Learning Theory for Transformers: An Operator-Learning Viewpoint

A THESIS SUBMITTED IN FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF
SYDNEY

PEILIN LIU

School of Mathematics and Statistics
Faculty of Science
The University of Sydney

Supervisor: Dr. Ding-Xuan Zhou

Associate Supervisors: Dr. Yiming Ying, Dr. Chenyu Wang

2 June 2026

Statement of Originality

This is to certify that the content of this thesis is my own work. This thesis has not been submitted for any other degree or purpose.

I certify that the intellectual content of this thesis is the product of my own work, and that all assistance received in preparing this thesis and all sources have been acknowledged.

Student Name: Peilin Liu

Signature: **Date:**

Authorship Attribution Statement

Chapter 2 of the thesis is based on [49], which has been accepted at Neural Computation. My contribution involved addressing technical ideas and drafting the manuscript. Professor Ding-Xuan Zhou supervised this work and gave feedback on the final manuscript. Chapter 3 of the thesis is based on [51], which has been accepted at Neural Networks. My contribution involved addressing technical ideas and drafting the manuscript. Professor Ding-Xuan Zhou supervised this work and gave feedback on the final manuscript. Chapter 4 of the thesis is based on [50], which is under review at Journal of Machine Learning Research. My contribution involved addressing technical ideas and drafting the manuscript. Professor Ding-Xuan Zhou supervised this work and gave feedback on the final manuscript. I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Student Name: Peilin Liu
Signature: **Date:**

As the supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor Name: Ding-Xuan Zhou
Signature: **Date:**

Artificial Intelligence Attribution Statement

The student used Claude Opus 4.6 for the purposes of text enhancement. The use of this generative AI tool includes spelling corrections, minor sentence restructuring, and clarity enhancement. The author confirms that where text was modified by generative AI, the content was reviewed for possible errors, inaccuracies, and bias. The author takes full responsibility

for the submitted thesis, confirms the work is their own, and has used generative AI in accordance with University guidelines and policies.

Student Name: Peilin Liu

Signature: **Date:**

Abstract

Large language models (LLMs) have reshaped the foundations of artificial intelligence research and the modes of interaction between human cognition and machine intelligence. Their influence extends further still, transforming the scientific tools through which we interrogate and model the physical world. Underlying most of these achievements and breakthroughs is a dominant architecture: the Transformer. Although the Transformer was proposed nearly a decade ago, established mathematical frameworks remain insufficient to explain the complex phenomena observed in practice with Transformer-based networks, particularly large language models. This thesis offers a principled theoretical foundation for understanding the remarkable capabilities these models exhibit, grounded in a central argument that the Transformer performs *operator learning* during pretraining over vast text corpora.

Around this central argument, we establish theoretical frameworks for in-context learning and scaling laws in Chapters 2 and 4, respectively. In Chapter 2, we consider a distributional regression task in which the Transformer, operating on a fixed query set, serves as the learning algorithm. The principal contribution of this chapter is the formalization of an input sequence to the Transformer as a realization of i.i.d. sampling with sequence-length many draws, which renders the attention mechanism a mapping from a probability space to a function space. Under certain regularity assumptions, we obtain quantitative convergence rates for both approximation and generalization, which, however, suffer from the curse of dimensionality and remain unable to explain the empirical scaling laws observed in experiment results. To further address the problem arising in Chapter 2, we dive into the algorithms of large language models, and identify a key distinction between language modeling and other data modalities such as images and speech: the input dimension d is a variable parameter of the model. In Chapter 3, we take a first step toward learning with variable input dimension, allowing d to tend to infinity. We consider a functional approximation problem using Fourier

Neural Operator–based networks and obtain a dimension-independent approximation rate when the input space is a Korobov space and the target functional is controlled by moduli of smoothness.

Finally, with Chapter 4, we combine the two streams discussed in Chapter 2 and 3 to deliver a comprehensive understanding on learning mechanisms of the Transformer structure. Compared with the distribution regression in Chapter 2, a different two-staged sampling is considered in Chapter 4 for modeling context-augmented representation in natural language processing (NLP). With an additional condition inspired by the techniques in Chapter 3, we achieve dimension-independent convergence rates for both approximation and generalization analysis of efficient Transformers. Especially, we obtain a nonparametric rate $O\left(N^{-\frac{2\xi}{2\xi+\frac{\gamma}{\gamma-1}}}\right)$ where N controls the diversity of the input training corpora, ξ controls the regularity of the regression operator and $\gamma > 1$ controls the capacity of the input probability class. This bound reveals a similar tradeoff in classical learning theory between the bless of regularity ξ and the curse of complexity $\frac{\gamma}{\gamma-1}$: when γ intends to 1, more probability distributions can be taken into the input class and $\frac{\gamma}{\gamma-1}$ goes to infinity; When ξ is larger, the regression operator is smoother, leading to faster convergence in both approximation and generalization. Our analysis reveals the nature of pretraining and in-context learning mechanisms of efficient Transformer structures in an operator learning framework. Transformers maps each context distribution to a response function for queries and with more samples from context distribution, they can recover information as much as possible to get a better response function with fixed and pretrained weights without any update.

Acknowledgements

Over the past five years, I have moved from Hong Kong to Sydney, across two cities of very different cultural character, and through a period marked by the aftermath of the pandemic and the rapid rise of artificial intelligence. Much has changed in the world, and much has changed in me. Yet throughout this time, certain things have remained constant: a curiosity toward the unknown, a fascination with the unfamiliar, and a desire to trace things back to their essential roots. These qualities have sustained me through uncertainty and have guided me to this point.

This research reported in this thesis was supported by the award of a Postgraduate Research Scholarship in Data Science to the PhD Candidate, and it was written during a period of profound change, both personal and collective. In these years, I have often found myself asking what it means to pursue enduring and meaningful work in a world transformed by accelerating technology and uncertainty. What has carried me forward is the hope of pursuing fundamental research, work substantial enough to break through the wall that stands fixed across my mind, and to follow, however imperfectly, a sense of direction toward the unknown.

Along the way, I have been fortunate to share this journey with many people whose presence has left a lasting mark on my life. In Hong Kong, I was deeply grateful for the friendship of Yang Guang, Yang Zonghao, Wu Sanyou, and Tang Wentao in Ngau Tau Kok. The meals, conversations, hikes, fireworks, and music we shared remain among my warmest memories, from evenings at Juxian Zhuang and walks to Victoria Peak and Sai Kung to watching the New Year fireworks over Victoria Harbour and singing together at Miriam Yeung's concert. In Sydney, life gradually settled into a quieter and gentler rhythm, and I was equally fortunate for the companionship of Hengrui and Jason. Our games of tennis, our wanderings through the city in search of good food, and our many conversations about research, technology, and life became an important source of balance and joy. I also remember with particular fondness the nights at IMAX and the period when AI agents were drawing

everyone's attention, when we learned together and compared notes on building our own workflows, all of us sensing the first distant signs of an enormous wave. I sincerely hope I can carry these memories with me for many years to come.

I am especially grateful to those who have taught and guided me. The time I spent studying with Chen Hai more than a decade ago remains one of my most treasured experiences. At a time when access to information was far more limited, those lessons opened a window onto the breadth and wonder of the wider world. During my doctoral years, I have benefited enormously from the guidance of Professor Ding-Xuan Zhou. Under his supervision, I learned from the ground up how to conduct research, and through his guidance I came to see more clearly both the discipline required for serious scholarship and my own shortcomings, which I hope to continue addressing in the years ahead. I am also deeply grateful to Professor Yiming Ying, from whom I learned much about the broader landscape and practical realities of academic life. Those insights have been of immeasurable value to me.

Years from now, I will no doubt have forgotten many things. But I hope I will continue to remember the people, moments, and lessons that have shaped these five years. From Hong Kong to Sydney, from 2021 to 2026, this journey has carried me through places, friendships, and transformations that I could not have anticipated when it began. Whatever lies ahead, I hope I will continue to follow that north star toward the unknown, and I hope as well that I can hold on, for as long as possible, to the many moments these years have given me. At this moment, when technological explosion makes "everything" seem almost within reach, I would like to close with the opening lines of Dickens' A Tale of Two Cities:

"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only."

Thesis Outcome

- **Peilin Liu**, Yuqing Liu, Xiang Zhou, and Ding-Xuan Zhou. "Approximation of functionals on Korobov spaces with Fourier Functional Networks." *Neural Networks* 182 (2025): 106922.
- **Peilin Liu**, and Ding-Xuan Zhou. "Generalization analysis of transformers in distribution regression." *Neural Computation* 37, no. 2 (2025): 260-293.
- Yang Ma, Dongang Wang, **Peilin Liu**, Lynette Masters, Michael Barnett, Weidong Cai, Chenyu Wang. "Symmetry Awareness Encoded Deep Learning Framework for Brain Imaging Analysis." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 742-752. Cham: Springer Nature Switzerland, 2024.
- Yang Ma, Dongang Wang, **Peilin Liu**, Michael Barnett, Ding-Xuan Zhou, Weidong Cai, and Chenyu Wang. "Multi-Scale Visual Prompting for Robust Visual Question Answering in Medical Imaging." In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3921-3925. IEEE, 2025.
- **Peilin Liu**, and Ding-Xuan Zhou. "Ghost in the Kernel: In-Context Learning with Efficient Transformers via Domain Generalization." *Under Review at the Journal of Machine Learning Research.*
- Dongang Wang, **Peilin Liu**, Hengrui Wang, Heidi Beadnall, Kain Kyle, Linda Ly, Mariano Cabezas, Geng Zhan, Ryan Sullivan, Weidong Cai, Wanli Ouyang, Fernando Calamante, Michael Barnett, Chenyu Wang. "How Much Data are Enough? Investigating Dataset Requirements for Patch-Based Brain MRI Segmentation Tasks." *arXiv preprint arXiv:2404.03451 (2024).*

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Transformers and NLP: The Road to Machine Intelligence

Transformers [101] have fundamentally reshaped the NLP research landscape over the past decade, yielding remarkable breakthroughs in language modeling [14, 75, 76], code generation [33] and, most notably, intimating a viable path toward the autonomous self-evolution of machine intelligence [66]. These achievements distinguish Transformer-based network architectures from preceding families of neural network structures like fully connected neural networks, convolutional neural networks (CNNs) [42], and long-short term memory (LSTM) [26], which have demonstrated notable limitations in scaling to the massive pretraining regimes that underpin modern language models [14, 75, 76]. To investigate the origins of the Transformer's exceptional capacity, it is necessary to examine the historical development of natural language processing, particularly those involving large-scale corpus processing.

Natural language processing has undergone several paradigmatic shifts since its emergence in the mid-twentieth century. Early approaches were primarily rule-based, relying on handcrafted grammars and symbolic systems to parse and generate languages [104]. These methods were proved inadequate when confronted with the polysemy and complexity of natural languages. The subsequent emergence of statistical methods marked a decisive turning point: researchers employed probabilistic models, such as hidden Markov models for sequence labelling and n-gram language models for ***next-word prediction*** [55], across a range of core NLP tasks, demonstrating that data-driven approaches could substantially outperform their rule-based predecessors. The advent of neural network-based methods further transformed

the field, beginning with distributed ***word embedding*** techniques such as Word2Vec [9] and GloVe [69], which demonstrated that rich semantic relationships could be encoded within continuous vector spaces through unsupervised learning over large corpora. Leveraging these dense representations as input, deeper architectures such as LSTMs enabled end-to-end representation learning with better performances on sequential modelling. Nevertheless, these recurrent architectures suffered from two critical limitations: the inherently sequential nature of their computation hindered efficient parallelization, while the gradient vanishing problem continued to impede the effective capture of ***long-range dependencies***, even with gating mechanisms. These cumulative techniques and their limitations have collectively driven the research community toward a new paradigm, which this thesis characterizes through high-dimensional domain generalization tasks with Transformer-based architectures. This paradigm suggests that the remarkable power of LLMs stems from the intricate interplay among ***intensive pretraining tasks***, ***scalable token embedding representations*** [1], and ***the Transformer architecture itself***. From this perspective, to elucidate the underlying mechanism of LLMs, we formulate a two-staged regression pretraining task in Chapter 2, develop a high-dimensional learning framework in Chapter 3, and investigate the interaction between them with an efficient Transformer architecture in Chapter 4.

Before presenting our contributions to the theoretical understanding of Transformer-based language models, we first introduce the standard Transformer architecture and two core concepts in LLM applications: scaling law [36, 27] and in-context learning [105]. The original Transformer was designed for machine translation tasks with both an encoder and a decoder, which were later inherited separately by BERT [14] and GPT [75]. In this thesis, we focus our analysis on the Transformer encoder only. The standard Transformer consists of blocks of attention mechanisms and shallow networks to process sequential inputs. Let the input sequence $Q = [x_1, \cdots, x_n]^T$ with token vectors $x_i \in \mathbb{R}^d$ for $1 \leq i \leq n$. Then Q is an input sequence of length n with feature dimension d. The softmax attention is defined as, for

[1] In practice, words and tokens constitute distinct units in natural language processing; however, for convenience, we do not distinguish between the two concepts in this thesis.

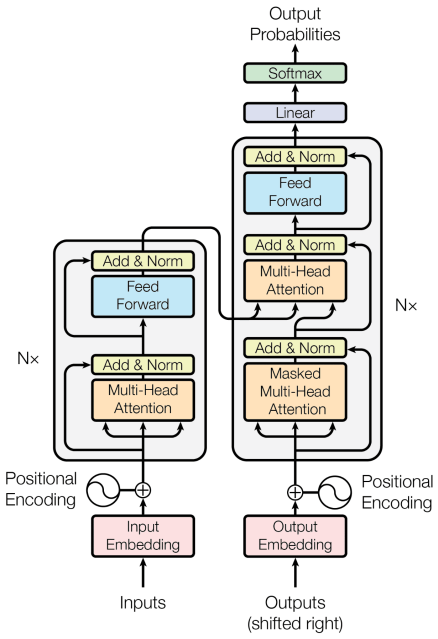


FIGURE 1.1. Encoder and Decoder Structure in Transformer [101]

$1 \leq i \leq n$,

$$\mathrm{SoftmaxAttn}(x_i|Q) = \frac{\sum_{j=1}^{n} \mathrm{sim}(x_i, x_j)(W_v x_j)}{\sum_{j=1}^{n} \mathrm{sim}(x_i, x_j)} \in \mathbb{R}^d \tag{1.1}$$

with $\mathrm{sim}(x_i, x_j) = \exp\left(\frac{\langle W_q x_i, W_k x_j \rangle}{\sqrt{d'}}\right)$ where $W_v \in \mathbb{R}^{d \times d}, W_q \in \mathbb{R}^{d' \times d}, W_k \in \mathbb{R}^{d' \times d}$ are parameter matrices for *value*, *query*, and *key* token vectors respectively. Intuitively, the attention mechanism $\mathrm{SoftmaxAttn}$ takes the input sequence Q as *context* and produces a refined context-aware representation $\mathrm{SoftmaxAttn}(x_i|Q)$ for each query token x_i in Q. Composing with a shallow neural network, we can obtain a Transformer block: the Transformer can be expressed as

$$\mathrm{Encoder}(x_i|Q) = W_2\, \sigma\Big(W_1\Big(\mathrm{SoftmaxAttn}(x_i|Q)\Big) + b_1\Big) + b_2 \in \mathbb{R}^d$$

with $W_1 \in \mathbb{R}^{d'' \times d}$, $W_2 \in \mathbb{R}^{d'' \times d}$ and $b_1 \in \mathbb{R}^{d''}$, $b_2 \in \mathbb{R}^d$. From the above definition, we can see the differences of the Transformer structure compared with CNNs and LSTM. Transformers model dependencies between tokens through a similarity function, rather than through fixed kernel weights as in CNNs, and are permutation-invariant in the absence of

positional encoding, enabling the model to capture long-range interactions without being affected by distance. With these architectures and intensive pretraining stages, Transformers have been shown to be scalable: a larger training-time compute budget yields consistently improved model performance, a phenomenon commonly referred to as the scaling law [27]. Another capability emerging from large-scale pretraining stages is in-context learning where a pretrained Transformer with fixed parameter weights can generate better predictions when provided with additional demonstrations or a more detailed prompt. Both the scaling law and in-context learning have been instrumental in driving the consistent improvement of large language models, collectively establishing a reinforcing cycle in which increased compute and data scale yields not only better performance but also greater capacity to leverage context information at inference time. Explaining the origin of in-context learning from a mathematical perspective is the central key to understanding what distinguishes Transformers from other architectures.

1.2 Learning Theory for Transformers

In this thesis, we establish a systematic framework for understanding the underlying mechanism of LLMs, particularly focusing on the interaction among pretraining tasks, high-dimensional representations, and Transformer-based architectures. In general, we adopt an operator-learning viewpoint to characterize the generalization capacity of pretrained Transformers across diverse tasks, a setting that classical learning theory is insufficient to explain. There are three questions that we seek to answer within our theoretical framework, which ultimately lead us to the essence of Transformer-based models:

- How should "context" be formally understood within a data generation process?
- How can scalable high-dimensional token representations be effectively handled as input?
- What constitutes "in-context learning" and what is its relationship to pretraining objectives with Transformers?

For the first question, "context" may be one of the most frequently used words with LLMs and has become the core idea in developing agentic frameworks. In the above formulation, the context of x_i refers to input sequence Q. In practice, a model with a longer context window means a better performance. To characterize this phenomenon in a rigorous mathematical language, we consider Q to be a realization of n i.i.d. samples from a probability distribution $\mathcal{P}$ and formulate the data generalization process as a two-staged sampling in Chapter 2. This assumption makes the connection between each token x_i and context Q more clear that Q provides a discrete approximation to the underlying distribution from which x_i's are sampled. With more and more samplings, probability distribution $\mathcal{P}$ can be recovered from context Q.

For the second question, it's a significant point to keep in mind that the data structure in NLP differs fundamentally from that of images and other sensory modalities: there's no isolated "camera" for natural languages. For image data like natural images, magnetic resonance imaging or computed tomography, one important step is imaging that transforms a physical object into its digital representation. With these digital forms as inputs, we could apply algorithms for object detection, segmentation and medical question-answering. Language, however, as a high-level abstraction, doesn't have such an isolated digital translation. While a raw text is first discretized into token sequences by a tokenizer, the resulting vocabulary constitutes an arbitrary symbolic system rather than a physical measurement. The subsequent mapping from these discrete tokens into continuous vector space in the form of token embedding, which is learned in the pretraining stage of LLMs and thus makes the embedding dimension d of the input a tunable parameter. In Chapter 3, we demonstrate that neural networks exploit latent data structure to achieve parameter-efficient approximations with scalable input dimension d. This result establishes that neural networks can harness the expressive power of scalable high-dimensional representations while circumventing the curse of dimensionality by exploiting the latent feature space.

For the third question, in-context learning has always been the key difference between Transformers and other network structures, since its emergence from large-scale pretrained language models. Understanding in-context learning and its relationship to the pretraining stage is essential, as it illuminates what occurs during pretraining with Transformers. However,

the existing literature has largely focused on constructing training samples for in-context learning, overlooking the fact that in-context learning is a capacity that emerges from a pretrained model rather than a learning process in its own right. It is this relationship to the pretraining stage that holds the key to understanding the phenomenon of in-context learning. In Chapter 4, we combine the ideas from Chapter 2 and Chapter 3 to model the pretraining task of linear Transformers as learning an operator $\Phi : \mathcal{B}_{\mathcal{X}} \to \mathcal{H}_k \otimes \mathbb{R}^d$ where $\mathcal{B}_{\mathcal{X}}$ is the class of the context distributions as defined in Chapter 2 and $\mathcal{H}_k \otimes \mathbb{R}^d$ is a vector-valued function space. For each input (x_i, Q), we first take the empirical distribution δ_Q generated by the elements in Q and maps δ_Q to $\Phi(\delta_Q) \in \mathcal{H}_k \otimes \mathbb{R}^d$. Then for each token x_i in Q, we generate a context-augmented output $\Phi(\delta_Q)(x_i) \in \mathbb{R}^d$. With this viewpoint, a pretraining stage is designed as a two-staged sampling regression task for learning the target operator Φ with a linear Transformer $\hat{\Phi}$. With $\hat{\Phi}$ held fixed, providing more samplings from $\mathcal{P}$ enables $\hat{\Phi}(\delta_Q)$ to produce a progressively better approximation to $\Phi(\mathcal{P})$, which is exactly the phenomenon of in-context learning.

