Deep Nonparametric Estimation of Operators between Infinite Dimensional Spaces

Hao Liu

Department of Mathematics, Hong Kong Baptist University

Haizhao Yang*

Department of Mathematics, Purdue University
Minshuo Liu and Tuo Zhao

the H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology
Wenjing Liao*

Department of Mathematics, Georgia Institute of Technology

October 5, 2021

Abstract

Learning operators between infinitely dimensional spaces is an important learning task arising in wide applications in machine learning, imaging science, mathematical modeling and simulations, etc. This paper studies the nonparametric estimation of these operators using deep neural networks. Non-asymptotic upper bounds for the estimation error of the empirical risk minimizer are derived. Under the assumption that the target operator has low local dimensionality or is supported near a low-dimensional manifold, our error bounds achieve attractive rates in the number of training samples without the curse of dimensionality. Our assumptions cover most scenarios in real applications and our results are sharper than existing results in the literature. We further investigate the influence of network structures (e.g., network width, depth, and sparsity) on the generalization error of the neural regression estimator and propose a general suggestion on the choice of structures to maximize the learning efficiency quantitatively.

1 Introduction

Learning nonlinear operators from a Banach space to another via nonparametric estimation has been an important topic with broad applications. For example, in reduced order modeling, a data-driven approach desires to map a full model trajectory to a reduced model trajectory or vice versa

*co-corresponding author.

[50]. In solving parametric partial differential equations (PDEs), it is desired to learn a map from the parametric function space to the PDE solution space [31, 38, 44]. In forward and inverse scattering problems [32, 64], it is interesting to learn an operator mapping the observed data function space to the parametric function space that models the underlying PDE. In density functional theory, it is desired to learn an nonlinear operator mapping a potential function to a density function [20]. In phase retrieval [16], an operator from the observed data function space to the reconstructed image function space is learned. Other image processing problems, e.g., image super-resolution [51], image denoising [63], image inpainting [52], are similar to the deep learning-based phase retrieval, where an operator from a function space to another function space is learned.

As a powerful tool of nonparametric regression, deep learning [23] has made astonishing breakthroughs in various applications, including computer vision [35], natural language processing [24], speech recognition [27], healthcare [46], as well as nonlinear operator learning [31, 69, 20, 21, 32, 12, 44, 36, 7, 38, 49]. A most typical method in operator learning is to first approximate Banach spaces with finite dimensional vector spaces and apply deep neural networks to learn the map between these vector spaces [31, 69, 20, 21, 32]. Though empirical successes have been demonstrated in learning nonlinear operators in many applications following this approach, it is computationally expensive to train these algorithms and the training procedure has to be repeated when the size of vector spaces is changed. Another approach based on the neural network approximation theorem [12] can allieviate this issue to a certain extent by avoiding the discretization of the output Banach space of the operator. This apporach was first proposed in [12] with two-layer neural networks and recently revisited with deeper neural networks in [44] with successful applications [39, 8]. However, the methods in [12, 44, 39, 8] are still mesh-dependent due to the requirement of a fixed number of sample points for the input function of the operator. More recently, discretization-invariant (mesh-independent) operator learning for problems essentially enjoying sparsity structures was proposed in [1, 7, 38, 49] by taking the advantage of graph kernel networks, principle component analysis (PCA) of Banach spaces, and kernel integral operators, etc. After training, discretization-invariant approaches can be applied to different problem sizes without retraining and, hence, are efficient in the application.

Although operator learning via nonparametric regression based on deep learning has been successful in many applications, its theory is still in its infancy, especially when the operator is from an infinite dimensional Banach space to another. The successes of deep neural networks are largely due to their universal approximation power [15, 28] showing the existence of a neural network with a small size fulfilling the learning task. Quantitative approximation theories for function approximation, provably better than traditional tools, have been extensively studied with various network architectures and activation functions, e.g., for continuous functions [65, 55, 57, 58, 59, 67], for smooth functions [66, 68, 43], and for functions with integral representations [3, 18, 19, 61]. In theory, deep neural networks can approximation high-dimensional functions with a dimenson-independent approximation rate [3, 18, 19, 61, 57, 58, 68, 56]. However, in the context of operator approximation, deep learning theory is very limited. Probably the first result is the universal approximation theorem for operators in [12]. More recently, quantitative approximation results for operator approximation were proposed in [7, 36] based on the function approximation theory in

[65]. Note that the approximation results in [65] are not optimal and lack the flexibility to choose arbitrary width and depth of neural networks. Therefore, in this paper, we will develop new operator approximaiton theory based on nearly optimal function approximation results for arbitrary network width and depth. The flexibility of choosing arbitrary width and depth makes it possible to have an explicit guideline to balance the approximation error and statistical variance to achieve a better generalization error in operator learning.

We will also develop novel statistical theory for deep nonparametric regression of operators between infinite dimensional Banach spaces. The core questions to be answered are: how the generalization error scales when the number of samples increases and whether the scaling is dimension-independent without the curse of dimensionality. The statistical theory of estimating high-dimensional functions via neural networks has been a popular research topic recently [26, 33, 29, 5, 53, 9, 11, 34, 48, 22, 40, 30]. These works have proved that deep nonparametric regression can achieve the optimal minimax rate of nonparametric regression established in [62], achieving a theoretical guarantee to lessen the curse of dimensionality when the target function has low complexity or the function domain has low-dimensional structures. In more sophisticated cases when a mathematical modeling problem is transferred to a special regression problem, e.g., solving high-dimensional PDEs and identifying the governing equation of spatialtemporal data, the generalization analysis of deep learning has been proposed in [6, 60, 45, 47, 42, 41, 17, 25]. All these results focus on the regression problem when the target function is a mapping from a finite dimensional domain to a one-dimensional or finite dimensional domain, and, therefore, cannot be similarly generalized to mappings from an infinite dimensional domain to another. To the best of our knowledge, the only existing work on the generalization error analysis of operator learning for Banach spaces is [36] for the algorithm in [44], which is not completely discretization-invariant. Furthermore, the generalization error in [36] is a posteriori depending on the properties of neural networks fitting the target operator. In this paper, we will establish a priori generalization error of discretization-invariant operator learning algorithms for operators between Banach spaces. As we shall see later, operator learning from a finite dimensional vector space to another is also a special case of our analysis. Therefore, the theoretical result established in this paper can facilitate the understanding of most operator learning algorithms in the literature.

In this paper, our contributions are summarized as follows:

1. We derive an upper bound on the generalization error of a general framework of learning operators between infinite dimensional spaces by deep neural networks. The framework considered here first approximates the input and output space by finite dimensional spaces using some encoders and decoders. Then a transformation between the dimension reduced spaces are learned using deep neural networks. The upper bound using two network architectures are derived: one has constraints on the number of nonzero weight parameters and parameter magnitude. The other one does not have such constraints and one has flexibility to choose the depth and width. Our upper bound consists of two parts: the error from learning the transformation by deep neural networks, and the dimension reduction error from encoders and decoders. Our result holds for general encoders and decoders under mild assumptions.

2. Our analysis is general including a wide rage of popular choices of encoders and decoder in the

numerical implementation, such as those derived from Legendre polynomials, trigonometric bases and principal component analysis. The generalization errors of these examples will be specified.

3. We discuss two scenarios to mitigate the curse of dimensionality. The first scenario is when the image of the input space after encoding is on a low-dimensional manifold embedded in a high-dimensional space. We show that the convergence rate depends on the intrinsic dimension of the manifold, instead of the ambient dimension. The second scenario is when the operator itself has low complexities: the composition of the operator with certain encoder and decoder is a multi index model. We show that the convergence rate depends on the intrinsic dimension of the composed operator.

We will organize this paper as follows. In Section 2, we introduce our notations and the framework that is considered in this paper. Our main results with general encoders and decoders are presented in Section 3. We discuss the applications of our main results on encoders and decoders derived from function space basis and PCA in Section 4 and 5, respectively. To mitigate the curse of dimensionality, we discuss the application of our results on two scenarios with low-dimensional structures in Section 6. The proofs of all results are put in Section 7. We conclude this paper in Section 8.

2 A general framework

2.1 Preliminaries

We first briefly introduce some definitions and notations on a Hilbert space, encoders, decoders and feedforward neural networks used in this paper. A Hilbert space is a Banach space equipped with an inner product. It is separable if it admits a countable orthonormal basis. Let $\mathcal{H}$ be a separable Hilbert space. An encoder for $\mathcal{H}$ is an operator $E_{\mathcal{H}} : \mathcal{H} \to \mathbb{R}^d$ for some positive integer d. The associated decoder is an operator $D_{\mathcal{H}} : \mathbb{R}^d \to \mathcal{H}$. The composition $\Pi_{\mathcal{H}} = D_{\mathcal{H}} \circ E_{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$ is a projection. For any $u \in \mathcal{H}$, we define the projection error as$\|\Pi_{\mathcal{H}}(u) - u\|_{\mathcal{H}}$.

In this paper, we consider the ReLU Feedforward Neural Network (FNN) in the form of

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

where W_l's are weight matrices, $\boldsymbol{b}_l$'s are biases, and $\mathrm{ReLU}(a) = \max\{a, 0\}$ is the rectified linear unit activation (ReLU) applied element-wise.

We consider two classes of network architectures. The first class is defined as

$$\begin{aligned}\mathcal{F}_{\mathrm{NN}}(L,p,K,\kappa,M) = \{\Gamma =&[f_1, f_2, ..., f_{d_{\mathcal{Y}}}]^\top : \text{ for each } k = 1, ..., d_{\mathcal{Y}},\\ & f_k(\mathbf{x}) \text{ is in the form of (1) with } L \text{ layers, width bounded by } p,\\ & \|f_k\|_\infty \le M,\ \|W_l\|_{\infty,\infty} \le \kappa,\ \|\boldsymbol{b}_l\|_\infty \le \kappa,\ \sum_{l=1}^{L} \|W_l\|_0 + \|\boldsymbol{b}_l\|_0 \le K\},\end{aligned} \tag{2}$$

where $\|f\|_\infty = \sup_{\mathbf{x}} |f(\mathbf{x})|$, $\|W\|_{\infty,\infty} = \max_{i,j} |W_{i,j}|$, $\|\boldsymbol{b}\|_\infty = \max_i |b_i|$ for any function f, matrix W, and vector $\boldsymbol{b}$ with $\|\cdot\|_0$ denoting the number of nonzero elements of its argument. The function class given by this first network architecture has an upper bound on all weight parameters and a cardinality constraint: the magnitude of all weight parameters are upper bounded by κ, and the total number of nonzero parameters are no more than K.

In the second class of network architecture, we drop the magnitude and cardinality constraints for practical concerns. The second network architecture is parameterized by L, p, M only:

$$\begin{aligned}\mathcal{F}_{\rm NN}(L,p,M) = \{\Gamma = &[f_1, f_2, ..., f_{d_\mathcal{Y}}]^\top : \text{ for each } k = 1, ..., d_\mathcal{Y}, \\ & f_k(\mathbf{x}) \text{ is in the form of (1) with } L \text{ layers, width bounded by } p, \\ & \|f_k\|_\infty \leq M\}. \end{aligned} \tag{3}$$

All theoretical results in this paper can be applied for both network architectures.

2.2 Problem setup and a learning framework

Let $\mathcal{X}$ and $\mathcal{Y}$ be two separable Hilbert spaces and $\Psi : \mathcal{X} \to \mathcal{Y}$ be an unknown operator. Our goal is to approximate the operator Ψ from a finite number of samples $\mathcal{S} = \{u_i, v_i\}_{i=1}^{2n}$ in the following setting.

Setting 1. Let $\mathcal{X}, \mathcal{Y}$ be two separable Hilbert spaces and ρ be a probability measure on $\mathcal{X}$. Let $\mathcal{S} = \{u_i, v_i\}_{i=1}^{2n}$ be the given data where the u_i's are i.i.d. samples from ρ and the v_i's are generated according to model:

$$v_i = \Psi(u_i) + \widetilde{\epsilon}_i, \tag{4}$$

where the $\widetilde{\epsilon}_i$'s are i.i.d. samples from a probability measure on $\mathcal{X}$, independently of the u_i's.

The push forward measure of ρ under Ψ is denoted by $\Psi_\# \rho$, such that for any $\Omega \subset \mathcal{Y}$,

$$\Psi_\#\rho(\Omega) = \rho\left(\{u : \Psi(u) \in \Omega\}\right).$$

The object of interest in this learning task is an operator between infinite dimensional spaces. Without additional assumptions, the estimation error of Ψ based on a finite number of samples may not converge to zero due to the curse of dimensionality, especially when the dimension is infinitely large. In this paper, we exploit the low-dimensional structures of this problem arising from practical applications, and prove a nonparametric estimation error for deep neural networks. In particular, we consider three low-dimensional structures: (1) The measure ρ is concentrated on a low-dimensional linear set in $\mathcal{X}$ (Section XXX); (2) The measure ρ is concentrated on a low-dimensional nonlinear set in $\mathcal{X}$ (Section XXX); (3) The operator Ψ has a low complexity in the sense that it only depends on few parameters (Section XXX). Hao: put this paragraph somewhere else?

Our learning framework follows the idea of model reduction [Cite XXX]. It consists of encoding and decoding in both the $\mathcal{X}$ and $\mathcal{Y}$ spaces, and deep learning of a transformation between the encoded vectors for the elements in $\mathcal{X}$ and $\mathcal{Y}$. We first encode the elements in $\mathcal{X}$ and $\mathcal{Y}$ to finite dimensional vectors by an encoding operator. For fixed positive integers $d_\mathcal{X}$ and $d_\mathcal{Y}$, let

$E_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}^{d_{\mathcal{X}}}$ and $D_{\mathcal{X}} : \mathbb{R}^{d_{\mathcal{X}}} \to \mathcal{X}$ be the encoder and decoder of $\mathcal{X}$, and $E_{\mathcal{Y}} : \mathcal{Y} \to \mathbb{R}^{d_{\mathcal{Y}}}$ and $D_{\mathcal{Y}} : \mathbb{R}^{d_{\mathcal{Y}}} \to \mathcal{Y}$ be the encoder and decoder of $\mathcal{Y}$ such that

$$D_{\mathcal{X}} \circ E_{\mathcal{X}} \approx I \quad \text{and} \quad D_{\mathcal{Y}} \circ E_{\mathcal{Y}} \approx I.$$

The empirical counterparts of these encoders and decoders are denoted by $E_{\mathcal{X}}^n, D_{\mathcal{X}}^n, E_{\mathcal{Y}}^n, D_{\mathcal{Y}}^n$.

The simplest encoder in a function space is the discretization operator. When $\mathcal{X}$ is a function space containing functions defined on a compact subset of $\mathbb{R}^D$, we can discretize the domain with a fixed grid size, and take the encoder as the sampling operator on this regular grid. However, the discretization operator may not reveal the low-dimensional structures in the functions of interest, and therefore may not effectively reduce the dimension.

A popular choice of encoders in applications is the basis encoder, such as the Fourier transform with trigonometric basis, and PCA with data-driven basis, etc. Given an orthonormal basis of $\mathcal{X}$ and a positive integer $d_{\mathcal{X}}$, the basis encoder maps an element in $\mathcal{X}$ to $d_{\mathcal{X}}$ coefficients associated with a fixed set of $d_{\mathcal{X}}$ bases. For any coefficient vector $\mathbf{a} \in \mathbb{R}^{d_{\mathcal{X}}}$, the decoder $D_{\mathcal{X}}(\mathbf{a})$ gives rise to a linear combination of these $d_{\mathcal{X}}$ bases weighted by $\mathbf{a}$. See Section XXX for the details. The trigonometric basis and orthonormal polynomials are commonly used bases in applications. These bases are a priori given, independently of the training data. In this case, the basis operator can be viewed as a deterministic encoder, which are given independently of the training data. The empirical encoder and decoder are the same as the oracle encoder and decoder, such that $E_{\mathcal{X}}^n = E_{\mathcal{X}}$ and $D_{\mathcal{X}}^n = D_{\mathcal{X}}$.

PCA is an effective dimension reduction technique, when the u_i's exhibit a low-dimensional linear structure. The PCA encoder encodes an element in $\mathcal{X}$ to the $d_{\mathcal{X}}$ coefficients associated with the top $d_{\mathcal{X}}$ eigenbasis of a trace operator. The decoder gives a linear combination of the eigenbasis weighted by the given coefficient vector. In practice, one needs to estimate this trace operator from the training data and obtain an empirical estimation of $E_{\mathcal{X}}$ and $D_{\mathcal{X}}$, which are denoted by $E_{\mathcal{X}}^n$ and $D_{\mathcal{X}}^n$, respectively. The PCA encoder is data-driven, and we expect $E_{\mathcal{X}}^n \approx E_{\mathcal{X}}$, $D_{\mathcal{X}}^n \approx D_{\mathcal{X}}$ when the sample size n is sufficiently large. The encoding and decoding operator in $\mathcal{Y}$ can be defined analogously.

The operator $D_{\mathcal{X}} \circ E_{\mathcal{X}}$ is the projection operator associated with the encoder $E_{\mathcal{X}}$. We have the following projections and their empirical counterparts:

$$\Pi_{\mathcal{X},d_{\mathcal{X}}} = D_{\mathcal{X}} \circ E_{\mathcal{X}}, \quad \Pi_{\mathcal{X},d_{\mathcal{X}}}^n = D_{\mathcal{X}}^n \circ E_{\mathcal{X}}^n,$$
$$\Pi_{\mathcal{Y},d_{\mathcal{Y}}} = D_{\mathcal{Y}} \circ E_{\mathcal{Y}}, \quad \Pi_{\mathcal{Y},d_{\mathcal{Y}}}^n = D_{\mathcal{Y}}^n \circ E_{\mathcal{Y}}^n.$$

We summarize the notations in Table 1.

After the empirical encoders $E_{\mathcal{X}}^n, E_{\mathcal{Y}}^n$ and decoders $D_{\mathcal{X}}^n, D_{\mathcal{Y}}^n$ are computed, our objective is to learn a transformation $\Gamma : \mathbb{R}^{d_{\mathcal{X}}} \to \mathbb{R}^{d_{\mathcal{Y}}}$ such that

$$D_{\mathcal{Y}}^n \circ \Gamma \circ E_{\mathcal{X}}^n \approx \Psi. \tag{5}$$

We learn Γ using a two-stage algorithm. Given the training data $\mathcal{S} = \{u_i, v_i\}_{i=1}^{2n}$, we split the data into two subsets $\mathcal{S}_1 = \{u_i, v_i\}_{i=1}^{n}$ and $\mathcal{S}_2 = \{u_i, v_i\}_{i=n+1}^{2n}$ [1], where $\mathcal{S}_1$ is used to compute the encoders and decoders and $\mathcal{S}_2$ is used to learn the transformation Γ between the encoded vectors. Our two-stage algorithm follows

[1]The data can be split unevenly as well.

Notation	Description	Notation	Description
$\mathcal{X}$	Input space	$\mathcal{Y}$	Output space
$\Psi : \mathcal{X} \to \mathcal{Y}$	An unknown operator	$\mathcal{S} = \{u_i, v_i\}_{i=1}^{2n}$	Given data set
ρ	A probability measure on $\mathcal{X}$	$\Psi_{\#}\rho$	Push forward measure of ρ under Ψ
$E_{\mathcal{X}}, D_{\mathcal{X}}$	Encoder and decoder of $\mathcal{X}$	$E_{\mathcal{Y}}, D_{\mathcal{Y}}$	Encoder and decoder of $\mathcal{Y}$
$E_{\mathcal{X}}^n, D_{\mathcal{X}}^n$	Empirical estimations of $E_{\mathcal{X}}, D_{\mathcal{X}}$	$E_{\mathcal{Y}}^n, D_{\mathcal{Y}}^n$	Empirical estimations of $E_{\mathcal{Y}}, D_{\mathcal{Y}}$
$d_{\mathcal{X}}$	Encoding dimension of $\mathcal{X}$	$d_{\mathcal{Y}}$	Encoding dimension of $\mathcal{Y}$
$\Pi_{\mathcal{X},d_{\mathcal{X}}}$	Projection $D_{\mathcal{X}} \circ E_{\mathcal{X}}$	$\Pi_{\mathcal{Y},d_{\mathcal{Y}}}$	Projection $D_{\mathcal{Y}} \circ E_{\mathcal{Y}}$
$\Pi_{\mathcal{X},d_{\mathcal{X}}}^n$	Empirical projection $D_{\mathcal{X}}^n \circ E_{\mathcal{X}}^n$	$\Pi_{\mathcal{Y},d_{\mathcal{Y}}}^n$	Empirical projection $D_{\mathcal{Y}}^n \circ E_{\mathcal{Y}}^n$
$\\|\Pi_{\mathcal{X},d_{\mathcal{X}}}(u) - u\\|_{\mathcal{X}}$	Encoding error for u in $\mathcal{X}$	$\\|\Pi_{\mathcal{Y},d_{\mathcal{Y}}}(v) - v\\|_{\mathcal{Y}}$	Encoding error for v in $\mathcal{Y}$
$\mathcal{F}_{\mathrm{NN}}$	Neural network class	Γ_{NN}	Neural network estimator in (6)

Table 1: Notations used in this paper.

Stage 1: Compute the empirical encoders and decoders $E_{\mathcal{X}}^n, D_{\mathcal{X}}^n, E_{\mathcal{Y}}^n, D_{\mathcal{Y}}^n$ based on $\mathcal{S}_1$. In the case of deterministic encoders, we skip Stage 1 and let $E_{\mathcal{X}}^n = E_{\mathcal{X}},\ D_{\mathcal{X}}^n = D_{\mathcal{X}},\ E_{\mathcal{Y}}^n = E_{\mathcal{Y}},\ D_{\mathcal{Y}}^n = D_{\mathcal{Y}}$.

Stage 2: Learn Γ with $\mathcal{S}_2$ by solving the following optimization problem

$$\Gamma_{\mathrm{NN}} \in \underset{\Gamma \in \mathcal{F}_{\mathrm{NN}}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=n+1}^{2n} \|\Gamma \circ E_{\mathcal{X}}^n(u_i) - E_{\mathcal{Y}}^n(v_i)\|_2^2 \tag{6}$$

for some $\mathcal{F}_{\mathrm{NN}}$ class with a proper choice of parameters.

Our estimator of Ψ is given as

$$\Psi_{\mathrm{NN}} := D_{\mathcal{Y}}^n \circ \Gamma_{\mathrm{NN}} \circ E_{\mathcal{X}}^n,$$

and the mean squared error is defined as

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{u \sim \rho} \|\Psi_{\mathrm{NN}}(u) - \Psi(u)\|_{\mathcal{Y}}^2. \tag{7}$$

3 Main results

The main results of this paper provide nonparametric statistical guarantees on the mean squared generalization error for the estimation of Lipchitz operators.

3.1 Assumptions

We first make some assumptions on the measure ρ and the operator Ψ.

Assumption 1 (Compactly supported measure)**.** The probability distribution ρ is supported on a compact set $\Omega_{\mathcal{X}} \subset \mathcal{X}$. There exists $R_{\mathcal{X}} > 0$ such that, for any $u \in \Omega_{\mathcal{X}}$, we have

$$\|u\|_{\mathcal{X}} \le R_{\mathcal{X}}. \tag{8}$$

Assumption 2 (Lipschitz operator)**.** There exists $L_\Psi > 0$ such that

$$\|\Psi(u_1) - \Psi(u_2)\|_{\mathcal{Y}} \le L_\Psi \|u_1 - u_2\|_{\mathcal{X}}, \quad \text{for any } u_1, u_2 \in \Omega_{\mathcal{X}}.$$

Assumption 1 and 2 assume that ρ is compactly supported and Ψ is Lipschitz continuous. We denote the image of $\Omega_{\mathcal{X}}$ under the transformation Ψ as $\Omega_{\mathcal{Y}} = \{v \in \mathcal{Y} : v = \Psi(u) \text{ for some } u \in \Omega_{\mathcal{X}}\}$. Assumption 1 and 2 imply that $\Omega_{\mathcal{Y}}$ is bounded: there exists a constant $R_{\mathcal{Y}} > 0$ such that for any $v \in \Omega_{\mathcal{Y}}$, we have $\|v\|_{\mathcal{Y}} \leq R_{\mathcal{Y}}$.

