserstein distance: Breaking the curse of dimensionality," *arXiv preprint arXiv:2010.11970*, 2020.

[135] J. Wang, M. Chen, T. Zhao, W. Liao, and Y. Xie, "A manifold two-sample test study: Integral probability metric with neural networks," *arXiv preprint arXiv:2205.02043*, 2022.

[136] S. Park, C. Yun, J. Lee, and J. Shin, "Minimum width for universal approximation," *arXiv preprint arXiv:2006.08859*, 2020.

[137] Y. Lu and J. Lu, "A universal approximation theorem of deep neural networks for expressing probability distributions," *Advances in neural information processing systems*, vol. 33, pp. 3094–3105, 2020.

[138] J. Huang, Y. Jiao, Z. Li, S. Liu, Y. Wang, and Y. Yang, "An error analysis of generative adversarial networks for learning distributions," *Journal of Machine Learning Research*, vol. 23, no. 116, pp. 1–43, 2022.

[139] P. Zhang, Q. Liu, D. Zhou, T. Xu, and X. He, "On the discrimination-generalization tradeoff in gans," *arXiv preprint arXiv:1711.02771*, 2017.

[140] H. Jiang, Z. Chen, M. Chen, F. Liu, D. Wang, and T. Zhao, "On computation and generalization of gans with spectrum control," *arXiv preprint arXiv:1812.10912*, 2018.

[141] T. Liang, "How well can generative adversarial networks learn densities: A nonparametric view," *arXiv preprint arXiv:1712.08244*, 2017.

[142] G. Luise, M. Pontil, and C. Ciliberto, "Generalization properties of optimal transport gans with latent distribution learning," *arXiv preprint arXiv:2007.14641*, 2020.

[143] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 6240–6249.

[144] M. Chae, D. Kim, Y. Kim, and L. Lin, “A likelihood approach to nonparametric estimation of a singular distribution using deep generative models,” *arXiv preprint arXiv:2105.04046*, 2021.

[145] S. Singh, A. Uppal, B. Li, C.-L. Li, M. Zaheer, and B. Póczos, “Nonparametric density estimation under adversarial losses,” in *Advances in Neural Information Processing Systems*, 2018, pp. 10 225–10 236.

[146] A. Uppal, S. Singh, and B. Póczos, “Nonparametric density estimation & convergence of gans under besov ipm losses,” *arXiv preprint arXiv:1902.03511*, 2019.

[147] R. Nickl and B. M. Pötscher, “Bracketing metric entropy rates and empirical central limit theorems for function classes of besov-and sobolev-type,” *Journal of Theoretical Probability*, vol. 20, no. 2, pp. 177–199, 2007.

[148] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*, Springer, 2014, pp. 818–833.

[149] B. Zhou, D. Bau, A. Oliva, and A. Torralba, “Interpreting deep visual representations via network dissection,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2131–2145, 2018.

[150] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, “Mode regularized generative adversarial networks,” *arXiv preprint arXiv:1612.02136*, 2016.

[151] Y. Cao, G. W. Ding, K. Y.-C. Lui, and R. Huang, “Improving gan training via binarized representation entropy (bre) regularization,” *arXiv preprint arXiv:1805.03644*, 2018.

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

[153] S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai, “Gradient descent finds global minima of deep neural networks,” *arXiv preprint arXiv:1811.03804*, 2018.

[154] Z. Allen-Zhu, Y. Li, and Z. Song, “A convergence theory for deep learning via over-parameterization,” *arXiv preprint arXiv:1811.03962*, 2018.

[155] S. S. Du, X. Zhai, B. Poczos, and A. Singh, “Gradient descent provably optimizes over-parameterized neural networks,” *arXiv preprint arXiv:1810.02054*, 2018.

[156] Y. Li and Y. Liang, “Learning overparameterized neural networks via stochastic gradient descent on structured data,” in *Advances in Neural Information Processing Systems*, 2018, pp. 8157–8166.

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

[158] Z. Allen-Zhu, Y. Li, and Y. Liang, “Learning and generalization in overparameterized neural networks, going beyond two layers,” in *Advances in neural information processing systems*, 2019, pp. 6155–6166.

[159] S. S. Du and W. Hu, “Width provably matters in optimization for deep linear neural networks,” *arXiv preprint arXiv:1901.08572*, 2019.