Chapter 2

Two-Staged Sampling Process

2.1 Introduction

Transformers [101, 121, 53, 7, 74] have undeniably become a fundamental component of modern deep learning models, extending the influence beyond the realms of natural language processing (NLP) and computer vision (CV). Transformer-based large models like GPT 4 [67], demonstrate remarkable capabilities to process multimodal inputs with texts and images, and scientific research tools like AlphaFold [35] are created to explore the patterns hidden in complex biological data. With the rapid developments of deep learning methods, numerous techniques for Transformers have been proposed to enhance the performance of LLMs across diverse applications. For example, techniques such as prompt tuning [45, 31] and the integration of adapter modules [29, 30] are employed to adapt a pretrained LLM to new tasks at a low computational cost; As the size of the training text corpora increases dramatically, the network complexity can be efficiently scaled up using mixture of expert methods [16, 87, 32] for superior performance. Despite the impressive success in practical applications, there remains a deficiency in theoretical frameworks to demonstrate the reasons why Transformer-based models and those techniques work efficiently across diverse domains.

In recent years, mathematical theories around deep fully connected networks (FNNs) [109, 83] and CNNs [118, 120, 57] have been established to investigate their approximation and generalization abilities. However, for transformer-based networks, due to the complex input data structures and network architectures, it is challenging to build a theoretical framework to study the transformer structures and the phenomena observed in practical applications. In

this chapter, we establish a rigorous mathematical framework to demonstrate the learning capabilities of Transformers from a viewpoint of distribution regression, and also provide theoretical foundations and justifications for those practical techniques with Transformers.

In the history of NLP, modeling problems in NLP with probabilistic tools is a classical approach. Here, we utilize a two-stage sampling process in distribution regression to formulate problems. In our distribution regression model, the inputs are distribution samples on the space $(P(\Omega), \gamma_k)$, where $P(\Omega)$ denotes the set of all Borel probability measures defined on a compact subset Ω of $\mathbb{R}^d$ and γ_k is a kernel embedding distance, also known as the maximum mean discrepancy (MMD) [63]. However, we assume that the distribution samples cannot be observed directly and that our observations are the data generated by a two-stage sampling process. In the first stage of sampling, a dataset $D = \{(\mu_i, y_i)\}_{i=1}^{m_1}$ is $i.i.d.$ sampled from a meta Borel distribution ρ on $\mathcal{U} \times \mathcal{Y}$, where $\mathcal{U} = P(\Omega)$ and $\mathcal{Y} = \mathbb{R}$ is the output space. In the second stage of sampling, the dataset is $\hat{D} = \{(\{x_{i,j}\}_{j=1}^{m_{2,i}}, y_i)\}_{i=1}^{m_1}$, where $\{x_{i,j} \in \Omega\}_{j=1}^{m_{2,i}}$ are $i.i.d.$ sampled from the probability measure μ_i, one of the first stage samples. We denote the empirical distribution of μ_i by $\hat{\mu}_i^{m_{2,i}} = \frac{1}{m_{2,i}} \sum_{j=1}^{m_{2,i}} \delta_{x_{i,j}}$, where δ_x is a Dirac measure. Then the second-stage sampling dataset can be denoted by $\hat{D} = \{(\hat{\mu}_i^{m_{2,i}}, y_i)\}_{i=1}^{m_1}$. By choosing an appropriate hypothesis space $\mathcal{H}$, the distribution regression scheme can be described as

$$\varphi_{\hat{D},\mathcal{H}} := \arg\min_{\varphi \in \mathcal{H}} \frac{1}{m_1} \sum_{i=1}^{m_1} \left(\varphi(\hat{\mu}_i^{m_{2,i}}) - y_i\right)^2. \tag{2.1}$$

The learning algorithm (2.1) for distribution regression is similar with the setting of domain generalization problems [5, 28]. Both learning algorithms use the empirical distributions as prediction inputs. However, the key difference is that hypothesis functions in domain generalization also takes a sample as an input besides the target empirical distribution, and the learning scheme is formally defined as $\varphi'_{\hat{D},\mathcal{H}} := \arg\min_{\varphi' \in \mathcal{H}} \frac{1}{m_1} \sum_{i=1}^{m_1} \frac{1}{m_{2,i}} \sum_{j=1}^{m_{2,i}} \left(\varphi'(\hat{\mu}_i^{m_{2,i}}, x_{i,j}) - y_{i,j}\right)^2$ where $\mathcal{H}$ is a reproducing kernel Hilbert space defined on $P(\Omega) \times \Omega$, which is inconsistent with the sequential modelling case in NLP.

Recently, some progress has been achieved in two-stage distribution regression with neural networks [89, 110], none of which, however, matches the architecture of Transformers. Prior to the era of the prominence of neural networks, a well-known approach was based on kernel mean embedding techniques and kernel ridge regression [98, 17, 111]. All these works consider a regularized empirical risk minimization algorithm and that the regression function belongs to a function space characterized by the integral operator induced by a kernel function. Under assumptions on regularization of the function space and integral operator techniques, some nice generalization bounds are obtained. Afterward, the study of distribution regression focused on the application of neural networks. [89] and [110] proposed network architectures with FNNs and deep CNNs to learn the two-stage distribution regression respectively. These works metrize $P(\Omega)$ with a Wasserstein distance W_p and learn functionals with only polynomial features. Yet, their methods don't integrate information of the input domain $(P(\Omega), W_p)$ into the network structure and suffer from a potential information loss by only encoding polynomial features. To this end, we propose a two-stage distribution regression framework with Transformer-based network structures, which combines both advantages of the classical kernel embedding techniques and the neural network methods.

In this work, we investigate the learning capabilities of Transformer-based networks in a two-stage distribution regression framework. Our main contributions in this chapter are listed as follows:

- We first utilize a two-stage sampling process to understand the processing of Transformers with natural languages. We also propose a novel operator called attention operator to study the behavior of the attention layers in Transformers. We also prove that the attention operator can embed distributions into function representations, without any loss of information.
- We then introduce the architecture of Transformer encoders based on our novel attention operator and establish a rigorous distribution regression framework for Transformer-based networks. The approximation rate and generalization bound for the distribution regression problem are obtained, which exhibits the remarkable expressivity of Transformers in learning more diverse features than FNNs and CNNs.

- We provide a theoretical intuition for the choice of query sets in practical applications. There exists a universal choice for various tasks, though it may suffer from the curse of dimension in high-dimensional cases. This result justifies some task-specific tuning methods and cross-modal alignment tricks.
- We establish theoretical insights for the design of FNN layers based on approximation and generalization analysis of Transformer encoders. We show that task-specific features are learned by the trainable FNN layer with a fixed attention layer, which provides theoretical foundations for adapter tuning with pretrained LLMs. We also illustrate that the complexity of the FNN layer should scale up with the training size of the text corpora to achieve a great generalization performance.

The remainder of the chapter is organized as follows. Section 2.2 provides the motivation and basic definitions of the learning problem. Within this section, subsection 2.2.1 introduces the basic structure of the vanilla Transformer [101] and gives the formal definition of our self-attention operator. Subsection 2.2.2 demonstrates the definitions of $(P(\Omega), \gamma_k)$ and the structure of Transformer encoders for distribution regression. Subsequently, Section 2.3 contains the main results. First, we show in subsection 2.3.1 an approximation rate of a functional class induced by Barron functionals, by the Transformer-based network, then establish an oracle inequality and finally obtain a generalization bound for distribution regression in subsection 2.3.2. Based on the aforementioned theoretical results, we explain the principles behind the successful practical strategies, such as prompt tuning, adapters, and efficient scaling in Section 2.4. Finally, Section 2.5 presents the proof details of the main results.

2.2 Motivation and Definitions

In this section, we start with the attention operator, a fundamental component in our network structure, which is motivated by the self-attention layers in the vanilla Transformer. The

section then continues with the basic settings for the generalization analysis of distribution regression, focusing on the metric of the input space and the definition of our novel Transformer encoder for distribution regression.

2.2.1 Attention operator

The original Transformer was proposed for machine translation, consisting of an encoder and a decoder. However, with recent developments, it has become quite common to apply only a decoder (e.g., GPT [77]) or an encoder (e.g., BERT [14]) in practical applications. In this work, we focus our analysis on a shallow Transformer encoder, which could be described as a composition of a self-attention layer and a position-wise fully connected layer. Now we present a mathematical definition of the Transformer introduced by [101]. Let $Q \in \mathbb{R}^{n\times d}$ and $Q^T = (x_1, x_2, \ldots, x_n)$ with $x_i \in \mathbb{R}^d$ for $1 \le i \le n$, indicating that Q is an n-length input sequence with feature dimension d. The single-head attention is defined as

$$\mathrm{SoftmaxAttn}(x_i) = \sum_{j=1}^{n} \frac{\exp\left(\frac{\langle W_q x_i, W_k x_j\rangle}{\sqrt{d_{in}}}\right)}{\sum_{j'=1}^{n}\exp\left(\frac{\langle W_q x_i, W_k x_{j'}\rangle}{\sqrt{d_{in}}}\right)}(W_v x_j)$$

for each row vector x_i^T in the input sequence Q, where $W_k \in \mathbb{R}^{d_{in}\times d}, W_q \in \mathbb{R}^{d_{in}\times d}, W_v \in \mathbb{R}^{d\times d}$ are parameter matrices. Then the output of a self-attention layer is defined as

$$\mathrm{Softmax\,Attn}(Q) = \begin{bmatrix} \mathrm{Softmax\,Attn}^T(x_1) \\ \vdots \\ \mathrm{Softmax\,Attn}^T(x_n) \end{bmatrix} \in \mathbb{R}^{n\times d} \tag{2.2}$$

and it is followed by a position-wise FNN in the form of

$$\mathrm{FNN}(V) = \begin{bmatrix} \sigma(v_1^T W_1 + b_1)W_2 + b_2 \\ \vdots \\ \sigma(v_n^T W_1 + b_1)W_2 + b_2 \end{bmatrix} \in \mathbb{R}^{n\times d} \text{ for } V = \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix} \in \mathbb{R}^{n\times d}$$

where $W_1 \in \mathbb{R}^{d\times d'_{in}}, W_2 \in \mathbb{R}^{d'_{in}\times d}$ are connection matrices and $b_1 \in \mathbb{R}^{d'_{in}}, b_2 \in \mathbb{R}^d$ are bias vectors. So an encoder of the Transformer can be expressed as

$$\text{Encoder}(x_i) = \sigma\left(\left(\frac{1}{n}\sum_{j=1}^{n} k_{attn}(x_i, x_j)(x_j^T W_v)\right) W_1 + b_1\right) W_2 + b_2 \in \mathbb{R}^{1\times d}$$

where

$$k_{attn}(x_i, x_j) := \frac{\exp\left(\frac{\langle W_q x_i, W_k x_j\rangle}{\sqrt{d_{in}}}\right)}{\sum_{j'=1}^{n}\exp\left(\frac{\langle W_q x_i, W_k x_{j'}\rangle}{\sqrt{d_{in}}}\right)}. \tag{2.3}$$

Hence the output of a Transformer encoder is

$$\text{Encoder}(Q) = \begin{bmatrix} \sigma\left(\left(\frac{1}{n}\sum_{j=1}^{n} k_{attn}(x_1, x_j)(x_j^T W_v)\right) W_1 + b_1\right) W_2 + b_2 \\ \vdots \\ \sigma\left(\left(\frac{1}{n}\sum_{j=1}^{n} k_{attn}(x_n, x_j)(x_j^T W_v)\right) W_1 + b_1\right) W_2 + b_2 \end{bmatrix} \in \mathbb{R}^{n\times d}$$

for $Q^T = \begin{bmatrix} x_1, \cdots, x_n \end{bmatrix} \in \mathbb{R}^{d\times n}$.

By examining each row of the output, it becomes clear that the encoder of Transformers has the structure of a fully connected layer following a self-attention operation. The self-attention operation is a weighted sum of the input features $\{W_v x_j\}_{j=1}^n$, where the weights are determined by a specific kernel function. Therefore, given a feature mapping $f : \mathbb{R}^d \to \mathbb{R}$ for the input sequence $Q \in \mathbb{R}^{n\times d}$ and a kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the self-attention can be written as

$$\text{k-Attn}(x_i) = \frac{1}{n}\sum_{j=1}^{n} k(x_i, x_j) f(x_j).$$

In the original Transformer, $k(x_i, x_j) = k_{attn}(x_i, x_j)$ and $f(x) = W'_v x$ with $W'_v \in \mathbb{R}^d$. Moreover, with the empirical distribution $\hat{\mu}^n = \frac{1}{n}\sum_{j=1}^{n}\delta_{x_j}$, we have

$$\text{k-Attn}(x_i) = \int k(x_i, x) f(x)\, d\hat{\mu}^n.$$

Then it follows that $\text{k-Attn}(\cdot)$ is an empirical form of $\int k(\cdot, x) f(x)\, d\mu$ where μ is the probability distribution that $\{x_j\}_{j=1}^n$ are drawn from. Consider Q as a realization of the distribution

μ, i.e., a collection of samples drawn from μ, and then, we define the attention operator on $P(\Omega)$ as

$$\mathrm{attn}(\mu) = \int_{\Omega} k(\cdot, x) f(x)\, d\mu \tag{2.4}$$

with some conditions on k and f, specified in the subsequent section. In contrast to $\mathrm{SoftmaxAttn}(Q) \in \mathbb{R}^{n\times d}$ for $Q \in \mathbb{R}^{n\times d}$, the operator attn maps a probability measure to a function. However, a form similar to $\mathrm{SoftmaxAttn}(Q)$ can be obtained through a discretization of the function $\mathrm{attn}(\mu)$. To make the attention operator well-defined, in the next subsection we will introduce a metric on $P(\Omega)$, conditions on kernel k and feature mapping f, and the form of functionals to be learned.

2.2.2 Metric space of $P(\Omega)$ and network structure

The choice of distances between probability measures is fundamental and has found many applications in deep learning. Many well-known distances share the following similar form. Let Ω be a compact subset of $\mathbb{R}^d$ and $P(\Omega)$ be the set of all Borel probability measures on Ω. For $\mathcal{P}, \mathcal{Q} \in P(\Omega)$, the distance $\gamma_{\mathcal{F}}$ between $\mathcal{P}$ and $\mathcal{Q}$ is defined as

$$\gamma_{\mathcal{F}}(\mathcal{P}, \mathcal{Q}) = \sup_{g\in\mathcal{F}} \left| \int_{\Omega} g\, d\mathcal{P} - \int_{\Omega} g\, d\mathcal{Q} \right|$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions on Ω. One can easily observe that $\gamma_{\mathcal{F}}$ satisfies all the conditions for a metric, except for one that $\mathcal{P} = \mathcal{Q}$ if $\gamma_{\mathcal{F}}(\mathcal{P}, \mathcal{Q}) = 0$. But with an appropriate choice of $\mathcal{F}$, $\gamma_{\mathcal{F}}$ can be made a metric on $P(\Omega)$, for example, let $C(\Omega)$ be the space of continuous functions on Ω, and take $\mathcal{F} = \{g \in C(\Omega) : \|g\|_{\infty} \le 1\}$ where $\|g\|_{\infty} = \sup_{x\in\Omega} |g(x)|$. It's particularly worth mentioning that the Wasserstein distance W_1 considered in [110, 89] is also an instance when $\mathcal{F} = \{g \in C(\Omega) : |g|_{C^{0,1}} \le 1\}$ with the Lipschitz semi-norm $|g|_{C^{0,1}} := \sup_{x\neq y\, \in\Omega} \frac{|g(x)-g(y)|}{\|x-y\|_2}$.

In this work, we consider $\mathcal{F}$ to be the unit ball of a reproducing kernel Hilbert space $\mathcal{H}_k$ with a reproducing kernel k on $\Omega \times \Omega$, that is, $\mathcal{F} = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \le 1\}$, and denote $\gamma_k := \gamma_{\{f\in\mathcal{H}_k : \|f\|_{\mathcal{H}_k}\le 1\}}$. Throughout the chapter, we always assume that k is a Mercer kernel,

that is, a symmetric, continuous and positive semi-definite kernel, on $\Omega \times \Omega$. Because $\int_\Omega \sqrt{k(x,x)}\, d\mathcal{P}'(x) < \infty$ for any $\mathcal{P}' \in P(\Omega)$, for any $\mathcal{P}, \mathcal{Q} \in P(\Omega)$, it can be inferred from [21] that

$$\gamma_k(\mathcal{P}, \mathcal{Q}) = \left\| \int_\Omega k(x, \cdot)\, d\mathcal{P}(x) - \int k(x, \cdot)\, d\mathcal{Q}(x) \right\|_{\mathcal{H}_k} := \|k_{\mathbf{P}}(\mathcal{P}) - k_{\mathbf{P}}(\mathcal{Q})\|_{\mathcal{H}_k}$$

where $k_{\mathbf{P}}(\mathcal{P}) := \int_\Omega k(x, \cdot)\, d\mathcal{P}(x)$. Yet, with an arbitrary Mercer kernel k, γ_k is not always a metric on $P(\Omega)$, in other words, not always satisfying the condition that $\mathcal{P} = \mathcal{Q}$ if $\gamma_k(\mathcal{P}, \mathcal{Q}) = 0$. Many studies have explored conditions on k under which γ_k becomes a metric on $P(\Omega)$. Here, we just present some conditions useful for the analysis later.

DEFINITION 2.1. *Let k be a Mercer kernel on $\Omega \times \Omega$ where Ω is a compact subset of $\mathbb{R}^d$.*

- *k is said to be universal if $\mathcal{H}_k$ is dense in $C(\Omega)$.*
- *k is said to integrally strictly positive definite if $\int_\Omega \int_\Omega k(x, y)\, d\mu(x)\, d\mu(y) > 0$ for all non-zero signed finite Borel measures μ defined on Ω.*

In fact, the two definitions above are shown [95] to be equivalent, but the second form of double integrals can be more useful in the proof later. It's also shown in [21] that for universal kernels, $k_{\mathbf{P}}$ defines an injective mapping from $P(\Omega)$ to $\mathcal{H}_k$, which implies that γ_k is a metric on $P(\Omega)$. Because a lot of popular kernels in the application are universal, including Gaussian kernels, Laplacian kernels, inverse multiquadrics, Matérn kernels, it's natural to consider the attention operators induced by universal kernels. Now, we give the formal definition of our attention operator:

DEFINITION 2.2. *Let $\Omega \subset \mathbb{R}^d$ be compact and $P(\Omega)$ be the set of all Borel probability measures defined on Ω. Suppose that k is a universal kernel on $\Omega \times \Omega$ and $f : \Omega \to \mathbb{R}$ is a continuous function with $c_f \leq |f(x)| \leq C_f$ for all $x \in \Omega$, where $c_f, C_f > 0$ are two constants. Then the attention operator* $\mathrm{attn} : (P(\Omega), \gamma_k) \to (\mathcal{H}_k, \|\cdot\|_{\mathcal{H}_k})$ *induced by k and f is defined as*

$$\mathrm{attn}(\mathcal{P}) = \int_\Omega k(x, \cdot) f(x)\, d\mathcal{P}.$$

It is easy to see that for $\mathcal{P} \in (P(\Omega), \gamma_k)$,

$$\|\mathrm{attn}(\mathcal{P})\|_{\mathcal{H}_k} = \left(\int_\Omega \int_\Omega k(x,y) f(x) f(y)\, d\mathcal{P}(x)\, d\mathcal{P}(y) \right)^{\frac{1}{2}} \leq r$$

with $r := C_f \|k\|_\infty$, where $\|k\|_\infty := \sup_{x \in \Omega} \sqrt{k(x,x)}$. Let $\mathcal{G}_{k,f}$ denote the image of attn in $\mathcal{H}_k$. Then $\mathcal{G}_{k,f}$ is contained in the closed ball $B_r := \{g \in \mathcal{H}_k : \|g\|_{\mathcal{H}_k} \leq r\}$. The attention operator can be viewed as an embedding from distributions to function representations. The following theorem shows some nice properties of this distribution embedding.