We next make some natural assumptions on the empirical encoders and decoders:

Assumption 3 (Lipchitz encoders and decoders)**.** The empirical encoders and decoders $E^n_{\mathcal{X}}, D^n_{\mathcal{X}}, E^n_{\mathcal{Y}}, D^n_{\mathcal{Y}}$ satisfy:

$$E^n_{\mathcal{X}}(0) = \mathbf{0},\ D^n_{\mathcal{X}}(\mathbf{0}) = 0,\ E^n_{\mathcal{Y}}(0) = \mathbf{0},\ D^n_{\mathcal{Y}}(\mathbf{0}) = 0.$$

They are also Lipschitz: there exist $L_{E^n_{\mathcal{X}}}, L_{D^n_{\mathcal{X}}}, L_{E^n_{\mathcal{Y}}}, L_{D^n_{\mathcal{Y}}} > 0$ such that, for any $u_1, u_2 \in \mathcal{X}$ and any $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^{d_{\mathcal{X}}}$, we have

$$\|E^n_{\mathcal{X}}(u_1) - E^n_{\mathcal{X}}(u_2)\|_2 \leq L_{E^n_{\mathcal{X}}}\|u_1 - u_2\|_{\mathcal{X}}, \quad \|D^n_{\mathcal{X}}(\mathbf{a}_1) - D^n_{\mathcal{X}}(\mathbf{a}_2)\|_{\mathcal{X}} \leq L_{D^n_{\mathcal{X}}}\|\mathbf{a}_1 - \mathbf{a}_2\|_2,$$

and for any $v_1, v_2 \in \mathcal{Y}$ and any $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^{d_{\mathcal{Y}}}$, we have

$$\|E^n_{\mathcal{Y}}(v_1) - E^n_{\mathcal{Y}}(v_2)\|_2 \leq L_{E^n_{\mathcal{Y}}}\|v_1 - v_2\|_{\mathcal{Y}}, \quad \|D^n_{\mathcal{Y}}(\mathbf{a}_1) - D^n_{\mathcal{Y}}(\mathbf{a}_2)\|_{\mathcal{Y}} \leq L_{D^n_{\mathcal{Y}}}\|\mathbf{a}_1 - \mathbf{a}_2\|_2.$$

Assumption 3 implies that $E^n_{\mathcal{X}}(u)$ and $E^n_{\mathcal{Y}}(v)$ are bounded for any $u \in \Omega_{\mathcal{X}}$ and $v \in \Omega_{\mathcal{Y}}$. For any $u \in \Omega_{\mathcal{X}}$, we have $\|E^n_{\mathcal{X}}(u)\|_2 \leq \|E^n_{\mathcal{X}}(u) - E^n_{\mathcal{X}}(0)\|_2 + \|E^n_{\mathcal{X}}(0)\|_2 \leq L_{E^n_{\mathcal{X}}} R_{\mathcal{X}}$. Similarly, for any $v \in \Omega_{\mathcal{Y}}$, we have $\|E^n_{\mathcal{Y}}(v)\|_2 \leq L_{E^n_{\mathcal{Y}}} R_{\mathcal{Y}}$.

Assumption 4 (Noise)**.** The random noise $\widetilde{\epsilon}$ satisfies

(i) $\widetilde{\epsilon}$ is independent of u.

(ii) $\mathbb{E}[\widetilde{\epsilon}] = 0$.

(iii) There exists $\widetilde{\sigma} > 0$ such that $\|\widetilde{\epsilon}\|_{\mathcal{Y}} \leq \widetilde{\sigma}$.

(iv) For any given $\mathcal{S}_1$, the conditional expectation satisfies

$$\mathbb{E}_{\widetilde{\epsilon}}\left[E^n_{\mathcal{Y}}(\Psi(u) + \widetilde{\epsilon}) - E^n_{\mathcal{Y}}(\Psi(u))|\mathcal{S}_1\right] = \mathbf{0}, \text{ for any } u \in \Omega_{\mathcal{X}},$$

where $E^n_{\mathcal{Y}}$ is the empirical encoder computed with $\mathcal{S}_1$.

Assumption 4(i)-(iv) are natural assumptions on noise. Assumption 4(i) is about the independent of the input and the noise, which is commonly used in nonparametric regression [XXXXX]. Assumption 4(ii)-(iii) together with Assumption 3 imply that the encoded vectors of noise are bounded: $\|E^n_{\mathcal{Y}}(\widetilde{\epsilon})\|_\infty \leq L_{E^n_{\mathcal{Y}}}\widetilde{\sigma}$. We denote $\sigma = L_{E^n_{\mathcal{Y}}}\widetilde{\sigma}$ such that $\|E^n_{\mathcal{Y}}(\widetilde{\epsilon})\|_\infty \leq \sigma$. Assumption 4(iv) requires that, if we condition on $\mathcal{S}_1$ based on which we compute the empirical encoder $E^n_{\mathcal{Y}}$, the perturbation on the encoded vector resulted from noise has a zero expectation. Assumption 4(iv) is guaranteed for all linear encoders as long as Assumption 4(ii) holds:

$$\mathbb{E}_{\widetilde{\epsilon}}\left[E^n_{\mathcal{Y}}(\Psi(u) + \widetilde{\epsilon}) - E^n_{\mathcal{Y}}(\Psi(u))|\mathcal{S}_1\right] = \mathbb{E}_{\widetilde{\epsilon}}\left[E^n_{\mathcal{Y}}(\widetilde{\epsilon})|\mathcal{S}_1\right] = \mathbf{0}.$$

Basis encoders, including the PCA encoder, are linear encoders, so they all satisfy Assumption 4(iv).

Assumption 3 and 4 are assumptions on empirical encoders and decoders, in which the subscript of the Lipschitz constants has a superscript n. When the oracle encoders and decoders, such as $E_{\mathcal{X}}$ and $D_{\mathcal{X}}$, are given, we set $E_{\mathcal{X}}^n = E_{\mathcal{X}}$ and $D_{\mathcal{X}}^n = D_{\mathcal{X}}$, and denote the Lipschitz constants by $L_{E_{\mathcal{X}}}$ and $L_{D_{\mathcal{X}}}$. The same notations are used when $E_{\mathcal{Y}}$ and $D_{\mathcal{Y}}$ are given.

3.2 Generalization error for general encoders and decoders

Our main result is an upper bound of the generalization error in (7) for general encoders and decoders. Our results can be applied to the two network architectures defined in (2) and (3). Our first theorem gives an upper bound of the generalization error with the network architecture defined in (2).

Theorem 1. In Setting 1, suppose Assumption 1 – 4 hold. Let Γ_{NN} be the minimizer of (6) with the network architecture $\mathcal{F}(L,p,K,\kappa,M)$ in (2), where

$$\begin{aligned} &L = O(\log n + \log d_{\mathcal{Y}}),\ p = O\left(d_{\mathcal{Y}}^{-\frac{d_{\mathcal{X}}}{2+d_{\mathcal{X}}}} n^{\frac{d_{\mathcal{X}}}{2+d_{\mathcal{X}}}}\right),\ K = O\left(d_{\mathcal{Y}}^{-\frac{d_{\mathcal{X}}}{2+d_{\mathcal{X}}}} n^{\frac{d_{\mathcal{X}}}{2+d_{\mathcal{X}}}} \log n\right), \\ &M = \sqrt{d_{\mathcal{Y}}} L_{E_{\mathcal{Y}}^n} R_{\mathcal{Y}},\ \kappa = \max\left\{1, \sqrt{d_{\mathcal{Y}}} L_{E_{\mathcal{Y}}^n} R_{\mathcal{Y}}, \sqrt{d_{\mathcal{X}}} L_{E_{\mathcal{X}}^n} R_{\mathcal{X}}, L_{E_{\mathcal{Y}}^n} L_{D_{\mathcal{X}}^n} L_{\Psi}\right\}. \end{aligned} \tag{9}$$

Then we have

$$\begin{aligned} &\mathbb{E}_{\mathcal{S}} \mathbb{E}_{u\sim\rho} \|D_{\mathcal{Y}}^n \circ \Gamma_{\mathrm{NN}} \circ E_{\mathcal{X}}^n(u) - \Psi(u)\|_{\mathcal{Y}}^2 \\ \le\ & C_1(\sigma^2 + R_{\mathcal{Y}}^2) d_{\mathcal{Y}}^{\frac{4+d_{\mathcal{X}}}{2+d_{\mathcal{X}}}} n^{-\frac{2}{2+d_{\mathcal{X}}}} \log^3 n + C_2(\sigma^2 + R_{\mathcal{Y}}^2) d_{\mathcal{Y}}^2 (\log d_{\mathcal{Y}}) n^{-1} \\ &+ C_3 \mathbb{E}_{\mathcal{S}_1} \mathbb{E}_{u\sim\rho} \|\Pi_{\mathcal{X},d_{\mathcal{X}}}^n(u) - u\|_2^2 + 2\mathbb{E}_{\mathcal{S}_1} \mathbb{E}_{v^*\sim\Psi_{\#}\rho} \|\Pi_{\mathcal{Y},d_{\mathcal{Y}}}^n(v^*) - v^*\|_{\mathcal{Y}}^2, \end{aligned} \tag{10}$$

where C_1, C_2 are constants depending on $d_{\mathcal{X}}, R_{\mathcal{X}}, R_{\mathcal{Y}}, L_{E_{\mathcal{X}}^n}, L_{E_{\mathcal{Y}}^n}, L_{D_{\mathcal{X}}^n}, L_{D_{\mathcal{Y}}^n}, L_{\Psi}$ and $C_3 = 16 L_{D_{\mathcal{Y}}^n}^2 L_{E_{\mathcal{Y}}^n}^2 L_{\Psi}^2$.

Our second theorem gives an upper bound of the generalization error with the network architecture defined in (3).

Theorem 2. In Setting 1, suppose Assumption 1 – 4 hold. Let Γ_{NN} be the minimizer of (6) with the network architecture $\mathcal{F}(L,p,M)$ defined in (3) with

$$L = O(\widetilde{L}\log \widetilde{L}),\ p = O\left(\widetilde{p}\log \widetilde{p}\right), M = \sqrt{d_{\mathcal{Y}}} L_{E_{\mathcal{Y}}^n} R_{\mathcal{Y}}, \tag{11}$$

where $\widetilde{L}, \widetilde{p} > 0$ are positive integers satisfying

$$\widetilde{L}\widetilde{p} = \left\lceil d_{\mathcal{Y}}^{-\frac{d_{\mathcal{X}}}{4+2d_{\mathcal{X}}}} n^{\frac{d_{\mathcal{X}}}{4+2d_{\mathcal{X}}}} \right\rceil. \tag{12}$$

Then we have

$$\begin{aligned} &\mathbb{E}_{\mathcal{S}} \mathbb{E}_{u\sim\rho} \|D_{\mathcal{Y}}^n \circ \Gamma_{\mathrm{NN}} \circ E_{\mathcal{X}}^n(u) - \Psi(u)\|_{\mathcal{Y}}^2 \\ \le\ & C_4(\sigma^2 + R_{\mathcal{Y}}^2) d_{\mathcal{Y}}^{\frac{4+d_{\mathcal{X}}}{2+d_{\mathcal{X}}}} n^{-\frac{2}{2+d_{\mathcal{X}}}} \log^6 n \\ &+ C_3 \mathbb{E}_{\mathcal{S}_1} \mathbb{E}_{u\sim\rho} \|\Pi_{\mathcal{X},d_{\mathcal{X}}}^n(u) - u\|_{\mathcal{X}}^2 + 2\mathbb{E}_{\mathcal{S}_1} \mathbb{E}_{v^*\sim\Psi_{\#}\rho} \|\Pi_{\mathcal{Y},d_{\mathcal{Y}}}^n(v^*) - v^*\|_{\mathcal{Y}}^2. \end{aligned} \tag{13}$$

where C_4 is a constant depending on $d_{\mathcal{X}}, R_{\mathcal{X}}, R_{\mathcal{Y}}, L_{E_{\mathcal{X}}^n}, L_{E_{\mathcal{Y}}^n}, L_{D_{\mathcal{X}}^n}, L_{\Psi}$ and $C_3 = 16 L_{D_{\mathcal{Y}}^n}^2 L_{E_{\mathcal{Y}}^n}^2 L_{\Psi}^2$.

Theorem 1 is proved in Section 7.1 and Theorem 2 is proved in Section 7.2. For both network architectures, the upper bound in (10) and (13) consists of a network estimation error and the projection errors in the $\mathcal{X}$ and $\mathcal{Y}$ space.

- The first two terms in (10) and the first term in (13) represent the network estimation error for the transformation $\Gamma : \mathbb{R}^{d_\mathcal{X}} \to \mathbb{R}^{d_\mathcal{Y}}$ which maps the encoded vector $E^n_\mathcal{X}(u)$ for u in $\mathcal{X}$ to the encoded vector $E^n_\mathcal{Y}(\Phi(u))$ for $\Phi(u)$ in $\mathcal{Y}$. This error decays exponentially as the sample size n increases with an exponent depending on the dimension $d_\mathcal{X}$ of the encoding space. The dimension $d_\mathcal{X}$ appears in the exponent and $d_\mathcal{Y}$ appears as a constant factor. This is because that the transformation Γ has $d_\mathcal{Y}$ outputs and each output is a function from $\mathbb{R}^\mathcal{X}$ to $\mathbb{R}$. Therefore the rate is only cursed by the input dimension $d_\mathcal{X}$.

- The last two terms in (10) and (13) are projection errors in the $\mathcal{X}$ and $\mathcal{Y}$ space, respectively. If the measure ρ is concentrated near a $d_\mathcal{X}$-dimensional subset in $\mathcal{X}$, both projection errors can be made small if the encoder and decoder are properly chosen as the projection onto this $d_\mathcal{X}$-dimensional subspace.

We next compare the difference between the network architectures in Theorem 1 and Theorem 2. Denote the network architecture in Theorem 1 and Theorem 2 by $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. The architecture $\mathcal{F}_1$ has the depth and width scaling properly with respect to each other, and an upper bound on all weight parameters and a cardinality constraint. The cardinality constraint is nonconvex and therefore not practical for solving this optimization problem (6). The architecture $\mathcal{F}_2$ has more flexibility in the choice of depth and width as long as (12) is satisfied. The cardinality is removed for practical concerns. When we set $\widetilde{L} = O(\log n), \widetilde{p} = O(n^{\frac{d_\mathcal{X}}{4+2d_\mathcal{X}}} \log^{-1} n)$ in $\mathcal{F}_2$, both networks have a depth of $O(\log n)$, while the width of $\mathcal{F}_1$ is the square of that of $\mathcal{F}_2$, i.e., $\mathcal{F}_1$ is wider than $\mathcal{F}_2$. The comparison between $\mathcal{F}_1$ and $\mathcal{F}_2$ is summarized in Table 2.

	$\mathcal{F}_1$	$\mathcal{F}_2$
General comparison		
Network architecture with a given n	Fixed L and p	One has the flexibility to choose L and p as long as (12) is satisfied
No constraints on cardinality	✗	✓
No constraints on the magnitude of weight parameters	✗	✓
Set $\widetilde{L} = O(\log n), \widetilde{p} = O(n^{\frac{d_\mathcal{X}}{4+2d_\mathcal{X}}} \log^{-1} n)$ in $\mathcal{F}_2$		
L	$O(\log n)$	$O(\log n)$
p	$O\left(d_\mathcal{Y}^{-\frac{d_\mathcal{X}}{2+d_\mathcal{X}}} n^{\frac{d_\mathcal{X}}{2+d_\mathcal{X}}}\right)$	$O\left(d_\mathcal{Y}^{-\frac{d_\mathcal{X}}{4+2d_\mathcal{X}}} n^{\frac{d_\mathcal{X}}{4+2d_\mathcal{X}}}\right)$

Table 2: Comparison of the network architectures in Theorem 1 and 2.

In the rest of this paper, we focus on the network architecture in Theorem 2 without specifications and discuss its applications in various scenarios. Theorem 1 can also be applied in each case with a similar upper bound.

4 Generalization error for basis encoders and decoders

In this section, we discuss the applications of Theorem 2 when the encoder is chosen to be the deterministic basis encoder with a given orthonormal basis of the Hilbert space. Popular choices of orthonormal bases include orthogonal polynomials (e.g., Legendre polynomials [13, 14]) and trigonometric functions [10, 37]. We will assume $\mathcal{X}$ and $\mathcal{Y}$ are subsets of Hölder spaces, implying that ρ and $\Psi_{\#}\rho$ concentrate on the sets spanned by low-degree polynomial (or low-frequency trigonometric) basis functions.

4.1 Basis encoders and the generalization error

Let $\mathcal{H}$ be a separable Hilbert space equipped with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, and $\{\phi_k\}_{k=1}^{\infty}$ be an orthonormal basis of $\mathcal{H}$ such that $\langle\phi_{k_1},\phi_{k_2}\rangle_{\mathcal{H}}=0$ whenever $k_1\neq k_2$ and $\|\phi_k\|_{\mathcal{H}}=1$ for any k. For any $u\in\mathcal{H}$, we have

$$u=\sum_{k=1}^{\infty}\langle u,\phi_k\rangle_{\mathcal{H}}\phi_k. \tag{14}$$

For a fixed positive integer d representing the encoding dimension, we define the encoder of $\mathcal{H}$ as

$$E_{\mathcal{H},d}(u)=\left[\langle u,\phi_1\rangle_{\mathcal{H}},...,\langle u,\phi_d\rangle_{\mathcal{H}}\right]^{\top}\in\mathbb{R}^d,\quad\text{for any } u\in\mathcal{H}, \tag{15}$$

which gives rise to the coefficients associated with a fixed set of d basis functions in the decomposition (14). The decoder $D_{\mathcal{H},d}$ is defined as

$$D_{\mathcal{H},d}(\mathbf{a})=\sum_{k=1}^{d}a_k\phi_k,\ \text{for any } \mathbf{a}\in\mathbb{R}^d. \tag{16}$$

The basis encoder and decoder naturally satisfy the Lipchitz property with Lipschitz constant 1 (see a proof in Section 7.3).

Lemma 1. The encoder $E_{\mathcal{H},d}$ and decoder $D_{\mathcal{H},d}$ defined in (15) and (16) satisfy

$$\|E_{\mathcal{H},d}(u)-E_{\mathcal{H},d}(\widetilde{u})\|_2\leq\|u-\widetilde{u}\|_{\mathcal{H}}, \tag{17}$$

$$\|D_{\mathcal{H},d}(\mathbf{a})-D_{\mathcal{H},d}(\widetilde{\mathbf{a}})\|_{\mathcal{H}}=\|\mathbf{a}-\widetilde{\mathbf{a}}\|_2, \tag{18}$$

for any $u,\widetilde{u}\in\mathcal{H}$ and $\mathbf{a},\widetilde{\mathbf{a}}\in\mathbb{R}^d$.

Remark 1. All encoders in the form of (15) are linear operators and therefore satisfy Assumption 4(iv) as long as Assumption 4(ii) holds.

We next consider the generalization error for basis encoders and decoders with the network architecture in Theorem 2. Substituting the Lipschitz constants of all encoders and decoders by 1 in Theorem 2, we obtain the network parameters

$$L=O(\widetilde{L}\log\widetilde{L}),\ p=O\left(\widetilde{p}\log\widetilde{p}\right),M=\sqrt{d_{\mathcal{Y}}}R_{\mathcal{Y}} \tag{19}$$

and $\widetilde{L},\widetilde{p}>0$ are integers and satisfy (12). The generalization error is given in (13).

Popular choices of orthonormal bases are orthogonal polynomials and trigonometric functions. We next study the generalization error when Legendre polynomials or trigonometric functions are used for encoding. In the rest of this section, we assume $\mathcal{X} = \mathcal{Y} \subset L^2(\Omega)$, where Ω is a compact subset of $\mathbb{R}^D$ and the inner product is

$$\langle u_1, u_2 \rangle = \int_\Omega u_1(\mathbf{x}) u_2(\mathbf{x}) d\mathbf{x}.$$

4.2 Legendre polynomials.

On the interval $[-1, 1]$, one-dimensional Legendre polynomials $\{\widetilde{P}_k\}_{k=0}^\infty$ are defined recursively as

$$\begin{cases} \widetilde{P}_0(x) = 1, \\ \widetilde{P}_1(x) = x, \\ \widetilde{P}_{k+1}(x) = \frac{1}{k+1}\left[(2k+1)x\widetilde{P}_k(x) - k\widetilde{P}_{k-1}(x)\right]. \end{cases}$$

The Legendre polynomials satisfy

$$\int_{-1}^1 \widetilde{P}_k(x)\widetilde{P}_l(x)dx = \frac{2}{2k+1}\delta_{kl},$$

where δ_{kl} is the Kronecker delta which equals to 1 if $k = l$ and equals to 0 otherwise. We define the normalized Legendre polynomials as

$$P_k(x) = \sqrt{\frac{2k+1}{2}}\widetilde{P}_k(x). \tag{20}$$

In the Hilbert space $L^2([-1, 1]^D)$ where we set $\Omega = [-1, 1]^D$, the D-variate normalized Legendre polynomials are defined as

$$\phi_{L,\mathbf{k}} = \prod_{j=1}^D P_{k_j}(x_j), \tag{21}$$

where $\mathbf{k} = (k_1, ..., k_D)^\top$. The orthonormal basis of Legendre polynomials in $L^2([-1, 1]^D)$ is $\{\phi_{L,\mathbf{k}}\}_{\mathbf{k}\in\mathbb{N}_0^D}$.

The encoder with Legendre polynomials can be naturally defined as the expansion coefficients associated with low-order polynomials. Specifically, we fix $d^* > 0$ and define the encoder and decoder according to (15) and (16) using low-order polynomials in the basis

$$\Phi_L^r := \{\phi_{L,\mathbf{k}} : \|\mathbf{k}\|_\infty \leq d^*\}.$$

Before we state our assumption on $\mathcal{X}$ and $\mathcal{Y}$, we first define the Hölder space:

Definition 1 (Hölder space)**.** Let $s > 0$. A function $f : [-1, 1]^D \to \mathbb{R}$ belongs to Hölder space $\mathcal{H}^s([-1, 1]^D)$ if

$$\|f\|_{\mathcal{H}^s} := \max_{|\mathbf{k}|<\lceil s-1\rceil} \sup_{\mathbf{x}\in[-1,1]^D} |\partial^{\mathbf{k}} f(\mathbf{x})| + \max_{|\mathbf{k}|=\lceil s-1\rceil} \sup_{\mathbf{x}_1\neq\mathbf{x}_2\in[-1,1]^D} \frac{|\partial^{\mathbf{k}} f(\mathbf{x}_1) - \partial^{\mathbf{k}} f(\mathbf{x}_2)|}{\|\mathbf{x}_1 - \mathbf{x}_2\|_2^{s-\lceil s-1\rceil}} < \infty.$$

We make the following assumption on $\mathcal{X}$ and $\mathcal{Y}$:

Assumption 5. Assume $\mathcal{X} = \mathcal{Y} = \mathcal{H}^s([-1,1]^D)$. Assume there exist $C_{\mathcal{H},\mathcal{X}}$ and $C_{\mathcal{H},\mathcal{Y}}$ such that for any $u \in \Omega_{\mathcal{X}}$ and $v \in \Omega_{\mathcal{Y}}$

$$\|u\|_{\mathcal{H}^s} < C_{\mathcal{H},\mathcal{X}}, \quad \|v\|_{\mathcal{H}^s} < C_{\mathcal{H},\mathcal{Y}}.$$

Corollary 1. Consider Setting 1. Let $r_{\mathcal{X}}$ and $r_{\mathcal{Y}}$ be two positive integers. Suppose Assumption 1–5 hold. Assume the encoders and decoders are chosen as in (15) and (16) with basis functions $\Phi_L^{r_{\mathcal{X}}}$ and $\Phi_L^{r_{\mathcal{Y}}}$ for $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\Gamma_{\rm NN}$ be the minimizer of (6) with the network architecture $\mathcal{F}(L,p,M)$ where L,p,M are set as in (19). We have

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{u\sim\rho}\|D_{\mathcal{Y}}^n \circ \Gamma_{\rm NN} \circ E_{\mathcal{X}}^n(u) - \Psi(u)\|_{\mathcal{Y}}^2 \le C_4(\sigma^2 + R_{\mathcal{Y}}^2) d_{\mathcal{Y}}^{\frac{4+d_{\mathcal{X}}}{2+d_{\mathcal{X}}}} n^{-\frac{2}{2+d_{\mathcal{X}}}} \log^6 n + C_5 L_{\Psi}^2 d_{\mathcal{X}}^{-\frac{2s}{D}} + C_6 d_{\mathcal{Y}}^{-\frac{2s}{D}}.$$

where $d_{\mathcal{X}} = r_{\mathcal{X}}^D, d_{\mathcal{Y}} = r_{\mathcal{Y}}^D$, C_4 is a constant depending on $d_{\mathcal{X}}, R_{\mathcal{X}}, R_{\mathcal{Y}}, L_{\Psi}$, and C_5, C_6 are constant depending on $D, C_{\mathcal{H},\mathcal{X}}, C_{\mathcal{H},\mathcal{Y}}, L_{\Psi}$.