[160] E. S. Kim *et al.*, “The battle trial: Personalizing therapy for lung cancer,” *Cancer discovery*, vol. 1, no. 1, pp. 44–53, 2011.

[161] J. K. Lunceford and M. Davidian, “Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study,” *Statistics in medicine*, vol. 23, no. 19, pp. 2937–2960, 2004.

[162] V. F. Farias and A. A. Li, “Learning preferences with side information,” *Management Science*, vol. 65, no. 7, pp. 3131–3149, 2019.

[163] A. Sharma, J. M. Hofman, and D. J. Watts, “Estimating the causal impact of recommendation systems from observational data,” in *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 2015, pp. 453–470.

[164] J. J. Heckman and E. J. Vytlacil, “Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation,” *Handbook of econometrics*, vol. 6, pp. 4779–4874, 2007.

[165] D. B. Rubin, “Estimating causal effects of treatments in randomized and nonrandomized studies.,” *Journal of Educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[166] J. J. Heckman, “Sample selection bias as a specification error (with an application to the estimation of labor supply functions),” National Bureau of Economic Research, Tech. Rep., 1977.

[167] W. Cao, A. A. Tsiatis, and M. Davidian, “Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data,” *Biometrika*, vol. 96, no. 3, pp. 723–734, 2009.

[168] J. M. Robins, A. Rotnitzky, and L. P. Zhao, “Estimation of regression coefficients when some regressors are not always observed,” *Journal of the American statistical Association*, vol. 89, no. 427, pp. 846–866, 1994.

[169] T. Kitagawa and A. Tetenov, “Who should be treated? Empirical welfare maximization methods for treatment choice,” *Econometrica*, vol. 86, no. 2, pp. 591–616, 2018.

[170] J. L. Hill, “Bayesian nonparametric modeling for causal inference,” *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.

[171] Y.-Q. Zhao, D. Zeng, E. B. Laber, R. Song, M. Yuan, and M. R. Kosorok, “Doubly robust learning for estimating individualized treatment with censored data,” *Biometrika*, vol. 102, no. 1, pp. 151–168, 2015.

[172] E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small, “Non-parametric methods for doubly robust estimation of continuous treatment effects,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 4, pp. 1229–1245, 2017.

[173] A. Richardson, M. G. Hudgens, P. B. Gilbert, and J. P. Fine, “Nonparametric bounds and sensitivity analysis of treatment effects,” *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, vol. 29, no. 4, p. 596, 2014.

[174] K. C. G. Chan, S. C. P. Yam, and Z. Zhang, “Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting,” *Journal of the Royal Statistical Society. Series B, Statistical methodology*, vol. 78, no. 3, p. 673, 2016.

[175] M. Frölich, M. Huber, and M. Wiesenfarth, “The finite sample performance of semi-and non-parametric estimators for treatment effects and policy evaluation,” *Computational Statistics & Data Analysis*, vol. 115, pp. 91–102, 2017.

[176] E. H. Kennedy, “Optimal doubly robust estimation of heterogeneous causal effects,” *arXiv preprint arXiv:2004.14497*, 2020.

[177] Y. Lee, E. Kennedy, and N. Mitra, “Doubly robust nonparametric instrumental variable estimators for survival outcomes,” *arXiv preprint arXiv:2007.12973*, 2020.

[178] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik, “Nonparametric tests for treatment effect heterogeneity,” *The Review of Economics and Statistics*, vol. 90, no. 3, pp. 389–405, 2008.

[179] D. Benkeser, M. Carone, M. V. D. Laan, and P. Gilbert, “Doubly robust nonparametric inference on the average treatment effect,” *Biometrika*, vol. 104, no. 4, pp. 863–880, 2017.

[180] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Scholkopf, and L. Bottou, “Discovering causal signals in images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6979–6987.

[181] T. T. Pham and Y. Shen, “A deep causal inference approach to measuring the effects of forming group loans in online non-profit microfinance platform,” *arXiv preprint arXiv:1706.02795*, 2017.

[182] K. Chalupka, P. Perona, and F. Eberhardt, “Visual causal feature learning,” *arXiv preprint arXiv:1412.2309*, 2014.

[183] W. van Amsterdam, J. Verhoeff, P. de Jong, T. Leiner, and M. Eijkemans, “Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning,” *npj Digital Medicine*, vol. 2, no. 1, pp. 1–6, 2019.