THEOREM 2.3. *Let k', k be two universal kernels defined on $\Omega \times \Omega$, and f defined in Definition 2.2. Then the attention operator* attn *induced by k and f is an injective and continuous mapping from $(P(\Omega), \gamma_{k'})$ to $(\mathcal{H}_k, \|\cdot\|_{\mathcal{H}_k})$.*

The kernel k' that metrizes $P(\Omega)$ can be a different universal kernel from the one inducing the attention operator. It has little effect on the properties of the attention operator. However, if we take $k'(x,y)$ to be $f(x)k(x,y)f(y)$ mentioned in the proof of Theorem 2.3, the attention operator is an isometry between $(P(\Omega), \gamma_{k'})$ and $(\mathcal{G}_{k,f}, \|\cdot\|_{\mathcal{H}_k})$, which shows that the attention operator can represent an embedding without any loss of information. But for simplicity of notations, we take both kernels to be the same universal kernel.

Next, we define the Transformers based on our novel attention operator. As mentioned in the Introduction, $\mathrm{SoftmaxAttn}(Q)$ can be regarded as a discretization of the function $\mathrm{attn}(\mathcal{P})$. Here, for any set of distinct points $\mathbf{T} = \{t_1, \ldots, t_{|\mathbf{T}|}\} \subset \Omega$, we introduce a sampling operator $[\cdot]_{\mathbf{T}} : \mathcal{H}_k \to \mathbb{R}^{|\mathbf{T}|}$ to discretize $\mathrm{attn}(\mathcal{P})$ where $|\mathbf{T}|$ denotes the size of the set $\mathbf{T}$, such that for any $g \in \mathcal{H}_k$, $[g]_{\mathbf{T}} = [g(t_j)]_{1 \leq j \leq |\mathbf{T}|} \in \mathbb{R}^{|\mathbf{T}|}$. Then we give the following precise definition of our Transformer encoder.

DEFINITION 2.4. *Let Ω be a compact subset of $\mathbb{R}^d$, k be a universal kernel, and* attn *be the attention operator defined in Definition 2.2. With distribution inputs from $(P(\Omega), \gamma_k)$, the Transformer encoder H_{n_1,n_2} of type (n_1, n_2) is defined by:*

$$H_{n_1,n_2}(\mathcal{P}) = c^T \sigma(A[\mathrm{attn}(\mathcal{P})]_{\mathbf{T}} + b) + b_0 \tag{2.5}$$

where $\mathbf{T}$ *is a set of* n_2 *points in* Ω*,* $A \in \mathbb{R}^{n_1 \times n_2}$ *a parameter matrix,* $b \in \mathbb{R}^{n_1}$ *a bias vector,* $c \in \mathbb{R}^{n_1}$*,* $b_0 \in \mathbb{R}$ *and* σ *is ReLU activation function given by* $\sigma(x) = \max\{0, x\}$*.*

REMARK 2.1. *With the definitions of the attention operator and the Transformer encoder above, there are some points we wish to clarify. First, the boundedness from below condition of the continuous feature mapping* f *can be removed in some applications, where a feature mapping is usually learned by a neural network. For example,* $f(x) = \sum_{j=1}^{n} c_j(\sigma(a_j \cdot x + b_j) + \epsilon_j)$ *with all* $\epsilon_j > 0$*, then we have*

$$\int_\Omega k(x, \cdot) f(x)\, d\mathcal{P} = \sum_{j=1}^{n} c_j \int_\Omega k(x, \cdot)(\sigma(a_j \cdot x + b_j) + \epsilon_j)\, d\mathcal{P}$$

and with the universality of shallow nets, there's no loss of the expressivity. Together with Theorem 2.3, we can observe that the composition structure of attention layers and FNN layers is crucial to the success of Transformers in embedding probability measures into function representations and learning diverse feature representations, which is also consistent with the Transformer's powerful data compression capabilities in practical applications.

REMARK 2.2. *Note that when the input is an empirical distribution* $\hat{\mu}^n = \frac{1}{n}\sum_{j=1}^{n} \delta_{x_j}$*, and* $\mathbf{T}' = \{x_j\}_{j=1}^n$*,* $[\mathrm{attn}(\hat{\mu}^n)]_{\mathbf{T}'}$ *is exactly the self-attention. For a functional* H_{n_1,n_2} *defined on* $(P(\Omega), \gamma_k)$*, generally speaking,* n_2 *controls the degree of discretization and* n_1 *controls the accuracy of approximation to the target functional. With a larger* n_2*, there is less information loss from the original distribution. In the application, this may also explain why engineers always manage to increase the length of the input sequence (i.e., the number of tokens) for the self-attention module in LLMs.*

2.3 Main Results on Transformer-based Network in Distribution Regression

This section provides the main results on learning capabilities of Transformer-based networks in distribution regression. First we define a functional class induced by Barron functionals and an example to illustrate its powerful expressivity, and then demonstrate the approximation rate of the defined functional class by a Transformer encoder. With the approximation rate and

an estimation of covering numbers, we obtain a generalization bound by an oracle inequality in the final subsection.

The attention operator embeds each Borel probability measure into a function in the closed ball B_r in $\mathcal{H}_k$. Then within the framework of distribution regression, we require a functional that maps functions in B_r to $\mathbb{R}$. Here, we consider a functional class produced by Barron functionals [3] and the definition is given below.

For a real-valued functional Φ on a Hilbert space $(\mathcal{H}, \|\cdot\|)$, we say that Φ is represented by a Fourier distribution $\tilde{F}$ on some domain $A \subset \mathcal{H}$ where $\tilde{F}$ is a complex-valued measure $\tilde{F}(d\omega) = e^{i\theta(\omega)}F(d\omega)$ if $\Phi(g) = \int_{\mathcal{H}} e^{i\langle g,\omega\rangle}\tilde{F}(d\omega)$ for all $g \in A$. Here $F(d\omega)$ denotes that magnitude distribution and $\theta(\omega)$ denotes the phase at the frequency ω.

DEFINITION 2.5. *For each $r, C > 0$, let $\Gamma_{r,C}(\mathcal{H})$ be the set of functionals Φ on $B_r := \{g \in \mathcal{H} : \|g\|_{\mathcal{H}} \leq r\}$ such that there's a Fourier distribution $\tilde{F}$ representing Φ on B_r satisfying $\int_{\mathcal{H}} \|\omega\|_{\mathcal{H}} F(d\omega) \leq C$. Every functional in $\Gamma_{r,C}(\mathcal{H})$ is called a Barron functional.*

We shall assume that the target function (regression function) in distribution regression has the form

$$\Phi(\mathrm{attn}(\mathcal{P})) = \Phi\left(\int_{\Omega} k(x,\cdot)f(x)\,d\mathcal{P}\right) \text{ where } \Phi \in \Gamma_{r,C}(\mathcal{H}_k).$$

To demonstrate the powerful expressivity of the Barron functional class, we provide an example here.

EXAMPLE 1. *Consider the ridge functional $\Phi(x) = g(\langle a, x\rangle_{\mathcal{H}_k})$ with $\|a\|_{\mathcal{H}_k} = 1$ and some univariate continuous function g with Fourier transform $\hat{g}$ on $\mathbb{R}$ satisfying $\int_{\mathbb{R}} |t||\hat{g}(t)|\,dt < \infty$. Then $\Phi(x) = \int e^{it\langle a,x\rangle_{\mathcal{H}_k}}\hat{g}(t)\,dt$, which means that a Fourier distribution $\tilde{F}$ supported on the set of $\{ta : t \in \mathbb{R}\}$, represents Φ on B_r for any $r > 0$. Φ is a Barron functional on B_r, as long as g is a Barron functional on $\mathbb{R}$. It is shown in [3] that g is a Barron functional on $[-r, r]$ when the second derivative of g is continuous. In this case, Φ is a Barron functional. When the input space is the embedding of all Borel probability measures and $\mathcal{H}_k$ is the corresponding*

RKHS, we obtain that

$$\begin{aligned}\Phi(\mathrm{attn}(\mathcal{P})) &= g(\langle a, \mathrm{attn}(\mathcal{P})\rangle_{\mathcal{H}_k}) \\ &= g\left(\left\langle a, \int k(x,\cdot) f(x)\, d\mathcal{P}\right\rangle_{\mathcal{H}_k}\right) \\ &= g\left(\int a(x) f(x)\, d\mathcal{P}\right). \end{aligned} \tag{2.6}$$

Since the definitions of many statistics (e.g., moments) are closely related with integration w.r.t. probability measures and $\mathcal{H}_k$ is dense in $C(\Omega)$, (2.6) is a nice tool to capture the relation between statistics of distributions and respond variables.

Beyond statistics of distributions, we may also consider a multivariate case and have distribution projections of the feature random variables $f(X), X \sim \mathcal{P}$ to retain as much information as desired. Note that the above case can be extended to ridge functionals with multiple features, i.e.,

$$\Phi(x) = g(\langle a_1, x\rangle_{\mathcal{H}_k}, \ldots, \langle a_{d'}, x\rangle_{\mathcal{H}_k})$$

with $\|a_j\|_{\mathcal{H}_k} = 1$ for all $1 \le j \le d'$ and some continuous function g defined on $\mathbb{R}^{d'}$. Similarly, if the partial derivatives of g of order $\lfloor d'/2 \rfloor + 2$ are continuous in $\mathbb{R}^{d'}$, then g is a Barron functional on $[-r, r]^{d'}$, which implies that Φ is also a Barron functional. This form of feature embedding with the attention operator is much more flexible than the functions with polynomial features considered in [89, 110]. We don't need to design a specific network structure for feature functions, but can still learn feature functions from a dense subset of $C(\Omega)$. Moreover, the above ridge functional form is just one case of the class of Barron functionals.

2.3.1 Approximation rate of Barron functional

We establish an approximation theory for a class of functionals $\Phi_{k,f}$ defined on $(P(\Omega), \gamma_k)$ by exploiting the proposed Transformer encoder. The functional $\Phi_{k,f}$ has the composition

form of a Barron functional and the attention operator:

$$\Phi_{k,f}(\mathcal{P}) := \Phi(\mathrm{attn}(\mathcal{P})) = \Phi\left(\int_{\Omega} k(x,\cdot) f(x)\, d\mathcal{P}\right) \tag{2.7}$$

with $\Phi \in \Gamma_{r,C}(\mathcal{H}_k)$ and $r = C_f \|k\|_\infty$.

In [3], we have the following approximation lemma for Barron functionals defined on a Hilbert space $\mathcal{H}$. A sigmoidal function ϕ on $\mathbb{R}$ means a bounded measurable function satisfying $\lim_{t\to\infty} \phi(t) = 1$ and $\lim_{t\to-\infty} \phi(t) = 0$.

LEMMA 1. *For $r, C > 0$, $\Phi \in \Gamma_{r,C}(\mathcal{H}_k)$, $\beta > 0, n \in \mathbb{N}$, sigmoidal function ϕ on $\mathbb{R}$, probability measure ν on B_r, there is a function $\Psi_n(g) = \sum_{p=1}^n c_p \phi(\langle a_p, g\rangle_{\mathcal{H}_k} + b_p) + \Phi(0)$ with $\|a_p\|_{\mathcal{H}_k} \le \frac{\beta}{r}, |b_p| \le \beta$ and $\|c\|_1 \le 2rC$, such that*

$$\int_{B_r} (\Phi(x) - \Psi_n(x))^2 \,\mu(dg) \le (2rC)^2 \left(\frac{1}{n^{1/2}} + \eta_\beta\right)^2$$

where $\eta_\beta = \inf_{0<\epsilon\le\frac{1}{2}}\{2\epsilon + \sup_{|z|\ge\epsilon} \left|\phi(\beta z) - 1_{\{z>0\}}\right|\}$.

However, the approximation form Ψ_n cannot be directly applied with popular network structures, because it involves functions $\{a_p\} \subset \mathcal{H}_k$ as parameters. One idea is to utilize $\sum_q a_{p,q} k(t_q, \cdot)$ with $a_{p,q} \in \mathbb{R}$ and $t_q \in \Omega$ to approximate each a_p, and then, by kernel trick, the inner product $\langle a_p, g\rangle_{\mathcal{H}_k}$ can be approximated by a linear combination the function values $\{g(t_q)\}$ at the sampled points, which exactly matches the form of our Transformer encoder. This idea was studied recently in [122]. Then by using the Transformer encoder with a Gaussian kernel $k = \exp(-\|x-y\|^2/\alpha^2)$ ($\alpha > 0$), we have the following result on $L^2_{\rho_\mathcal{U}}$ approximation rates of the functional $\Phi_{k,f}$ with the same Gaussian kernel, where we denote $\rho_\mathcal{U}$ as the marginal distribution ρ on $\mathcal{U} = P(\Omega)$ and $(L^2_{\rho_\mathcal{U}}, \|\cdot\|_\rho)$ as the space of square integrable functions with respect to $\rho_\mathcal{U}$.

THEOREM 2.6. *Let k be a Gaussian kernel $k(x,y) = \exp\{-\|x-y\|^2/\alpha^2\}$ with $\alpha > 0$ and $\Omega = [-1,1]^d$. For every functional $\Phi_{k,f}$ defined by (2.7) for any $r, C > 0, n \in \mathbb{N}$, there exists*

a Transformer encoder h_n in the hypothesis $\mathcal{H}_{R,n}$ such that

$$\|\Phi_{k,f} - h_n\|_\rho \leq \frac{(4C_f + r)C}{n^{\frac{1}{2}}}.$$

The hypothesis space $\mathcal{H}_{R,n}$ is a class of functions in Definition 2.4 of type $(2n, s(n))$ with $s(n) := \lceil C_{k,d}(\log n)^d \rceil$ such that

$$\mathcal{H}_{R,n} = \left\{ H_{n,s(n)} : \|c\|_1 \leq Rn^{\frac{1}{2}}, |A_{p,q}| \leq R(\log n)^{-\frac{d}{2}} n^{R\log n}, \|b\|_\infty \leq R,\ b_0 = \Phi(0) \right\}$$

where $A = (A_{p,q})$, $C_{k,d}$ is a constant depending on d and kernel k, and R depends only on r, C, d and kernel k. The total number of free parameters of Transformer encoder is $O(n(\log n)^d)$.

The above theorem gives rates of approximating a class of functionals by a Transformer encoder, with the complexity bound on the total number of free parameters and the parameter bounds on connection matrices and bias vectors. These will be useful later to derive an estimation of covering numbers and generalization bounds in distribution regression.

2.3.2 Distribution regression with Transformers

This section conducts generalization analysis of the empirical risk minimization (ERM) algorithm for distribution regression with Transformer encoders. Now we propose the ERM algorithm for two-stage distribution regression. Take $\mathcal{Z} = \mathcal{U} \times \mathcal{Y}$, where $\mathcal{U} = P(\Omega)$ is the input space of all Borel probability measures on $\Omega = [-1, 1]^d$ and $\mathcal{Y} = [-M, M]$ is the output space with $M > 0$. The regression function φ_ρ on $\mathcal{U}$ is defined as

$$\varphi_\rho(\mu) = \int_{\mathcal{Y}} y \, d\rho(y|\mu)$$

where $\rho(\cdot|\mu)$ is the conditional distribution at μ induced by ρ, and it minimizes the mean squared error for $\varphi : \mathcal{U} \to \mathcal{Y}$,

$$\mathcal{E}(\varphi) = \int_{\mathcal{Z}} (\varphi(\mu) - y)^2 \, d\rho.$$

For establishing a learning theory for the ERM algorithm, it is crucial: whether $\mathcal{H}_{R,n}$ is well defined as a compact hypothesis space. To answer the question, we denote $C(P(\Omega))$ with norm $\|\varphi\|_\infty := \sup_{\mu\in P(\Omega)} |\varphi(\mu)|$, to be the Banach space of continuous functions on $(P(\Omega), \gamma_k)$. Then we have the following theorem:

THEOREM 2.7. *The hypothesis $\mathcal{H}_{R,n}$ of Transformer encoders is a compact subset of $C(P(\Omega))$.*

With Theorem 2.7, the covering number $\mathcal{N}(\mathcal{H}_{R,n}, \epsilon, \|\cdot\|_\infty)$ of $\mathcal{H}_{R,n}$ as a subset of $C(P(\Omega))$ makes sense, where $\mathcal{N}(\mathcal{H}_{R,n}, \epsilon, \|\cdot\|_\infty)$ denotes the minimum number of balls with radius $\epsilon > 0$ whose union covers $\mathcal{H}_{R,n}$ in the space $C(P(\Omega))$. The estimation of the covering number plays a vital role in deriving a generalization bound for distribution regression. The related details will be presented in Section 2.5.

Recall that for the second stage dataset $\hat{D} = \{\big(\{x_{i,j}\}_{j=1}^{m_{2,i}}, y_i\big)\}_{i=1}^{m_1}$ in two-stage distribution regression, the empirical target functional from the ERM algorithm with the hypothesis space $\mathcal{H}_{R,n}$ is the functional defined as

$$\varphi_{\hat{D},R,n} = \arg\min_{\varphi\in\mathcal{H}_{R,n}} \mathcal{E}_{\hat{D}}(\varphi)$$

with

$$\mathcal{E}_{\hat{D}}(\varphi) := \frac{1}{m_1}\sum_{i=1}^{m_1}\big(\varphi(\hat{\mu}_i^{m_{2,i}}) - y_i\big)^2,$$

where the existence of the minimizer $\varphi_{\hat{D},R,n}$ is guaranteed by the compactness of $\mathcal{H}_{R,n}$.

We now define the truncation operator π_M on the space $C(P(\Omega))$ as

$$\pi_M(\varphi)(\mu) = \begin{cases} M, & \text{if } \varphi(\mu) > M, \\ -M, & \text{if } \varphi(\mu) < -M, \\ \varphi(\mu), & \text{if } -M \le \varphi(\mu) \le M. \end{cases}$$

Since the regression function φ_ρ is bounded by M, the truncated empirical target functional

$$\pi_M\varphi_{\hat{D},R,n}$$

is considered as the final estimator.

REMARK 2.3. *The distribution regression framework considered here involves a two-stage sampling process, which makes our Transformer encoder (2.5) compatible with practical applications, particularly when dealing with sequential inputs like sentences in NLP. From the two-stage sampling process, $m_{2,i}$ controls the length of the sequential input i, which can be understood as the number of tokens in the input sequence i in the case of NLP. The two-stage sampling process allows us to study the generalization capabilities of the model under the practical data structure, while also taking into account the approximation ability to abstract functionals.*

To derive the excess generalization error $\mathcal{E}(\pi_M\varphi_{\hat{D},R,n}) - \mathcal{E}(\varphi)$, we introduce the empirical error of the first-stage sample

$$\mathcal{E}_D(\varphi) := \frac{1}{m_1}\sum_{i=1}^{m_1}(\varphi(\mu_i) - y_i)^2$$

and then we can obtain a decomposition of the excess generalization error in the following lemma which can be easily seen from the fact that $\mathcal{E}_{\hat{D}}(\pi_M\varphi_{\hat{D},R,n}) \leq \mathcal{E}_{\hat{D}}(h)$.

LEMMA 2. *For any $h \in \mathcal{H}_{R,n}$ and $\varphi_{\hat{D},R,n}$ defined in (2.1), we have*

$$\begin{aligned}\mathcal{E}\left(\pi_M\varphi_{\hat{D},R,n}\right) - \mathcal{E}\left(\varphi_\rho\right) \leq \mathcal{E}\left(\pi_M\varphi_{\hat{D},R,n}\right) - \mathcal{E}_D\left(\pi_M\varphi_{\hat{D},R,n}\right) + \mathcal{E}_D\left(\pi_M\varphi_{\hat{D},R,n}\right) \\ - \mathcal{E}_{\hat{D}}\left(\pi_M\varphi_{\hat{D},R,n}\right) + \mathcal{E}_{\hat{D}}(h) - \mathcal{E}_D(h) + \mathcal{E}_D(h) - \mathcal{E}(h) + \mathcal{E}(h) - \mathcal{E}\left(\varphi_\rho\right)\end{aligned}$$

which can be bounded by the summation

$$\mathcal{I}_1(D,\mathcal{H}_{R,n}) + \mathcal{I}_2(D,\mathcal{H}_{R,n}) + \left|\mathcal{I}_3(\hat{D},\mathcal{H}_{R,n})\right| + \left|\mathcal{I}_4(\hat{D},\mathcal{H}_{R,n})\right| + R(\mathcal{H}_{R,n})$$

in which

$$\begin{aligned}\mathcal{I}_1(D,\mathcal{H}_{R,n}) &= \left\{\mathcal{E}\left(\pi_M\varphi_{\hat{D},R,n}\right) - \mathcal{E}\left(\varphi_\rho\right)\right\} - \left\{\mathcal{E}_D\left(\pi_M\varphi_{\hat{D},R,n}\right) - \mathcal{E}_D\left(\varphi_\rho\right)\right\} \\ \mathcal{I}_2(D,\mathcal{H}_{R,n}) &= \left\{\mathcal{E}_D(h) - \mathcal{E}_D\left(\varphi_\rho\right)\right\} - \left\{\mathcal{E}(h) - \mathcal{E}\left(\varphi_\rho\right)\right\} \\ \mathcal{I}_3(\hat{D},\mathcal{H}_{R,n}) &= \mathcal{E}_D\left(\pi_M\varphi_{\hat{D},R,n}\right) - \mathcal{E}_{\hat{D}}\left(\pi_M\varphi_{\hat{D},R,n}\right) \\ \mathcal{I}_4(\hat{D},\mathcal{H}_{R,n}) &= \mathcal{E}_{\hat{D}}(h) - \mathcal{E}_D(h), \quad R(\mathcal{H}_{R,n}) = \mathcal{E}(h) - \mathcal{E}\left(\varphi_\rho\right).\end{aligned}$$

Based on the two-stage error decomposition, the two-stage oracle inequality for distribution regression in the hypothesis space $\mathcal{H}_{R,n}$ of our proposed Transformer encoder is established in the following theorem to be proved in Section 2.5.