Corollary 1 is proved in Section 7.4.

4.3 Trigonometric functions.

Trigonometric functions and the Fourier transform have been widely used in various application where the computation is converted from the spacial domain to the frequency domain. Let $\{T_k(x)\}_{k=1}^{\infty}$ be one-dimensional trigonometric functions defined on $[-1,1]$ such that

$$\begin{cases} T_1 = 1/2, \\ T_{2k} = \sin(k\pi x) \quad \text{for } k > 1, \\ T_{2k+1} = \cos(k\pi x) \quad \text{for } k > 1. \end{cases} \tag{22}$$

In the Hilbert space $L^2([-1,1]^D)$, the trigonometric basis is given as $\{\phi_{T,\mathbf{k}}\}_{\mathbf{k}\in\mathbb{N}^D}$ with

$$\phi_{T,\mathbf{k}}(\mathbf{x}) = \prod_{j=1}^{D} T_{k_j}(x_j). \tag{23}$$

Fix $r > 0$. We can compute the encoders and decoders according to (15) and (16) using low-frequency elements in the basis

$$\Phi_T^r = \{\phi_{T,\mathbf{k}} : \|\mathbf{k}\|_\infty \le r\}. \tag{24}$$

Note that Φ^r has r^D basis functions. Denote the set of periodic functions on $[-1,1]^D$ with period 2 by $\mathcal{P}$.

We make the following assumption on $\mathcal{X}$ and $\mathcal{Y}$:

Assumption 6. Assume $\mathcal{X} = \mathcal{Y} = \mathcal{P} \cap \mathcal{H}^s([-1,1]^D)$. Assume there exist $C_{\mathcal{H}_P,\mathcal{X}}$ and $C_{\mathcal{H}_P,\mathcal{Y}}$ such that for any $u \in \Omega_{\mathcal{X}}$ and $v \in \Omega_{\mathcal{Y}}$

$$\|u\|_{\mathcal{H}^s} < C_{\mathcal{H}_P,\mathcal{X}}, \quad \|v\|_{\mathcal{H}^s} < C_{\mathcal{H}_P,\mathcal{Y}}.$$

Corollary 2. Consider Setting 1. Let $r_{\mathcal{X}}$ and $r_{\mathcal{Y}}$ be two positive odd integers. Suppose Assumption 1–4 and 6 hold. Assume the encoders and decoders are chosen as in (15) and (16) with basis functions $\Phi_T^{r_{\mathcal{X}}}$ and $\Phi_T^{r_{\mathcal{Y}}}$ for $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\Gamma_{\rm NN}$ be the minimizer of (6) with the network architecture $\mathcal{F}(L,p,M)$ where L,p,M are set as in (19). We have

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{u\sim\rho}\|D_{\mathcal{Y}}^n\circ\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u)-\Psi(u)\|_{\mathcal{Y}}^2\leq C_4(\sigma^2+R_{\mathcal{Y}}^2)d_{\mathcal{Y}}^{\frac{4+d_{\mathcal{X}}}{2+d_{\mathcal{X}}}}n^{-\frac{2}{2+d_{\mathcal{X}}}}\log^6 n+C_7L_{\Psi}^2d_{\mathcal{X}}^{-\frac{2s}{D}}+C_8d_{\mathcal{Y}}^{-\frac{2s}{D}}.$$

where $d_{\mathcal{X}}=r_{\mathcal{X}}^D, d_{\mathcal{Y}}=r_{\mathcal{Y}}^D$, C_4 is a constant depending on $d_{\mathcal{X}},R_{\mathcal{X}},R_{\mathcal{Y}},L_{\Psi}$ and C_7,C_8 are constants depending on $D,C_{\mathcal{H}_P,\mathcal{X}},C_{\mathcal{H}_P,\mathcal{Y}},L_{\Psi}$.

Corollary 2 is proved in Section 7.5.

5 Generalization error for PCA encoders and decoders

The basis encoders and decoders in Section 4 are a priori given, independently of the training data. When the given data are near a low-dimensional subspace, PCA is an effective tool for dimension reduction. In this section, we consider the PCA encoder, where the orthonormal basis is estimated from the training data.

5.1 PCA encoders and decoders

Let γ be a probability measure on a separable Hilbert space $\mathcal{H}$. Define the covariance operator as

$$C_{\mathcal{H}}=\mathbb{E}_{u\sim\gamma}u\otimes u \tag{25}$$

where $\otimes$ denotes the outer product $(f\otimes g)(h)=\langle g,h\rangle_{\mathcal{H}}f$ for any $f,g,h\in\mathcal{H}$, where $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ denotes the inner product in $\mathcal{H}$. Let $\{\lambda_k\}_{k=1}^{\infty}$ be the eigenvalues of $C_{\mathcal{H}}$ in a non-increasing order, and ϕ_k be the eigenfunction associated with λ_k. For any $u\in\mathcal{H}$, we have

$$u=\sum_{j=1}^{\infty}\langle u,\phi_j\rangle_{\mathcal{H}}\phi_j.$$

For a fixed positive integer d, we define the encoder operator $E_{\mathcal{H},d}:\mathcal{H}\to\mathbb{R}^d$ as

$$E_{\mathcal{H},d}(u)=[\langle u,\phi_1\rangle,\langle u,\phi_2\rangle,...,\langle u,\phi_d\rangle]^\top,\ \text{for any } u\in\mathcal{H}$$

which gives rise to the coefficients of u associated with the first d principal components, i.e., the eigenfunctions corresponding to the d largest eigenvalues. Denote the linear subspace spanned by the first d eigenfunctions of $C_{\mathcal{H}}$ by $V_{\mathcal{H},d}$. The decoder $D_{\mathcal{H},d}:\mathbb{R}^d\to V_{\mathcal{H},d}$ is defined as

$$D_{\mathcal{H},d}(\mathbf{a})=\sum_{j=1}^{d}a_j\phi_j,\ \text{for any } \mathbf{a}=[a_1,...,a_d]^\top\in\mathbb{R}^d.$$

Given n i.i.d samples $\{u_i\}_{i=1}^n$ from γ, the empirical covariance operator is

$$C_{\mathcal{H}}^n=\frac{1}{n}\sum_{i=1}^{n}u_i\otimes u_i. \tag{26}$$

Let $\{\lambda_k^n\}_{k=1}^\infty$ be the eigenvalues of $C_{\mathcal{H}}^n$ in a non-increasing order, and ϕ_k^n be the eigenfunction associated with λ_k^n. We define the empirical encoder $E_{\mathcal{H},d}^n : \mathcal{H} \to \mathbb{R}^d$ as

$$E_{\mathcal{H},d}^n(u) = [\langle u, \phi_1^n \rangle, \langle u, \phi_2^n \rangle, ..., \langle u, \phi_d^n \rangle]^\top \text{ for any } u \in \mathcal{H}.$$

The empirical decoder is

$$D_{\mathcal{H},d}^n(\mathbf{a}) = \sum_{j=1}^d a_j \phi_d^n \text{ for any } \mathbf{a} \in \mathbb{R}^d. \tag{27}$$

The linear subspace spanned by the first d eigenfunctions of $C_{\mathcal{H}}^n$ are denoted by $V_{\mathcal{H},d}^n$.

The PCA encoders and decoders $E_{\mathcal{H},d}, D_{\mathcal{H},d}, E_{\mathcal{H},d}^n, D_{\mathcal{H},d}^n$ are Lipchitz operators with Lipchitz constant 1 (see a proof in Section 7.6).

Lemma 2. Let $\mathcal{H}$ be a separable Hilbert space and γ be a probability measure defined on $\mathcal{H}$. Let $d > 0$ be a positive integer. Denote the empirical estimation of $E_{\mathcal{H},d}$ and $D_{\mathcal{H},d}$ by $E_{\mathcal{H},d}^n$ and $D_{\mathcal{H},d}^n$, respectively. Then for any $u, \widetilde{u} \in \mathcal{H}$, we have

$$\begin{aligned} \|E_{\mathcal{H},d}(u) - E_{\mathcal{H},d}(\widetilde{u})\|_2 &\leq \|u - \widetilde{u}\|_{\mathcal{X}}, \\ \|E_{\mathcal{H},d}^n(u) - E_{\mathcal{H},d}^n(\widetilde{u})\|_2 &\leq \|u - \widetilde{u}\|_{\mathcal{X}}. \end{aligned}$$

For any $\mathbf{a}, \widetilde{\mathbf{a}} \in \mathbb{R}^d$, we have

$$\begin{aligned} \|D_{\mathcal{H},d}(\mathbf{a}) - D_{\mathcal{H},d}(\widetilde{\mathbf{a}})\|_{\mathcal{H}} &= \|\mathbf{a} - \widetilde{\mathbf{a}}\|_2, \\ \|D_{\mathcal{H},d}^n(\mathbf{a}) - D_{\mathcal{H},d}^n(\widetilde{\mathbf{a}})\|_{\mathcal{H}} &= \|\mathbf{a} - \widetilde{\mathbf{a}}\|_2. \end{aligned}$$

5.2 Generalization error for PCA encoders and decoders

For encoders and decoders computed by PCA, we choose

$$\begin{aligned} &E_{\mathcal{X}} = E_{\mathcal{X},d_{\mathcal{X}}},\ D_{\mathcal{X}} = D_{\mathcal{X},d_{\mathcal{X}}},\ E_{\mathcal{X}}^n = E_{\mathcal{X},d_{\mathcal{X}}}^n,\ D_{\mathcal{X}}^n = D_{\mathcal{X},d_{\mathcal{X}}}^n, \\ &E_{\mathcal{Y}} = E_{\mathcal{Y},d_{\mathcal{Y}}},\ D_{\mathcal{Y}} = D_{\mathcal{Y},d_{\mathcal{Y}}},\ E_{\mathcal{Y}}^n = E_{\mathcal{Y},d_{\mathcal{Y}}}^n,\ D_{\mathcal{Y}}^n = D_{\mathcal{Y},d_{\mathcal{Y}}}^n. \end{aligned}$$

In this settings, we consider the clean data set such that $v_i = \Psi(u_i)$. The following theorem gives a bound on the generalization error of the learned transformation:

Theorem 3. Consider Setting 1. Suppose the encoders and decoders are derived from PCA, and Assumption 1 and 2 hold. Let Γ_{NN} be the minimizer of (6) with the network architecture $\mathcal{F}(L, p, M)$ where L, p, M are set as in (19). We have

$$\begin{aligned} &\mathbb{E}_{\mathcal{S}} \mathbb{E}_{u\sim\rho} \|D_{\mathcal{Y}}^n \circ \Gamma_{\mathrm{NN}} \circ E_{\mathcal{X}}^n(u) - \Psi(u)\|_{\mathcal{Y}}^2 \\ \leq & C_4(\sigma^2 + R_{\mathcal{Y}}^2) d_{\mathcal{Y}}^{\frac{4+d_{\mathcal{X}}}{2+d_{\mathcal{X}}}} n^{-\frac{2}{2+d_{\mathcal{X}}}} \log^6 n + 4\left(8R_{\mathcal{X}}^2 L_{\Psi}^2 \sqrt{d_{\mathcal{X}}} + R_{\mathcal{Y}}^2 \sqrt{d_{\mathcal{Y}}}\right) n^{-\frac{1}{2}} \\ &+ 16 L_{\Psi}^2 \mathbb{E}_{u\sim\rho} \|\Pi_{\mathcal{X},d_{\mathcal{X}}}(u) - u\|_2^2 + 2\mathbb{E}_{v^*\sim\Psi_{\#}\rho} \|\Pi_{\mathcal{Y},d_{\mathcal{Y}}}(v^*) - v^*\|_{\mathcal{Y}}^2 \end{aligned}$$

where C_4 is a constant depending on $d_{\mathcal{X}}, R_{\mathcal{X}}, R_{\mathcal{Y}}, L_{\Psi}$.

Theorem 3 is proved in Section 7.7. Since the encoders and decoders derived from PCA are data-driven, we expect the corresponding projection errors are smaller than those by using basis encoders and decoders.

6 Mitigate the curse of dimensionality

In Theorem 2, $E_{\mathcal{X}}$ encodes each u into a vector in $\mathbb{R}^{d_{\mathcal{X}}}$, which is the input to the network $\Gamma_{\rm NN}$. The convergence rate is cursed by the input dimension $d_{\mathcal{X}}$. When $d_{\mathcal{X}}$ is large, Theorem 2 gives a slow rate. In this section, we discuss two scenarios that we can mitigate the curse of dimensionality and get a faster rate: (1) when $E_{\mathcal{X}}(\Omega_{\mathcal{X}})$, the image of $\Omega_{\mathcal{X}}$ under $E_{\mathcal{X}}$, has low-dimensional structures and (2) when $E_{\mathcal{Y}} \circ \Psi \circ D_{\mathcal{X}}$ has low complexities. The main difference between the two scenarios is the object on which we impose low-dimensional property assumptions. The first scenario focuses on the input set $\Omega_{\mathcal{X}}$ and the second scenario focuses on the operator Ψ. In the rest of this section, we give details on each scenario with a faster convergence rate.

6.1 Generalization error when $E_{\mathcal{X}}(\Omega_{\mathcal{X}})$ locates on a low-dimensional manifold

Although $\mathcal{X}$ is an infinitely dimensional function space, in practice, the functions of interest are only a small subset of $\mathcal{X}$ with low-complexities, leading to low-dimensional structures of $\Omega_{\mathcal{X}}$. In the following, we give an example on such a low-dimensional structure.

Example 1. Let $\mathcal{X} = L^2([-1,1])$ and $0 < \widetilde{d}_{\mathcal{X}} < d_{\mathcal{X}}$ be an integer. Let T_k's be trigonometric functions defined in (22) and g_k's are some real valued functions for $\widetilde{d}_{\mathcal{X}} < k \leq d_{\mathcal{X}}$. Set

$$\Omega_{\mathcal{X}} = \left\{ u : u = \sum_{k=1}^{d_{\mathcal{X}}} a_k T_k \text{ with } a_k \in \mathbb{R} \text{ for } 1 \leq k \leq \widetilde{d}_{\mathcal{X}}, \text{ and } a_k = g_k(a_1, ..., a_{\widetilde{d}_{\mathcal{X}}}) \text{ for } k > \widetilde{d}_{\mathcal{X}} \right\}.$$

Such a $\Omega_{\mathcal{X}}$ has an intrinsic dimension $\widetilde{d}_{\mathcal{X}}$. Consider $E_{\mathcal{X}}$ derived using the basis $\{T_k\}_{k=1}^{d_{\mathcal{X}}}$. Then for any $u \in \Omega_{\mathcal{X}}$, we have that $E_{\mathcal{X}}(u) \in \mathbb{R}^{d_{\mathcal{X}}}$ locates on a $\widetilde{d}_{\mathcal{X}}$-dimensional manifold embedded in $\mathbb{R}^{d_{\mathcal{X}}}$. An illustration of the manifold with $d_{\mathcal{X}} = 3, \widetilde{d}_{\mathcal{X}} = 2$ and $g_3 = a_1^2 + a_2$ is shown in Figure 1.

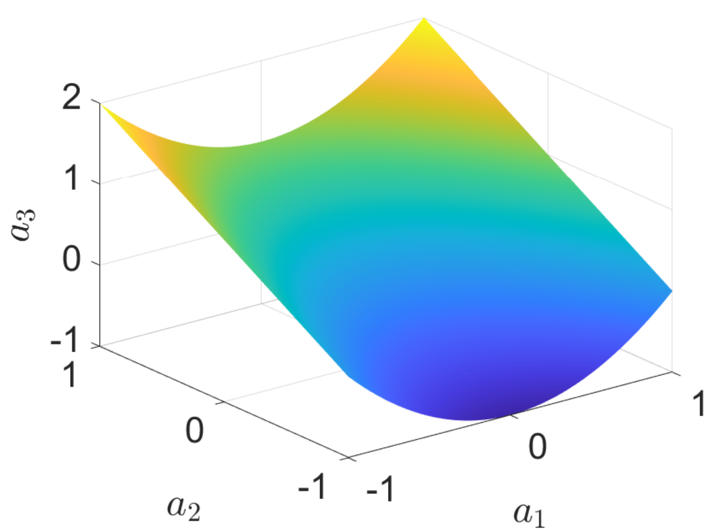

Figure 1: An illustration of Example 1 with $d_{\mathcal{X}} = 3, \widetilde{d}_{\mathcal{X}} = 2$ and $g_3 = a_1^2 + a_2$. Here $E_{\mathcal{X}}(\Omega_{\mathcal{X}})$ is a two-dimensional manifold embedded in $\mathbb{R}^3$.

In this subsection, we show that the curse of dimensionality can be mitigated by exploiting the low-dimensional structures of $E_{\mathcal{X}}(\Omega_{\mathcal{X}})$. Specifically, we consider the encoder $E_{\mathcal{X}} : \mathcal{X} \to \mathcal{M}$ for some low-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^{d_{\mathcal{X}}}$. We will show that when such an $E_{\mathcal{X}}$ is known, under appropriate assumptions, the convergence rate of the generalization error depends on the intrinsic dimension of $\mathcal{M}$, instead of $d_{\mathcal{X}}$. We first make the following assumptions on $E_{\mathcal{X}}$ and $\mathcal{M}$:

Assumption 7. Assume

(i) There exists an $E_{\mathcal{X}}$ such that $E_{\mathcal{X}}(\Omega_{\mathcal{X}})$ is on a manifold $\mathcal{M}$.

(ii) $\mathcal{M}$ is a $\widetilde{d}_{\mathcal{X}}$-dimensional compact smooth Riemannian manifold isometrically embedded in $\mathbb{R}^{d_{\mathcal{X}}}$ with $\widetilde{d}_{\mathcal{X}} \leq d_{\mathcal{X}}$.

(iii) The reach of $\mathcal{M}$ is $\tau > 0$.

Recall that our network Γ_{NN} is an estimation of $E_{\mathcal{Y}} \circ \Psi \circ D_{\mathcal{X}}$. Under Assumption 7, the latter one is a function from $\mathcal{M}$ to $\mathbb{R}^{d_{\mathcal{Y}}}$. Therefore training Γ_{NN} is equivalent to learning a function defined on $\mathcal{M}$. The following theorem considers the network architecture (2) and gives a generalization error for this scenario.

Theorem 4. Consider Setting 1. Suppose Assumption 1–4 and 7 hold and the encoder $E_{\mathcal{X}}$ in Assumption 7 is given. Let Γ_{NN} be the minimizer of (6) with the network architecture $\mathcal{F}(L, p, K, \kappa, M)$, where

$$\begin{aligned} &L = O(\log n),\ p = O\left(n^{\frac{\widetilde{d}_{\mathcal{X}}}{2+\widetilde{d}_{\mathcal{X}}}} + d_{\mathcal{X}}\right),\ K = O\left(d_{\mathcal{X}} n^{\frac{\widetilde{d}_{\mathcal{X}}}{2+\widetilde{d}_{\mathcal{X}}}} \log n\right),\ M = \sqrt{d_{\mathcal{Y}}} L_{E_{\mathcal{Y}}} R_{\mathcal{Y}} \\ &\kappa = \max\left\{1, \sqrt{d_{\mathcal{Y}}} L_{E_{\mathcal{Y}}} R_{\mathcal{Y}}, R_{\mathcal{X}}, L_{E_{\mathcal{Y}}} L_{D_{\mathcal{X}}} L_{\Psi}, \tau^2, \sqrt{d_{\mathcal{X}}}\right\}. \end{aligned} \tag{28}$$

We have

$$\begin{aligned} &\mathbb{E}_{\mathcal{S}} \mathbb{E}_{u \sim \rho} \| D_{\mathcal{Y}} \circ \Gamma_{\mathrm{NN}} \circ E_{\mathcal{X}}(u) - \Psi(u) \|_{\mathcal{Y}}^2 \\ &\qquad \leq C_9 (\sigma^2 + R_{\mathcal{Y}}^2) d_{\mathcal{Y}}^{\frac{4+\widetilde{d}_{\mathcal{X}}}{2+\widetilde{d}_{\mathcal{X}}}} n^{-\frac{2}{2+\widetilde{d}_{\mathcal{X}}}} \log^3 n + C_{10} (\sigma^2 + R_{\mathcal{Y}}^2) d_{\mathcal{Y}}^2 (\log d_{\mathcal{Y}}) n^{-1} \\ &\qquad\quad + C_3 \mathbb{E}_{\mathcal{S}_1} \mathbb{E}_{u \sim \rho} \| \Pi_{\mathcal{X}, d_{\mathcal{X}}}(u) - u \|_2^2 + 2 \mathbb{E}_{\mathcal{S}_1} \mathbb{E}_{v^* \sim \Psi_{\#}\rho} \| \Pi_{\mathcal{Y}, d_{\mathcal{Y}}}^n (v^*) - v^* \|_{\mathcal{Y}}^2 \end{aligned} \tag{29}$$

where C_9, C_{10} are constants depending on $\widetilde{d}_{\mathcal{X}}, \log d_{\mathcal{X}}, R_{\mathcal{X}}, R_{\mathcal{Y}}, L_{E_{\mathcal{X}}}, L_{E_{\mathcal{Y}}^n}, L_{D_{\mathcal{X}}^n}, L_{D_{\mathcal{Y}}^n}, L_{\Psi}, \tau$, the surface area of $\mathcal{M}$, and $C_3 = 16 L_{D_{\mathcal{Y}}^n}^2 L_{E_{\mathcal{Y}}^n}^2 L_{\Psi}^2$.

Theorem 4 is proved in Section 7.8. The convergence rate in Theorem 4 has an exponent depending on $\widetilde{d}_{\mathcal{X}}$, instead of $d_{\mathcal{X}}$. Theorem 4 shows that when the encoded vectors locate on a low-dimensional manifold, deep neural networks are adaptive to the these structures.

6.2 Generalization error when $E_{\mathcal{Y}} \circ \Psi \circ D_{\mathcal{X}}$ has low complexities

This second scenario considers Ψ with low complexities. While Ψ is an operator between infinite dimensional function spaces, it may only depend on the projection of the input along certain directions. In the following we give a simple example:

Example 2. Let $\mathcal{X} = L^2([-1,1]), \Omega \subset \mathcal{X}$ be a compact set and $0 < \widetilde{d}_{\mathcal{X}} < d_{\mathcal{X}}$ be an integer. Let T_k's be trigonometric functions defined in (22). For any $u \in \Omega_{\mathcal{X}}$, it can be expressed as $u = \sum_{k=1}^{\infty} a_k T_k$. Denote $\mathbf{a}_u = \begin{bmatrix} a_1 & \cdots & a_{d_{\mathcal{X}}} \end{bmatrix}^\top$. Consider the operator defined as follows

$$\Psi(u) = \sum_{k=1}^{d_{\mathcal{Y}}} g_k(V_k^\top \mathbf{a}_u) T_k$$

where $V_k \in \mathbb{R}^{d_{\mathcal{X}} \times \widetilde{d}_{\mathcal{X}}}$ and $g_k : \mathbb{R}^{\widetilde{d}_{\mathcal{X}}} \to \mathbb{R}$ is some real valued function for $k = 1, ..., d_{\mathcal{Y}}$. We set $E_{\mathcal{X}}, D_{\mathcal{X}}$ as encoder and decoder derived using basis $\{T_k\}_{k=1}^{d_{\mathcal{X}}}$, and $E_{\mathcal{Y}}, D_{\mathcal{Y}}$ as encoder and decoder derived using basis $\{T_k\}_{k=1}^{d_{\mathcal{Y}}}$. Then learning Ψ reduces to learning g_k's and V_k's. In neural networks, V_k can be realized by a single layer. Therefore, the major task is to learn g_k whose input dimension is $\widetilde{d}_{\mathcal{X}}$. An illustration of the estimator is shown in Figure 2.