[184] F. Johansson, U. Shalit, and D. Sontag, “Learning representations for counterfactual inference,” in *International Conference on Machine Learning*, 2016, pp. 3020–3029.

[185] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy, “Deep IV: A flexible approach for counterfactual prediction,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017, pp. 1414–1423.

[186] B. Lim, “Forecasting treatment responses over time using recurrent marginal structural networks,” in *Advances in Neural Information Processing Systems*, 2018, pp. 7483–7493.

[187] Z. Gao and Y. Han, “Minimax optimal nonparametric estimation of heterogeneous treatment effects,” *arXiv preprint arXiv:2002.06471*, 2020.

[188] A. Beygelzimer and J. Langford, “The offset tree for learning with partial labels,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 129–138.

[189] D. G. Horvitz and D. J. Thompson, “A generalization of sampling without replacement from a finite universe,” *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 663–685, 1952.

[190] C. M. Cassel, C. E. Särndal, and J. H. Wretman, “Some results on generalized difference estimation and generalized regression estimation for finite populations,” *Biometrika*, vol. 63, no. 3, pp. 615–620, 1976.

[191] M. Dudík, J. Langford, and L. Li, “Doubly robust policy evaluation and learning,” *arXiv preprint arXiv:1103.4601*, 2011.

[192] S. Athey and S. Wager, “Efficient policy learning,” *arXiv preprint arXiv:1702.02896*, 2017.

[193] Z. Zhou, S. Athey, and S. Wager, “Offline multi-action policy learning: Generalization and optimization,” *arXiv preprint arXiv:1810.04778*, 2018.

[194] N. Kallus, “Balanced policy evaluation and learning,” in *Advances in Neural Information Processing Systems*, 2018, pp. 8895–8906.

[195] B. Zhang, A. A. Tsiatis, M. Davidian, M. Zhang, and E. Laber, “Estimating optimal treatment regimes from a classification perspective,” *Stat*, vol. 1, no. 1, pp. 103–114, 2012.

[196] Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok, “Estimating individualized treatment rules using outcome weighted learning,” *Journal of the American Statistical Association*, vol. 107, no. 499, pp. 1106–1118, 2012.

[197] N. Kallus, “More efficient policy learning via optimal retargeting,” *Journal of the American Statistical Association*, pp. 1–13, 2020.

[198] A. Ward, Z. Zhou, N. Bambos, E. Wang, and D. Scheinker, “Anesthesiologist surgery assignments using policy learning,” in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*, IEEE, 2019, pp. 1–6.

[199] A. Bennett and N. Kallus, “Efficient policy learning from surrogate-loss classification reductions,” *arXiv preprint arXiv:2002.05153*, 2020.

[200] N. Kallus and A. Zhou, “Policy evaluation and optimization with continuous treatments,” in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 1243–1251.

[201] M. Demirer, V. Syrgkanis, G. Lewis, and V. Chernozhukov, “Semi-parametric efficient policy learning with continuous actions,” *arXiv preprint arXiv:1905.10116*, 2019.

[202] N. Kallus and M. Santacatterina, “Kernel optimal orthogonality weighting: A balancing approach to estimating effects of continuous treatments,” *arXiv preprint arXiv:1910.11972*, 2019.

[203] K. Colangelo and Y.-Y. Lee, “Double debiased machine learning nonparametric inference with continuous treatments,” *arXiv preprint arXiv:2004.03036*, 2020.

[204] D. J. Foster and V. Syrgkanis, “Orthogonal statistical learning,” *arXiv preprint arXiv:1901.09036*, 2019.

[205] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media, 2013.

[206] M. H. Farrell, T. Liang, and S. Misra, “Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands,” *arXiv preprint arXiv:1809.09953*, 2018.

[207] M. Chen, H. Jiang, W. Liao, and T. Zhao, “Nonparametric regression on low-dimensional manifolds using deep relu networks: Function approximation and statistical recovery,” *arXiv preprint arXiv:1908.01842*, 2019.

[208] T. Salimans and D. P. Kingma, “Weight normalization: A simple reparameterization to accelerate training of deep neural networks,” *Advances in neural information processing systems*, vol. 29, 2016.

[209] T. Van Laarhoven, “L2 regularization versus batch and weight normalization,” *arXiv preprint arXiv:1706.05350*, 2017.