THEOREM 2.8. *Consider the distribution regression framework with the first stage sample size $m_1 \in \mathbb{N}$, and the second stage sample size $\min\{m_{2,i} : 1 \leq i \leq m_1\} = m_2$. Then for $n \geq 3$, any $h \in \mathcal{H}_{R,n}$ and $\epsilon > 0$, we have*

$$\begin{aligned}
&\operatorname{Prob}\left\{\left\|\pi_M\varphi_{\hat{D},R,n} - \varphi_\rho\right\|_\rho^2 > 2\left\|h - \varphi_\rho\right\|_\rho^2 + 8\epsilon\right\} \\
\leq &\mathcal{N}\left(\mathcal{H}_{R,n}, \frac{\epsilon}{16M}, \|\cdot\|_\infty\right) \exp\left\{-\frac{3m_1\epsilon}{2048M^2}\right\} \\
&+ \exp\left\{-\frac{m_1\epsilon^2}{2\left(3M + \|h\|_\infty\right)^2\left(\left\|h - \varphi_\rho\right\|_\rho^2 + \frac{2}{3}\epsilon\right)}\right\} \\
&+ 4m_1 s(n) \exp\left\{-\frac{m_2\epsilon^2}{128\max\{\|h\|_\infty^2, M^2\}C_f^2 C_5^2 n^{4R\log n}(\log n)^d}\right\}
\end{aligned}$$

where $C_5 := 2C_{k,d}R^2$ is a constant depending on d, R and kernel k.

Based on the the oracle inequality for distribution regression, the excess generalization error can be bounded in the following theorem to be proved in Section 2.5, where we assume that the regression function φ_ρ belongs to the functional class defined in (2.7).

THEOREM 2.9. *Suppose that the regression function φ_ρ has the form (2.7) with $\Phi(0) = 0$. If the total number N of free parameters of Transformer encoders and the second stage sample size m_2 are chosen by*

$$N = \left\lfloor \mathcal{A}_7 m_1^{\frac{1}{2}}\left(\log(\mathcal{A}_3 m_1^{\frac{1}{2}})\right)^d \right\rfloor \text{ and } m_2 = \left\lceil \mathcal{A}_5\left(\mathcal{A}_3 m_1^{\frac{1}{2}}\right)^{8R\log\left(\mathcal{A}_3 m_1^{\frac{1}{2}}\right)} \right\rceil,$$

then for the truncated estimator produced by the distribution regression framework with Transformer encoders,

$$\mathbb{E}\{\mathcal{E}(\pi_M\varphi_{\hat{D},R,n}) - \mathcal{E}(\varphi_\rho)\} \leq \mathcal{A}_6 m_1^{-\frac{1}{2}}\left(\log\left(\mathcal{A}_3 m_1^{\frac{1}{2}}\right)\right)^{d+2}$$

where $\mathcal{A}_3, \mathcal{A}_4, \mathcal{A}_5, \mathcal{A}_6, \mathcal{A}_7$ are constants depending only on C_f, C, M, d and α.

REMARK 2.4. *In the proof of Theorem 2.9, we choose n to scale up with $m_1^{\frac{1}{2}}$ for a balance between the approximation error and the estimation error. Recall that n controls the complexity of the FNN layer: a larger n increases the hypothesis complexity of the Transformer class, leading to better approximation and more diverse feature representations. Besides, to achieve a nice generalization performance, the hypothesis complexity of the FNN layer also scales up with the first stage sample size m_1. For the case of NLP, m_1 measures the diversity of semantics in the datatset D. Especially when training an LLM with a dataset size often exceeding hundreds of terabytes of text data, m_1 becomes incredibly large. This poses challenges for efficiently scaling up the complexity of the FNN layer to achieve a balance. We will discuss this point in subsection 2.4.3 below.*

2.4 Discussion

In this section, we propose the attention operator for modeling attention mechanisms in Transformers, and apply the two-stage sampling process to understand the training process of Transformers in practical applications. We analyze the expressivity and demonstrate the generalization capacity of the proposed Transformer structures. In this section, we exploit the established framework and theoretical results to develop deep understanding of various tricks and techniques with Transformers in practical applications.

2.4.1 Kernel normalization

Recall the original self-attention module (2.3) in [101]

$$k_{attn}\left(x_i, x_j\right) := \frac{\exp\left(\frac{\langle W_q x_i, W_k x_j\rangle}{\sqrt{d_{in}}}\right)}{\sum_{j'=1}^{n}\exp\left(\frac{\langle W_q x_i, W_k x_{j'}\rangle}{\sqrt{d_{in}}}\right)}.$$

If we consider the normalization factor $\sum_{j'=1}^{n} \exp\left(\frac{\langle W_q x_i, W_k x_{j'}\rangle}{\sqrt{d_{in}}}\right)$ independently (since it depends on the input sequence Q), then its probability embedding can be reformulated as

$$\overline{\mathrm{attn}}_k(\mu) = \int_\Omega \frac{k(x,\cdot)}{\int_\Omega k(y,\cdot)\, d\mu(y)} f(x)\, d\mu(x), \mu \in P(\Omega),$$

where k is a Laplace or Gaussian kernel. It can be further written as

$$\overline{\mathrm{attn}}_k(\mu) = \frac{\int_\Omega k(x,\cdot) f(x)\, d\mu(x)}{\int_\Omega k(y,\cdot)\, d\mu(y)}. \tag{2.8}$$

If we allow multiple features represented by different FNNs (further discussed in subsection 2.4.3 below), then $\overline{\mathrm{attn}}_k$ can be approximated by our attention operator attn (2.4) induced by the same k and f.

Let $\Omega = [-1,1]^d$. We define f' as $f'(y) = f_1(y) + f_1(-y)$ for $y \in [-1,1]$ where

$$f_1(y) = \frac{1}{2}\left[\sigma(y+1) - 2\sigma(y-1) + \sigma(y-3)\right].$$

Then it's easily verified that $f' = 1$ on $[-1,1]$. For $x \in [-1,1]^d$, take $f_1(x) := \frac{1}{d}\mathbf{1}^T f'(x) = 1$ in $[-1,1]^d$ where f' applies element-wise on x and $\mathbf{1} = (1, \cdots, 1)^T \in \mathbb{R}^d$. In other words, the normalization function can be represented by our attention operator with the feature function f' constructed by the above FNN.

Note that $\overline{\mathrm{attn}}_k$ can be written as a composition of $g_1\left(\mathrm{attn}_{k,f}(\mu), g_2(\mathrm{attn}_{k,f'}(\mu))\right)$, where $\mathrm{attn}_{k,f}$ denotes the attention operator induced by the kernel k and feature f, $g_1(y_1, y_2) := y_1 y_2$ and $g_2(y_3) := 1/y_3$, and k is a Laplace or Gaussian kernel, which implies $0 < \mathrm{attn}_{k,f}(\mu) \leq r$ and $0 < c' < \mathrm{attn}_{k,f'}(\mu) \leq \|k\|_\infty^2$. Then g_1, g_2 can both be approximated by fully connected neural networks [109]. In conclusion, we show that $\overline{\mathrm{attn}}_k$ can be approximated by our attention operator attn with multiple features.

In our framework, we are mainly concerned with the function space and the properties of the attention operator induced by k, f. Therefore, a normalization function is of less importance in our analysis, since it can be separated from the attention operator as in (2.8). However, in practical applications, a normalization function has two advantages. First, it introduces asymmetric dependency as an inductive bias into the attention mechanism, which means that

$k_{\mathrm{attn}}(x_i, x_j)$ may not be equal to $k_{\mathrm{attn}}(x_j, x_i)$. The inductive bias is very useful for modeling certain data structures such as natural languages, because words often have hierarchical and directional relationships. Second, a normalization function makes the sum of kernel weights always equal to 1, which can be useful for auto-regressive tasks [77, 14, 78]. By normalizing the weights, the function effectively balances the contribution of each component and enhances the stability and consistency of the model.

2.4.2 Discretization subsets T

Distribution regression is a special case of operator learning [92, 93]. To construct a computable Transformer encoder, we apply two techniques: the two-stage sampling process and the kernel discretization w.r.t. $\mathbf{T}$. The elements in $\mathbf{T}$ are often called queries in NLP. In our proof, we choose the set of queries to be the uniform mesh on Ω. Then for any function in $\mathcal{H}_k$, we can apply the kernel trick to replace function parameters by function values at each query. Although it is a universal choice to recover information from any function parameters in $\mathcal{H}_k$, the uniform mesh also introduces the term $(\log n)^d$ that still rules out practical use of the framework in high-dimensional cases.

However, there are several cases in practice showing that with an appropriately chosen query set of (much) smaller size, Transformer encoders still perform well on various tasks. The most common is self-attention, as mentioned in Remark 2.2. The choice of query sets is adaptive to each input in self-attention, and more precisely, a set of second-stage samples (independent of the dimension). It would be interesting to investigate the underlying mechanism of self-attention within our framework. Another example is that the query set $\mathbf{T}$ can be learned from training data using stochastic gradient descent, which is usually applied in prompt tuning [52] and Q-former for cross-modal alignment [46]. The basic idea behind these techniques can be understood with our theoretical framework: When handling specific learning tasks or aiming to compress data for more refined feature representations (e.g., low-dimensional features), we often do not require a high-resolution uniform mesh as a universal choice to retain as much information as possible. Instead, we just need a set of queries, with a significantly smaller

size, which performs well for a particular task or certain data structure. These queries can be obtained by optimizing the corresponding loss functions.

2.4.3 FNN for learning features

Note that the attention operator primarily facilitates the efficient compression of probability distributions, whereas the FNN learns the pertinent features of the Barron functional (shown in Theorem 2.6). This theoretical discovery also aligns with practical engineering experiences in fine-tuning LLMs, such as adapter modules [29]. Fine-tuning is often required to adapt the existing models to new tasks or datasets [23]. When applied with pretrained LLMs, engineers always fix the parameters of the original network and add only some FNN modules with a few trainable parameters into the existing model, e.g., adapters [29]. In this way, the training for adaptation to new tasks or datasets can be dramatically decreased for pretrained LLMs with billions of parameters, and the well-trained small modules can be directly plugged into other models to transfer features learned for new tasks. In our theory, features for target functional are entirely learned by the FNN layers and the attention operator merely embeds probability measures into function representations. In other words, for different target functionals, all trainable parameters are contained within the FNN layer, while the attention operator retains no trainable parameters. This provides a theoretical foundation for the successful application of adapters.

Another topic arising from our generalization analysis (Theorem 2.9) of distribution regression with Transformer encoders, is how to efficiently scale up the complexity of the FNN layer. Recall that in order to balance the approximation error and estimation error, we take $n = O(m_1^{\frac{1}{2}})$ where m_1 is the number of first-stage samples (i.e., probability measures in $P(\Omega)$). In our interpretation of NLP, m_1 quantifies the complexity of semantics within a dataset. When training LLMs with increasingly large corpora of texts, m_1 becomes extremely large. Then to achieve better generalization performance, we need to scale up the complexity of the FNN layer dramatically. However, this poses challenges in training process. A popular solution is called Mixture of Experts (MoE) [16, 87, 32]. Roughly speaking, the basic idea is that for each query, we may choose one out of a pool of FNNs. A gating network decides which FNN

to activate based on the input query. This enables efficient scaling with Transformers, since we can increase the number of FNNs in the pool and just activate one FNN each time in the training. It would also be interesting to investigate the MoE mechanism further within the distribution regression framework.

2.5 Proof of Main Results on Transformer-based Networks in Distribution Regression

2.5.1 Proof of Theorem 2.3

Recall

$$\mathrm{attn}(\mu) = \int_\Omega k(x,\cdot) f(x)\, d\mu$$

for $\mu \in (P(\Omega), \gamma_{k'})$ where $\gamma_{k'}$ is a metric on $P(\Omega)$ induced by the universal kernel k'.

For $\mathcal{P}, \mathcal{Q} \in (P(\Omega), \gamma_{k'})$, we have

$$\begin{aligned}
&\|\mathrm{attn}(\mathcal{P}) - \mathrm{attn}(\mathcal{Q})\|_{\mathcal{H}_k}^2 \\
&= \left\| \int_\Omega k(x,\cdot) f(x)\, d\mathcal{P} - \int_\Omega k(x,\cdot) f(x)\, d\mathcal{Q} \right\|_{\mathcal{H}_k}^2 \\
&= \int_\Omega \int_\Omega k(x,y) f(x) f(y)\, d\mathcal{P}(x) d\mathcal{P}(y) \\
&\quad + \int_\Omega \int_\Omega k(x,y) f(x) f(y)\, d\mathcal{Q}(x) d\mathcal{Q}(y) \\
&\quad - 2 \int_\Omega \int_\Omega k(x,y) f(x) f(y)\, d\mathcal{P}(x)\, d\mathcal{Q}(y) \\
&= \left\| \int_\Omega f(\cdot) k(x,\cdot) f(x)\, d\mathcal{P} - \int_\Omega f(\cdot) k(x,\cdot) f(x)\, d\mathcal{Q} \right\|_{\mathcal{H}_{\tilde{k}}}^2 \\
&= \left\| \int_\Omega \tilde{k}(x,\cdot)\, d\mathcal{P} - \int_\Omega \tilde{k}(x,\cdot)\, d\mathcal{Q} \right\|_{\mathcal{H}_{\tilde{k}}} = \left\| \tilde{k}_{\mathbf{P}}(\mathcal{P}) - \tilde{k}_{\mathbf{P}}(\mathcal{Q}) \right\|_{\mathcal{H}_{\tilde{k}}}, \qquad (2.9)
\end{aligned}$$

where $\tilde{k}(x,y) := f(x)k(x,y)f(y)$. It's easy to see that $\tilde{k}$ is a Mercer kernel on $\Omega \times \Omega$. For any finite nonzero signed Borel measure $\mathcal{P}$, define $\mathcal{P}_f(E) = \int_E f\, d\mathcal{P}$ for any Borel

set E. Then $\mathcal{P}_f$ is also a finite nonzero signed Borel measure, since f is measurable and $0 < c_f \leq |f(x)| \leq C_f$ for all $x \in \Omega$. It follows by the universality of Mercer kernel k that

$$\int_\Omega \int_\Omega \tilde{k}(x,y)\, d\mathcal{P}(x)\, d\mathcal{P}(y) = \int_\Omega \int_\Omega k(x,y) f(x) d\mathcal{P}(x) f(y) d\mathcal{P}(y) \tag{2.10}$$

$$= \int_\Omega \int_\Omega k(x,y) d\mathcal{P}_f(x) d\mathcal{P}_f(y) > 0 \tag{2.11}$$

which implies that $\tilde{k}$ is an integrally strictly pd kernel. Therefore, $\gamma_{\tilde{k}}$ is also a metric on $P(\Omega)$ and then attn is an injective mapping defined on $P(\Omega)$.

Any $g \in C(\Omega)$ can be approximated by $\sum_{p=1}^N \alpha_p \tilde{k}(\cdot, t_p)$ to an arbitrary accuracy when N is large enough, because

$$g(x) - \sum_{p=1}^N \alpha_p f(x) k(x, t_p) f(t_p) = f(x) \left(\frac{g(x)}{f(x)} - \sum_{p=1}^N \tilde{\alpha}_p k(x, t_p) \right)$$

with $\alpha_p = \frac{\tilde{\alpha}_p}{f(t_p)}$ and $g/f \in C(\Omega)$ can be approximated by $\sum_{p=1}^N \tilde{\alpha}_p k(x, t_p)$ to an arbitrary accuracy when N is large enough. Thus $\tilde{k}$ is also universal. By Lemma 5 below, all universal kernels defined on the compact metric space Ω metrize the same topology on $P(\Omega)$ as , i.e., the weak topology on $P(\Omega)$, which is the weakest topology such that the map $\mu \mapsto \int_\Omega g\, d\mu$ is continuous for all $f \in C(\Omega)$. Then for the universal kernels $k', \tilde{k}$, $(P(\Omega), \gamma_{k'})$ and $(P(\Omega), \gamma_{\tilde{k}})$ share the same topology on $P(\Omega)$. It's also easy to observe that the kernel embedding $\tilde{k}_{\mathbf{P}}$ is an isometry between $(P(\Omega), \gamma_{\tilde{k}})$ and $\left(\tilde{k}_{\mathbf{P}}(P(\Omega)), \|\cdot\|_{\mathcal{H}_{\tilde{k}}}\right)$, which follows that $\tilde{k}_{\mathbf{p}}$ is a continuous mapping from $(P(\Omega), \gamma_{k'})$ to $(\mathcal{H}_{\tilde{k}}, \|\cdot\|_{\mathcal{H}_{\tilde{k}}})$. Then it can be concluded by (2.9) that attn is also a continuous mapping from $(P(\Omega), \gamma_{k'})$ to $(\mathcal{H}_k, \|\cdot\|_{\mathcal{H}_k})$. ■

2.5.2 Proof of Theorem 2.6

First we have the following error decomposition

$$\|\Phi_{k,f} - H_{n_1,n_2}\|_\rho \leq \|\Phi_{k,f} - H_{n_1}\|_\rho + \|H_{n_1} - H_{n_1,n_2}\|_\rho \tag{2.12}$$

where $H_{n_1}(\mu)$ has the form of $\sum_{p=1}^{n_1} c_p\phi(\langle a_p, g\rangle_{\mathcal{H}_k} + b_p) + b_0$ with $c_p, b_p, b_0 \in \mathbb{R}$, $a_p \in \mathcal{H}_k$ for all $1 \leq p \leq n_1$ and a certain sigmoidal function ϕ. In the following we present upper bounds for the two terms on RHS of (2.12) respectively.

Step 1. An upper bound for $\|\Phi_{k,f} - H_{n_1}\|_\rho^2$ is derived with the help of Lemma 1. We specify the sigmoidal function and the probability measure in our case as follows. Let $\phi(x)$ be the sigmoidal function $\sigma\left(x + \frac{1}{2}\right) - \sigma\left(x - \frac{1}{2}\right)$. Let $\beta \geq 1$. Then for ReLU neural networks,

$$\eta_\beta = \inf_{0<\epsilon\leq\frac{1}{2}} \left\{ 2\epsilon + \sup_{|z|\geq\epsilon} \left| \sigma\left(\beta z + \frac{1}{2}\right) - \sigma\left(\beta z - \frac{1}{2}\right) - 1_{\{z>0\}} \right| \right\} \tag{2.13}$$

with $\left|\sigma\left(\beta z + \frac{1}{2}\right) - \sigma\left(\beta z - \frac{1}{2}\right) - 1_{\{z>0\}}\right| \leq \frac{1}{2} - \beta\epsilon$ for $\epsilon \leq |z| \leq \frac{1}{2\beta}$ and is 0 for $|z| \geq \frac{1}{2\beta}$. Take $\epsilon = \frac{1}{2\beta}$ then we have $\eta_\beta \leq \frac{1}{\beta}$.