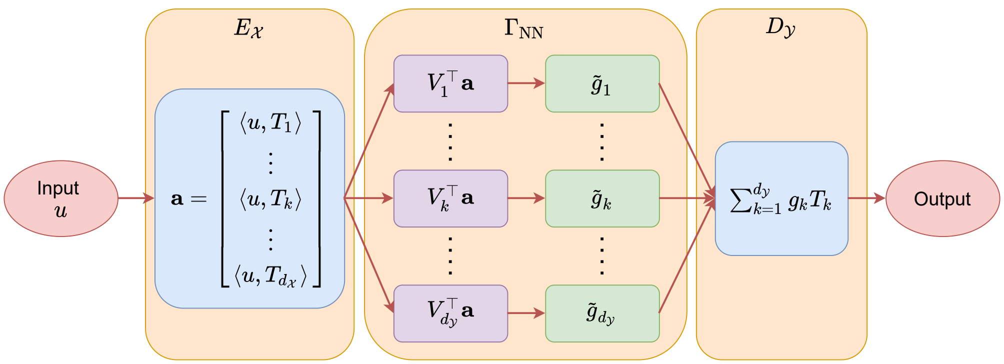

Figure 2: An illustration of Example 2. The $\widetilde{g}_k$'s are network approximations of g_k's.

In this subsection, we show that the curse of dimensionality can be mitigated by exploiting the low complexities of Ψ. We first make an assumption on Ψ:

Assumption 8. Let $0 < \widetilde{d}_{\mathcal{X}} \le d_{\mathcal{X}}$ be an integer. Assume there exist $E_{\mathcal{X}}, D_{\mathcal{X}}, E_{\mathcal{Y}}, D_{\mathcal{Y}}$ such that for any $\mathbf{a} \in E_{\mathcal{X}}(\Omega_{\mathcal{X}})$, we have

$$E_{\mathcal{Y}} \circ \Psi \circ D_{\mathcal{X}}(\mathbf{a}) = \begin{bmatrix} g_1(V_1^\top \mathbf{a}) & \cdots & g_{d_{\mathcal{Y}}}(V_{d_{\mathcal{Y}}}^\top \mathbf{a}) \end{bmatrix}^\top, \tag{30}$$

where $V \in \mathbb{R}^{d_{\mathcal{X}} \times \widetilde{d}_{\mathcal{X}}}$ and $g_k : \mathbb{R}^{\widetilde{d}_{\mathcal{X}}} \to \mathbb{R}$ is some real valued functions for $k = 1, ..., d_{\mathcal{Y}}$.

In statistics, the functions g_k's in Assumption 8 are known as single index model for $\widetilde{d}_{\mathcal{X}} = 1$, and are known as multi index model for $\widetilde{d}_{\mathcal{X}} > 1$. With Assumption 8, the following theorem gives a faster rate on the generalization error:

Theorem 5. Consider Setting 1. Suppose Assumption 1–4 and 8 hold and the encoders and decoders $E_{\mathcal{X}}, D_{\mathcal{X}}, E_{\mathcal{Y}}, D_{\mathcal{Y}}$ in Assumption 8 is given. Let Γ_{NN} be the minimizer of (6) with the network architecture $\mathcal{F}(L, p, M)$, where

$$L = O(\widetilde{L} \log \widetilde{L}), \;\; p = O\left(\widetilde{p} \log \widetilde{p}\right), M = \sqrt{d_{\mathcal{Y}}} L_{E_{\mathcal{Y}}^n} R_{\mathcal{Y}} \tag{31}$$

and $\widetilde{L},\widetilde{p}>0$ are integers and satisfy $\widetilde{L}\widetilde{p}=\left\lceil d_{\mathcal{Y}}^{-\frac{\widetilde{d}_{\mathcal{X}}}{4+2\widetilde{d}_{\mathcal{X}}}} n^{\frac{\widetilde{d}_{\mathcal{X}}}{4+2\widetilde{d}_{\mathcal{X}}}}\right\rceil$.

We have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\mathbb{E}_{u\sim\rho}\|D_{\mathcal{Y}}^n\circ\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u)-\Psi(u)\|_{\mathcal{Y}}^2\\
\leq& C_{11}(\sigma^2+R_{\mathcal{Y}}^2)d_{\mathcal{Y}}^{\frac{4+\widetilde{d}_{\mathcal{X}}}{2+\widetilde{d}_{\mathcal{X}}}} n^{-\frac{2}{2+\widetilde{d}_{\mathcal{X}}}}\log^6 n + C_3\mathbb{E}_{\mathcal{S}_1}\mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_{\mathcal{X}}}^n(u)-u\|_{\mathcal{X}}^2+2\mathbb{E}_{\mathcal{S}_1}\mathbb{E}_{v^*\sim\Psi_{\#}\rho}\|\Pi_{\mathcal{Y},d_{\mathcal{Y}}}^n(v^*)-v^*\|_{\mathcal{Y}}^2.
\end{aligned} \tag{32}
$$

where C_{11} depending on $\widetilde{d}_{\mathcal{X}},\log d_{\mathcal{X}},R_{\mathcal{X}},R_{\mathcal{Y}},L_{E_{\mathcal{X}}},L_{E_{\mathcal{Y}}^n},L_{D_{\mathcal{X}}^n},L_{D_{\mathcal{Y}}^n},L_{\Psi},\tau$, the surface area of $\mathcal{M}$, and $C_3=16L_{D_{\mathcal{Y}}^n}^2L_{E_{\mathcal{Y}}^n}^2L_{\Psi}^2$.

Theorem 5 can be proved exactly in the same way as that of Theorem 2, except we need to add an additional layer next to the input layer to realize V_k's. The proof is omitted here.

7 Proof of main results

7.1 Proof of Theorem 1

We decompose the error as

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\mathbb{E}_{u\sim\rho}\|D_{\mathcal{Y}}^n\circ\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u)-\Psi(u)\|_{\mathcal{Y}}^2\\
\leq&\underbrace{2\mathbb{E}_{\mathcal{S}}\mathbb{E}_{u\sim\rho}\|D_{\mathcal{Y}}^n\circ\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u)-D_{\mathcal{Y}}^n\circ E_{\mathcal{Y}}^n\circ\Psi(u)\|_{\mathcal{Y}}^2}_{\rm I}\\
&+\underbrace{2\mathbb{E}_{\mathcal{S}}\mathbb{E}_{u\sim\rho}\|D_{\mathcal{Y}}^n\circ E_{\mathcal{Y}}^n\circ\Psi(u)-\Psi(u)\|_{\mathcal{Y}}^2}_{\rm II}
\end{aligned} \tag{33}
$$

Here I is the network estimation error, II is the empirical projection error, which can be rewritten as

$$
{\rm II}=2\mathbb{E}_{\mathcal{S}}\mathbb{E}_{v^*\sim\Psi_{\#}\rho}\|\Pi_{\mathcal{Y},d_{\mathcal{Y}}}^n(v^*)-v^*\|_{\mathcal{Y}}^2. \tag{34}
$$

In the remaining of this subsections, we derive an upper bound of I. Term I can be bounded as

$$
\begin{aligned}
{\rm I}=&2\mathbb{E}_{\mathcal{S}}\mathbb{E}_{u\sim\rho}\|D_{\mathcal{Y}}^n\circ\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u)-D_{\mathcal{Y}}^n\circ E_{\mathcal{Y}}^n\circ\Psi(u)\|_{\mathcal{Y}}^2\\
\leq&2L_{D_{\mathcal{Y}}^n}^2\mathbb{E}_{\mathcal{S}}\mathbb{E}_{u\sim\rho}\|\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u)-E_{\mathcal{Y}}^n\circ\Psi(u)\|_2^2.
\end{aligned} \tag{35}
$$

We first fix $\mathcal{S}_1$. Then

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}_2}\mathbb{E}_{u\sim\rho}\|\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u)-E_{\mathcal{Y}}^n\circ\Psi(u)\|_2^2\\
=&\underbrace{2\,\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\|\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u_i)-E_{\mathcal{Y}}^n\circ\Psi(u_i)\|_2^2}_{\rm T_1}\\
&+\underbrace{\mathbb{E}_{\mathcal{S}_2}\mathbb{E}_{u\sim\rho}\|\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u)-E_{\mathcal{Y}}^n\circ\Psi(u)\|_2^2-\mathbb{E}_{\mathcal{S}_2}\frac{2}{n}\sum_{i=n+1}^{2n}\|\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u_i)-E_{\mathcal{Y}}^n\circ\Psi(u_i)\|_2^2}_{\rm T_2}.
\end{aligned} \tag{36}
$$

In the above decomposition, the term T_1 consists of the bias of using neural network to approximate the transformation Γ and the projection error of $\Pi^n_{\mathcal{X},d_{\mathcal{X}}}$. The term T_2 captures the variance. We next derive bounds for T_1 and T_2 in order.

Upper bound of T_1. The term T_1 is the expected mean squared error of the learned transformation with respect to $\mathcal{S}_2$. We will derive an upper bound using network approximation error and network architecture's covering number. The network approximation error is used to capture the bias. Since the transformation Γ_{NN} is learned using noisy data, we use network architecture's covering number to bound the stocastic error induced by noise.

Define $\Gamma^n_d : \mathbb{R}^{d_{\mathcal{X}}} \to \mathbb{R}^{d_{\mathcal{Y}}}$ as

$$\Gamma^n_d = E^n_{\mathcal{Y}} \circ \Psi \circ D^n_{\mathcal{X}}. \tag{37}$$

The transformation Γ^n_d is the target transformation to be estimated by Γ_{NN}. One can prove that Γ^n_d is Lipschitz (see a proof in Section 7.9).

Lemma 3. Assume Assumption 2 and 3. Γ^n_d is Lipschitz with Lipschitz constant $L_{E^n_{\mathcal{Y}}} L_{D^n_{\mathcal{X}}} L_{\Psi}$.

Denote

$$\boldsymbol{\epsilon}_i = E^n_{\mathcal{Y}}(v_i) - E^n_{\mathcal{Y}}(\Psi(u_i)). \tag{38}$$

According to Assumption 4(iv), we have $\mathbb{E}\boldsymbol{\epsilon}_i = \mathbf{0}$. Using Lemma 3, we decompose T_1 as

$$\begin{aligned}
\mathrm{T}_1 =& 2\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\|\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i) - E_{\mathcal{Y}}^n\circ\Psi(u_i)\|_2^2 \\
=& 2\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\|\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i) - E_{\mathcal{Y}}^n\circ\Psi(u_i) - \boldsymbol{\epsilon}_i + \boldsymbol{\epsilon}_i\|_2^2 \\
=& 2\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\|\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i) - E_{\mathcal{Y}}^n\circ\Psi(u_i) - \boldsymbol{\epsilon}_i\|_2^2 \\
&+ 4\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\langle\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i) - E_{\mathcal{Y}}^n\circ\Psi(u_i) - \boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_i\rangle + 2\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\|\boldsymbol{\epsilon}_i\|_2^2 \\
=& 2\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\|\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i) - E_{\mathcal{Y}}^n(v_i)\|_2^2 + 4\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\langle\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i), \boldsymbol{\epsilon}_i\rangle - 2\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\|\boldsymbol{\epsilon}_i\|_2^2 \\
=& 2\mathbb{E}_{\mathcal{S}_2}\inf_{\Gamma\in\mathcal{F}_{\mathrm{NN}}}\frac{1}{n}\sum_{i=n+1}^{2n}\|\Gamma\circ E_{\mathcal{X}}^n(u_i) - E_{\mathcal{Y}}^n(v_i)\|_2^2 \\
&+ 4\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\langle\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i), \boldsymbol{\epsilon}_i\rangle - 2\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\|\boldsymbol{\epsilon}_i\|_2^2 \qquad \text{by the definition of } \Gamma_{\mathrm{NN}} \text{ in (6)} \\
\leq& 2\inf_{\Gamma\in\mathcal{F}_{\mathrm{NN}}}\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\|\Gamma\circ E_{\mathcal{X}}^n(u_i) - E_{\mathcal{Y}}^n(v_i)\|_2^2 + 4\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\langle\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i), \boldsymbol{\epsilon}_i\rangle \\
&- 2\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\|\boldsymbol{\epsilon}_i\|_2^2 \\
=& 2\inf_{\Gamma\in\mathcal{F}_{\mathrm{NN}}}\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\left[\|\Gamma\circ E_{\mathcal{X}}^n(u_i) - 2E_{\mathcal{Y}}^n\circ\Psi(u_i) - \boldsymbol{\epsilon}_i\|_2^2 - \|\boldsymbol{\epsilon}_i\|_2^2\right] \\
&+ 4\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\langle\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i), \boldsymbol{\epsilon}_i\rangle \\
=& 2\inf_{\Gamma\in\mathcal{F}_{\mathrm{NN}}}\mathbb{E}_{\mathcal{S}_2}\|\Gamma\circ E_{\mathcal{X}}^n(u_i) - E_{\mathcal{Y}}^n\circ\Psi(u_i)\|_2^2 + 4\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\langle\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i), \boldsymbol{\epsilon}_i\rangle. \qquad (39)
\end{aligned}$$

In (39), the first term is the neural network approximation error, the second term is the stochastic error from noise. We will give a bound of T_1 using the global covering number of $\mathcal{F}_{\mathrm{NN}}$ which is defined as follows:

Definition 2 (Global cover)**.** Let $\mathcal{F}$ be a class of functions. The function set $\mathcal{S}_f$ is a global δ-cover of $\mathcal{F}$ under the measure $\|\cdot\|$ if for any $f \in F$, there exits $f^* \in \mathcal{S}_f$ such that

$$\|f(\mathbf{x}_k) - f^*(\mathbf{x}_k)\| \leq \varepsilon.$$

Definition 3 (Global covering number)**.** Let F be a class of functions. For any $\delta > 0$, the global covering number of $\mathcal{F}$ is defined as

$$\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|) = \min|\mathcal{S}_f|$$

such that $\mathcal{S}_f$ is a global δ-cover of $\mathcal{F}$, where $|\mathcal{S}_f|$ denotes the cardinality of $\mathcal{S}_f$.

Based on the decomposition in (39), the following lemma gives an upper bound of T_1 (see a proof in Section 7.10):

Lemma 4. Under the conditions of Theorem 1, there exists a network architecture $\mathcal{F}_{\mathrm{NN}}(L, p, K, \kappa, M)$ such that for any $\varepsilon_1 \in (0, 1)$, we have

$$\begin{aligned}\mathrm{T}_1 \leq & 8d_{\mathcal{Y}}\varepsilon_1^2 + 64d_{\mathcal{Y}}\sigma^2 \frac{\log \mathcal{N}(\delta, \mathcal{F}_{\mathrm{NN}}, \|\cdot\|_\infty) + 2}{n} + 16\sqrt{2}d_{\mathcal{Y}}\sigma\delta\sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}_{\mathrm{NN}}, \|\cdot\|_\infty) + 2}{n}} \\ & + 8d_{\mathcal{Y}}\sigma\delta + 8L_{E_{\mathcal{Y}}^n}^2 L_\Psi^2 \mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_{\mathcal{X}}}^n(u) - u\|_{\mathcal{X}}^2. \end{aligned} \tag{40}$$

Such a network architecture has

$$\begin{aligned} & L = O(\log \varepsilon_1),\ p = O\left(\varepsilon_1^{-d_{\mathcal{X}}}\right),\ K = O\left(\varepsilon_1^{-d_{\mathcal{X}}} \log \varepsilon_1\right), \\ & \kappa = \max\left\{1, \sqrt{d_{\mathcal{Y}}}L_{E_{\mathcal{Y}}^n}R_{\mathcal{Y}}, \sqrt{d_{\mathcal{X}}}L_{E_{\mathcal{X}}^n}R_{\mathcal{X}}, L_{E_{\mathcal{Y}}^n}L_{D_{\mathcal{X}}^n}L_\Psi\right\},\ M = \sqrt{d_{\mathcal{Y}}}L_{E_{\mathcal{Y}}^n}R_{\mathcal{Y}}. \end{aligned} \tag{41}$$

Upper bound of T_2. The term T_2 is the difference between the population risk and the empirical risk of the network estimator Γ_{NN}, while there is a factor 2 ahead of the empirical risk. To derive an upper bound, we evenly split the empirical risk into two parts, one of which is bounded using its forth moment. Utilizing a global covering of $\mathcal{F}_{\mathrm{NN}}(L, p, K, \kappa, M)$ and Bernstein-type inequalities, we establish a $1/n$ convergence of T_2 with the global covering number as a factor. The upper bound is presented in the following lemma (see a proof in Section 7.11).

Lemma 5. Under the conditions of Theorem 1, we have

$$\mathrm{T}_2 \leq \frac{104 d_{\mathcal{Y}} L_{E_{\mathcal{Y}}^n}^2 R_{E_{\mathcal{Y}}}^2}{3n} \log \mathcal{N}\left(\frac{\delta}{4d_{\mathcal{Y}}L_{E_{\mathcal{Y}}^n}R_{\mathcal{Y}}}, \mathcal{F}_{\mathrm{NN}}, \|\cdot\|_\infty\right) + 6\delta. \tag{42}$$

Substituting (40) and (42) into (35) gives rise to

$$\begin{aligned}\mathrm{I} \leq & 2L_{D_{\mathcal{Y}}^n}^2 \mathbb{E}_{\mathcal{S}_1}\mathbb{E}_{u\sim\rho}\|\Gamma_{\mathrm{NN}} \circ E_{\mathcal{X}}^n(u) - E_{\mathcal{Y}}^n \circ \Psi(u)\|_2^2 \\ = & 2L_{D_{\mathcal{Y}}^n}^2 \mathbb{E}_{\mathcal{S}_1}\mathrm{T}_1 + 2L_{D_{\mathcal{Y}}^n}^2 \mathbb{E}_{\mathcal{S}_1}\mathrm{T}_2 \\ \leq & 16d_{\mathcal{Y}}L_{D_{\mathcal{Y}}^n}^2\varepsilon_1^2 + 128d_{\mathcal{Y}}\sigma^2 L_{D_{\mathcal{Y}}^n}^2 \frac{\log \mathcal{N}(\delta, \mathcal{F}_{\mathrm{NN}}, \|\cdot\|_\infty) + 2}{n} + 32\sqrt{2}d_{\mathcal{Y}}\sigma L_{D_{\mathcal{Y}}^n}^2\delta\sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}_{\mathrm{NN}}, \|\cdot\|_\infty) + 2}{n}} \\ & + 16d_{\mathcal{Y}}\sigma L_{D_{\mathcal{Y}}^n}^2\delta + 16L_{D_{\mathcal{Y}}^n}^2 L_{E_{\mathcal{Y}}^n}^2 L_\Psi^2 \mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_{\mathcal{X}}}^n(u) - u\|_{\mathcal{X}}^2 \\ & + \frac{208 d_{\mathcal{Y}} L_{D_{\mathcal{Y}}^n}^2 L_{E_{\mathcal{Y}}^n}^2 R_{\mathcal{Y}}^2}{3n} \log \mathcal{N}\left(\frac{\delta}{4d_{\mathcal{Y}}L_{E_{\mathcal{Y}}^n}R_{\mathcal{Y}}}, \mathcal{F}_{\mathrm{NN}}, \|\cdot\|_\infty\right) + 12L_{D_{\mathcal{Y}}^n}^2\delta \\ \leq & 16d_{\mathcal{Y}}L_{D_{\mathcal{Y}}^n}^2\varepsilon_1^2 + \frac{672 d_{\mathcal{Y}}\sigma^2 L_{D_{\mathcal{Y}}^n}^2 + 208 d_{\mathcal{Y}} L_{D_{\mathcal{Y}}^n}^2 L_{E_{\mathcal{Y}}^n}^2 R_{\mathcal{Y}}^2}{3n} \log \mathcal{N}(\frac{\delta}{4d_{\mathcal{Y}}L_{E_{\mathcal{Y}}^n}R_{\mathcal{Y}}}, \mathcal{F}_{\mathrm{NN}}, \|\cdot\|_\infty) \\ & + (16d_{\mathcal{Y}}\sigma + 12)L_{D_{\mathcal{Y}}^n}^2\delta + 16L_{D_{\mathcal{Y}}^n}^2 L_{E_{\mathcal{Y}}^n}^2 L_\Psi^2 \mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_{\mathcal{X}}}^n(u) - u\|_{\mathcal{X}}^2 \end{aligned} \tag{43}$$

when $\delta < 1$. The covering number can be bounded by the following lemma:

Lemma 6 ([11] Lemma 6)**.** Let $\mathcal{F}_{\rm NN}(L,p,K,\kappa,M)$ be a class of network: $[-B,B]^{d_\mathcal{X}} \to [-M,M]^{d_\mathcal{Y}}$. For any $\delta>0$, the δ-covering number of $\mathcal{F}_{\rm NN}(L,p,K,\kappa,M)$ is bounded by

$$\mathcal{N}(\delta,\mathcal{F}_{\rm NN}(L,p,K,\kappa,M),\|\cdot\|_\infty) \le \left(\frac{2L^2(pB+2)\kappa^L p^{L+1}}{\delta}\right)^{d_\mathcal{Y}K}. \tag{44}$$

Combining (41) and (44) gives

$$\log\mathcal{N}(\delta,\mathcal{F}_{\rm NN}(L,p,K,\kappa,M),\|\cdot\|_\infty) \le C_{12}d_\mathcal{Y}\left(\varepsilon_1^{-d_\mathcal{X}}\log^3\varepsilon_1^{-1}+\log\delta+\log d_\mathcal{Y}\right), \tag{45}$$

where C_{12} is a constant depending on $d_\mathcal{X}, R_\mathcal{X}, R_\mathcal{Y}, L_{E_\mathcal{X}^n}, L_{E_\mathcal{Y}^n}, L_{D_\mathcal{X}^n}$ and L_Ψ. Substituting (45) into (43) yields

$$\begin{aligned}\mathrm{I} \le& 16d_\mathcal{Y}L_{D_\mathcal{Y}^n}^2\varepsilon_1^2 + C_{12}d_\mathcal{Y}^2L_{D_\mathcal{Y}^n}^2\frac{672\sigma^2+208L_{E_\mathcal{Y}^n}^2R_\mathcal{Y}^2}{3n}\left(\varepsilon_1^{-d_\mathcal{X}}\log^3\varepsilon_1^{-1}+\log\delta+\log d_\mathcal{Y}\right)\\ &+(16d_\mathcal{Y}\sigma+12)L_{D_\mathcal{Y}^n}^2\delta + 16L_{D_\mathcal{Y}^n}^2L_{E_\mathcal{Y}^n}^2L_\Psi^2\mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_\mathcal{X}}^n(u)-u\|_\mathcal{X}^2.\end{aligned} \tag{46}$$

Setting

$$\varepsilon_1 = d_\mathcal{Y}^{\frac{1}{2+d_\mathcal{X}}}n^{-\frac{1}{2+d_\mathcal{X}}}, \delta=n^{-1},$$

we have

$$\begin{aligned}\mathrm{I} \le& C_1(\sigma^2+R_\mathcal{Y}^2)d_\mathcal{Y}^{\frac{4+d_\mathcal{X}}{2+d_\mathcal{X}}}n^{-\frac{2}{2+d_\mathcal{X}}}\log^3 n + C_2(\sigma^2+R_\mathcal{Y}^2)d_\mathcal{Y}^2(\log d_\mathcal{Y})n^{-1}\\ &+16L_{D_\mathcal{Y}^n}^2L_{E_\mathcal{Y}^n}^2L_\Psi^2\mathbb{E}_{\mathcal{S}_1}\mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_\mathcal{X}}^n(u)-u\|_2^2\end{aligned} \tag{47}$$

for some C_2, C_3 depending on $d_\mathcal{X}, R_\mathcal{X}, R_\mathcal{Y}, L_{E_\mathcal{X}^n}, L_{E_\mathcal{Y}^n}, L_{D_\mathcal{X}^n}, L_\Psi$. The resulting network architecture $\mathcal{F}(L,p,K,\kappa,M)$ has

$$\begin{aligned}&L=O(\log n+\log d_\mathcal{Y}),\ p=O\left(d_\mathcal{Y}^{-\frac{d_\mathcal{X}}{2+d_\mathcal{X}}}n^{\frac{d_\mathcal{X}}{2+d_\mathcal{X}}}\right),\ K=O\left(d_\mathcal{Y}^{-\frac{d_\mathcal{X}}{2+d_\mathcal{X}}}n^{\frac{d_\mathcal{X}}{2+d_\mathcal{X}}}\log n\right),\\ &\kappa=\max\left\{1,\sqrt{d_\mathcal{Y}}L_{E_\mathcal{Y}^n}R_\mathcal{Y},\sqrt{d_\mathcal{X}}L_{E_\mathcal{X}^n}R_\mathcal{X},\sqrt{d_\mathcal{X}}L_{E_\mathcal{X}^n}L_{E_\mathcal{Y}^n}L_{D_\mathcal{X}^n}L_\Psi R_\mathcal{X}\right\},\ M=\sqrt{d_\mathcal{Y}}L_{E_\mathcal{Y}^n}R_\mathcal{Y}.\end{aligned} \tag{48}$$

Putting all ingredients together. Putting (46) and (34) together gives rise to

$$\begin{aligned}&\mathbb{E}_\mathcal{S}\mathbb{E}_{u\sim\rho}\|D_\mathcal{Y}^n\circ\Gamma_{\rm NN}\circ E_\mathcal{X}^n(u)-\Psi(u)\|_\mathcal{Y}^2\\ \le&\mathrm{I}+\mathrm{II}\\ \le& C_1(\sigma^2+R_\mathcal{Y}^2)d_\mathcal{Y}^{\frac{4+d_\mathcal{X}}{2+d_\mathcal{X}}}n^{-\frac{2}{2+d_\mathcal{X}}}\log^3 n + C_2(\sigma^2+R_\mathcal{Y}^2)d_\mathcal{Y}^2(\log d_\mathcal{Y})n^{-1}\\ &+16L_{D_\mathcal{Y}^n}^2L_{E_\mathcal{Y}^n}^2L_\Psi^2\mathbb{E}_{\mathcal{S}_1}\mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_\mathcal{X}}^n(u)-u\|_2^2+2\mathbb{E}_{\mathcal{S}_1}\mathbb{E}_{v^*\sim\Psi_\#\rho}\|\Pi_{\mathcal{Y},d_\mathcal{Y}}^n(v^*)-v^*\|_\mathcal{Y}^2,\end{aligned} \tag{49}$$

where C_1, C_2 are constants depending on $d_\mathcal{X}, R_\mathcal{X}, R_\mathcal{Y}, L_{E_\mathcal{X}^n}, L_{E_\mathcal{Y}^n}, L_{D_\mathcal{X}^n}, L_\Psi$.