[210] A. B. Tsybakov *et al.*, “Optimal aggregation of classifiers in statistical learning,” *The Annals of Statistics*, vol. 32, no. 1, pp. 135–166, 2004.

[211] Y. Wang and A. Singh, “Noise-adaptive margin-based active learning and lower bounds under tsybakov noise condition,” in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[212] D. E. Koulouriotis and A. Xanthopoulos, “Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems,” *Applied Mathematics and Computation*, vol. 196, no. 2, pp. 913–922, 2008.

[213] V. Kuleshov and D. Precup, “Algorithms for multi-armed bandit problems,” *arXiv preprint arXiv:1402.6028*, 2014.

[214] M. Chen, H. Jiang, W. Liao, and T. Zhao, “Efficient approximation of deep relu networks for functions on low dimensional manifolds,” in *Advances in Neural Information Processing Systems*, 2019, pp. 8172–8182.

[215] V. Chernozhukov, D. Chetverikov, and K. Kato, “Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors,” *The Annals of Statistics*, vol. 41, no. 6, pp. 2786–2819, 2013.

[216] L. Wasserman, “Stein’s method and the bootstrap in low and high dimensions: A tutorial,” 2014.

[217] M. Chen, L. Yang, M. Wang, and T. Zhao, “Dimensionality reduction for stationary time series via stochastic nonconvex optimization,” *Advances in Neural Information Processing Systems*, 2018.

[218] T. Liu, M. Chen, M. Zhou, S. S. Du, E. Zhou, and T. Zhao, “Towards understanding the importance of shortcut connections in residual networks,” *Advances in Neural Information Processing Systems*, 2019.

[219] Y. Wang, M. Chen, T. Zhao, and M. Tao, “Large learning rate tames homogeneity: Convergence and balancing effect,” *arXiv preprint arXiv:2110.03677*, 2021.

[220] M. Chen, Y. Li, E. Wang, Z. Yang, Z. Wang, and T. Zhao, “Pessimism meets invariance: Provably efficient offline mean-field multi-agent RL,” *Advances in Neural Information Processing Systems*, 2021.

[221] R. M. Dudley, “The sizes of compact subsets of Hilbert space and continuity of Gaussian processes,” *Journal of Functional Analysis*, vol. 1, no. 3, pp. 290–330, 1967.

[222] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. Cambridge, MA, USA: MIT press, 2018.

[223] H. Liu, H. Yang, M. Chen, T. Zhao, and W. Liao, “Deep nonparametric estimation of operators between infinite dimensional spaces,” *arXiv preprint arXiv:2201.00217*, 2022.

[224] J. Lu, Z. Shen, H. Yang, and S. Zhang, “Deep network approximation for smooth functions,” *SIAM Journal on Mathematical Analysis*, vol. 53, no. 5, pp. 5465–5506, 2021.

[225] Z. Shen, H. Yang, and S. Zhang, “Deep network approximation characterized by number of neurons,” *arXiv preprint arXiv:1906.05497*, 2019.

[226] M. Anthony and P. Bartlett, *Neural network learning: Theoretical foundations*. cambridge university press, 1999.

[227] M. J. Wainwright, *High-dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge University Press, 2019, vol. 48.

[228] W. Liao and M. Maggioni, “Adaptive geometric multiscale approximations for intrinsically low-dimensional data,” *Journal of Machine Learning Research*, vol. 20, no. 98, pp. 1–63, 2019.

[229] A. Maurer, “A vector-contraction inequality for rademacher complexities,” in *International Conference on Algorithmic Learning Theory*, Springer, 2016, pp. 3–17.

[230] C. McDiarmid, “On the method of bounded differences,” *Surveys in Combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.

[231] P. Massart, “Some applications of concentration inequalities to statistics,” in *Annales de la Faculté des sciences de Toulouse: Mathématiques*, vol. 9, 2000, pp. 245–303.

VITA

Minshuo Chen obtained his Bechelor's degree in 2015 from Zhejiang University, with honor from Chu Kochen Honor's College (advanced class of engineering education). Afterwards, he finished a Master's degree at UCLA in 2017 and joined Georgia Tech as a Ph.D. student with the Machine Learning program. In the past five years, he was working under the supervision of Dr. Tuo Zhao and Dr. Wenjing Liao.