By Theorem 2.3, $\mathrm{attn} : (P(\Omega), \gamma_k) \to (\mathcal{H}_k, \|\cdot\|_k)$ is continuous. Then Borel probability measure $\rho_{\mathcal{U}}$ on $P(\Omega)$ defines another Borel probability measure $\tilde{\mu}$ on $\mathcal{H}_k$ by $\tilde{\mu}(\mathcal{B}) := \rho_{\mathcal{U}}(\mathrm{attn}^{-1}(\mathcal{B}))$ where $\mathcal{B}$ is a Borel set in $\mathcal{H}_k$. Note that $\mathcal{G}_{k,f}$, the image of the attention operator, is contained in the closed ball B_r. Then we can denote the restriction of $\tilde{\mu}$ on B_r by ν.

Recall that $\Phi_{k,f}$ has the form of $\Phi_{k,f}(\mathcal{P}) = \Phi\left(\int_\Omega k(x,\cdot)f(x)\,d\mathcal{P}\right)$ with $\Phi \in \Gamma_{r,C}(\mathcal{H}_k)$. Then by taking $\beta = n_1^{\frac{1}{2}}$ and the probability measure ν, and applying Lemma 1, we get a functional defined on B_r by $\widetilde{H}_{n_1}(g) = \sum_{p=1}^{n_1} c_p\phi\left(\langle a_p, g\rangle_{\mathcal{H}_k} + b_p\right)$ with $\|a_p\|_{\mathcal{H}_k} \leq \frac{n_1^{1/2}}{r}$, $|b_p| \leq n_1^{1/2}$ for all p and $\|c\|_1 \leq 2rC$ such that

$$\int_{\mathcal{H}_k} (\Phi(g) - \widetilde{H}_{n_1}(g))^2 \tilde{\mu}(dg) = \int_{B_r} (\Phi(g) - \widetilde{H}_{n_1}(g))^2\,\nu(dg) \leq \frac{(4rC)^2}{n_1},$$

which follows that

$$\int_{P(\Omega)} (\Phi_{k,f}(\mathcal{P}) - H_{n_1}(\mathcal{P}))^2\,d\mu \leq \frac{(4rC)^2}{n_1} \tag{2.14}$$

with $H_{n_1}(\mathcal{P}) = \widetilde{H}_{n_1}\left(\int_\Omega k(x,\cdot)f(x)\,d\mathcal{P}\right)$.

Step 2. We discretize the weight parameters $\{a_p\}$ in the network by kernel trick.

Observe that

$$H_{n_1}(\mathcal{P}) = \sum_{p=1}^{n_1} c_p \phi\left(\left\langle a_p, \int_\Omega k(x,\cdot) f(x)\, d\mathcal{P}\right\rangle_{\mathcal{H}_k} + b_p\right) \tag{2.15}$$

and that

$$\left\langle a_p, \int_\Omega k(x,\cdot) f(x)\, d\mathcal{P}\right\rangle_{\mathcal{H}_k} = \int_\Omega \langle a_p, k(x,\cdot)\rangle_{\mathcal{H}_k} f(x) d\mathcal{P} = \int_\Omega a_p(x) f(x)\, d\mathcal{P} \tag{2.16}$$

with $\|a_p\|_{\mathcal{H}_k} \leq \frac{n_1^{1/2}}{r}$ for all $1 \leq p \leq k$.

For $(a'_p)_{p=1}^{n_1} \subset \mathcal{H}_k$, we have from the Lipschitz continuity of ϕ that

$$\begin{aligned}
&\left|\sum_{p=1}^{n_1} c_p \phi\left(\int_\Omega a_p(x) f(x)\, d\mathcal{P} + b_p\right) - \sum_{p=1}^{n_1} c_p \phi\left(\int_\Omega a'_p(x) f(x)\, d\mathcal{P} + b_p\right)\right| \\
&\leq \|c\|_1 \max_p \left|\phi\left(\int_\Omega a_p(x) f(x)\, d\mathcal{P} + b_p\right) - \phi\left(\int_\Omega a'_p(x) f(x)\, d\mathcal{P} + b_p\right)\right| \\
&\leq 2\|c\|_1 \max_p \left|\int_\Omega [a_p(x) - a'_p(x)] f(x)\, d\mathcal{P}\right| \\
&\leq 2\|c\|_1 C_f \max_p \sup_{x\in\Omega} \left|a_p(x) - a'_p(x)\right|.
\end{aligned}$$

To apply kernel trick and get discrete approximations to network parameters $(a_p)_{p=1}^{n_1}$, we take $n \in \mathbb{N}$ and

$$\mathbf{T} = \left\{(t^{(1)}, \ldots, t^{(d)}) : t^{(j)} \in \left\{-1 + \frac{2l}{n}\right\}_{l=0}^{n} \text{ for } 1 \leq j \leq d\right\}$$

to be the uniform mesh on $\Omega = [-1,1]^d$.

We choose

$$a_p^{n_2}(x) := \sum_{q=1}^{n_2} a_p(t_q) u_q(x) \tag{2.17}$$

with $n_2 = |\mathbf{T}|$ and

$$(u_q)_{q=1}^{n_2} = [(k(x_i, x_j))_{x_i, x_j \in \mathbf{T}}]^{-1} (k(\cdot, x_i))_{x_i \in \mathbf{T}}$$

to be the so-called nodal functions satisfying

$$\left|a_p(x) - a_p^{n_2}(x)\right| \leq \|a_p\|_{\mathcal{H}_k} \exp\left(-\frac{c_k n_2^{1/d}}{2\sqrt{d}}\right), \forall x \in \Omega. \tag{2.18}$$

This can be found in [122] with a constant c_k depending on the kernel k.

Step 3. Let $H_{n_1,n_2}(\mathcal{P}) = \sum_{p=1}^{n_1} c_p \phi\left(\left\langle a_p^{n_2}, \int_\Omega k(x,\cdot) f(x)\, d\mathcal{P}\right\rangle_{\mathcal{H}_k} + b_p\right)$. Then we can obtain that

$$\|H_{n_1} - H_{n_1,n_2}\|_\rho^2 = \int \left(H_{n_1}(\mathcal{P}) - H_{n_1,n_2}(\mathcal{P})\right)^2 d\rho \leq 16 C^2 C_f^2 n_1 \exp\left(-\frac{c_k n_2^{1/d}}{\sqrt{d}}\right).$$

Here we have used the bound $\|c\|_1 \leq 2rC$ and $\|a_p\|_{\mathcal{H}_k} \leq \frac{\sqrt{n_1}}{r}$.

Combining this error bound with the estimate in **Step 1** and $r = C_f$ ($\|k\|_\infty = 1$ for Gaussian kernels) results in

$$\|\Phi_{k,f} - H_{n_1,n_2}\|_\rho \leq \frac{4C_f C}{n_1^{1/2}} + 4CC_f n_1^{1/2} \exp\left(-\frac{c_k n_2^{1/d}}{2\sqrt{d}}\right).$$

Let $n_2 = \lceil C_{k,d} (\log n_1)^d \rceil$ with $C_{k,d} = \left(\frac{2\sqrt{d}}{c_k}\right)^d$. Then we have the RHS of the above bounded as

$$\|\Phi_{k,f} - h_n\|_\rho \leq \frac{(8C_f)C}{n^{1/2}} \text{ where } h_n := H_{n_1,n_2} \text{ with } n_1 = n \text{ and } n_2 = \lceil C_{k,d} (\log n)^d \rceil.$$

Insert the linear expressions of $\{a_p^{n_2}\}$ in (2.17) into H_{n_1,n_2} and it can be obtained that

$$H_{n_1,n_2}(\mathcal{P}) = \sum_{p=1}^{n_1} c_p \phi\left(\sum_{q=1}^{n_2} A_{p,q} \int_\Omega k(x,t_q) f(x)\, d\mathcal{P} + b_p\right)$$

where $\{A_{p,q}\}$ is determined by the values of a_p on $\{t_q\}$ and K_T and furthermore, $a_{p,q}$ can be bounded by $\sqrt{n_2} \left\|K_T^{-1}\right\|_2 n_1^{1/2}$ with $K_{\mathbf{T}} = (k(x_i,x_j))_{x_i,x_j \in \mathbf{T}}$. In conclusion, with $O(n(\log n)^d)$ parameters, we can achieve the error bound $\|\Phi_{k,f} - h_n\|_\rho^2 = O(n^{-1})$ with $\|c\|_1 \leq 2C_f C$, $|A_{p,q}| \leq \sqrt{2C_{k,d}(\log n)^d} \left\|K_T^{-1}\right\|_2 n^{1/2}$ and $|b_p| \leq n^{1/2}$ for all p, q.

Let $h_n(\mathcal{P}) := c^T\sigma\left(A[\mathrm{attn}(\mathcal{P})]_{\mathbf{T}_{s(n)}} + b\right)$ with $c \in \mathbb{R}^{2n}, A \in \mathbb{R}^{2n\times s(n)}, b \in \mathbb{R}^{2n}$ and $\mathbf{T}_{s(n)}$ the set of $s(n)$ fixed points. Recall that $\phi(x) = \sigma(x+\frac{1}{2}) - \sigma(x-\frac{1}{2})$. Now we define

$$\mathcal{H}'_{R,n} = \left\{h_n : \|c\|_1 \le 4rC, |A_{p,q}| \le \sqrt{2C_{k,d}(\log n)^d}\left\|K_T^{-1}\right\|_2 n^{1/2} \text{ and } |b_p| \le 2n^{1/2} \text{ for all } p,q\right\}.$$

Since σ is homogeneous,

$$\mathcal{H}'_{R,n} = \left\{h_n : \|c\|_1 \le 8rCn^{1/2}, |A_{p,q}| \le \sqrt{\frac{1}{2}C_{k,d}(\log n)^d}\left\|K_T^{-1}\right\|_2 \text{ and } |b_p| \le 1 \text{ for all } p,q\right\}$$

By Example 1 in [116], for the case of the Gaussian kernel, an upper bound can be derived for $\left\|K_T^{-1}\right\|_2$ that

$$\left\|K_T^{-1}\right\|_2 \le C_1(\log n)^{-d}n^{C_2\log n}$$

where $C_1 = (\alpha\sqrt{\pi})^{-d}C_{k,d}^{-1}$ and $C_2 = d\pi^2\alpha^2C_{k,d}^{2/d}$. Let $R = \max\{\sqrt{\frac{1}{2}C_{k,d}}C_1, C_2, 8rC, 1\}$. Then take the hypothesis space to be

$$\mathcal{H}_{R,n} = \{h_n : \|c\|_1 \le Rn^{1/2}, |A_{p,q}| \le R(\log n)^{-d/2}n^{R\log n} \text{ and } |b_p| \le R \text{ for all } p,q\}.$$

We see the conclusion of Theorem 2.6. ■

2.5.3 Proof of Theorem 2.7

Since $C(P(\Omega))$ is a metric space, it suffices to prove that the hypothesis space is a sequentially compact subset. Let $\{h^{(j)}\}$ be a countable collection of functions in the hypothesis space $\mathcal{H}_{R,n}$. By Lemma 5 in the appendix, it suffices to show that the functions $\{h^{(j)}\}$ are equi-bounded and equi-continuous.

Equi-continuity: For $\mathcal{P}, \mathcal{Q} \in (P(\Omega), \gamma_k)$, we have

$$\begin{aligned}
\left|h^{(j)}(\mathcal{P}) - h^{(j)}(\mathcal{Q})\right| &\leq \|c^{(j)}\|_1 \max_p \left| \sum_{q=1}^{n_2} A_{p,q}^{(j)} \int_\Omega k(x,t_q) f(x)\, d\mathcal{P} - \sum_{q=1}^{n_2} A_{p,q}^{(j)} \int_\Omega k(x,t_q) f(x)\, d\mathcal{Q} \right| \\
&\leq \|c^{(j)}\|_1 \max_p \left\{ \left(\sum_{q=1}^{n_2} \left|A_{p,q}^{(j)}\right| \right) \right\} \max_q \left| \int_\Omega k(x,t_q) f(x)\, d(\mathcal{P}-\mathcal{Q}) \right| \\
&\leq \left\|c^{(j)}\right\|_1 \|k\|_\infty \max_p \left\{ \left(\sum_{q=1}^{n_2} \left|A_{p,q}^{(j)}\right| \right) \right\} \left\| \int_\Omega k(x,\cdot) f(x)\, d(\mathcal{P}-\mathcal{Q}) \right\|_{\mathcal{H}_k}
\end{aligned}$$

By Theorem 2.3, attn is a continuous mapping from $(P(\Omega), \gamma_k)$ to $(\mathcal{H}_k, \|\cdot\|_{\mathcal{H}_k})$, which implies that $\{h^{(j)}\}$ are equi-continuous.

Equi-boundedness: For a collection of functions $\{h^{(j)}\} \subset \mathcal{H}_{R,n}$ with $h^{(j)} : (P(\Omega), \gamma_k) \to (\mathbb{R}, |\cdot|)$, it's sufficient to show that $\{h^{(j)}\}$ is uniformly bounded. For any $n \in \mathbb{N}$, it's easy to obtain that

$$\begin{aligned}
\sup_{\mathcal{P} \in (P(\Omega), \gamma_k)} \left|h^{(j)}(\mathcal{P})\right| &\leq \|c^{(j)}\|_1 \max_p \left[\left(\sum_{q=1}^{n_2} \left|A_{p,q}^{(j)}\right| \right) C_f \|k\|_\infty^2 + \left|b_p^{(j)}\right| \right] \\
&\leq R^2 n^{\frac{1}{2}} \left(C_f s(n) (\log n)^{-d/2} n^{R \log n} + 1 \right)
\end{aligned}$$

Therefore, $\mathcal{H}_{R,n}$ is a compact subset of $C(P(\Omega))$. This completes the proof of Theorem 2.7.

■

Our generalization analysis needs an estimate of the covering numbers of the hypothesis space $\mathcal{H}_{R,n}$, which is given by the following lemma.

LEMMA 3. *For $n \geq 3$, the covering number $\mathcal{N}(\mathcal{H}_{R,n}, \epsilon, \|\cdot\|_\infty)$ induced by the Transformer encoders can be bounded as*

$$\log \mathcal{N}(\mathcal{H}_{R,n}, \epsilon, \|\cdot\|_\infty) \leq R_1 n (\log n)^d \log\left(\frac{\tilde{R}}{\epsilon}\right) + R_2 n (\log n)^{d+2} \tag{2.19}$$

where $R_1 := 6C_{k,d}$, $R_2 := 2(8R + 3d)$ and $\tilde{R} := 6R^2(1 + 2C_{k,d}) \max\{1, 2C_f(C_{k,d} + 1)\}$.

For $h \in \mathcal{H}_{R,n}$, denote $\sigma_h(\mathcal{P}) := \sigma(A[\int_\Omega k(x,\cdot)f(x)\,d\mathcal{P}]_{\mathbf{T}_{s(n)}} + b)$ with the parameters A and b in h. Then we have

$$\begin{aligned}
\|\sigma_h\|_\infty &\le \|A\|_\infty \sup_{\mathcal{P}\in(P(\Omega),\gamma_k)} \left\| \int_\Omega k(x,\cdot)f(x)\,d\mathcal{P} \right\|_\infty + \|b\|_\infty \\
&\le s(n)R(\log n)^{-d/2} n^{R\log n} \|k\|_\infty^2 C_f + R \\
&\le 2C_{k,d}\|k\|_\infty^2 C_f R(\log n)^{d/2} n^{R\log n} + R \\
&\le (2C_f C_{k,d}+1)R(\log n)^{d/2} n^{R\log n}.
\end{aligned}$$

Let $\hat{h}$ be another function in $\mathcal{H}_{R,n}$ with parameters $\hat{c}, \hat{A}_{p,q}$ and $\hat{b}$ such that $\|c-\hat{c}\|_\infty \le \epsilon$, $\left|A_{p,q} - \hat{A}_{p,q}\right| \le \epsilon$ and $\left\|b-\hat{b}\right\|_\infty \le \epsilon$. Then we have that

$$\begin{aligned}
\|\sigma_h - \sigma_{\hat{h}}\|_\infty &\le \left\|A-\hat{A}\right\|_\infty \sup_{\mathcal{P}\in(P(\Omega),\gamma_k)} \left\| \int_\Omega k(x,\cdot)f(x)\,d\mathcal{P} \right\|_\infty + \left\|b-\hat{b}\right\|_\infty \\
&\le (s(n)C_f+1)\epsilon
\end{aligned}$$

which results in a bound on the final output,

$$\begin{aligned}
\left\|h-\hat{h}\right\|_\infty &= \left\|c^T\sigma_h - \hat{c}^T\sigma_{\hat{h}}\right\|_\infty \le \left\|(c-\hat{c})^T\sigma_h\right\|_\infty + \left\|\hat{c}^T(\sigma_h-\sigma_{\hat{h}})\right\|_\infty \\
&\le 2\epsilon n(2C_fC_{k,d}+1)R(\log n)^{d/2}n^{R\log n} + Rn^{1/2}(s(n)C_f+1)\epsilon \\
&\le 2C_3R\,\epsilon\, n(\log n)^{d/2}n^{R\log n} + 2C_3R\,\epsilon\, n^{1/2}s(n) \\
&\le C_4(\log n)^d n^{2R\log n}\epsilon
\end{aligned}$$

for $n \ge 3$ where $C_3 := \max\{2, 4C_f(C_{k,d}+1)\}$ and $C_4 := 2C_3R(1+2C_{k,d})$. Let $\tilde{\epsilon} = C_4(\log n)^d n^{2R\log n}\epsilon$. Then the $\tilde{\epsilon}$-covering number of $\mathcal{H}_{R,n}$ can be bounded

$$\mathcal{N}(\mathcal{H}_{R,n},\tilde{\epsilon},\|\cdot\|_\infty) \leq \left\lceil \frac{2Rn^{1/2}}{\epsilon} \right\rceil^{2n} \left\lceil \frac{2R(\log n)^{-d/2} n^{R\log n}}{\epsilon} \right\rceil^{2ns(n)} \left\lceil \frac{2R}{\epsilon} \right\rceil^{2n}$$
$$\leq \left(\frac{\tilde{R}}{\tilde{\epsilon}}\right)^{4n+2ns(n)} (\log n)^{6dns(n)} n^{16Rn(\log n)s(n)}$$

with $\tilde{R} := 3RC_4$. Then by taking logarithms on both sides, we obtain

$$\log \mathcal{N}(\mathcal{H}_{R,n},\tilde{\epsilon},\|\cdot\|_\infty) \leq [4n+2ns(n)] \log\left(\frac{\tilde{R}}{\tilde{\epsilon}}\right) + 6dns(n)\log(\log n) + 16Rn(\log n)s(n)\log n$$
$$\leq R_1 n(\log n)^d \log\left(\frac{\tilde{R}}{\tilde{\epsilon}}\right) + R_2 n(\log n)^{d+2}$$

where $R_1 := 12C_{k,d}$ and $R_2 := 4(8R+3d)$. ■

2.5.4 Proof of Theorem 2.8

We apply the following concentration inequalities given in [110] for $\mathcal{I}_1(D,\mathcal{H}_{R,n})$ and $\mathcal{I}_2(D,\mathcal{H}_{R,n})$:

$$Prob\left\{\mathcal{I}_1(D,\mathcal{H}_{R,n})) > \frac{1}{2}\left(\mathcal{E}(\pi_M\varphi_{\hat{D},R,n}) - \mathcal{E}(\varphi_\rho)\right) + \epsilon\right\}$$
$$\leq \mathcal{N}\left(\mathcal{H}_{R,n}, \frac{\epsilon}{16M}, \|\cdot\|_\infty\right) \exp\left\{-\frac{3m_1\epsilon}{2048M^2}\right\} \tag{2.20}$$

and

$$Prob\{\mathcal{I}_2(D,\mathcal{H}_{R,n})) > \epsilon\} \leq \exp\left\{-\frac{m_1\epsilon^2}{2(3M+\|h\|_\infty)^2\left(R(\mathcal{H})+\frac{2}{3}\epsilon\right)}\right\}. \tag{2.21}$$

For $\mathcal{I}_3(D,\mathcal{H}_{R,n})$, since $\left\|\pi_M\varphi_{\hat{D},R,n}\right\|_\infty \leq M$ and $|y_i| \leq M$, there holds

$$\begin{aligned}\left|I_3(\hat{D},\mathcal{H}_{R,n})\right| &= \left|\frac{1}{m_1}\sum_{i=1}^{m_1}\left(\pi_M\varphi_{\hat{D},R,n}(\hat{\mu}_i^{m_2})-y_i\right)^2-\left(\pi_M\varphi_{\hat{D},R,n}(\mu_i)-y_i\right)^2\right| \\ &\leq \frac{1}{m_1}\sum_{i=1}^{m_1}4M\left|\varphi_{\hat{D},R,n}(\hat{\mu}_i^{m_2})-\varphi_{\hat{D},R,n}(\mu_i)\right|\end{aligned}$$

where $\varphi_{\hat{D},R,n}(\mu) = c_\varphi^T\left(\sigma\left(A_\varphi\left[\int_\Omega k(x,\cdot)f(x)\,d\mu\right]_{\mathbf{T}_{s(n)}}+b_\varphi\right)\right)$ and $\mathbf{T}_{s(n)}$ denotes a set of $s(n)$ distinct points in Ω.