7.2 Proof of Theorem 2

The main idea of the proof of Theorem 2 is the same as that of Theorem 1, except two differences in bounding T_1 and T_2 in (36). For T_1, we use a different lemma on the approximation error of deep neural networks, which focuses on the architecture of $\mathcal{F}_{\rm NN}(L,p,M)$. For T_2, we derive an upper bound using the uniform convering numbers, a different kind of covering numbers. The first part of our proof is the same as that of Theorem 1 up to (39), which is omitted here. In the following, we bound T_1 and T_2 in order.

Upper bound of T_1**.** We will derive a bound of T_1 using uniform covering number of $\mathcal{F}_{\mathrm{NN}}$, which is similar to that in Lemma 4. We first give the definitions of local cover and uniform covering number.

Definition 4 (Local δ-cover)**.** Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}$. Given a sequence $X = \{\mathbf{x}_k\}_{k=1}^m \in \left(\mathbb{R}^d\right)^m$, for any $\delta > 0$, the function set $\mathcal{S}_f(X)$ is a local δ-cover of F if for any $f \in \mathcal{F}$, there exits $f^* \in \mathcal{S}_f$ such that

$$\|f(\mathbf{x}_k) - f^*(\mathbf{x}_k)\|_\infty \leq \delta, \quad \forall 1 \leq k \leq m.$$

Definition 5 (Uniform covering number)**.** Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d$ to $\mathbb{R}$. Given a sequence $X = \{\mathbf{x}_k\}_{k=1}^m \in \left(\mathbb{R}^d\right)^m$, 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

$$\mathcal{N}(\delta, \mathcal{F}, m) = \max_{X \in (\mathbb{R}^D)^m} \min_{\mathcal{S}_f(X)} \{|\mathcal{S}_f(X)| : \mathcal{S}_f(X) \text{ is a local } \delta\text{-cover of } \mathcal{F}\}. \tag{50}$$

The following lemma is an analog of Lemma 4 and gives an upper bound of T_1 using network architecture $\mathcal{F}_{\mathrm{NN}}(L, p, M)$ (see a proof in Section 7.12).

Lemma 7. Under the conditions of Theorem 2, there exists a network architecture $\mathcal{F}_{\mathrm{NN}}(L, p, M)$ such that for any $\varepsilon_1 \in (0, 1)$, we have

$$\begin{aligned} \mathrm{T}_1 \leq & 8d_{\mathcal{Y}}\varepsilon_1^2 + 64 d_{\mathcal{Y}}\sigma^2 \frac{\log \mathcal{N}(\delta, \mathcal{F}_{\mathrm{NN}}, n) + 2}{n} + 16\sqrt{2} d_{\mathcal{Y}}\sigma\delta \sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}_{\mathrm{NN}}, n) + 2}{n}} \\ & + 8 d_{\mathcal{Y}}\sigma\delta + 8 L_{E_{\mathcal{Y}}^n}^2 L_\Psi^2 \mathbb{E}_{u\sim\rho} \|\Pi_{\mathcal{X}, d_{\mathcal{X}}}^n(u) - u\|_{\mathcal{X}}^2. \end{aligned} \tag{51}$$

Such a network architecture has

$$L = O(\widetilde{L} \log \widetilde{L}), \ \ p = O\left(\widetilde{p} \log \widetilde{p}\right), M = \sqrt{d_{\mathcal{Y}}} L_{E_{\mathcal{Y}}^n} R_{\mathcal{Y}}, \tag{52}$$

where $\widetilde{L}, \widetilde{p} > 0$ are integers satisfying $\widetilde{L}\widetilde{p} = \left\lceil \varepsilon_1^{-d_{\mathcal{X}}/2} \right\rceil$. The constant hidden in $O(\cdot)$ depends on $d_{\mathcal{X}}, L_{E_{\mathcal{Y}}^n}, L_{D_{\mathcal{X}}^n}, L_\Psi, B$ and M.

Upper bound of T_2**.** Using the covering number defined in Definition 5, we have the following bound of T_2.

Lemma 8. Under the conditions of Theorem 2, we have

$$\mathrm{T}_2 \leq \frac{104 d_{\mathcal{Y}} R_{\mathcal{Y}}^2}{3n} \log \mathcal{N}\left(\frac{\delta}{4 d_{\mathcal{Y}} L_{E_{\mathcal{Y}}^n} R_{\mathcal{Y}}}, \mathcal{F}_{\mathrm{NN}}, 2n\right) + 6\delta. \tag{53}$$

Lemma 8 is proved in Section 7.13 using techniques similar to those in the proof of Lemma 5. Substituting (51) and (53) into (35) gives rise to

$$
\begin{aligned}
\mathrm{I} \leq & 2L_{D_\mathcal{Y}}^2 \mathbb{E}_{\mathcal{S}_1}\mathbb{E}_{u\sim\rho}\|\Gamma_{\rm NN}\circ E_\mathcal{X}^n(u) - E_\mathcal{Y}^n\circ\Psi(u)\|_2^2 \\
= & 2L_{D_\mathcal{Y}^n}^2 \mathbb{E}_{\mathcal{S}_1}\mathrm{T}_1 + 2L_{D_\mathcal{Y}^n}^2 \mathbb{E}_{\mathcal{S}_1}\mathrm{T}_2 \\
\leq & 16 d_\mathcal{Y} L_{D_\mathcal{Y}^n}^2 \varepsilon_1^2 + 128 d_\mathcal{Y}\sigma^2 L_{D_\mathcal{Y}^n}^2 \frac{\log\mathcal{N}(\delta,\mathcal{F}_{\rm NN},n)+2}{n} + 32\sqrt{2} d_\mathcal{Y}\sigma L_{D_\mathcal{Y}^n}^2\delta\sqrt{\frac{\log\mathcal{N}(\delta,\mathcal{F}_{\rm NN},n)+2}{n}} \\
& + 16 d_\mathcal{Y}\sigma L_{D_\mathcal{Y}^n}^2\delta + 16 L_{D_\mathcal{Y}^n}^2 L_{E_\mathcal{Y}^n}^2 L_\Psi^2 \mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_\mathcal{X}}^n(u)-u\|_\mathcal{X}^2 \qquad (54) \\
& + \frac{208 d_\mathcal{Y} L_{D_\mathcal{Y}^n}^2 L_{E_\mathcal{Y}^n}^2 R_\mathcal{Y}^2}{3n}\log\mathcal{N}\left(\frac{\delta}{4 d_\mathcal{Y} L_{E_\mathcal{Y}^n} R_\mathcal{Y}},\mathcal{F}_{\rm NN},2n\right) + 12 L_{D_\mathcal{Y}^n}^2\delta \\
\leq & 16 d_\mathcal{Y} L_{D_\mathcal{Y}^n}^2 \varepsilon_1^2 + \frac{672 d_\mathcal{Y}\sigma^2 L_{D_\mathcal{Y}^n}^2 + 208 d_\mathcal{Y} L_{D_\mathcal{Y}^n}^2 L_{E_\mathcal{Y}^n}^2 R_\mathcal{Y}^2}{3n}\log\mathcal{N}(\frac{\delta}{4 d_\mathcal{Y} L_{E_\mathcal{Y}^n} R_\mathcal{Y}},\mathcal{F}_{\rm NN},2n) \\
& + (16 d_\mathcal{Y}\sigma + 12) L_{D_\mathcal{Y}^n}^2\delta + 16 L_{D_\mathcal{Y}^n}^2 L_{E_\mathcal{Y}^n}^2 L_\Psi^2 \mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_\mathcal{X}}^n(u)-u\|_\mathcal{X}^2 \qquad (55)
\end{aligned}
$$

The covering number in (55) can be bounded using the following two lemmas:

Lemma 9 (Theorem 12.2 of [2]). Let F be a class of functions from some domain Ω to $[-M,M]$. Denote the pseudo-dimension of F by $\mathrm{Pdim}(F)$. For any $\delta>0$, we have

$$
\mathcal{N}(\delta,F,m) \leq \left(\frac{2eMm}{\delta\mathrm{Pdim}(F)}\right)^{\mathrm{Pdim}(F)} \qquad (56)
$$

for $m>\mathrm{Pdim}(F)$.

Lemma 10 (Theorem 7 of [4]). Let $\mathcal{F}_{\rm NN}$ be the class of networks with L layers, width bounded by p and bounded output. There exists a universal constant C such that

$$
\mathrm{Pdim}(\mathcal{F}_{\rm NN}) \leq Cp^2L^2\log(p^2L). \qquad (57)
$$

Combing Lemma 9 and 10, we have

$$
\log\mathcal{N}\left(\frac{\delta}{4 d_\mathcal{Y} L_{E_\mathcal{Y}^n} R_\mathcal{Y}},\mathcal{F}_{\rm NN},2n\right) \leq C_{13} d_\mathcal{Y} p^2 L^2\log(p^2L)\left(\log M + \log\delta^{-1} + \log n\right) \qquad (58)
$$

for some universal constant C_{13}. Substituting (52) into (58) gives rise to

$$
\log\mathcal{N}\left(\frac{\delta}{4 d_\mathcal{Y} L_{E_\mathcal{Y}^n} R_\mathcal{Y}},\mathcal{F}_{\rm NN},2n\right) \leq C_{13} d_\mathcal{Y}\varepsilon_1^{-d_\mathcal{X}}\log^5(\varepsilon_1^{-1})\left(\log\delta^{-1}+\log n\right). \qquad (59)
$$

Substituting (59) into (55) yields

$$
\begin{aligned}
\mathrm{I} \leq & 16 d_\mathcal{Y} L_{D_\mathcal{Y}^n}^2 \varepsilon_1^2 + C_{13} d_\mathcal{Y}^2 L_{D_\mathcal{Y}^n}^2 \frac{672\sigma^2 + 208 L_{E_\mathcal{Y}^n}^2 R_\mathcal{Y}^2}{3n}\varepsilon_1^{-d_\mathcal{X}}\log^5(\varepsilon_1^{-1})\left(\log\delta^{-1}+\log n\right) \\
& + (16 d_\mathcal{Y}\sigma + 12) L_{D_\mathcal{Y}^n}^2\delta + 16 L_{D_\mathcal{Y}^n}^2 L_{E_\mathcal{Y}^n}^2 L_\Psi^2 \mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_\mathcal{X}}^n(u)-u\|_\mathcal{X}^2. \qquad (60)
\end{aligned}
$$

Setting

$$\varepsilon_1 = d_{\mathcal{Y}}^{\frac{1}{2+d_{\mathcal{X}}}} n^{-\frac{1}{2+d_{\mathcal{X}}}}, \delta = n^{-1},$$

we have

$$\mathrm{I} \leq C_4(\sigma^2 + R_{\mathcal{Y}}^2) d_{\mathcal{Y}}^{\frac{4+d_{\mathcal{X}}}{2+d_{\mathcal{X}}}} n^{-\frac{2}{2+d_{\mathcal{X}}}} \log^6 n + 16 L_{D_{\mathcal{Y}}^n}^2 L_{E_{\mathcal{Y}}^n}^2 L_{\Psi}^2 \mathbb{E}_{u\sim\rho} \|\Pi_{\mathcal{X},d_{\mathcal{X}}}^n(u) - u\|_{\mathcal{X}}^2, \tag{61}$$

where C_4 is a constant depending on $d_{\mathcal{X}}, R_{\mathcal{X}}, R_{\mathcal{Y}}, L_{E_{\mathcal{X}}^n}, L_{E_{\mathcal{Y}}^n}, L_{D_{\mathcal{X}}^n}, L_{\Psi}$. The resulting network architecture $\mathcal{F}(L, p, M)$ has

$$L = O(\widetilde{L}\log\widetilde{L}),\ p = O\left(\widetilde{p}\log\widetilde{p}\right), M = \sqrt{d_{\mathcal{Y}}} L_{E_{\mathcal{Y}}^n} R_{\mathcal{Y}} \tag{62}$$

where $\widetilde{L}\widetilde{p} = d_{\mathcal{Y}}^{-\frac{d_{\mathcal{X}}}{4+2d_{\mathcal{X}}}} n^{\frac{d_{\mathcal{X}}}{4+2d_{\mathcal{X}}}}$.

7.2.1 Putting all ingredients together

Putting (61) and (34) together gives rise to

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\mathbb{E}_{u\sim\rho}\|D_{\mathcal{Y}}^n \circ \Gamma_{\mathrm{NN}} \circ E_{\mathcal{X}}^n(u) - \Psi(u)\|_{\mathcal{Y}}^2 \\
\leq& \mathrm{I} + \mathrm{II} \\
\leq& C_4(\sigma^2 + R_{\mathcal{Y}}^2) d_{\mathcal{Y}}^{\frac{4+d_{\mathcal{X}}}{2+d_{\mathcal{X}}}} n^{-\frac{2}{2+d_{\mathcal{X}}}} \log^6 n \\
&+ 16 L_{D_{\mathcal{Y}}^n}^2 L_{E_{\mathcal{Y}}^n}^2 L_{\Psi}^2 \mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_{\mathcal{X}}}^n(u) - u\|_{\mathcal{X}}^2 + 2\mathbb{E}_{\mathcal{S}_1}\mathbb{E}_{v^*\sim\Psi_{\#}\rho}\|\Pi_{\mathcal{Y},d_{\mathcal{Y}}}^n(v^*) - v^*\|_{\mathcal{Y}}^2,
\end{aligned} \tag{63}$$

which finishes the proof.

7.3 Proof of Lemma 1

Proof of Lemma 1. We first prove (17):

$$\begin{aligned}
\|E_{\mathcal{H},d}(u) - E_{\mathcal{H},d}(\widetilde{u})\|_2^2 &= \left\|[\langle u-\widetilde{u},\phi_1\rangle_{\mathcal{H}}, ..., \langle u-\widetilde{u},\phi_d\rangle_{\mathcal{H}}]^\top\right\|_2^2 \\
&= \sum_{k=1}^{d} |\langle u-\widetilde{u},\phi_k\rangle_{\mathcal{H}}|^2 \\
&\leq \sum_{k=1}^{\infty} |\langle u-\widetilde{u},\phi_k\rangle_{\mathcal{H}}|^2 \\
&= \|u-\widetilde{u}\|_{\mathcal{H}}^2.
\end{aligned}$$

For (18), we have

$$\|D_{\mathcal{H},d}(\mathbf{a}) - D_{\mathcal{H},d}(\widetilde{\mathbf{a}})\|_{\mathcal{H}}^2 = \left\|\sum_{k=1}^{d}(a_k - \widetilde{a}_k)\phi_k\right\|_{\mathcal{H}}^2 = \|\mathbf{a} - \widetilde{\mathbf{a}}\|_2^2,$$

since $\{\phi_k\}_{k=1}^d$ is an orthonormal set. $\square$

7.4 Proof of Corollary 1

Proof of Corollary 1. We only need to derive upper bounds of $\mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_{\mathcal{X}}}(u)-u\|_2^2$ and $\mathbb{E}_{v\sim\Psi_{\#}\rho}\|\Pi_{\mathcal{Y},d_{\mathcal{Y}}}(v)-v\|_{\mathcal{Y}}^2$. Then Corollary 1 is a direct result of Theorem 1 by substituting $L_{E_{\mathcal{X}}}, L_{D_{\mathcal{X}}}, L_{E_{\mathcal{Y}}}, L_{D_{\mathcal{Y}}}$ by 1. Our proof relies on the following lemma

Lemma 11 (Theorem 4.5(ii) of [54])**.** Let $s>0$. For any $f\in\mathcal{H}^s([0,1]^D)$ with $\|f\|_{\mathcal{H}^s}<\infty$, there exists $\widetilde{f}\in\mathrm{span}(\Phi_T^r)$ such that

$$\|f-\widetilde{f}\|_\infty \le \frac{C}{r^s}, \tag{64}$$

where C is a constant depending on D and $\|f\|_{\mathcal{H}^s}$.

We first derive an upper bound of $\mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_{\mathcal{X}}}(u)-u\|_2^2$. For any $u\in\Omega_{\mathcal{X}}$, according to Lemma 11, there exists $\widetilde{u}\in\mathrm{span}(\Phi^{r_{\mathcal{X}}})$ such that

$$\|u-\widetilde{u}\|_\infty \le C_7 r_{\mathcal{X}}^{-s},$$

where C_7 is a constant depending on D and $C_{\mathcal{H}_P,\mathcal{X}}$. We deduce that

$$\begin{aligned}\|\Pi_{\mathcal{X},d_{\mathcal{X}}}(u)-u\|_{\mathcal{X}}^2 =& \min_{\bar{u}\in\mathrm{span}(\Phi_T^{r_{\mathcal{X}}})}\|\bar{u}-u\|_{\mathcal{X}}^2\\ \le&\|\widetilde{u}-u\|_{\mathcal{X}}^2\\ \le&\int_{[-1,1]^D}|\widetilde{u}-u|^2 d\boldsymbol{x}\\ \le&2^D C_7 r_{\mathcal{X}}^{-2s}\\ =&2^D C_7 d_{\mathcal{X}}^{-\frac{2s}{D}},\end{aligned}$$

where in the last equality $d_{\mathcal{X}}=r_{\mathcal{X}}^D$ is used. Therefore

$$\mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_{\mathcal{X}}}(u)-u\|_2^2 \le C_5 d_{\mathcal{X}}^{-\frac{2s}{D}}.$$

where C_5 is a constant depending on D and $C_{\mathcal{H}_P,\mathcal{X}}$. Similarly, one can show

$$\mathbb{E}_{v\sim\Psi_{\#}\rho}\|\Pi_{\mathcal{Y},d_{\mathcal{Y}}}(v)-v\|_{\mathcal{Y}}^2 \le C_6 d_{\mathcal{Y}}^{-\frac{2s}{D}},$$

where C_6 is a constant depending on D and $C_{\mathcal{H}_P,\mathcal{Y}}$. The theorem is proved. ☐

7.5 Proof of Corollary 2

Proof of Corollary 2. Our proof relies on the following lemma

Lemma 12 (Theorem 4.3(ii) of [54])**.** Let $s>0$. For any $f\in\mathcal{P}\cap\mathcal{H}^s([0,1]^D)$ with $\|f\|_{\mathcal{H}^s}<\infty$, there exists $\widetilde{f}\in\mathrm{span}(\Phi_T^r)$ such that

$$\|f-\widetilde{f}\|_\infty \le \frac{C}{r^s}, \tag{65}$$

where C is a constant depending on D and $\|f\|_{\mathcal{H}^s}$.

Corollary 2 can be proved by following the proof of Corollary 2 in which Lemma 11 is replaced by Lemma 12. ☐

7.6 Proof of Lemma 2

Proof of Lemma 2. This can be proved similarly as [7, Lemma B.3].

For $E^n_{\mathcal{H},d}$, let $\{\phi^n_k\}_{k=d+1}^\infty$ be an orthonormal basis of $\mathcal{H}\backslash V^n_{\mathcal{H},d}$. Then $\{\phi^n_k\}_{k=1}^\infty$ is an orthonormal basis of $\mathcal{H}$. We have

$$\begin{aligned}
\|E^n_{\mathcal{H},d}(u)-E^n_{\mathcal{H},d}(\widetilde{u})\|_2^2 =& \left\|[\langle\phi^n_1,u\rangle_{\mathcal{H}},...,\langle\phi^n_d,u\rangle_{\mathcal{H}}]^\top-[\langle\phi^n_1,\widetilde{u}\rangle_{\mathcal{H}},...,\langle\phi^n_d,\widetilde{u}\rangle_{\mathcal{H}}]^\top\right\|_2^2\\
=&\left\|[\langle\phi^n_1,(u-\widetilde{u})\rangle_{\mathcal{H}},...,\langle\phi^n_d,(u-\widetilde{u})\rangle_{\mathcal{H}}]^\top\right\|_2^2\\
=&\sum_{k=1}^d(\langle\phi^n_k,(u-\widetilde{u})\rangle_{\mathcal{H}})^2\\
\leq&\sum_{k=1}^\infty(\langle\phi^n_k,(u-\widetilde{u})\rangle_{\mathcal{H}})^2\\
=&\left\|\sum_{k=1}^\infty\langle\phi^n_k,(u-\widetilde{u})\rangle_{\mathcal{H}}\phi^n_k\right\|_{\mathcal{H}}^2\\
=&\|u-\widetilde{u}\|_{\mathcal{H}}^2.
\end{aligned}$$

We next prove the equality of $D^n_{\mathcal{H},d}$.

$$\begin{aligned}
\|D^n_{\mathcal{H},d}(\mathbf{a})-D^n_{\mathcal{H},d}(\widetilde{\mathbf{a}})\|_{\mathcal{H}} =& \left\|\sum_{k=1}^d a_k\phi^n_k-\sum_{k=1}^d\widetilde{a}_k\phi^n_k\right\|_{\mathcal{H}}^2\\
=&\left\|\sum_{k=1}^d(a_k-\widetilde{a}_k)\phi^n_k\right\|_{\mathcal{H}}^2\\
=&\sum_{k=1}^d(a_k-\widetilde{a}_k)^2\\
=&\|\mathbf{a}-\widetilde{\mathbf{a}}\|_2^2.
\end{aligned}$$

□

7.7 Proof of Theorem 3

Lemma 2 implies that $E^n_{\mathcal{X}}, D^n_{\mathcal{X}}, E^n_{\mathcal{Y}}, D^n_{\mathcal{Y}}$ are Lipschitz with Lipschitz constant 1. Substituting their Lipschitz constant into (13), we get

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\mathbb{E}_{u\sim\rho}\|D^n_{\mathcal{Y}}\circ\Gamma_{\mathrm{NN}}\circ E^n_{\mathcal{X}}(u)-\Psi(u)\|_{\mathcal{Y}}^2\\
\leq& C_4(\sigma^2+R_{\mathcal{Y}}^2)d_{\mathcal{Y}}^{\frac{4+d_{\mathcal{X}}}{2+d_{\mathcal{X}}}}n^{-\frac{2}{2+d_{\mathcal{X}}}}\log^6 n+C_3\mathbb{E}_{\mathcal{S}_1}\mathbb{E}_{u\sim\rho}\|\Pi^n_{\mathcal{X},d_{\mathcal{X}}}(u)-u\|_{\mathcal{X}}^2+2\mathbb{E}_{\mathcal{S}_1}\mathbb{E}_{v^*\sim\Psi_{\#}\rho}\|\Pi^n_{\mathcal{Y},d_{\mathcal{Y}}}(v^*)-v^*\|_{\mathcal{Y}}^2.
\end{aligned} \tag{66}$$

The last two terms in (66) can be bounded using the following lemma:

Lemma 13 (Theorem 3.4 of [7])**.** Let $\mathcal{H}$ be a seperable Hilbert space and γ be a probabillity measure defined on it. Let $V_{\mathcal{H},d}$ and $V_{\mathcal{H},d}^n$ be defined as in Section 5.1. Then

$$\mathbb{E}_{\{u_k\}_{k=1}^n\sim\gamma}\mathbb{E}_{u\sim\gamma}\|\Pi_{V_{\mathcal{H},d}^n}(u)-u\|_{\mathcal{H}}^2 \leq \sqrt{\frac{Cd}{n}}+\mathbb{E}_{u\sim\rho}\|\Pi_{V_{\mathcal{H},d}}(u)-u\|_{\mathcal{H}}^2, \tag{67}$$

with $C=\mathbb{E}_{u\sim\gamma}\|C_{\mathcal{H}}^n-C_{\mathcal{H}}\|_{\mathrm{HS}}^2$, where $C_{\mathcal{H}}$ and $C_{\mathcal{H}}^n$ are the covariance operator and its empirical estimation defined in Section 5.1, and $\|\cdot\|_{\mathrm{HS}}$ is the Hilbert-Schmidt norm.