Let $g_t(x) := k(x,t)f(x)$ for $t \in \mathbf{T}_{s(n)}$. It can be derived that $|g_t(x)| \leq C_f$ for any $x \in \Omega$. For each $t \in \mathbf{T}_{s(n)}$, we conclude that

$$Prob\{|\mathbb{E}(g_t(X_i)) - S_{m_2}(g_t(X_{i,j}))| > \epsilon\} \leq 2\exp\left\{-\frac{m_2\epsilon^2}{8C_f^2}\right\},$$

which follows that

$$Prob\left\{\sup_{t\in\mathbf{T}_{s(n)}}|\mathbb{E}(g_t(X_i)) - S_{m_2}(g_t(X_{i,j}))| > \epsilon\right\} \leq 2s(n)\exp\left\{-\frac{m_2\epsilon^2}{8C_f^2}\right\}.$$

Since

$$\begin{aligned}\sup_{\varphi\in\mathcal{H}_{R,n}}|\varphi(\mu_i)-\varphi(\hat{\mu}_i^{m_2})| &\leq \sup_{\varphi\in\mathcal{H}_{R,n}}\|c_\varphi\|_1\,\|A\|_\infty\sup_{t\in T_n}|\mathbb{E}(g_t(X_i)) - S_{m_2}(g_t(X_{i,j}))| \\ &\leq Rn^{1/2}s(n)R(\log n)^{-d/2}n^{R\log n}\epsilon \\ &\leq C_5 n^{2R\log n}(\log n)^{d/2}\epsilon\end{aligned}$$

where $C_5 := 2C_{k,d}R^2$, it can be concluded that

$$Prob\left\{\sup_{\varphi\in\mathcal{H}_{R,n}}|\varphi(\mu_i)-\varphi(\hat{\mu}_i^{m_2})| > C_5 n^{2R\log n}(\log n)^{d/2}\epsilon\right\} \leq 2s(n)\exp\left\{-\frac{m_2\epsilon^2}{8C_f^2}\right\}$$

which results in the probability concentration of $\mathcal{I}_3(\hat{D},\mathcal{H}_{R,n})$ as

$$Prob\left\{\left|I_3(\hat{D},\mathcal{H}_{R,n}) > \epsilon\right|\right\} \leq 2m_1 s(n)\exp\left\{-\frac{m_2\epsilon^2}{128M^2C_f^2C_5^2n^{4R\log n}(\log n)^d}\right\}$$

Similarly, it can be obtained that

$$Prob\left\{\left|I_4(\hat{D},\mathcal{H}_{R,n})>\epsilon\right|\right\}\leq 2m_1 s(n)\exp\left\{-\frac{m_2\epsilon^2}{32(M^2+\|h\|_\infty^2)C_f^2C_5^2n^{4R\log n}(\log n)^d}\right\}$$

Finally, combine all the probability concentration inequalities. We can derive

$$\begin{aligned}
&Prob\left\{\left\|\pi_M\varphi_{\hat{D},R,n}-\varphi_\rho\right\|_\rho^2>2\left\|h-\varphi_\rho\right\|_\rho^2+8\epsilon\right\}\\
\leq&\mathcal{N}\left(\mathcal{H}_{R,n},\frac{\epsilon}{16M},\|\cdot\|_\infty\right)\exp\left\{-\frac{3m_1\epsilon}{2048M^2}\right\}\\
&+\exp\left\{-\frac{m_1\epsilon^2}{2\left(3M+\|h\|_\infty\right)^2\left(\left\|h-\varphi_\rho\right\|_\rho^2+\frac{2}{3}\epsilon\right)}\right\}\\
&+4m_1s(n)\exp\left\{-\frac{m_2\epsilon^2}{128\max\{\|h\|_\infty^2,M^2\}C_f^2C_5^2n^{4R\log n}(\log n)^d}\right\}.
\end{aligned}$$

This proves Theorem 2.8. ■

2.5.5 Proof of Theorem 2.9

By the construction of the approximation with sigmoidal functions in Theorem 2.3, we have $\|h\|_\infty\leq 2C_fC$.

By inserting the estimations for the approximation error and the covering number into Theorem 2.8, it can be obtained that

$$\begin{aligned}
&\mathrm{Prob}\left\{\left\|\pi_M\varphi_{\hat{D},R,n}-\varphi_\rho\right\|_\rho^2>2C_*^2n^{-1}+8\epsilon\right\}\\
&\leq\exp\left\{R_1n(\log n)^d\log\left(\frac{16M\tilde{R}}{\epsilon}\right)+R_2n(\log n)^{d+2}-\frac{3m_1\epsilon}{2048M^2}\right\}\\
&+\exp\left\{-\frac{m_1\epsilon^2}{2(3M+2C_fC)^2\left(C_*^2n^{-1}+\frac{2}{3}\epsilon\right)}\right\}\\
&+\exp\left\{\log(8C_{k,d}m_1)+d\log(\log n)-\frac{m_2\epsilon^2}{128(2C_fC+M)^2C_f^2C_5^2n^{4R\log n}(\log n)^d}\right\}.
\end{aligned}$$

If $\epsilon \geq 2C_*^2 n^{-1}(\log n)^{d+2}$, then we have

$$
\begin{aligned}
&\text{Prob}\left\{\left\|\pi_M \varphi_{\hat{D},R,n} - \varphi_\rho\right\|_\rho^2 > 9\epsilon\right\} \\
&\leq \exp\left\{R_1 n(\log n)^d \log\left(\frac{8M\tilde{R}n}{C_*^2}\right) + R_2 n(\log n)^{d+2} - \frac{3m_1\epsilon}{2048M^2}\right\} \\
&+ \exp\left\{-\frac{3m_1\epsilon}{8(3M+2C_fC)^2}\right\} \\
&+ \exp\left\{\log(8C_{k,d}m_1) + d\log(\log n) - \frac{m_2\epsilon^2}{128(2C_fC+M)^2C_f^2C_5^2 n^{4R\log n}(\log n)^d}\right\} \\
&\leq \exp\left\{\mathcal{A}_1 n(\log n)^{d+2} - \frac{3m_1\epsilon}{2048M^2}\right\} + \exp\left\{-\frac{3m_1\epsilon}{8(3M+2C_fC)^2}\right\} \\
&+ \exp\left\{\log(8C_{k,d}(\log n)^d m_1) - \frac{m_2\epsilon^2}{\mathcal{A}_2 n^{4R\log n}(\log n)^d}\right\}
\end{aligned}
$$

where $\mathcal{A}_1 := R_1\left(\log\left(\frac{8M\tilde{R}}{C_*^2}\right)+1\right) + R_2$ and $\mathcal{A}_2 := 128(2C_fC+M)^2C_f^2C_5^2$. If we choose the neural network parameter $n = [\mathcal{A}_3 m_1^{1/2}]$ with $\mathcal{A}_3 := (\frac{3C_*^2}{2048\mathcal{A}_1M^2})^{1/2}$, then we have that when $1 \leq \log(8C_{k,d}m_1) \leq \frac{3\mathcal{A}_4^{-1}\mathcal{A}_3^{-1}C_*^2}{4096M^2}m_1^{1/2}$,

$$
\begin{aligned}
\text{Prob}\left\{\left\|\pi_M \varphi_{\hat{D},R,n} - \varphi_\rho\right\|_\rho^2 > 9\epsilon\right\} &\leq \exp\left\{\frac{3m_1\epsilon}{4096M^2} - \frac{3m_1\epsilon}{2048M^2}\right\} + \exp\left\{-\frac{3m_1\epsilon}{8(3M+2C_fC)^2}\right\} \\
&+ \exp\left\{\log(8C_{k,d}m_1) + d\log\log(\mathcal{A}_3 m_1^{1/2}) - \frac{m_2\epsilon^2}{\mathcal{A}_2 n^{4R\log n}(\log n)^d}\right\} \\
&\leq \exp\left\{-\frac{3m_1\epsilon}{4096M^2}\right\} + \exp\left\{-\frac{3m_1\epsilon}{8(3M+2C_fC)^2}\right\} \\
&+ \exp\{\mathcal{A}_4\log(8C_{k,d}m_1) - \frac{m_2\epsilon^2}{\mathcal{A}_2 n^{4R\log n}(\log n)^d}\}
\end{aligned}
$$

where $\mathcal{A}_4 := 2d(1+\log(\log \mathcal{A}_3))$. Take $m_2 \geq \left\lceil \mathcal{A}_5 \left(\mathcal{A}_3 m_1^{\frac{1}{2}}\right)^{8R\log\left(\mathcal{A}_3 m_1^{\frac{1}{2}}\right)} \right\rceil$ with $\mathcal{A}_5 := \frac{3m_1}{4096C_*^2M^2}\mathcal{A}_2$. Then

$$\begin{aligned}\text{Prob}\left\{\left\|\pi_M\varphi_{\hat{D},R,n}-\varphi_\rho\right\|_\rho^2 > 9\epsilon\right\} &\leq \exp\left\{-\frac{3m_1\epsilon}{4096M^2}\right\}+\exp\left\{-\frac{3m_1\epsilon}{8(3M+2C_fC)^2}\right\}\\ &\quad+\exp\left\{\frac{3m_1\epsilon}{4096M^2}-\frac{3m_1\epsilon}{2048M^2}\right\}\\ &\leq 3\exp\left\{-\frac{3m_1\epsilon}{256(4M+2C_fC)^2}\right\}.\end{aligned}$$

Let $\tau = 9\epsilon$ and it can be obtained that

$$\text{Prob}\left\{\left\|\pi_M\varphi_{\hat{D},R,n}-\varphi_\rho\right\|_\rho^2 > \tau\right\} \leq 3\exp\left\{-\frac{m_1^{\frac{1}{2}}\tau}{768(4M+2C_fC)^2}\right\}$$

for $\tau \geq 18C_*^2 n^{-1}(\log n)^{d+2}$. Then by the formula for the expectation of the non-negative random variable ξ, $\mathbb{E}\xi = \int_0^\infty P(\xi > \tau)\,d\tau$, we get

$$\begin{aligned}&\mathbb{E}\{\mathcal{E}(\pi_M\varphi_{\hat{D},R,n})-\mathcal{E}(\varphi_\rho)\} = \int_0^\infty \text{Prob}\{\left\|\pi_M\varphi_{\hat{D},R,n}-\varphi_\rho\right\|_\rho^2 > \tau\}\,d\tau\\ &= \left(\int_0^{18C_*^2n^{-1}(\log n)^{d+2}} + \int_{18C_*^2n^{-1}(\log n)^{d+2}}\right)\text{Prob}\{\left\|\pi_M\varphi_{\hat{D},R,n}-\varphi_\rho\right\|_\rho^2 > \tau\}d\tau\\ &\leq 18C_*^2n^{-1}(\log n)^{d+2} + \int_0^\infty 3\exp\left\{-\frac{m_1^{\frac{1}{2}}\tau}{768(4M+2C_fC)^2}\right\}d\tau\\ &\leq \mathcal{A}_6 m_1^{-\frac{1}{2}}\left(\log\left(\mathcal{A}_3 m_1^{\frac{1}{2}}\right)\right)^{d+2}\end{aligned}$$

where $\mathcal{A}_6 := 36C_*^2\mathcal{A}_3^{-1} + 2304(4M+2C_fC)^2$. This proves the desired bound when the first stage sample size m_1 satisfies

$$1 \leq \log(8C_{k,d}m_1) \leq \frac{3C_*^2}{4096\mathcal{A}_4\mathcal{A}_3M^2}m_1^{\frac{1}{2}} \text{ and } \lfloor \mathcal{A}_3 m_1^{\frac{1}{2}}\rfloor \geq 3,$$

we can see that this restriction on m_1 is satisfied when $m_1 \geq \mathcal{A}_7$ where $\mathcal{A}_7$ is constant depending on $C_{k,d}, C_*, \mathcal{A}_3, \mathcal{A}_4$ and M. When $m_1 < \mathcal{A}_7$, we can also easily see that

$$\mathbb{E}\{\mathcal{E}(\pi_M \varphi_{\hat{D},R,n}) - \mathcal{E}(\varphi_\rho)\} \leq 4M^2 \leq 4M^2 \sqrt{\mathcal{A}_7} m_1^{-\frac{1}{2}} \left(\log\left(\mathcal{A}_3 m_1^{\frac{1}{2}}\right)\right)^{d+2}$$

and thereby the desired bound still holds. This completes the proof of Theorem 2.9. ■

Appendix A

Ascoli-Arzelà theorem

LEMMA 4. *let $\{f_j\}$ be a sequence of continuous functions from a separable topological spaces (X, T) into a metric space (Y, d_Y). Assume that the functions $\{f_j\}$ are equi-bounded and equi-continuous at each $x \in X$. Then, there exists a subsequence $\{f_{j'}\} \subset \{f_j\}$ and a continuous function $f : X \to Y$ such that $\{f_{j'}\} \to f$ pointwise in X. Moreover the convergence is uniform on compact subsets of X.*

Weak Topology of Probability Space

LEMMA 5. *[96]. Let (Ω, m) be a compact metric space. If k is universal, then γ_k metrizes the weak topology on $P(\Omega)$.*

We need to show the topology induced by γ_k is equivalent to the weak topology, i.e., for measures $\{\mathcal{P}_n\} \subset P(\Omega)$, $\mathcal{P}_n \xrightarrow{w} \mathcal{P}$ iff $\gamma_k(\mathcal{P}_n, \mathcal{P}) \to 0$ as $n \to \infty$. First, since k is universal, $\mathcal{H}_k$ is dense in $C(\Omega)$. Then for every $g \in C(\Omega)$ and every $\epsilon > 0$, there exists a $g' \in \mathcal{H}_k$ such that $\|g - g'\|_\infty \leq \epsilon$. Therefore, we have

$$\begin{aligned}
\left| \int_\Omega g \, d\mathcal{P}_n - \int_\Omega g \, d\mathcal{P} \right| &= \left| \int_\Omega (g - g') \, d\mathcal{P}_n + \int_\Omega g' \, d(\mathcal{P}_n - \mathcal{P}) + \int_\Omega (g' - g) \, d\mathcal{P} \right| \\
&\leq \int_\Omega |g - g'| \, d\mathcal{P}_n + \int_\Omega |g' - g| \, d\mathcal{P} + \|g'\|_{\mathcal{H}_k} \gamma_k(\mathcal{P}_n, \mathcal{P}) \\
&\leq 2\epsilon + \|g'\|_{\mathcal{H}_k} \gamma_k(\mathcal{P}_n, \mathcal{P}).
\end{aligned}$$

Since ϵ can be arbitrarily small, $\gamma_k(\mathcal{P}_n, \mathcal{P}) \to 0$ implies that $\mathcal{P}_n \xrightarrow{w} \mathcal{P}$. It's also trivial that $\mathcal{P}_n \xrightarrow{w} \mathcal{P}$ implies $\gamma_k(\mathcal{P}_n, \mathcal{P}) \to 0$, since $\|g\|_\infty \leq \|k\|_\infty \|g\|_{\mathcal{H}_k}$ for $g \in \mathcal{H}_k$.

■

Chapter 3

High-Dimensional Learning Framework

3.1 Introduction

Deep Learning [43] has achieved remarkable successes in processing big data from many practical domains with its superiority in approximation, expressivity, and generalization. Along with the achievements in speech recognition, computer vision, and natural language processing, it is worth noting the recent breakthroughs in scientific research by deep learning methods, e.g., AlphaFold [84] for predicting protein molecule structures. To address the diverse challenges emerging from various research fields, neural network architectures [79, 48, 47, 37] have been developed that possess the capability to process function-input cases with more complex data structures than speeches, images and texts. However, with their impressive performances in solving scientific problems in practice comes a paucity of theoretical understanding about how they work so well across different domains. Before introducing the latest applications and theoretical results about learning with data from function spaces by neural networks, we first recall some definitions and approximation results of neural networks defined on the Euclidean space $\mathbb{R}^d$.

Since the late 1980s, the approximation properties [3, 11, 39] have been well studied with the classical shallow neural networks of form

$$f_N(x) = \sum_{k=1}^{N} c_k \sigma(a_k \cdot x + b_k)$$

with $a_k \in \mathbb{R}^d, b_k, c_k \in \mathbb{R}$ and σ an activation function from $\mathbb{R}$ to $\mathbb{R}$. The fully connected multilayer neural network (FNN) of layer L is defined by applying shallow neural networks inductively,

$$f^{(l)}(x) = \sigma\left(A^{(l)} f^{(l-1)}(x) + b^{(l)}\right), \;\; l = 1, 2, \ldots, L$$

where the activation function σ acts element-wise, $d_l \in \mathbb{N}$ is the number of hidden neurons (width) in l-th layer, $A^{(l)}$ is a $d_l \times d_{l-1}$ matrix without special structures, $b^{(l)} \in \mathbb{R}^{d_l}$ and $f^{(0)}(x) = x$ with $d_0 = d$. The expressivity of FNNs was also well explored in [71, 60, 109, 70].

Later, when deep learning started with the mission to reduce redundant parameters and improve computational efficiency, convolutional neural networks (CNNs) were proposed with the mechanism of weight sharing. Convolutional kernels in CNNs induced by 1-D convolutions were formally defined by Toeplitz matrices in [120, 119]. For a 1-D convolutional filter $\omega = (\omega_k)_{k\in\mathbb{Z}}$ supported in $\{0, 1, \ldots, s\}$ and an input $x = (x_k)_{k\in\mathbb{Z}}$ supported in $\{1, 2, \ldots, d\}$, the 1-D convolution between ω and x is defined as

$$(\omega * x)_i = \sum_{k\in\mathbb{Z}} \omega_{i-k} x_k = \sum_{k=1}^{d} \omega_{i-k} x_k, \quad i \in \mathbb{Z} \tag{3.1}$$

which by considering possibly nonzero entries of $\omega * x$, gives a $(d + s) \times d$ Toeplitz matrix T^ω

$$T^\omega := \begin{bmatrix} \omega_0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ \omega_1 & \omega_0 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \omega_s & \omega_{s-1} & \ldots & \omega_0 & \ldots & 0 & 0 \\ 0 & \omega_s & \ldots & \omega_1 & \ddots & \vdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \ldots & \ldots & 0 & \omega_s & \ldots & \omega_1 & \omega_0 \\ \ldots & \ldots & \ldots & 0 & \omega_s & \cdots & \omega_1 \\ \vdots & \ldots & \ldots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & \ldots & \ldots & 0 & \omega_s \end{bmatrix}.$$

Similarly with the fully connected neural networks, a deep convolutional neural network (DCNN) is defined by the iteration

$$f^l(x) = \sigma(T^{(l)} f^{(l-1)}(x) + b^{(l)}), \quad l = 1, 2, \ldots, L$$

where $T^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ is a Toeplitz matrix, $b^{(l)} \in \mathbb{R}^{d_l}$, $d_l = d + sl$, and $f^{(0)}(x) = x$ with $d_0 = d$.

The FNNs and CNNs defined above show great superiority in tasks in speech recognition, computer vision, and natural language processing, but cannot be applied directly to the tasks of learning operators defined on function spaces with infinite dimensions.