For any $u\sim\rho$, we have $\|u\|_{\mathcal{X}}\leq R_{\mathcal{X}}$. Therefore

$$\mathbb{E}_{u\sim\rho}\|C_{\mathcal{X}}^n-C_{\mathcal{X}}\|_{\mathrm{HS}}^2\leq 4\mathbb{E}_{u\sim\rho}\|u\|_{\mathcal{X}}^4\leq 4R_{\mathcal{X}}^4$$

and Lemma 13 gives

$$\mathbb{E}_{\mathcal{S}_1}\mathbb{E}_{u\sim\rho}\|\Pi_{V_{\mathcal{X},d_{\mathcal{X}}}^n}(u)-u\|_{\mathcal{X}}^2\leq\sqrt{\frac{4R_{\mathcal{X}}^4 d_{\mathcal{X}}}{n}}+\mathbb{E}_{u\sim\rho}\|\Pi_{V_{\mathcal{X},d_{\mathcal{X}}}}(u)-u\|_{\mathcal{X}}^2. \tag{68}$$

Similarly, we can show that

$$\mathbb{E}_{\mathcal{S}_1}\mathbb{E}_{v\sim\Psi_{\#\rho}}\|\Pi_{V_{\mathcal{Y},d_{\mathcal{Y}}}^n}(v)-v\|_{\mathcal{Y}}^2\leq\sqrt{\frac{4R_{\mathcal{Y}}^4 d_{\mathcal{Y}}}{n}}+\mathbb{E}_{v\sim\Psi_{\#\rho}}\|\Pi_{V_{\mathcal{Y},d_{\mathcal{Y}}}}(v)-v\|_{\mathcal{Y}}^2. \tag{69}$$

Substituting (68) and (69) into (66) finishes the proof.

7.8 Proof of Theorem 4

Proof of Theorem 4. Theorem 4 can be proved by following the proof of Theorem 1 with the following changes:

- Replace $E_{\mathcal{X}}^n$ by $E_{\mathcal{X}}$.
- Replace Lemma 15 by the following one

  Lemma 14 (Theorem 1 of [11])**.** Let $s\geq 1$ be an integer. Suppose Assumption 7 holds. Assume for any $\mathbf{a}\in\mathcal{M}$, $\|\mathbf{a}\|_\infty\leq B$ for some $B>0$. These exists a FNN architecture $\mathcal{F}_{\mathrm{NN}}(L,p,K,\kappa,M)$ with $d_{\mathcal{Y}}=1$ such that for any $\varepsilon\in(0,1)$ and $f^*\in\mathcal{H}^s(\mathcal{M})$ with $\|f^*\|_{\mathcal{H}^s(\mathcal{M})}\leq R$, such an architecture gives rise to an FNN $\widetilde{f}$ with

  $$\|\widetilde{f}-f^*\|_\infty\leq\varepsilon.$$

  This architecture has

  $$L=O\left(\log\frac{1}{\varepsilon}+\log d_{\mathcal{X}}\right),\ p=O\left(\varepsilon^{-\frac{\widetilde{d}_{\mathcal{X}}}{s}}+d_{\mathcal{X}}\right),\ K=O\left(\varepsilon^{-\frac{\widetilde{d}_{\mathcal{X}}}{s}}\log\frac{1}{\varepsilon}+d_{\mathcal{X}}\log\frac{1}{\varepsilon}+d_{\mathcal{X}}\log d_{\mathcal{X}}\right),$$
  $$\kappa=\max\left\{1,B,\tau^2,\sqrt{d_{\mathcal{X}}}\right\},\ M=R.$$

  The constant hidden in $O(\cdot)$ depends on $s,\widetilde{d}_{\mathcal{X}},\log d_{\mathcal{X}},B,R,\tau$ and the surface area of $\mathcal{M}$.

□

7.9 Proof of Lemma 3

Proof of Lemma 3. Let $\mathbf{a}, \widetilde{\mathbf{a}} \in \mathbb{R}^{d_\mathcal{X}}$. We have

$$\begin{aligned}
\|\Gamma_d^n(\mathbf{a}) - \Gamma_d^n(\widetilde{\mathbf{a}})\|_2 =& \|E_\mathcal{Y}^n \circ \Psi \circ D_\mathcal{X}^n(\mathbf{a}) - E_\mathcal{Y}^n \circ \Psi \circ D_\mathcal{X}^n(\widetilde{\mathbf{a}})\|_2 \\
\leq& L_{E_\mathcal{Y}^n} \|\Psi \circ D_\mathcal{X}^n(\mathbf{a}) - \Psi \circ D_\mathcal{X}^n(\widetilde{\mathbf{a}})\|_2 \\
\leq& L_{E_\mathcal{Y}^n} L_\Psi \|D_\mathcal{X}^n(\mathbf{a}) - D_\mathcal{X}^n(\widetilde{\mathbf{a}})\|_\mathcal{Y} \\
\leq& L_{E_\mathcal{Y}^n} L_{D_\mathcal{X}^n} L_\Psi \|\mathbf{a} - \widetilde{\mathbf{a}}\|_2.
\end{aligned} \tag{70}$$

□

7.10 Proof of Lemma 4

We prove Lemma 4 by deriving bounds of both terms in (39). To derive an upper bound of the first term, we use the following lemma which shows that for any function f^* in the Hölder space $\mathcal{H}^\alpha$, when the network architecture is properly set, FNN can approximate f^* with arbitrary accuracy:

Lemma 15 (Theorem 1 of [65])**.** Let $s \geq 1$ be an integer. These exists a FNN architecture $\mathcal{F}_{\rm NN}(L, p, K, \kappa, M)$ with $d_\mathcal{Y} = 1$ such that for any $\varepsilon \in (0, 1)$ and $f^* \in \mathcal{H}^s([-B, B]^d)$ with $\|f^*\|_{\mathcal{H}^s} \leq R$, such an architecture gives rise to an FNN $\widetilde{f}$ with

$$\|\widetilde{f} - f^*\|_\infty \leq \varepsilon.$$

This architecture has

$$L = O\left(\log \frac{1}{\varepsilon}\right),\ p = O\left(\varepsilon^{-\frac{d}{s}}\right),\ K = O\left(\varepsilon^{-\frac{d}{s}} \log \frac{1}{\varepsilon}\right),\ \kappa = \max\{1, B, R\},\ M = R.$$

The constant hidden in $O(\cdot)$ depends on s, d, B, R.

According to Lemma 15, for any $\varepsilon_1 > 0$, there is a network architecture $\mathcal{F}_{\rm NN}(L, p, K, \kappa, M)$, such that for any Γ_d^n defined in (37), there exists a $\widetilde{\Gamma}_d^n \in \mathcal{F}_{\rm NN}(L, p, K, \kappa, M)$ with

$$\|\widetilde{\Gamma}_d^n - \Gamma_d^n\|_\infty \leq \varepsilon_1. \tag{71}$$

Such a network architecture has

$$\begin{aligned}
&L = O(\log \varepsilon_1),\ p = O\left(\varepsilon_1^{-d_\mathcal{X}}\right),\ K = O\left(\varepsilon_1^{-d_\mathcal{X}} \log \varepsilon_1\right), \\
&\kappa = \max\left\{1, \sqrt{d_\mathcal{Y}} L_{E_\mathcal{Y}^n} R_\mathcal{Y}, \sqrt{d_\mathcal{X}} L_{E_\mathcal{X}^n} R_\mathcal{X}, L_{E_\mathcal{Y}^n} L_{D_\mathcal{X}^n} L_\Psi\right\},\ M = \sqrt{d_\mathcal{Y}} L_{E_\mathcal{Y}^n} R_\mathcal{Y}.
\end{aligned} \tag{72}$$

We bound the first term in (39) as

$$\begin{aligned}
&\inf_{\Gamma \in \mathcal{F}_{\rm NN}} \mathbb{E}_{u\sim\rho} \|\Gamma \circ E_\mathcal{X}^n(u) - E_\mathcal{Y}^n \circ \Psi(u)\|_2^2 \\
\leq& \mathbb{E}_{u\sim\rho} \|\widetilde{\Gamma}_d^n \circ E_\mathcal{X}^n(u) - E_\mathcal{Y}^n \circ \Psi(u)\|_2^2 \\
\leq& 2\mathbb{E}_{u\sim\rho} \|\widetilde{\Gamma}_d^n \circ E_\mathcal{X}^n(u) - \Gamma_d \circ E_\mathcal{X}^n(u)\|_2^2 + 2\mathbb{E}_{u\sim\rho} \|\Gamma_d^n \circ E_\mathcal{X}^n(u) - E_\mathcal{Y}^n \circ \Psi(u)\|_2^2 \\
\leq& 2d_\mathcal{Y}\varepsilon_1^2 + 2\mathbb{E}_{u\sim\rho} \|\Gamma_d^n \circ E_\mathcal{X}^n(u) - E_\mathcal{Y}^n \circ \Psi(u)\|_2^2 \\
=& 2d_\mathcal{Y}\varepsilon_1^2 + 2\mathbb{E}_{u\sim\rho} \|E_\mathcal{Y}^n \circ \Psi \circ D_\mathcal{X}^n \circ E_\mathcal{X}^n(u) - E_\mathcal{Y}^n \circ \Psi(u)\|_2^2 \qquad \text{by the definition of } \Gamma_d \text{ in (37)} \\
\leq& 2d_\mathcal{Y}\varepsilon_1^2 + 2L_{E_\mathcal{Y}^n}^2 L_\Psi^2 \mathbb{E}_{u\sim\rho} \|D_\mathcal{X}^n \circ E_\mathcal{X}^n(u) - u\|_2^2 \\
=& 2d_\mathcal{Y}\varepsilon_1^2 + 2L_{E_\mathcal{Y}^n}^2 L_\Psi^2 \mathbb{E}_{u\sim\rho} \|\Pi_{\mathcal{X},d_\mathcal{X}}^n(u) - u\|_\mathcal{X}^2
\end{aligned} \tag{73}$$

We next bound the second term in (39). Let $\mathcal{F}^* = \{\Gamma_j^*\}_{j=1}^{\mathcal{N}(\delta,\mathcal{F}_{\rm NN},\|\cdot\|_\infty)}$ be a global δ-cover of $\mathcal{F}_{\rm NN}$, where $\mathcal{N}(\delta,\mathcal{F}_{\rm NN},\|\cdot\|_\infty)$ is the global covering number. Then there exists $\Gamma^* \in \mathcal{F}^*$ satisfying $\|\Gamma^* - \Gamma_{\rm NN}\|_\infty \le \delta$. Denote $\|\Gamma \circ E_{\mathcal{X}}^n\|_n^2 = \frac{1}{n}\sum_{i=n+1}^{2n} \|\Gamma \circ E_{\mathcal{X}}^n(u_i)\|_2^2$. We have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\langle \Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u_i), \boldsymbol{\epsilon}_i\rangle \\
=&\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\langle \Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u_i) - \Gamma^*\circ E_{\mathcal{X}}^n(u_i) + \Gamma^*\circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n \circ E_{\mathcal{X}}^n(u_i), \boldsymbol{\epsilon}_i\rangle \\
\le&\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\langle \Gamma^*\circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n \circ E_{\mathcal{X}}^n(u_i), \boldsymbol{\epsilon}_i\rangle + \mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\|\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u_i) - \Gamma^*\circ E_{\mathcal{X}}^n(u_i)\|_2\|\xi_i\|_2 \\
\le&\mathbb{E}_{\mathcal{S}_2}\frac{\|\Gamma^*\circ E_{\mathcal{X}}^n - \Gamma_d^n\circ E_{\mathcal{X}}^n\|_n}{\sqrt{n}}\frac{\sum_{i=n+1}^{2n}\langle \Gamma^*\circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n \circ E_{\mathcal{X}}^n(u_i), \boldsymbol{\epsilon}_i\rangle}{\sqrt{n}\|\Gamma^*\circ E_{\mathcal{X}}^n - \Gamma_d^n\circ E_{\mathcal{X}}^n\|_n} + d_{\mathcal{Y}}\sigma\delta \\
\le&\sqrt{2}\mathbb{E}_{\mathcal{S}_2}\frac{\|\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n\circ E_{\mathcal{X}}^n(u_i)\|_n + \sqrt{d_{\mathcal{Y}}}\delta}{\sqrt{n}}\left|\frac{\sum_{i=n+1}^{2n}\langle \Gamma^*\circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n \circ E_{\mathcal{X}}^n(u_i), \boldsymbol{\epsilon}_i\rangle}{\sqrt{n}\|\Gamma^*\circ E_{\mathcal{X}}^n - \Gamma_d^n\circ E_{\mathcal{X}}^n\|_n}\right| + d_{\mathcal{Y}}\sigma\delta,
\end{aligned}
\tag{74}
$$

where the first inequality follows from Cauchy–Schwarz inequality, the third inequality holds since

$$
\begin{aligned}
&\|\Gamma^*\circ E_{\mathcal{X}}^n - \Gamma_d^n\circ E_{\mathcal{X}}^n\|_n \\
=&\sqrt{\frac{1}{n}\sum_{i=n+1}^{2n}\|\Gamma^*\circ E_{\mathcal{X}}^n(u_i) - \Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u_i) + \Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n\circ E_{\mathcal{X}}^n(u_i)\|_2^2} \\
\le&\sqrt{\frac{2}{n}\sum_{i=n+1}^{2n}\|\Gamma^*\circ E_{\mathcal{X}}^n(u_i) - \Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u_i)\|_2^2 + \|\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n\circ E_{\mathcal{X}}^n(u_i)\|_2^2} \\
\le&\sqrt{\frac{2}{n}\sum_{i=n+1}^{2n} d_{\mathcal{Y}}\delta^2 + \|\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n\circ E_{\mathcal{X}}^n(u_i)\|_2^2} \\
\le&\sqrt{2}\|\Gamma_{\rm NN}\circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n\circ E_{\mathcal{X}}^n(u_i)\|_n + \sqrt{2d_{\mathcal{Y}}}\delta.
\end{aligned}
\tag{75}
$$

Denote $z_j = \frac{\sum_{i=n+1}^{2n}\left\langle \Gamma_j^* \circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n \circ E_{\mathcal{X}}^n(u_i), \boldsymbol{\epsilon}_i \right\rangle}{\sqrt{n}\|\Gamma_j^* \circ E_{\mathcal{X}}^n - \Gamma_d^n \circ E_{\mathcal{X}}^n\|_n}$. The expectation term in (74) can be bounded as

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}_2} \frac{\|\Gamma_{\mathrm{NN}} \circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n \circ E_{\mathcal{X}}^n(u_i)\|_n + \sqrt{d_{\mathcal{Y}}}\delta}{\sqrt{n}} \left| \frac{\sum_{i=n+1}^{2n} \langle \Gamma^* \circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n \circ E_{\mathcal{X}}^n(u_i), \boldsymbol{\epsilon}_i \rangle}{\sqrt{n}\|\Gamma^* \circ E_{\mathcal{X}}^n - \Gamma_d^n \circ E_{\mathcal{X}}^n\|_n} \right| \\
\leq & \mathbb{E}_{\mathcal{S}_2} \frac{\|\Gamma_{\mathrm{NN}} \circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n \circ E_{\mathcal{X}}^n(u_i)\|_n + \sqrt{d_{\mathcal{Y}}}\delta}{\sqrt{n}} \max_j |z_j| \\
= & \mathbb{E}_{\mathcal{S}_2} \left[\frac{\|\Gamma_{\mathrm{NN}} \circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n \circ E_{\mathcal{X}}^n(u_i)\|_n}{\sqrt{n}} \max_j |z_j| + \frac{\sqrt{d_{\mathcal{Y}}}\delta}{\sqrt{n}} \max_j |z_j| \right] \\
\leq & \mathbb{E}_{\mathcal{S}_2} \left[\sqrt{\frac{1}{n}\|\Gamma_{\mathrm{NN}} \circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n \circ E_{\mathcal{X}}^n(u_i)\|_n^2} \sqrt{\max_j |z_j|^2} + \frac{\sqrt{d_{\mathcal{Y}}}\delta}{\sqrt{n}} \sqrt{\max_j |z_j|^2} \right] \\
\leq & \sqrt{\frac{1}{n}\mathbb{E}_{\mathcal{S}_2}\|\Gamma_{\mathrm{NN}} \circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n \circ E_{\mathcal{X}}^n(u_i)\|_n^2} \sqrt{\mathbb{E}_{\mathcal{S}_2}\left[\max_j |z_j|^2\right]} + \frac{\sqrt{d_{\mathcal{Y}}}\delta}{\sqrt{n}} \sqrt{\mathbb{E}_{\mathcal{S}_2}\left[\max_j |z_j|^2\right]} \\
= & \left(\sqrt{\frac{1}{n}\mathbb{E}_{\mathcal{S}_2}\|\Gamma_{\mathrm{NN}} \circ E_{\mathcal{X}}^n(u_i) - \Gamma_d^n \circ E_{\mathcal{X}}^n(u_i)\|_n^2} + \frac{\sqrt{d_{\mathcal{Y}}}\delta}{\sqrt{n}} \right) \sqrt{\mathbb{E}_{\mathcal{S}_2}\left[\max_j |z_j|^2\right]}. \qquad (76)
\end{aligned}
$$

where the second inequality comes from Cauchy–Schwarz inequality, the third inequality comes from Jensen's inequality.

Since $\boldsymbol{\epsilon}_i \in [-\sigma, \sigma]^{d_{\mathcal{Y}}}$, each component of $\boldsymbol{\epsilon}_i$ is a sub-Gaussian variable with parameter σ. Therefore for given $u_{n+1}, ..., u_{2n}$, each z_j is a sub-gaussian variable with parameter $\sqrt{d_{\mathcal{Y}}}\sigma$. The last term is the maximum of a collection of squared sub-Gaussian variables and is bounded as

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}_2}\left[\max_j |z_j|^2 | u_{n+1}, ..., u_{2n}\right] = & \frac{1}{t}\log\exp\left(t\mathbb{E}_{\mathcal{S}_2}\left[\max_j |z_j|^2 | u_{n+1}, ..., u_{2n}\right]\right) \\
\leq & \frac{1}{t}\log\mathbb{E}_{\mathcal{S}_2}\left[\exp\left(t\max_j |z_j|^2 | u_{n+1}, ..., u_{2n}\right]\right) \\
\leq & \frac{1}{t}\log\mathbb{E}_{\mathcal{S}_2}\left[\sum_j \exp\left(t|z_j|^2 | u_{n+1}, ..., u_{2n}\right]\right) \\
\leq & \frac{1}{t}\log\mathcal{N}(\delta, \mathcal{F}_{\mathrm{NN}}, \|\cdot\|_\infty) + \frac{1}{t}\log\mathbb{E}_{\mathcal{S}_2}\left[\exp\left(t|z_1|^2 | u_{n+1}, ..., u_{2n}\right)\right]. \qquad (77)
\end{aligned}
$$

Since z_1 is sub-Gaussian with parameter σ^2, we have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}_2}\left[\exp\left(t|z_1|^2|u_{n+1},...,u_{2n}\right)\right] =&1+\sum_{k=1}^{\infty}\frac{t^k\mathbb{E}_{\mathcal{S}_2}\left[z_1^{2k}|u_{n+1},...,u_{2n}\right]}{k!}\\
=&1+\sum_{k=1}^{\infty}\frac{t^k}{k!}\int_0^{\infty}\mathbb{P}\left(|z_1|\geq\tau^{\frac{1}{2k}}|u_{n+1},...,u_{2n}\right)d\tau\\
\leq&1+2\sum_{k=1}^{\infty}\frac{t^k}{k!}\int_0^{\infty}\exp\left(-\frac{\tau^{1/k}}{2d_{\mathcal{Y}}\sigma^2}\right)d\tau\\
=&1+\sum_{k=1}^{\infty}\frac{2k(2td_{\mathcal{Y}}\sigma^2)^k}{k!}\Gamma_{\mathrm{G}}(k)\\
=&1+2\sum_{k=1}^{\infty}(2td_{\mathcal{Y}}\sigma^2)^k, \qquad (78)
\end{aligned}
$$

where Γ_{G} represents the Gamma function. Setting $t=(4d_{\mathcal{Y}}\sigma^2)^{-1}$ gives rise to

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}_2}\left[\max_j|z_j|^2|u_{n+1},...,u_{2n}\right]&\leq 4d_{\mathcal{Y}}\sigma^2\log\mathcal{N}(\delta,\mathcal{F}_{\mathrm{NN}},\|\cdot\|_{\infty})+4d_{\mathcal{Y}}\sigma^2\log 3\\
&\leq 4d_{\mathcal{Y}}\sigma^2\log\mathcal{N}(\delta,\mathcal{F}_{\mathrm{NN}},\|\cdot\|_{\infty})+6d_{\mathcal{Y}}\sigma^2. \qquad (79)
\end{aligned}
$$

Combining (79), (76), (74) gives

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}_2}\frac{1}{n}\sum_{i=n+1}^{2n}\langle\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i),\boldsymbol{\epsilon}_i\rangle\\
\leq&2\sqrt{2d_{\mathcal{Y}}}\sigma\left(\sqrt{\mathbb{E}_{\mathcal{S}_2}\|\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i)-\Gamma_d^n\circ E_{\mathcal{X}}^n(u_i)\|_n^2}+\sqrt{d_{\mathcal{Y}}}\delta\right)\sqrt{\frac{\log\mathcal{N}(\delta,\mathcal{F}_{\mathrm{NN}},\|\cdot\|_{\infty})+2}{n}}+d_{\mathcal{Y}}\sigma\delta.
\end{aligned}
\qquad (80)
$$

Substituting (73) and (80) into (39), we have

$$
\begin{aligned}
\mathrm{T}_1=&2\mathbb{E}_{\mathcal{S}_2}\|\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i)-E_{\mathcal{Y}}^n\circ\Psi(u_i)\|_n^2\\
\leq&4d_{\mathcal{Y}}\varepsilon_1^2+8\sqrt{2d_{\mathcal{Y}}}\sigma\left(\sqrt{\mathbb{E}_{\mathcal{S}_2}\|\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i)-\Gamma_d^n\circ E_{\mathcal{X}}^n(u_i)\|_n^2}+\sqrt{d_{\mathcal{Y}}}\delta\right)\sqrt{\frac{\log\mathcal{N}(\delta,\mathcal{F}_{\mathrm{NN}},\|\cdot\|_{\infty})+2}{n}}\\
&+4d_{\mathcal{Y}}\sigma\delta+4L_{E_{\mathcal{Y}}^n}^2L_{\Psi}^2\mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_{\mathcal{X}}}^n(u)-u\|_{\mathcal{X}}^2. \qquad (81)
\end{aligned}
$$

Denote

$$
\begin{aligned}
&\gamma=\mathbb{E}_{\mathcal{S}_2}\|\Gamma_{\mathrm{NN}}\circ E_{\mathcal{X}}^n(u_i)-E_{\mathcal{Y}}^n\circ\Psi(u_i)\|_n^2,\\
&a=2d_{\mathcal{Y}}\varepsilon_1^2+2d_{\mathcal{Y}}\sigma\delta+2L_{E_{\mathcal{Y}}^n}^2L_{\Psi}^2\mathbb{E}_{u\sim\rho}\|\Pi_{\mathcal{X},d_{\mathcal{X}}}^n(u)-u\|_{\mathcal{X}}^2+4\sqrt{2}d_{\mathcal{Y}}\sigma\delta\sqrt{\frac{\log\mathcal{N}(\delta,\mathcal{F}_{\mathrm{NN}},\|\cdot\|_{\infty})+2}{n}},\\
&b=2\sqrt{2d_{\mathcal{Y}}}\sigma\sqrt{\frac{\log\mathcal{N}(\delta,\mathcal{F}_{\mathrm{NN}},\|\cdot\|_{\infty})+2}{n}}.
\end{aligned}
$$

Inequality (81) implies

$$\gamma^2\leq a+2b\gamma,$$

from which we deduce that

$$(\gamma - b)^2 \le a + b^2 \Rightarrow \gamma^2 \le 2a + 4b^2.$$

Therefore,

$$\begin{aligned} \mathrm{T}_1 =& 2\gamma^2 \le 8d_{\mathcal{Y}}\varepsilon_1^2 + 64 d_{\mathcal{Y}}\sigma^2 \frac{\log \mathcal{N}(\delta, \mathcal{F}_{\mathrm{NN}}, \|\cdot\|_\infty) + 2}{n} + 16\sqrt{2} d_{\mathcal{Y}}\sigma\delta \sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}_{\mathrm{NN}}, \|\cdot\|_\infty) + 2}{n}} \\ &+ 8 d_{\mathcal{Y}}\sigma\delta + 8 L^2_{E^n_{\mathcal{Y}}} L^2_{\Psi} \mathbb{E}_{u\sim\rho} \|\Pi^n_{\mathcal{X}, d_{\mathcal{X}}}(u) - u\|^2_{\mathcal{X}}. \end{aligned} \tag{82}$$

7.11 Proof of Lemma 5

Our proof follows the proof of [11, Lemma 4.2]. Denote $g(u) = \|\Gamma_{\mathrm{NN}} \circ E^n_{\mathcal{X}}(u) - E^n_{\mathcal{Y}} \circ \Psi(u)\|_2^2$. We have $\|g\|_\infty \le 4 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}$. Then

$$\begin{aligned} \mathrm{T}_2 =& \mathbb{E}_{\mathcal{S}_2}\left[\mathbb{E}_{u\sim\rho}\left[g(u)|\mathcal{S}_2\right] - \frac{2}{n}\sum_{i=n+1}^{2n} g(u_i)\right] \\ =& 2\mathbb{E}_{\mathcal{S}_2}\left[\frac{1}{2}\mathbb{E}_{u\sim\rho}\left[g(u)|\mathcal{S}_2\right] - \frac{1}{n}\sum_{i=n+1}^{2n} g(u_i)\right] \\ =& 2\mathbb{E}_{\mathcal{S}_2}\left[\mathbb{E}_{u\sim\rho}[g(u)|\mathcal{S}_2] - \frac{1}{n}\sum_{i=n+1}^{2n} g(u_i) - \frac{1}{2}\mathbb{E}_{u\sim\rho}\left[g(u)|\mathcal{S}_2\right]\right]. \end{aligned} \tag{83}$$

A lower bound of $\frac{1}{2}\mathbb{E}_{u\sim\rho}\left[g(u)|\mathcal{S}_2\right]$ can be derived as

$$\mathbb{E}_{u\sim\rho}\left[g(u)|\mathcal{S}_2\right] = \mathbb{E}_{u\sim\rho}\left[\frac{4d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}}{4d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}} g(u)|\mathcal{S}_2\right] \ge \frac{1}{4d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}} \mathbb{E}_{u\sim\rho}\left[g^2(u)|\mathcal{S}_2\right] \tag{84}$$

Substituting (84) into (83) gives

$$\mathrm{T}_2 \le 2\mathbb{E}_{\mathcal{S}_2}\left[\mathbb{E}_{u\sim\rho}[g(u)|\mathcal{S}_2] - \frac{1}{n}\sum_{i=n+1}^{2n} g(u_i) - \frac{1}{8d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}} \mathbb{E}_{u\sim\rho}\left[g^2(u)|\mathcal{S}_2\right]\right]. \tag{85}$$

Define the set

$$\mathcal{R} = \left\{ g(u) = \|\Gamma_{\mathrm{NN}} \circ E^n_{\mathcal{X}}(u) - E^n_{\mathcal{Y}} \circ \Psi(u)\|_2^2 : \Gamma_{\mathrm{NN}} \in \mathcal{F}_{\mathrm{NN}} \right\}. \tag{86}$$

Denote $\bar{u}_i$ as an independent copy of u_i following the same distribution. We rewrite T_2 as

$$\begin{aligned} \mathrm{T}_2 \le& 2\mathbb{E}_{\mathcal{S}_2}\left[\sup_{g\in\mathcal{R}} \mathbb{E}_{\bar{u}\sim\rho} g(u) - \frac{1}{n}\sum_{i=n+1}^{2n} g(u_i) - \frac{1}{8d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}} \mathbb{E}_{u\sim\rho}\left[g^2(u)\right]\right] \\ \le& 2\mathbb{E}_{\mathcal{S}_2}\left[\sup_{g\in\mathcal{R}} \mathbb{E}_{\bar{u}\sim\rho}\left(\frac{1}{n}\sum_{i=n+1}^{2n}(g(\bar{u}_i) - g(u_i))\right) - \frac{1}{16d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}} \mathbb{E}_{u,\bar{u}\sim\rho}\left[g^2(u) + g^2(\bar{u})\right]\right] \\ \le& 2\mathbb{E}_{u,\bar{u},\xi}\left[\sup_{g\in\mathcal{R}} \frac{1}{n}\sum_{i=n+1}^{2n} \xi_i(g(u_i) - g(\bar{u}_i)) - \frac{1}{16d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}} 2}\left[g^2(u) + g^2(\bar{u})\right]\right], \end{aligned} \tag{87}$$

where ξ_i's are i.i.d. Rademacher random variables which equals to 1 or -1 with the same probability.

Let $\mathcal{R}^* = \{g_i^*\}_{i=1}^{\mathcal{N}(\delta,\mathcal{R},\|\cdot\|_\infty)}$ be a global δ-cover of $\mathcal{R}$. Then for any $g \in \mathcal{R}$, there exists $g^* \in \mathcal{R}^*$ such that $\|g - g^*\|_\infty \leq \delta$.

We next bound (87) using g^*'s. For the first term in (87), we have

$$\begin{aligned}\xi_i(g(u_i) - g(\bar{u}_i)) =& \xi_i\left(g(u_i) - g^*(u_i) + g^*(u_i) - g^*(\bar{u}_i) + g^*(\bar{u}_i) - g(\bar{u}_i)\right) \\ =& \xi_i\left(g(u_i) - g^*(u_i)\right) + \xi_i\left(g^*(u_i) - g^*(\bar{u}_i)\right) + \xi_i\left(g^*(\bar{u}_i) - g(\bar{u}_i)\right) \\ \leq& \xi_i\left(g^*(u_i) - g^*(\bar{u}_i)\right) + 2\delta. \end{aligned} \tag{88}$$

We lower bound $g^2(u) + g^2(\bar{u})$ as

$$\begin{aligned} g^2(u) + g^2(\bar{u}) =& \left(g^2(u) - (g^*)^2(u)\right) + \left((g^*)^2(u) + (g^*)^2(\bar{u})\right) - \left((g^*)^2(\bar{u}) - g^2(\bar{u})\right) \\ \geq& (g^*)^2(u) + (g^*)^2(\bar{u}) - |g(u) - g^*(u)|\,|g(u) + g^*(u)| - |g^*(\bar{u}) - g(\bar{u})|\,|g^*(\bar{u}) + g(\bar{u})| \\ \geq& (g^*)^2(u) + (g^*)^2(\bar{u}) - 16 d_{\mathcal{Y}} L_{E_{\mathcal{Y}}^n}^2 R_{\mathcal{Y}}^2 \delta. \end{aligned} \tag{89}$$

Substituting (88) and (89) into (87) gives rise to

$$\begin{aligned} \mathrm{T}_2 \leq& 2\mathbb{E}_{u,\bar{u},\xi}\left[\sup_{g^*\in\mathcal{R}^*} \frac{1}{n}\sum_{i=n+1}^{2n} \xi_i(g^*(u_i) - g^*(\bar{u}_i)) - \frac{1}{16 d_{\mathcal{Y}} L_{E_{\mathcal{Y}}^n}^2 R_{\mathcal{Y}}^2}\left[(g^*)^2(u) + (g^*)^2(\bar{u})\right]\right] + 6\delta \\ =& 2\mathbb{E}_{u,\bar{u},\xi}\left[\max_j \frac{1}{n}\sum_{i=n+1}^{2n} \xi_i(g_j^*(u_i) - g_j^*(\bar{u}_i)) - \frac{1}{16 d_{\mathcal{Y}} L_{E_{\mathcal{Y}}^n}^2 R_{\mathcal{Y}}^2}\left[(g_j^*)^2(u) + (g_j^*)^2(\bar{u})\right]\right] + 6\delta. \end{aligned} \tag{90}$$

Denote $h_j = (u_i, \bar{u}_i, \xi_i) = \xi_i(g_j^*(u_i) - g_j^*(\bar{u}_i))$. We have

$$\begin{aligned} &\mathbb{E}_{u,\bar{u},\xi}[h_j(u_i, \bar{u}_i, \xi_i)] = 0, \\ &\mathrm{Var}[h_j(u_i, \bar{u}_i, \xi_i)] = \mathbb{E}\left[h_j^2(u_i, \bar{u}_i, \xi_i)\right] \\ &\qquad = \mathbb{E}_{u,\bar{u},\xi}\left[\xi_i^2(g_j^*(u_i) - g_j^*(\bar{u}_i))^2\right] \\ &\qquad \leq 2\mathbb{E}_{u,\bar{u}}\left[(g_j^*)^2(u_i) + (g_j^*)^2(\bar{u}_i)\right]. \end{aligned}$$

Thus T_2 can be bounded as

$$\begin{aligned} &\mathrm{T}_2 \leq \widetilde{\mathrm{T}}_2 + 6\delta \\ &\text{with } \widetilde{\mathrm{T}}_2 = 2\mathbb{E}_{u,\bar{u},\xi}\left[\max_j \frac{1}{n}\sum_{i=n+1}^{2n} h_j(u_i, \bar{u}_i, \xi_i) - \frac{1}{32 d_{\mathcal{Y}} L_{E_{\mathcal{Y}}^n}^2 R_{\mathcal{Y}}^2}\frac{1}{n}\sum_{i=n+1}^{2n} \mathrm{Var}[h_j(u_i, \bar{u}_i, \xi_i)]\right]. \end{aligned}$$

Note that $\|h_j\|_\infty \leq 4 d_{\mathcal{Y}} L_{E_{\mathcal{Y}}^n}^2 R_{\mathcal{Y}}^2$. We next derive the moment generating function of h_j. For any

$0 < t < \frac{3}{4d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}}$, we have

$$\begin{aligned}
\mathbb{E}_{u,\bar{u},\xi}[\exp(t h_j(u_i,\bar{u}_i,\xi_i))] =& \mathbb{E}_{u,\bar{u},\xi}\left[1 + t h_j(u_i,\bar{u}_i,\xi_i) + \sum_{k=2}^{\infty} \frac{t^k h_j^k(u_i,\bar{u}_i,\xi_i)}{k!}\right] \\
\leq& \mathbb{E}_{u,\bar{u},\xi}\left[1 + t h_j(u_i,\bar{u}_i,\xi_i) + \sum_{k=2}^{\infty} \frac{(4d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}})^{k-2} t^k h_j^2(u_i,\bar{u}_i,\xi_i)}{2 \times 3^{k-2}}\right] \\
=& \mathbb{E}_{u,\bar{u},\xi}\left[1 + t h_j(u_i,\bar{u}_i,\xi_i) + \frac{t^2 h_j^2(u_i,\bar{u}_i,\xi_i)}{2} \sum_{k=2}^{\infty} \frac{(4d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}})^{k-2} t^{k-2}}{3^{k-2}}\right] \\
=& \mathbb{E}_{u,\bar{u},\xi}\left[1 + t h_j(u_i,\bar{u}_i,\xi_i) + \frac{t^2 h_j^2(u_i,\bar{u}_i,\xi_i)}{2} \frac{1}{1 - 4d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}} t/3}\right] \\
=& 1 + t^2 \operatorname{Var}[h_j(u_i,\bar{u}_i,\xi_i)] \frac{1}{2 - 8d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}} t/3} \\
\leq& \exp\left(\operatorname{Var}[h_j(u_i,\bar{u}_i,\xi_i)] \frac{3t^2}{6 - 8d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}} t}\right),
\end{aligned} \tag{91}$$

where the last inequality comes from $1 + x \leq \exp(x)$ for $x \geq 0$.

Then for $0 < t/n < \frac{3}{4d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}}$, we have

$$\begin{aligned}
&\exp\left(\frac{t\widetilde{\mathrm{T}}_2}{2}\right) \\
=& \exp\left(t \mathbb{E}_{u,\bar{u},\xi}\left[\max_j \frac{1}{n} \sum_{i=n+1}^{2n} h_j(u_i,\bar{u}_i,\xi_i) - \frac{1}{32 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}} \frac{1}{n} \sum_{i=n+1}^{2n} \operatorname{Var}[h_j(u_i,\bar{u}_i,\xi_i)]\right]\right) \\
\leq& \mathbb{E}_{u,\bar{u},\xi}\left[\exp\left(t \max_j \frac{1}{n} \sum_{i=n+1}^{2n} h_j(u_i,\bar{u}_i,\xi_i) - \frac{1}{32 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}} \frac{1}{n} \sum_{i=n+1}^{2n} \operatorname{Var}[h_j(u_i,\bar{u}_i,\xi_i)]\right)\right] \\
\leq& \mathbb{E}_{u,\bar{u},\xi}\left[\sum_j \exp\left(\frac{t}{n} \sum_{i=n+1}^{2n} h_j(u_i,\bar{u}_i,\xi_i) - \frac{t}{32 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}} \frac{1}{n} \sum_{i=n+1}^{2n} \operatorname{Var}[h_j(u_i,\bar{u}_i,\xi_i)]\right)\right] \\
\leq& \left[\sum_j \exp\left(\sum_{i=n+1}^{2n} \operatorname{Var}[h_j(u_i,\bar{u}_i,\xi_i)] \frac{3t^2/n^2}{6 - 8d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}} t/n} - \frac{1}{32 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}} \frac{t}{n} \operatorname{Var}[h_j(u_i,\bar{u}_i,\xi_i)]\right)\right] \\
=& \left[\sum_j \exp\left(\sum_{i=n+1}^{2n} \frac{t}{n} \operatorname{Var}[h_j(u_i,\bar{u}_i,\xi_i)] \left(\frac{3t/n}{6 - 8d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}} t/n} - \frac{1}{32 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}}\right)\right)\right],
\end{aligned} \tag{92}$$

where the first inequality follows from Jensen's inequality and the third inequality uses (91). Setting

$$\frac{3t/n}{6 - 8d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}} t/n} - \frac{1}{32 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}} = 0$$

gives $t = \frac{3n}{52d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}} < \frac{3n}{4d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}}$. Substituting our choice of t into (92) gives

$$\frac{t\widetilde{\mathrm{T}}_2}{2} \leq \log \sum_j \exp(0).$$

Therefore

$$\widetilde{\mathrm{T}}_2 \leq \frac{2}{t}\log \mathcal{N}(\delta, \mathcal{R}, \|\cdot\|_\infty) = \frac{104d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}}{3n}\log \mathcal{N}(\delta, \mathcal{R}, \|\cdot\|_\infty)$$

and

$$\mathrm{T}_2 \leq \frac{104d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}}{3n}\log \mathcal{N}(\delta, \mathcal{R}, \|\cdot\|_\infty) + 6\delta.$$

We next derive a relation between the covering number of $\mathcal{F}_{\mathrm{NN}}$ and $\mathcal{R}$. For any $g, \widetilde{g} \in \mathcal{R}$, we have

$$g(u) = \|\Gamma \circ E^n_{\mathcal{X}}(u) - E^n_{\mathcal{Y}} \circ \Psi(u)\|_2^2,\ \widetilde{g}(u) = \|\widetilde{\Gamma} \circ E^n_{\mathcal{X}}(u) - E^n_{\mathcal{Y}} \circ \Psi(u)\|_2^2$$

for some $\Gamma, \widetilde{\Gamma} \in \mathcal{F}_{\mathrm{NN}}$. We have

$$\begin{aligned}
\|g - \widetilde{g}\|_\infty =& \sup_u \left| \|\Gamma \circ E^n_{\mathcal{X}}(u) - E^n_{\mathcal{Y}} \circ \Psi(u)\|_2^2 - \|\widetilde{\Gamma} \circ E^n_{\mathcal{X}}(u) - E^n_{\mathcal{Y}} \circ \Psi(u)\|_2^2 \right| \\
=& \sup_u \left| \left\langle \Gamma \circ E^n_{\mathcal{X}}(u) - \widetilde{\Gamma} \circ E^n_{\mathcal{X}}(u), \Gamma \circ E^n_{\mathcal{X}}(u) + \widetilde{\Gamma} \circ E^n_{\mathcal{X}}(u) - 2E^n_{\mathcal{Y}} \circ \Psi(u) \right\rangle \right| \\
\leq& \sup_u \left\| \Gamma \circ E^n_{\mathcal{X}}(u) - \widetilde{\Gamma} \circ E^n_{\mathcal{X}}(u) \right\|_2 \left\| \Gamma \circ E^n_{\mathcal{X}}(u) + \widetilde{\Gamma} \circ E^n_{\mathcal{X}}(u) - 2E^n_{\mathcal{Y}} \circ \Psi(u) \right\|_2 \\
\leq& 4d_{\mathcal{Y}}L_{E^n_{\mathcal{Y}}}R_{\mathcal{Y}} \left\| \Gamma - \widetilde{\Gamma} \right\|_\infty.
\end{aligned}$$

As a result, we have

$$\mathcal{N}(\delta, \mathcal{R}, \|\cdot\|_\infty) \leq \mathcal{N}\left(\frac{\delta}{4d_{\mathcal{Y}}L_{E^n_{\mathcal{Y}}}R_{\mathcal{Y}}}, \mathcal{F}_{\mathrm{NN}}, \|\cdot\|_\infty \right).$$

and Lemma 5 is proved.

7.12 Proof of Lemma 7

The proof of Lemma 7 is the same as that of Lemma 4, except we make the following changes:

- Replace Lemma 15 by the following one

  Lemma 16 (Theorem 1.1 of [43])**.** Let $s \geq 1$ be an integer. These exists a FNN architecture $\mathcal{F}_{\mathrm{NN}}(L, p, M)$ with $d_{\mathcal{Y}} = 1$ such that for any integers $\widetilde{L}, \widetilde{p} > 0$ and $f^* \in \mathcal{H}^s([-B, B]^d)$ with $\|f^*\|_{\mathcal{H}^s} \leq R$, such an architecture gives rise to an FNN $\widetilde{f}$ with

  $$\|\widetilde{f} - f^*\|_\infty \leq C\widetilde{L}^{-\frac{2s}{d}}\widetilde{p}^{-\frac{2s}{d}}$$

  for some constant C depending on s, d, B, R. This architecture has

  $$L = O(\widetilde{L}\log \widetilde{L}),\ p = O\left(\widetilde{p}\log \widetilde{p}\right),\ M = R.$$

  The constant hidden in $O(\cdot)$ depends on s, d, B, R.

According to Lemma 16, for any $\varepsilon_1 > 0$, there is a network architecture $\mathcal{F}_{\rm NN}(L,p,M)$, such that for any $\Gamma_{\rm NN}^n$ defined in (37), there exists a $\widetilde{\Gamma}_d^n \in \mathcal{F}_{\rm NN}(L,p,M)$ with

$$\|\widetilde{\Gamma}_d^n - \Gamma_d^n\|_\infty \leq \varepsilon_1. \tag{93}$$

Such a network architecture has

$$L = O(\widetilde{L}\log \widetilde{L}),\ p = O\left(\widetilde{p}\log \widetilde{p}\right), M = \sqrt{d_{\mathcal{Y}}}L_{E_{\mathcal{Y}}^n}R_{\mathcal{Y}}, \tag{94}$$

where $\widetilde{L}, \widetilde{p} > 0$ are integers satisfying $\widetilde{L}\widetilde{p} = \varepsilon_1^{-d_{\mathcal{X}}/2}$. The constant hidden in $O(\cdot)$ depends on $d_{\mathcal{X}}, L_{E_{\mathcal{Y}}}^n, L_{D_{\mathcal{X}}^n}, L_\Psi, B$ and M.

- Replace the global δ-cover $\mathcal{F}^* = \{\Gamma_j^*\}_{j=1}^{\mathcal{N}(\delta,\mathcal{F}_{\rm NN},\|\cdot\|_\infty)}$ by a local δ-cover of $\mathcal{F}_{\rm NN}$: $\mathcal{F}^* = \{\Gamma_j^*\}_{j=1}^{\mathcal{N}(\delta,\mathcal{F}_{\rm NN},n)}$, where $\mathcal{N}(\delta,\mathcal{F}_{\rm NN},n)$ is the uniform covering number. Here the cover $\mathcal{F}^*$ depends on the samples $\{E_{\mathcal{X}}^n(u_i)\}_{i=n+1}^{2n}$. Then there exists $\Gamma^* \in \mathcal{F}^*$ satisfying $\|\Gamma^* \circ E_{\mathcal{X}}^n(u_i) - \Gamma_{\rm NN} \circ E_{\mathcal{X}}^n(u_i)\|_\infty \leq \delta$ for any $n+1 \leq i \leq 2n$.