Recently, [92, 93] proposed functional nets to learn functionals with certain smoothness and [89, 110] showed some results in distribution regression with neural networks. Although the recent works [93] achieve almost optimal rates in the respective learning tasks, the claimed network architectures declined to meet those applied in practice, e.g., Fourier Neural Operators (FNOs) and Physic-informed neural networks (PINNs). FNOs [48] are inspired by the integral form of solutions to partial differential equations and extensively used in scientific computing. Convergence rates of approximating solutions to some PDEs by FNOs are studied in [41], while approximating an operator with a more general smooth condition by FNOs has not been considered yet. Apart from the network architecture, the input function space also plays an important role in learning from functional data. For example, [92] took $L^p([-1, 1]^d)$ as the input space and obtained the approximation rate $O\left(\left(\frac{\log(M)}{\log(\log(M))}\right)^{-\beta\lambda/d}\right)$ with M nonzero parameters, when the functional is defined on the unit ball of the Hölder space $C^\beta[-1, 1]^d$ and is Lipschitz λ with $0 < \lambda \leq 1$. To alleviate the effect of the curse of dimension, the Korobov spaces are considered here. Compared with the works [62, 56] on a Korobov space of functions vanishing on the boundary of a cube, the Korobov space discussed in this chapter is a reproducing kernel Hilbert space (RKHS) of periodic functions, and the capacity and the efficiency of approximation by the Korobov space are intrinsically related to the kernel functions of which polynomial cases and exponential cases are taken into consideration.

In this chapter, we are interested in approximating a nonlinear continuous operator $F : H_{d,\alpha,\gamma} \to \mathbb{R}$ by a network consisting of the FNO and a DCNN, and $H_{d,\alpha,\gamma}$ is a Korobov space of periodic functions where a network with an FNO is defined. We establish a theory to show the ability of FNOs to extract features of functions from Korobov spaces. Then we construct a DCNN with multiple channels to realize the high-dimensional interpolation with a great reduction in the number of parameters from $O(d^2)$ to $O(\log_2 d)$. Finally, a convergence rate, which beats the curse of dimension, is achieved by our proposed network. The remainder of this chapter is organized as follows. In Section 2, the definitions of Korobov Spaces and Fourier Functional Networks will be introduced. The main results of this chapter will be established in Section 3, and the proofs of the main results are presented in Section 4.

3.2 Definitions

3.2.1 Korobov Space

The Korobov space $H_{d,\alpha,\gamma}$ of one-periodic functions is a separable Hilbert space with complex-valued functions which can be specified by values on $\mathbb{T}^d := [0,1]^d$. The parameter $\alpha \geq 0$ or $\alpha = \infty$ measures the smoothness of these functions. Throughout the chapter, we always assume $\alpha > 1$ in which case $H_{d,\alpha,\gamma}$ becomes a reproducing kernel Hilbert space of periodic functions.

Let $\{a_j\}$ and $\{b_j\}$ be two positive sequences such that $1 \geq a_1 \geq a_2 \geq \cdots > 0$ and $b := \inf_{j\in\mathbb{N}} b_j > 0$. Then $\gamma = \{\gamma_j\}_{j\in\mathbb{N}}$ can be expressed in terms of the two sequences $\{a_j\}$ and $\{b_j\}$ as $\gamma_j = (a_j, b_j)$. Though the definition of the Korobov space $H_{d,\alpha,\gamma}$ uses only the truncated sequences $\{a_j\}_{j=1}^d, \{b_j\}_{j=1}^d$, our main results stated in Theorems 3.6 and 3.7 below give the dimension-independent rates of approximation and allow d to be as large as possible.

For $h = [h_1, h_2, \ldots, h_d] \in \mathbb{Z}^d$, define the weight function as

$$\omega_\alpha(\gamma, h) = \prod_{j=1}^{d} \omega_\alpha(\gamma_j, h_j).$$

For the polynomial case with $\alpha < \infty$, let $b_j = 1$ for all j and

$$\omega_\alpha(\gamma_j, h_j) = \begin{cases} 1 & \text{if } h_j = 0 \\ |h_j|^\alpha / a_j & \text{if } h_j \neq 0. \end{cases}$$

For $\alpha = \infty$ in the exponential case, we fix $\omega > 1$, and take

$$\omega_\infty(\gamma_j, h_j) = \omega^{|h_j|^{b_j}/a_j}.$$

The Korobov space $H_{d,\alpha,\gamma}$ of complex-valued one-periodic functions is defined on $\mathbb{T}^d$ with a reproducing kernel

$$K_{\alpha,\gamma}(x, y) = \sum_{h \in \mathbb{Z}^d} \frac{\exp(2\pi i h \cdot (x - y))}{\omega_\alpha(\gamma, h)}.$$

Note that in both the polynomial case and exponential case, the kernel is bounded, that is, $\kappa := \sup_{x \in \mathbb{T}^d} \sqrt{K_{\alpha,\gamma}(x, x)} < \infty$.

The inner product on the RKHS can be easily seen by the periodicity of functions as

$$\langle f, g \rangle_H = \sum_{h \in \mathbb{Z}^d} \omega_\alpha(\gamma, h) \hat{f}(h) \overline{\hat{g}(h)}, \quad f, g \in H_{d,\alpha,\gamma}$$

where $\hat{f}$ is the Fourier series of f, given by

$$\hat{f}(h) = \int_{\mathbb{T}^d} f(x) \exp(-2\pi i \, h \cdot x) \, dx$$

and $\overline{\hat{g}(h)}$ is the complex conjugate of the Fourier series $\hat{g}$.

The norm in the RKHS $H_{d,\alpha,\gamma}$ is then given by

$$\|f\|_H = \left(\sum_{h \in \mathbb{Z}^d} \omega_\alpha(\gamma, h) \left| \hat{f}(h) \right|^2 \right)^{1/2}.$$

Since $\omega_\alpha(\gamma, h) \geq 1$, we have

$$\|f\|_{L_2(\mathbb{T}^d)} \leq \|f\|_H$$

for all $f \in H_{d,\alpha,\gamma}$. Thus, $H_{d,\alpha,\gamma} \subset L_2(\mathbb{T}^d)$ where $L_2(\mathbb{T}^d)$ denotes the space of square integrable one-periodic functions with norm $\|f\|_{L_2(\mathbb{T}^d)} = (\int_{\mathbb{T}^d} |f(x)|^2 dx)^{1/2}$.

REMARK 3.1. *Before exhibiting the approximation in the Korobov space $H_{d,\alpha,\gamma}$, we would address the essential assumption $\alpha > 1$ to control the smoothness. For example, taking $d = 1$, if α is an even integer, then for any $f \in H_{d,\alpha,\gamma}$, f is $\frac{\alpha}{2}$ times differentiable, and its k-th derivative is absolutely continuous for $k = 1, \ldots, \frac{\alpha}{2} - 1$ while the $\frac{\alpha}{2}$-th derivative belongs to $L_2(\mathbb{T}^d)$. In the exponential case with $\alpha = \infty$, the Korobov space $H_{d,\alpha,\gamma}$ consists of periodic functions that are analytic. For more details, refer to [64].*

As in learning theory [91], the Korobov space can be understood by an integral operator approach which will enable us to derive projections onto finite-dimensional subspaces for discretization in operator learning. Define the inclusion mapping $\mathrm{id} : H_{d,\alpha,\gamma} \to L_2(\mathbb{T}^d)$ given by $\mathrm{id}(f) = f$. Then $\|\mathrm{id}\| = 1$ because $\|f\|_{L_2(\mathbb{T}^d)} \leq \|f\|_H$ and the equality holds for constant functions. Moreover, the adjoint operator of id is the integral operator $\mathrm{id}^* : L_2(\mathbb{T}^d) \to H_{d,\alpha,\gamma}$ given by

$$\mathrm{id}^*(g) = \int_{\mathbb{T}^d} K_{\alpha,\gamma}(\,\cdot, x)\, g(x) dx.$$

Consider the positive and self-adjoint operator $T := \mathrm{id}^* \,\mathrm{id} : H_{d,\alpha,\gamma} \to H_{d,\alpha,\gamma}$. We have

$$\langle Tf, g\rangle_H = \langle \mathrm{id}(f), \mathrm{id}(g)\rangle_{L_2(\mathbb{T}^d)} = \langle f, g\rangle_{L_2(\mathbb{T}^d)}$$

which follows by the reproducing property that for any $f \in H_{d,\alpha,\gamma}$,

$$\begin{aligned} Tf(x) &= \langle Tf, K_{\alpha,\gamma}(x, \cdot)\rangle_H = \langle f, K_{\alpha,\gamma}(x, \cdot)\rangle_{L_2(\mathbb{T}^d)} \\ &= \sum_{h\in\mathbb{Z}^d} \omega_\alpha^{-1}(\gamma, h)\hat{f}(h) \exp(2\pi i h \cdot x). \end{aligned}$$

Here $\{\omega_\alpha^{-1/2}(\gamma, h) \exp(2\pi i h \cdot\,)\}_{h\in\mathbb{Z}^d}$ is a set of eigenvectors of T and an orthonormal basis of $H_{d,\alpha,\gamma}$. Since

$$\sum_{h\in\mathbb{Z}^d} \omega_\alpha^{-1}(\gamma, h) = \kappa^2 < \infty,$$

T is a trace class operator and id is a Hilbert-Schmidt operator and thus compact. Now we can state projections on the Korobov space for discretization in operator learning. This is done by keeping components with truncating the eigenvalues $\{\omega_\alpha^{-1}(\gamma, h)\}$.

For $\epsilon \in (0, 1)$, let

$$R(\epsilon, d) := \{h \in \mathbb{Z}_d : \omega_\alpha^{-1}(\gamma, h) > \epsilon^2\}$$

be the set of selected eigenvalues and

$$A_{n,d}(f)(x) := \sum_{h \in R(\epsilon,d)} \hat{f}(h) \exp(2\pi i h \cdot x)$$

with $n = |R(\epsilon, d)|$. According to a general procedure about such projections described in Lemma 9 in the Appendix, we see that $A_{n,d}$ achieves an L_2 approximation error ϵ, that is, $\|A_{n,d}(f) - f\|_{L_2(\mathbb{T}^d)} \leq \epsilon \|f\|_H$ for $f \in H_{d,\alpha,\gamma}$.

3.2.2 Fourier Neural Operator

One of the main findings of this chapter is the realization of the projection $A_{n,d}$ by deep FNOs induced by the activation function $\sigma : \mathbb{C} \to \mathbb{C}$ coupled by the rectified linear unit (ReLU) $\max(0, x)$ and defined as $\sigma(z) = \max(0, x) + i \max(0, y)$ for all $z = x + iy \in \mathbb{C}$. A core ingredient of an FNO layer is a finite impulse response (FIR) filter or kernel P. Here, we take the FIR range to be the frequency domain

$$[h]_m = \{h \in \mathbb{Z}^d : |h|_\infty \leq m\} \text{ with } m \in \mathbb{N}.$$

Then we define the truncated Fourier coefficients $\mathcal{F}_m : L_2(\mathbb{T}^d) \to \mathbb{C}^{[h]_m}$ and inverse transform $\mathcal{F}_m^{-1} : \mathbb{C}^{(2m+1)^d} \to L_2(\mathbb{T}^d)$ respectively as

$$\mathcal{F}_m(v)(h) = \int_{\mathbb{T}^d} v(x) \exp(-2\pi i\, h \cdot x)\, dx, \text{ for } h \in [h]_m$$

$$\mathcal{F}_m^{-1}(\hat{v})(x) = \sum_{h \in [h]_m} \hat{v}_h \exp(2\pi i\, h \cdot x), \text{ for } \hat{v} \in \mathbb{C}^{(2m+1)^d}.$$

An FIR kernel $P : [h]_m \to \mathbb{C}$ induces a filter operation on the space of periodic functions as $v \in L_2(\mathbb{T}^d) \to \mathcal{F}_m^{-1}\big(P \odot \mathcal{F}_m(v)\big)$ where $\odot$ is the Hadamard product given by

$$P \odot \mathcal{F}_m(v) = [P(h)\mathcal{F}_m(v)(h)]_{h\in[h]_m}.$$

DEFINITION 3.1 (Fourier Neural Operator). *A deep FNO network $\{v_l\}_{l=0}^L$ of depth $L \in \mathbb{N}$ with $v_0 \in L_2(\mathbb{T}^d)$ is defined as*

$$v_{l+1} = \sigma\Big(w_{l+1}v_l + c_{l+1} + \mathcal{F}_m^{-1}\big(P_{l+1} \odot \mathcal{F}_m(v_l)\big)\Big)$$

where $w_{l+1}, c_{l+1} \in \mathbb{C}$ are parameters and $P_{l+1} : [h]_m \to \mathbb{C}$ is a FIR kernel. For the filter sequence $\{P_l\}_{l=1}^L$, we define the kernel size $s \in \mathbb{N}$ to be $\max_{l\in L} |\operatorname{supp}(P_l)|$ where $\operatorname{supp}(P_l) = \{h \in [h]_m : P_l(h) \neq 0\}$ and $|\operatorname{supp}(P_l)|$ denotes its cardinality.

From Lemma 9 in the appendix, we know that the linear approximation $A_{n,d}$ is the optimal approximation to achieve an error bound ϵ with $n = |R(\epsilon, d)|$. We denote the radius of $R(\epsilon, d)$, to be the largest coordinate in norm, by

$$\tilde{R}(\epsilon, d) = \max_{h\in R(\epsilon,d)} |h_j| .$$

We define our L_2-approximation in the Korobov space $H_{d,\alpha,\gamma}$ by the FNO layers as a mapping $\Psi : H_{d,\alpha,\gamma} \to H_{d,\alpha,\gamma}$ of the form

$$\Psi(f) := \mathcal{L}_L \circ \mathcal{L}_{L-1} \circ \cdots \circ \mathcal{L}_1(f)$$

where for $1 \leq l \leq L-1$, $\mathcal{L}_l(v_{l-1}) = v_l$ as defined in the iteration in Definition 3.1 and $v_0 = f$. For $l = L$, we add an end-to-end skip connection into the structure and remove the activation function:

$$\mathcal{L}_L(v_{L-1}) = w'_L v_0 + w''_L v_{L-1} + c_L + \mathcal{F}_m^{-1}\big(P_L \odot \mathcal{F}_m(v_{L-1})\big).$$

Then we have the following proposition proved in the appendix to show the capacity of FNOs to extract features for the approximation.

PROPOSITION 1. *For the Korobov space $H_{d,\alpha,\gamma}$ with $\alpha > 1$ and any $\epsilon \in (0,1)$, there exists a deep FNO Ψ with the kernel size $1 \leq s \leq |R(\epsilon,d)|$, $\lceil \frac{|R(\epsilon,d)|}{s} \rceil$ layers and $m = \tilde{R}(\epsilon,d)$, such that for any $f \in B_{d,\alpha,\gamma}$, $\Psi(f) = A_{n,d}(f)$ and then $\|f - \Psi(f)\|_{L_2(\mathbb{T}^d)} \leq \epsilon$, where $B_{d,\alpha,\gamma} := \{f \in H_{d,\alpha,\gamma} : \|f\|_H \leq 1\}$.*

The FNO layers project the Korobov function space onto a finite-dimensional space with an estimation on L_2 error, instead of utilizing a fixed polynomial system on $L^2([-1,1]^d)$ in [92, 93]. Korobov spaces exhibit different underlying structures characterized by the weight functions $\omega_\alpha(\gamma, h)$, which necessitates the use of varying systems of n trigonometric polynomials for the n-th optimal approximations (Lemma 9). There is also an advantage of FNO layers over the fixed trigonometric systems to extract finite-dimensional features: the FNO layers with the same network structure can adapt to Korobov spaces with different structures and achieve the optimal approximations as described in Proposition 1.

For Ψ satisfying Proposition 1, the features extracted by the FNO Layers are defined as

$$\mathbf{F}_n(f) = W_{\mathbf{F}} \mathcal{F}'_m(\Psi(f)) \tag{3.2}$$

where $W_{\mathbf{F}}$ is a $2n \times 2(2m+1)^d$ matrix with $2n$ real weights and $\mathcal{F}'_m := \mathrm{vec} \circ \mathcal{F}_m$ with

$$\mathrm{vec}([x_j + iy_j]_{j=1}^t) = [x_1, y_1, \cdots, x_j, y_j, \cdots, x_t, y_t]^T \text{ for any } t \in \mathbb{N}.$$

The effect of the matrix $W_{\mathbf{F}}$ is to preserve information in certain frequency domain i.e., $h \in R(\epsilon, d)$ and filter out the other by taking

$$W_{\mathbf{F}}[\mathrm{Re}(\hat{f}(h)), \mathrm{Im}(\hat{f}(h))]_{h \in [h]_m}^T := [\mathrm{Re}(\hat{f}(h)), \mathrm{Im}(\hat{f}(h))]_{h \in R(\epsilon,d)}^T \in \mathbb{R}^{2n}.$$

REMARK 3.2. *Compared to recent advances in complex-valued neural networks [102, 20], the FNO layers $\mathcal{F}_m \circ \Psi$ also produce complex vectors. However, the input of our FNO layers consists of functions from Korobov spaces rather than complex vectors in $\mathbb{C}^d$. In this setting, the analytic properties of complex functions become less significant when studying the approximation of Lipschitz functionals. Although exploring the use of an FNO structure for approximating complex functions would be interesting, it lies beyond the scope of this chapter.*

3.2.3 Fourier Functional Network

Based on the extracted Fourier features, we now utilize a DCNN with multiple channels to learn an approximation to nonlinear operators. The convolution operation (3.1) and downsampling methods were extensively studied for a DCNN network structure in [120, 119]. Inspired by these works on DCNNs, we define multi-channel CNNs with downsampling operations here.

DEFINITION 3.2 (Convolution with Multiple Channels). *For channel size $c \in \mathbb{N}$, let $\omega := \{w^{(j)}\}_{j=1}^{c}$ be a collection of convolutional kernels supported in $\{0, 1, \ldots, s_c\}$. Then for an input sequence $x = (x_k)_{k\in\mathbb{Z}}$ with $x_k \in \mathbb{R}$ supported in $\{1, 2, \ldots, d\}$, the 1-D multi-channel convolution with a replication padding $[\cdot]$ between ω and x is a summation of the convolutions on each channel, defined as*

$$\omega * [x] = \sum_{j=1}^{c} \sigma\left(w^{(j)} * [x] + b_j\right)$$

where $b_j \in \mathbb{R}^{d+s_c+2}$,

$$[x] = (x_1, x_1, x_2, x_3, \ldots, x_{d-1}, x_d, x_d)$$

and $w^{(j)} * [x]$ is defined to be the vector restricting onto $\{1, ..., d + s\}$ of the sequence filter by (3.1) for $1 \le j \le c$.

DEFINITION 3.3 (Downsampled DCNN with Multiple Channels). *For $D \in \mathbb{N}$, the downsampling operator $\mathcal{D}_\nu : \mathbb{R}^D \to \mathbb{R}^{\mathbf{k}(D)}$ with a scaling parameter $\nu \le D$ is defined as $\mathcal{D}_\nu(v) = (v_{k\nu+1})_{1\le k\le \mathbf{k}(D)}, v \in \mathbb{R}^D$, with*

$$\mathbf{k}(D) := \max\left\{k : k \le \left\lfloor \frac{D}{\nu} \right\rfloor, k\nu + 1 < D\right\}.$$

Then a downsampled DCNN with multichannel kernels $\{\omega_l\}_{l=1}^{L}$ with kernel size s_c has widths $\{d_l\}_{l=0}^{L}$ defined iteratively by $d_0 = d'$ and $d_l = \mathbf{k}(d_{l-1} + s_c + 2)$ for $l = 1, ..., L$, and is a sequence of function vectors

$$v_l(x) = \mathcal{D}_\nu(\omega_l * [v_{l-1}(x)] + b_l) \text{ for } 1 \le l \le L$$

where $b_l \in \mathbb{R}^{d_{l-1}+s_c+2}$ and $v_0(x) = x$.

Motivated by an interpolation framework in [92], the CNN structure defined above can realize a high-dimensional interpolation function with a great reduction in the number of parameters as shown in Lemma 6 below, which plays a significant role in the functional approximation.

LEMMA 6. *Let $x = (x_1, x_2, \ldots, x_d) \in [-1, 1]^d$, then the function $\min(x) = \min\{x_1, \ldots, x_d\}$ can be represented by a DCNN of kernel size 2 and channel size 4 with $\lceil \log_2 d \rceil$ layers and $13\lceil \log_2 d \rceil$ total parameters.*

Then we shall present the definition of our novel Fourier Functional Network. The Fourier Functional Network consists of FNO layers followed by a multichannel DCNN, where FNO layers learn a vector representation for each input function from a Korobov space and the multichannel CNN performs an approximation to functions from a Euclidean space to $\mathbb{R}$.