7.13 Proof of Lemma 8

Denote $g(u) = \|\Gamma_{\rm NN} \circ E_{\mathcal{X}}^n(u) - E_{\mathcal{Y}}^n \circ \Psi(u)\|_2^2$ and let $\mathcal{S}_2' = \{u_i'\}_{i=n+1}^{2n}$ be an independent copy of $\mathcal{S}_2$. We have $\|g\|_\infty \leq 4d_{\mathcal{Y}}L_{E_{\mathcal{Y}}^n}^2 R_{\mathcal{Y}}^2$. Then

$$\begin{aligned}
\mathrm{T}_2 =& \mathbb{E}_{\mathcal{S}_2}\left[\mathbb{E}_{u\sim\rho}\left[g(u)\right] - \frac{2}{n}\sum_{i=n+1}^{2n} g(u_i)\right] \\
=& 2\mathbb{E}_{\mathcal{S}_2}\left[\frac{1}{2}\mathbb{E}_{\mathcal{S}_2'}\left[\frac{1}{n}\sum_{i=n+1}^{2n} g(u_i')\right] - \frac{1}{n}\sum_{i=n+1}^{2n} g(u_i)\right] \\
=& 2\mathbb{E}_{\mathcal{S}_2}\left[\mathbb{E}_{\mathcal{S}_2'}\left[\frac{1}{n}\sum_{i=n+1}^{2n} g(u_i')\right] - \frac{1}{n}\sum_{i=n+1}^{2n} g(u_i) - \frac{1}{2}\mathbb{E}_{\mathcal{S}_2'}\left[\frac{1}{n}\sum_{i=n+1}^{2n} g(u_i')\right]\right].
\end{aligned} \tag{95}$$

A lower bound of $\frac{1}{2}\mathbb{E}_{\mathcal{S}_2'}\left[\frac{1}{n}\sum\limits_{i=n+1}^{2n} g(u_i')\right]$ can be derived as

$$\mathbb{E}_{\mathcal{S}_2'}\left[\frac{1}{n}\sum_{i=n+1}^{2n} g(u_i')\right] = \mathbb{E}_{\mathcal{S}_2'}\left[\frac{1}{n}\sum_{i=n+1}^{2n} \frac{4d_{\mathcal{Y}}L_{E_{\mathcal{Y}}^n}^2 R_{\mathcal{Y}}^2}{4d_{\mathcal{Y}}L_{E_{\mathcal{Y}}^n}^2 R_{\mathcal{Y}}^2} g(u_i')\right] \geq \frac{1}{4d_{\mathcal{Y}}L_{E_{\mathcal{Y}}^n}^2 R_{\mathcal{Y}}^2}\mathbb{E}_{\mathcal{S}_2'}\left[\frac{1}{n}\sum_{i=n+1}^{2n} g^2(u_i')\right] \tag{96}$$

Substituting (96) into (95) gives

$$\mathrm{T}_2 \leq 2\mathbb{E}_{\mathcal{S}_2}\left[\mathbb{E}_{\mathcal{S}_2'}\left[\frac{1}{n}\sum_{i=n+1}^{2n} g(u_i')\right] - \frac{1}{n}\sum_{i=n+1}^{2n} g(u_i) - \frac{1}{8d_{\mathcal{Y}}L_{E_{\mathcal{Y}}^n}^2 R_{\mathcal{Y}}^2}\mathbb{E}_{\mathcal{S}_2'}\left[\frac{1}{n}\sum_{i=n+1}^{2n} g^2(u_i')\right]\right].$$

Define the set

$$\mathcal{R} = \left\{g(u) = \|\Gamma_{\rm NN} \circ E_{\mathcal{X}}^n(u) - E_{\mathcal{Y}}^n \circ \Psi(u)\|_2^2 : \Gamma_{\rm NN} \in \mathcal{F}_{\rm NN}\right\}.$$

We rewrite T_2 as

$$\begin{aligned}\mathrm{T}_2 \leq & 2\mathbb{E}_{\mathcal{S}_2}\left[\sup_{g\in\mathcal{R}}\left(\mathbb{E}_{\mathcal{S}_2'}\left[\frac{1}{n}\sum_{i=n+1}^{2n} g(u_i')\right]-\frac{1}{n}\sum_{i=n+1}^{2n} g(u_i)-\frac{1}{8d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}}\mathbb{E}_{\mathcal{S}_2'}\left[\frac{1}{n}\sum_{i=n+1}^{2n} g^2(u_i')\right]\right)\right]\\ \leq & 2\mathbb{E}_{\mathcal{S}_2}\left[\sup_{g\in\mathcal{R}}\left(\mathbb{E}_{\mathcal{S}_2'}\left(\frac{1}{n}\sum_{i=n+1}^{2n}(g(u_i')-g(u_i))\right)-\frac{1}{16d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}}\mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2'}\left[\frac{1}{n}\sum_{i=n+1}^{2n}\left(g^2(u_i')+g^2(u_i)\right)\right]\right)\right]\\ \leq & 2\mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}\left[\sup_{g\in\mathcal{R}}\left(\frac{1}{n}\sum_{i=n+1}^{2n}\xi_i(g(u_i)-g(u_i'))-\frac{1}{16d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}}\frac{1}{n}\sum_{i=n+1}^{2n}\left(g^2(u_i)+g^2(u_i')\right)\right)\right],\end{aligned} \tag{97}$$

465 where ξ_i's are i.i.d. Rademacher random variables which equals to 1 or -1 with the same probability.

466 Let $\mathcal{R}^*=\{g_i^*\}_{i=1}^{\mathcal{N}(\delta,\mathcal{R},2n)}$ be a local δ-cover of $\mathcal{R}$ with respect to the data set $\widetilde{\mathcal{S}}=\{u_i\}_{i=1}^n\cup\{u_i'\}_{i=1}^n$.
467 Then for any $g\in\mathcal{R}$, there exists $g^*\in\mathcal{R}^*$ such that $|g(u)-g^*(u)|\leq\delta, \forall u\in\widetilde{\mathcal{S}}$.

We next bound (97) using g^*'s. For the first term in (97), we have

$$\begin{aligned}\xi_i(g(u_i)-g(u_i')) =&\xi_i\left(g(u_i)-g^*(u_i)+g^*(u_i)-g^*(\bar{u}_i)+g^*(u_i')-g(u_i')\right)\\ =&\xi_i\left(g(u_i)-g^*(u_i)\right)+\xi_i\left(g^*(u_i)-g^*(u_i')\right)+\xi_i\left(g^*(u_i')-g(u_i')\right)\\ \leq&\xi_i\left(g^*(u_i)-g^*(u_i')\right)+2\delta.\end{aligned} \tag{98}$$

We lower bound $g^2(u_i)+g^2(u_i')$ as

$$\begin{aligned}g^2(u_i)+g^2(u_i') =& \left(g^2(u_i)-(g^*)^2(u_i)\right)+\left((g^*)^2(u_i)+(g^*)^2(u_i')\right)-\left((g^*)^2(u_i')-g^2(u_i')\right)\\ \geq&(g^*)^2(u_i)+(g^*)^2(u_i')-|g(u_i)-g^*(u_i)|\,|g(u_i)+g^*(u_i)|\\ &-\left|g^*(u_i')-g(u_i')\right|\left|g^*(u_i')+g(u_i')\right|\\ \geq&(g^*)^2(u_i)+(g^*)^2(u_i')-16d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}\delta.\end{aligned} \tag{99}$$

Substituting (98) and (99) into (97) gives rise to

$$\begin{aligned}\mathrm{T}_2 \leq&2\mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}\left[\sup_{g^*\in\mathcal{R}^*}\left(\frac{1}{n}\sum_{i=n+1}^{2n}\xi_i(g^*(u_i)-g^*(u_i'))-\frac{1}{16d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}}\frac{1}{n}\sum_{i=n+1}^{2n}\left((g^*)^2(u_i)+(g^*)^2(u_i')\right)\right)\right]\\ &+6\delta\\ =&2\mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}\left[\max_j\left(\frac{1}{n}\sum_{i=n+1}^{2n}\xi_i(g_j^*(u_i)-g_j^*(u_i'))-\frac{1}{16d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}}\frac{1}{n}\sum_{i=n+1}^{2n}\left((g_j^*)^2(u_i)+(g_j^*)^2(u_i')\right)\right)\right]\\ &+6\delta.\end{aligned}$$

Denote $h_j(u_i,u_i',\xi_i)=\xi_i(g_j^*(u_i)-g_j^*(u_i'))$. We have

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}[h_j(u_i,u_i',\xi_i)]=0,\\ &\mathrm{Var}[h_j(u_i,u_i',\xi_i)]=\mathbb{E}\left[h_j^2(u_i,u_i',\xi_i)\right]\\ &\qquad=\mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}\left[\xi_i^2(g_j^*(u_i)-g_j^*(u_i'))^2\right]\\ &\qquad\leq 2\mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2'}\left[(g_j^*)^2(u_i)+(g_j^*)^2(u_i')\right].\end{aligned}$$

Thus T_2 can be bounded as

$$
\begin{aligned}
&\mathrm{T}_2 \leq \widetilde{\mathrm{T}}_2 + 6\delta \\
&\text{with } \widetilde{\mathrm{T}}_2 = 2\mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}\left[\max_j\left(\frac{1}{n}\sum_{i=n+1}^{2n} h_j(u_i,u_i',\xi_i) - \frac{1}{32 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}}\frac{1}{n}\sum_{i=n+1}^{2n}\mathrm{Var}[h_j(u_i,u_i',\xi_i)]\right)\right].
\end{aligned}
$$

Note that $\|h_j\|_\infty \leq 4 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}$. We next derive the moment generating function of h_j. For any $0 < t < \frac{3}{4 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}}$, we have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}[\exp(t h_j(u_i,u_i',\xi_i))] =& \mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}\left[1 + t h_j(u_i,u_i',\xi_i) + \sum_{k=2}^{\infty}\frac{t^k h_j^k(u_i,u_i',\xi_i)}{k!}\right] \\
\leq& \mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}\left[1 + t h_j(u_i,u_i',\xi_i) + \sum_{k=2}^{\infty}\frac{(4 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}})^{k-2} t^k h_j^2(u_i,u_i',\xi_i)}{2\times 3^{k-2}}\right] \\
=& \mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}\left[1 + t h_j(u_i,u_i',\xi_i) + \frac{t^2 h_j^2(u_i,u_i',\xi_i)}{2}\sum_{k=2}^{\infty}\frac{(4 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}})^{k-2} t^{k-2}}{3^{k-2}}\right] \\
=& \mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}\left[1 + t h_j(u_i,u_i',\xi_i) + \frac{t^2 h_j^2(u_i,u_i',\xi_i)}{2}\frac{1}{1 - 4 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}} t/3}\right] \\
=& 1 + t^2\,\mathrm{Var}[h_j(u_i,u_i',\xi_i)]\frac{1}{2 - 8 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}} t/3} \\
\leq& \exp\left(\mathrm{Var}[h_j(u_i,u_i',\xi_i)]\frac{3t^2}{6 - 8 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}} t}\right),
\end{aligned} \tag{100}
$$

where the last inequality comes from $1 + x \leq \exp(x)$ for $x \geq 0$.

Then for $0 < t/n < \frac{3}{4 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}}$, we have

$$
\begin{aligned}
&\exp\left(\frac{t\widetilde{\mathrm{T}}_2}{2}\right) \\
=& \exp\left(t\mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}\left[\max_j \frac{1}{n}\sum_{i=n+1}^{2n} h_j(u_i,u_i',\xi_i) - \frac{1}{32 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}}\frac{1}{n}\sum_{i=n+1}^{2n}\mathrm{Var}[h_j(u_i,u_i',\xi_i)]\right]\right) \\
\leq& \mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}\left[\exp\left(t\max_j \frac{1}{n}\sum_{i=n+1}^{2n} h_j(u_i,u_i',\xi_i) - \frac{1}{32 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}}\frac{1}{n}\sum_{i=n+1}^{2n}\mathrm{Var}[h_j(u_i,u_i',\xi_i)]\right)\right] \\
\leq& \mathbb{E}_{\mathcal{S}_2,\mathcal{S}_2',\xi}\left[\sum_j \exp\left(\frac{t}{n}\sum_{i=n+1}^{2n} h_j(u_i,u_i',\xi_i) - \frac{t}{32 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}}\frac{1}{n}\sum_{i=n+1}^{2n}\mathrm{Var}[h_j(u_i,u_i',\xi_i)]\right)\right] \\
\leq& \left[\sum_j \exp\left(\sum_{i=n+1}^{2n}\mathrm{Var}[h_j(u_i,u_i',\xi_i)]\frac{3t^2/n^2}{6 - 8 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}} t/n} - \frac{1}{32 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}}\frac{t}{n}\mathrm{Var}[h_j(u_i,u_i',\xi_i)]\right)\right] \\
=& \left[\sum_j \exp\left(\sum_{i=n+1}^{2n}\frac{t}{n}\mathrm{Var}[h_j(u_i,u_i',\xi_i)]\left(\frac{3t/n}{6 - 8 d_{\mathcal{Y}} L^2_{E_{\mathcal{Y}}} R^2_{\mathcal{Y}} t/n} - \frac{1}{32 d_{\mathcal{Y}} L^2_{E^n_{\mathcal{Y}}} R^2_{\mathcal{Y}}}\right)\right)\right],
\end{aligned} \tag{101}
$$

where the first inequality follows from Jensen's inequality and the third inequality uses (100). Setting

$$\frac{3t/n}{6-8d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}t/n}-\frac{1}{32d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}}=0$$

gives $t=\frac{3n}{52d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}}<\frac{3n}{4d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}}$. Substituting our choice of t into (101) gives

$$\frac{t\widetilde{\mathrm{T}}_2}{2}\leq\log\sum_j\exp(0).$$

Therefore

$$\widetilde{\mathrm{T}}_2\leq\frac{2}{t}\log\mathcal{N}(\delta,\mathcal{R},2n)=\frac{104d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}}{3n}\log\mathcal{N}(\delta,\mathcal{R},2n)$$

and

$$\mathrm{T}_2\leq\frac{104d_{\mathcal{Y}}L^2_{E^n_{\mathcal{Y}}}R^2_{\mathcal{Y}}}{3n}\log\mathcal{N}(\delta,\mathcal{R},2n)+6\delta.$$

We next derive a relation between the covering number of $\mathcal{F}_{\mathrm{NN}}$ and $\mathcal{R}$. For any $g,\widetilde{g}\in\mathcal{R}$, we have

$$g(u)=\|\Gamma\circ E^n_{\mathcal{X}}(u)-E^n_{\mathcal{Y}}\circ\Psi(u)\|_2^2,\ \widetilde{g}(u)=\|\widetilde{\Gamma}\circ E^n_{\mathcal{X}}(u)-E^n_{\mathcal{Y}}\circ\Psi(u)\|_2^2$$

for some $\Gamma,\widetilde{\Gamma}\in\mathcal{F}_{\mathrm{NN}}$. We have

$$\begin{aligned}
\|g-\widetilde{g}\|_\infty=&\sup_u\left|\|\Gamma\circ E^n_{\mathcal{X}}(u)-E^n_{\mathcal{Y}}\circ\Psi(u)\|_2^2-\|\widetilde{\Gamma}\circ E^n_{\mathcal{X}}(u)-E^n_{\mathcal{Y}}\circ\Psi(u)\|_2^2\right|\\
=&\sup_u\left|\left\langle\Gamma\circ E^n_{\mathcal{X}}(u)-\widetilde{\Gamma}\circ E^n_{\mathcal{X}}(u),\Gamma\circ E^n_{\mathcal{X}}(u)+\widetilde{\Gamma}\circ E^n_{\mathcal{X}}(u)-2E^n_{\mathcal{Y}}\circ\Psi(u)\right\rangle\right|\\
\leq&\sup_u\left\|\Gamma\circ E^n_{\mathcal{X}}(u)-\widetilde{\Gamma}\circ E^n_{\mathcal{X}}(u)\right\|_2\left\|\Gamma\circ E^n_{\mathcal{X}}(u)+\widetilde{\Gamma}\circ E^n_{\mathcal{X}}(u)-2E^n_{\mathcal{Y}}\circ\Psi(u)\right\|_2\\
\leq&4d_{\mathcal{Y}}L_{E^n_{\mathcal{Y}}}R_{\mathcal{Y}}\left\|\Gamma-\widetilde{\Gamma}\right\|_\infty.
\end{aligned}$$

As a result, we have

$$\mathcal{N}(\delta,\mathcal{R},2n)\leq\mathcal{N}\left(\frac{\delta}{4d_{\mathcal{Y}}L_{E^n_{\mathcal{Y}}}R_{\mathcal{Y}}},\mathcal{F}_{\mathrm{NN}},2n\right).$$

and Lemma 8 is proved.

8 Conclusion

We study the generalization error of a general framework on learning maps between infinite dimensional spaces by two types of deep neural networks. The upper bound consists of network estimation error and projections error, and holds for general encoders and decoders under mild assumptions. The application of our results on some popular encoders and decoders are discussed, such as those derived from Legendre polynomials, trigonometric functions and PCA. To mitigate the curse of dimensionality, we also study two scenarios of low-dimensional structures. Our results show that in both scenarios, deep neural networks are adaptive to the low-dimensional structures and have a faster rate. Our results provide theoretical support on learning infinite dimensional maps by deep neural networks, and partially explains their empirical successes.

References

[1] A. Anandkumar, K. Azizzadenesheli, K. Bhattacharya, N. Kovachki, Z. Li, B. Liu, and A. Stuart. Neural operator: Graph kernel network for partial differential equations. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

[2] M. Anthony and P. Bartlett. Neural network learning: theoretical foundations, 1999.

[3] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993.

[4] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

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

[6] J. Berner, P. Grohs, and A. Jentzen. Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of black-scholes partial differential equations. *CoRR*, abs/1809.03062, 2018.

[7] K. Bhattacharya, B. Hosseini, N. B. Kovachki, and A. M. Stuart. Model reduction and neural networks for parametric pdes. *arXiv preprint arXiv:2005.03180*, 2020.

[8] S. Cai, Z. Wang, L. Lu, T. A. Zaki, and G. E. Karniadakis. DeepM&Mnet: Inferring the electroconvection multiphysics fields based on operator approximation by neural networks. *Journal of Computational Physics*, 436:110296, 2021.

[9] Y. Cao and Q. Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *CoRR*, abs/1905.13210, 2019.

[10] L. Q. Chen and J. Shen. Applications of semi-implicit fourier-spectral method to phase field equations. *Computer Physics Communications*, 108(2-3):147–158, 1998.

[11] M. Chen, H. Jiang, W. Liao, and T. Zhao. Nonparametric regression on low-dimensional manifolds using deep relu networks. *arXiv preprint arXiv:1908.01842*, 2019.

[12] T. Chen and H. Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917, 1995.

[13] A. Chkifa, A. Cohen, and C. Schwab. Breaking the curse of dimensionality in sparse polynomial approximation of parametric pdes. *Journal de Mathématiques Pures et Appliquées*, 103(2):400–428, 2015.

[14] A. Cohen and R. DeVore. Approximation of high-dimensional parametric pdes. *Acta Numerica*, 24:1–159, 2015.

[15] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[16] M. Deng, S. Li, A. Goy, I. Kang, and G. Barbastathis. Learning to synthesize: robust phase retrieval at low photon counts. *Light: Science & Applications*, 9(1):36, 2020.

[17] C. Duan, Y. Jiao, Y. Lai, X. Lu, and Z. Yang. Convergence rate analysis for deep ritz method. *arxiv:2103.13330*, 2021.

[18] W. E, C. Ma, and L. Wu. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019.

[19] W. E, C. Ma, and L. Wu. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 2021.

[20] Y. Fan, J. Feliu-Fabà, L. Lin, L. Ying, and L. Zepeda-Núñez. A multiscale neural network based on hierarchical nested bases. *Research in the Mathematical Sciences*, 6(2):21, 2019.

[21] Y. Fan, C. Orozco Bohorquez, and L. Ying. Bcr-net: A neural network based on the nonstandard wavelet form. *Journal of Computational Physics*, 384:1–15, 2019.

[22] M. H. Farrell, T. Liang, and S. Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181213, 2021.

[23] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[24] 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*, pages 6645–6649. IEEE, 2013.

[25] Y. Gu, J. Harlim, S. Liang, and H. Yang. Stationary density estimation of it diffusions using deep learning. *arxiv:2109.03992*, 2021.

[26] M. Hamers and M. Kohler. Nonasymptotic bounds on the l2 error of neural network regression estimates. *Annals of the Institute of Statistical Mathematics*, 58(1):131–151, 2006.

[27] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012.

[28] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[29] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *CoRR*, abs/1806.07572, 2018.

[30] Y. Jiao, G. Shen, Y. Lin, and J. Huang. Deep nonparametric regression on approximately low-dimensional manifolds. *arXiv: Statistics Theory*, 2021.

[31] Y. Khoo, J. Lu, and L. Ying. Solving parametric pde problems with artificial neural networks. *European Journal of Applied Mathematics*, 32(3):421–435, 2021.

[32] Y. Khoo and L. Ying. Switchnet: A neural network model for forward and inverse scattering problems. *SIAM Journal on Scientific Computing*, 41(5):A3182–A3201, 2019.

[33] M. Kohler and A. Krzyżak. Adaptive regression estimation with multilayer feedforward neural networks. *Nonparametric Statistics*, 17(8):891–913, 2005.

[34] M. Kohler, A. Krzyzak, and S. Langer. Estimation of a function of low local dimensionality by deep neural networks. *arxiv:1908.11140*, 2020.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[36] S. Lanthaler, S. Mishra, and G. E. Karniadakis. Error estimates for deeponets: A deep learning framework in infinite dimensions. *arXiv preprint arXiv:2102.09618*, 2021.

[37] D. Li, Z. Qiao, and T. Tang. Characterizing the stabilization size for semi-implicit fourier-spectral method to phase field equations. *SIAM Journal on Numerical Analysis*, 54(3):1653–1681, 2016.

[38] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

[39] C. Lin, Z. Li, L. Lu, S. Cai, M. Maxey, and G. E. Karniadakis. Operator learning for predicting multiscale bubble growth dynamics. *The Journal of Chemical Physics*, 154(10):104118, 2021.

[40] 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*, 2021.

[41] J. Lu and Y. Lu. A priori generalization error analysis of two-layer neural networks for solving high dimensional schrödinger eigenvalue problems. *arxiv:2105.01228*, 2021.

[42] J. Lu, Y. Lu, and M. Wang. A priori generalization analysis of the deep ritz method for solving high dimensional elliptic equations. *arxiv:2101.01708*, 2021.

[43] J. Lu, Z. Shen, H. Yang, and S. Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, to appear.

[44] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

[45] T. Luo and H. Yang. Two-layer neural networks for partial differential equations: Optimization and generalization theory. *ArXiv*, abs/2006.15733, 2020.

[46] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2017.

[47] S. Mishra and R. Molinaro. Estimates on the generalization error of physics informed neural networks (pinns) for approximating pdes. *arxiv:2006.16144*, 2020.

[48] R. Nakada and M. Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.

[49] N. H. Nelsen and A. M. Stuart. The random feature model for input-output maps between banach spaces. *arXiv preprint arXiv:2005.10224*, 2020.

[50] B. Peherstorfer and K. Willcox. Data-driven operator inference for nonintrusive projection-based model reduction. *Computer Methods in Applied Mechanics and Engineering*, 306:196–215, 2016.

[51] C. Qiao, D. Li, Y. Guo, C. Liu, T. Jiang, Q. Dai, and D. Li. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. *Nature Methods*, 18(2):194–202, 2021.

[52] Z. Qin, Q. Zeng, Y. Zong, and F. Xu. Image inpainting based on deep learning: A review. *Displays*, 69:102028, 2021.

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

[54] M. H. Schultz. Lˆ?-multivariate approximation theory. *SIAM Journal on Numerical Analysis*, 6(2):161–183, 1969.

[55] Z. Shen, H. Yang, and S. Zhang. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020.

[56] Z. Shen, H. Yang, and S. Zhang. Deep network approximation: Achieving arbitrary accuracy with fixed number of neurons. *arxiv:2107.02397*, 2021.

[57] Z. Shen, H. Yang, and S. Zhang. Deep network with approximation error being reciprocal of width to power of square root of depth. *Neural Computation*, 33(4):1005–1036, 03 2021.

[58] Z. Shen, H. Yang, and S. Zhang. Neural network approximation: Three hidden layers are enough. *Neural Networks*, 141:160–173, 2021.

[59] Z. Shen, H. Yang, and S. Zhang. Optimal approximation rate of ReLU networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, to appear.

[60] Y. Shin, J. Darbon, and G. E. Karniadakis. On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type pdes. *arxiv:2004.01806*, 2020.

[61] J. W. Siegel and J. Xu. Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *arxiv:2101.12365*, 2021.

[62] C. J. Stone. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4):1040 – 1053, 1982.

[63] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin. Deep learning on image denoising: An overview. *Neural Networks*, 131:251275, Nov 2020.

[64] Z. Wei and X. Chen. Physics-inspired convolutional neural network for solving full-wave inverse scattering problems. *IEEE Transactions on Antennas and Propagation*, 67(9):6138–6148, 2019.

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

[66] D. Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 639–649. PMLR, 06–09 Jul 2018.

[67] D. Yarotsky. Elementary superexpressive activations. *arXiv e-prints*, page arXiv:2102.10911, Feb. 2021.

[68] D. Yarotsky and A. Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13005–13015. Curran Associates, Inc., 2020.

[69] Y. Zhu and N. Zabaras. Bayesian deep convolutional encoderdecoder networks for surrogate modeling and uncertainty quantification. *Journal of Computational Physics*, 366:415–447, 2018.