DEFINITION 3.4. *Let $f \in L_2(\mathbb{T}^d)$. First we define for each $j \in \mathbb{N}$ that*

$$\Phi^{(j)}(f) = \Gamma_L \sigma(W_j \mathbf{F}_n(f) + b_j)$$

where $\mathbf{F}_n$ is defined by (3.2), W_j is a $(4n^2 + 2n) \times 2n$ matrix with $8n^2$ possibly nonzero weights, $b_j \in \mathbb{R}^{4n^2+2n}$, $\Gamma_L : \mathbb{R}^{4n^2+2n} \to \mathbb{R}$ is a DCNN of kernel size 2 and channel size 4 with scaling parameter $\nu = 2$ in Lemma 6. Then for $\mathcal{M} \in \mathbb{N}$, the Fourier Functional Network Φ is defined as

$$\Phi(f) = \sum_{j=1}^{\mathcal{M}} \zeta_j \Phi^{(j)}(f) \tag{3.3}$$

where $\zeta_j \in \mathbb{R}$ for $1 \le j \le \mathcal{M}$.

3.3 Main Results on Fourier Functional Networks

Our main results in this chapter are stated below and proved in the next section. Suppose $F : L_2(\mathbb{T}^d) \to \mathbb{R}$ is a continuous functional with modulus of continuity ω_F defined for $t > 0$ by

$$\omega_F(t) = \sup\{|F(f) - F(g)| : \|f - g\|_{L_2(\mathbb{T}^d)} \le t\}.$$

Then for the unit ball $B_{d,\alpha,\gamma}$, we consider the functional approximation on the subspace $H_{d,\alpha,\gamma}$ of $L_2(\mathbb{T}^d)$ by a Fourier Functional Network Φ. The following theorem demonstrates a general result of the approximation to a nonlinear continuous functional F.

THEOREM 3.5. *Let $d, n, M \in \mathbb{N}$ and $n \geq 2$. If $F : L_2(\mathbb{T}^d) \to \mathbb{R}$ is a continuous functional with the modulus of continuity ω_F, then for the unit ball $B_{d,\alpha,\gamma}$ of $H_{d,\alpha,\gamma} \subset L_2(\mathbb{T}^d)$, there exists a Fourier Functional Network Φ with the number of possibly nonzero parameters at most $2M$ such that*

$$\sup_{f \in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \leq \omega_F\left(\epsilon_{n,d}\right) + 4n\, \omega_F\left[C\big(\frac{n}{\sqrt{M}}\big)^{\frac{1}{n}}\right]$$

where $\epsilon_{n,d} := \sup_{f \in B_{d,\alpha,\gamma}} \|f - A_{n,d}(f)\|_{L_2(\mathbb{T}^d)}$ and C is a constant independent of n or M.

To derive asymptotics of the approximation, we need to investigate properties of the weight function $\omega_\alpha(\gamma, h)$ which yield the approximation bound for $\epsilon_{n,d}$. The polynomial and exponential cases are dealt with in the following lemmas from [65, 15].

LEMMA 7. ***Polynomial Case**: If $1 < \alpha < \infty$, we define the sum-exponent s_a of the sequence $\{a_j\}$ as*

$$s_a = \inf\left\{s > 0 : \sum_{j=1}^{\infty} a_j^s < \infty\right\}.$$

If $s_a < \infty$, then for any positive η, there exists a positive A_η depending on η, α and $\{a_j\}$ such that

$$|R(\epsilon, d)| \leq A_\eta \epsilon^{-(p^* + \eta)} \text{ for all } \epsilon \in (0, 1), \tag{3.4}$$

where $p^* = 2\max(s_a, \alpha^{-1})$ is so-called the exponent of strong tractability.

LEMMA 8. ***Exponential Case**: If $\alpha = \infty$, then for any sequences $\{a_j\}$ and $\{b_j\}$ with $B(d) := \sum_{j=1}^{d} \frac{1}{b_j}$ and $p_d^* := 1/B(d)$, we have the following exponential convergence:*

$$\sup_{f \in B_{d,\alpha,\gamma}} \|f - A_{n,d}(f)\|_{L_2(\mathbb{T}^d)} \leq A_0 \exp\left\{-\left(\frac{n}{A_1}\right)^{p_d^*}\right\} \tag{3.5}$$

where A_0, A_1 are constants, depending on d.

If $B := \sum_{j=1}^{\infty} \frac{1}{b_j} < \infty$ and $p^* = 1/B$, then the uniform exponential convergence can be obtained by (3.5) that

$$\sup_{f \in B_{d,\alpha,\gamma}} \|f - A_{n,d}(f)\|_{L_2(\mathbb{T}^d)} \leq A_0 \exp\left\{ - \left(\frac{n}{A_1} \right)^{p^*} \right\}. \tag{3.6}$$

Then by Lemma 7 and Lemma 8, the approximation rates, which beat the curse of dimension, are obtained from the general result in Theorem 3.5, and stated in the following:

THEOREM 3.6. ***(Polynomial Case)*** *Let $d \in \mathbb{N}, M \geq 8, \beta > 0$ and $1 < \alpha < \infty$. Suppose the parameters of the weight function $\omega_\alpha(\gamma, h)$ satisfy that $s_a < \infty$. Let $p^* = 2\max(s_a, \alpha^{-1})$. If $\omega_F(r) \leq \beta r^\lambda$ for some $\lambda \in (0, 1]$, then there exists a Fourier Functional Network with the number of possibly nonzero parameters at most $2M$ such that for any positive number η, with $p = p^* + \eta$, we have*

$$\sup_{f \in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| = O\left(\left(\frac{\log M}{\log(\log M)} \right)^{-\lambda/p} \right).$$

THEOREM 3.7. ***(Exponential Case)*** *Let $d, M \in \mathbb{N}, \beta > 0$ and $\alpha = \infty$. Suppose the parameters of the weight function $\omega_\alpha(\gamma, h)$ satisfy that $B = \sum_{j=1}^{\infty} \frac{1}{b_j} < \infty$. Let $p^* = 1/B$. If $\omega_F(r) \leq \beta\, r^\lambda$ for some $\lambda \in (0, 1]$ and $M \geq N'_{\lambda,d,p^*} \in \mathbb{N}$ (to be given in the proof), then there exists a Fourier Functional Network with the number of possibly nonzero parameters $2M$ such that*

$$\sup_{f \in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| = O\left(\exp\left(-\tau (\log M)^{\frac{p^*}{p^*+1}} \right) \right)$$

where τ is a constant depending on d, λ, p^*.

3.4 Proof of Main Results on Fourier Functional Networks

3.4.1 Error Decomposition

The approximation of the nonlinear functional F will be analyzed by an error decomposition procedure. For any $f \in B_{d,\alpha,\gamma}$, we have the following error decomposition

$$\begin{aligned} |F(f) - \Phi(f)| &\leq |F(f) - F \circ A_{n,d}(f)| + |F \circ A_{n,d}(f) - \Phi(f)| \\ &\leq \omega_F\big(\|f - A_{n,d}(f)\|_{L_2(\mathbb{T}^d)} \big) + |\phi_F \circ (\mu_n \circ A_{n,d})(f) - \Phi(f)| \end{aligned} \tag{3.7}$$

with $\phi_F = F \circ \mu_n^{-1}$ where

$$\mu_n : (A_{n,d}(B_{d,\alpha,\gamma}), \|\cdot\|_{L_2(\mathbb{T}^d)}) \to (\mathbb{R}^{2n}, \|\cdot\|_2)$$

is an isometric isomorphism given by

$$\mu_n \left(\sum_{h \in R(\epsilon,d)} \hat{f}(h) \exp(2\pi i h \cdot \,) \right) = [\mathrm{Im}(\hat{f}(h)), \mathrm{Re}(\hat{f}(h))]_{h \in R(\epsilon,d)}.$$

For the RHS of (3.7), the first term can be bounded by the approximation error of the projection operator $A_{n,d}$. For the second term, note that $\phi_F : \mathbb{R}^{2n} \to \mathbb{R}$ is a real-valued function and we can estimate the modulus of continuity ϕ_F by ω_ϕ. In fact, for any $u, v \in \mathbb{R}^{2n}$, we have

$$\begin{aligned} |\phi_F(u) - \phi_F(v)| &= \left| F \circ \mu_n^{-1}(u) - F \circ \mu_n^{-1}(v) \right| \\ &\leq \omega_F \left(\left\| \mu_n^{-1}(u) - \mu_n^{-1}(v) \right\|_{L_2(\mathbb{T}^d)} \right). \end{aligned}$$

Since $\|g\|_{L_2(\mathbb{T}^d)} = \left(\sum_{h \in \mathbb{Z}^d} |\hat{g}(h)|^2 \right)^{1/2}$ for $g \in L_2(\mathbb{T}^d)$, the RHS of the above inequality becomes

$$\omega_F \left(\left(\sum_{k=1}^{n} (u_k^{\mathrm{Im}} - v_k^{\mathrm{Im}})^2 + (u_k^{\mathrm{Re}} - v_k^{\mathrm{Re}})^2 \right)^{1/2} \right) = \omega_F(\|u - v\|_2)$$

which follows that $\omega_\phi(r) \leq \omega_F(r)$.

Having obtained the smoothness condition of ϕ_F, we construct a multichannel DCNN by Lemma 6 to approximate ϕ_F in Theorem 3.5.

3.4.2 Proof of Theorem 3.5

Proof. Similarly with Proposition 2 in [92], to construct an approximation to ϕ_F by a multichannel CNN in our Fourier functional network, we define the piecewise linear interpolation $H : \mathbb{R}^{2n} \to \mathbb{R}$ as

$$H(y) = \sum_{\xi \in \mathcal{G}} \phi_F(\xi) \psi\left(\frac{N}{2}(y - \xi)\right) \tag{3.8}$$

where $\mathcal{G} = \left\{-1 + \frac{2}{N} i : i = 0, \cdots, N\right\}^{2n}$ and

$$\psi(y) = \sigma\left(\min\left\{\min_{k \neq j}\left(1 + y_k - y_j\right), \min_k\left(1 + y_k\right), \min_k\left(1 - y_k\right)\right\}\right). \tag{3.9}$$

We first denote a simplex in $\mathbb{R}^t$ by

$$\triangle_{\eta,\rho} = \left\{\mathbf{y} \in \mathbb{R}^{2n} : 0 \leq y_{\rho(1)} - n_{\rho(1)} \leq \cdots \leq y_{\rho(2n)} - n_{\rho(2n)} \leq 1\right\},$$

where $\eta = (\eta_1, \cdots, \eta_{2n}) \in \mathbb{Z}^{2n}, \mathbf{y} = (y_1, \cdots, y_{2n}), \rho \in \mathcal{P}_{2n}$, the set of all permutations of $2n$ elements. Then $\{\triangle_{\eta,\rho}\}_{\eta \in \mathbb{Z}^{2n}, \rho \in \mathcal{P}_{2n}}$ is a partition of $\mathbb{R}^{2n}$. It's easy to check that $\psi(\mathbf{0}) = 1$, and $\psi(\mathbf{y}) = 0$ for $\mathbf{y} \in \mathbb{Z}^{2n} \backslash \{\mathbf{0}\}$ and ψ is linear in each simplex $\triangle_{\eta,\rho}$ for $\eta \in \mathbb{Z}, \rho \in \mathcal{P}_{2n}$.

By Lemma 6, we construct the linear interpolation ψ by multichannel CNNs with much fewer parameters. Since ψ is in the form of

$$\sigma(\min\{a_i : i = 1, \cdots, 4n^2 + 2n\}) = \min\{\sigma(a_i) : i = 1, \cdots, 4n^2 + 2n\},$$

then the piecewise linear interpolation $\psi : \mathbb{R}^{2n} \to \mathbb{R}$ can be represented by a DCNN of kernel size 2 and channel size 4 with $\lceil \log_2(4n^2 + 2n) \rceil$ layers and $13\lceil \log_2(4n^2 + 2n) \rceil$ in the form of

$$\psi(y) = -\Gamma_L(\sigma(W_2 y + b_2)) \tag{3.10}$$

where $L = \log_2(4n^2 + 2n)$, W_2 is a $(4n^2 + 2n) \times (2n)$ matrix with $8n^2$ nonzero weights $b_2 = \mathbf{1}$ such that

$$W_2 y + b_2 = [1 + y_1, \cdots, 1 + y_{2n}, 1 - y_1, \cdots, 1 - y_{2n},$$
$$1 + y_1 - y_2, \cdots, 1 + y_1 - y_{2n}, \cdots, 1 + y_{2n} - y_1, \cdots, 1 + y_{2n} - y_{2n-1}]$$

Let $\psi_\xi = \psi\left(\frac{N}{2}(y - \xi)\right)$ and then it follows that

$$\psi_\xi = -\Gamma_L(\sigma(W_{2,N} y + b_{2,\xi}))$$

where $W_{2,N} = \frac{N}{2} W_2$ and $b_{2,\xi} = \mathbf{1} - W_{2,N}\xi$. Then H can be written into

$$H(y) = \sum_{\xi \in \mathcal{G}} (-\phi_F(\xi)) \Gamma_L(\sigma(W_{2,N} y + b_{2,\xi})) \tag{3.11}$$

which can viewed as a linear combination of the outputs of the multichannel CNN with different input features. This structured network has the depth $\lceil \log_2(4n^2 + 2n) \rceil$ and number of nonzero weights

$$M \leq 2(N+1)^{2n} + 8n^2 + 13\lceil \log_2(4n^2 + 2n) \rceil \leq c_1 n^2 (N+1)^{2n} \tag{3.12}$$

for some absolute constant c_1.

Next, we show how to approximate ϕ_F by H and estimate the approximation error. For any $y \in [-1, 1]^{2n}$, there exists a simplex $\triangle$ containing y. Then we denote the restriction of H on $\triangle$ by $H_\triangle$. Then for $r > 0$,

$$\begin{aligned} \omega_{H_\triangle}(r) &= \sup\{|H_\triangle(y_1) - H_\triangle(y_2)| : \|y_1 - y_2\|_2 \leq r, y_1, y_2 \in \triangle\} \\ &\leq \sup_{y \in \triangle} \|\nabla H_\triangle(y)\|_2 \, r \leq \sqrt{2n} \sup_{y \in \triangle} |\nabla H_\triangle(y)|_\infty \, r \end{aligned} \tag{3.13}$$

where ∇ is the gradient operator. Since $H_{\triangle}$ is linear in $\triangle$ and interpolates ϕ_F on every node, we can obtain that

$$|\partial_j H_{\triangle}(y)| = \left| \frac{N}{2}(\phi_F(\alpha_j) - \phi_F(\beta_j)) \right| \leq \frac{N}{2}\omega_\phi\left(\frac{2}{N}\right)$$

for $1 \leq j \leq 2n$, where α_j, β_j are the vertices of $\triangle$ with the same coordinates except for the j-th coordinate. Since H coincides with ϕ_F on every vertex of $\triangle$, it follows that

$$\begin{aligned}
\sup_{y\in[-1,1]^{2n}} |\phi_F(y) - H(y)| &\leq \omega_\phi\left(\frac{\sqrt{2n}}{N}\right) + \omega_{H_{\triangle}}\left(\frac{\sqrt{2n}}{N}\right) \\
&\leq \omega_\phi\left(\frac{\sqrt{2n}}{2}\cdot\frac{2}{N}\right) + n\omega_\phi\left(\frac{2}{N}\right) \\
&\leq \omega_\phi\left(\lfloor\sqrt{2n}\rfloor\frac{2}{N}\right) + n\omega_\phi\left(\frac{2}{N}\right) \\
&\leq \sqrt{2n}\omega_\phi\left(\frac{2}{N}\right) + n\omega_\phi\left(\frac{2}{N}\right) \\
&\leq 4n\omega_\phi\left(\frac{2}{N}\right) \leq 4n\omega_F(2/N).
\end{aligned}$$

Since $N \geq \frac{M^{1/2n}}{c_1 n^{1/n}}$ from (3.12), we have that

$$\sup_{y\in[-1,1]^{2n}} \|\phi_F(y) - H(y)\| \leq 4n\omega_F\left(\frac{c_2 n^{\frac{1}{n}}}{M^{\frac{1}{2n}}}\right) \tag{3.14}$$

where $c_2 = 2c_1$.

Let our Fourier functional network $\Phi_{\mathcal{M}}$ with $\mathcal{M} = (N+1)^{2n}$ to be

$$H(\mathbf{F}_n(f)) = \sum_{\xi\in\mathcal{G}}(-\phi_F(\xi))\Gamma_L(\sigma(W_{2,N}\mathbf{F}_n(f) + b_{2,\xi})). \tag{3.15}$$

Then from (3.7) and Proposition 1, it can be obtained that

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \leq \omega_F\left(\epsilon_{n,d}\right) + 4n\omega_F\left(\frac{Cn^{\frac{1}{n}}}{M^{\frac{1}{2n}}}\right). \tag{3.16}$$

Take M to be the largest number satisfying (3.12), then the total number of parameters $4n+M$ can be bounded by $2M$. The proof of Theorem 3.5 is complete.

■

3.4.3 Proof of Theorem 3.6

Proof. With the assumption on the weight function $\omega_\alpha(\gamma, h)$ in the polynomial case, we have $\epsilon_{n,d} \leq C_\eta n^{-1/(p^*+\eta)}$ with $p^* = 2\max(s_a, \alpha^{-1})$ and then for any $p = p^* + \eta > p^*$, it follows that

$$\begin{aligned}\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| &\leq \beta \left(C_\eta n^{-1/p}\right)^\lambda + 4\beta n \left(C n^{1/n} M^{-1/(2n)}\right)^\lambda \\ &\leq \beta C_\eta^\lambda n^{-\lambda/p} + 4\beta C^\lambda n^{\lambda/n} n M^{-\lambda/(2n)}.\end{aligned}$$

Since $n^{1/n} \leq 3$ for all $n \in \mathbb{N}$, by taking $C_{\beta,\lambda,\eta} = 16\beta(3C^\lambda + C_\eta^\lambda)$ we have

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \leq C_{\beta,\lambda,\eta} \left(n^{-\lambda/p} + nM^{-\lambda/(2n)}\right).$$

To obtain a convergence rate with respect to the number of possibly nonzero parameters $2M$, we need to build some relations between n and M. We choose n to be the integer such that

$$C_{\lambda,p} n \log n \leq \log M < C_{\lambda,p}(n+1)\log(n+1)$$

where $C_{\lambda,p} = 2(p^{-1} + \lambda^{-1}) > 2$. It follows that

$$-\frac{\lambda}{p}\log n \geq \log n - \frac{\lambda}{2n}\log M$$

and therefore, $n^{-\lambda/p} \geq nM^{-\lambda/(2n)}$. Then we have

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \leq 2C_{\beta,\lambda,\eta} n^{-\lambda/p}.$$

Since $n \geq 2$, we have $n \leq \frac{\log M}{C_{\lambda,p}\log n} \leq \log M$, which follows that

$$\log M < C_{\lambda,p}(n+1)\log(n+1) \leq 4C_{\lambda,p} n \log n \leq 4C_{\lambda,p} n \log(\log M).$$

Finally, we can obtain the result that

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \leq 2C_{\beta,\lambda,\eta}(4C_{\lambda,p})^{\lambda/p} \left(\frac{\log M}{\log(\log(M))}\right)^{-\lambda/p}.$$

It's easy to see $4n \leq M$ by the relation between n and M. Then the total number is less than $2M$. This completes the proof of Theorem 3.6. ■

3.4.4 Proof of Theorem 3.7

Proof. In the exponential case, $\epsilon_{n,d}$ has an exponential convergence rate as $C_0 \exp\left(-\frac{n^{p^*}}{C_1}\right)$ where C_0, C_1 are constants depending on dimension d. Then we have

$$\begin{aligned}\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| &\leq \beta\left(C_0 \exp\left(-\frac{n^{p^*}}{C_1}\right)\right)^{\lambda} + 4\beta n\left(Cn^{1/n}M^{-1/(2n)}\right)^{\lambda} \\ &\leq \beta C_0^{\lambda} \exp\left(-\frac{\lambda n^{p^*}}{C_1}\right) + 4\beta C^{\lambda} n^{\lambda/n} n M^{-\lambda/(2n)}.\end{aligned}$$

Let $C'_{\beta,\lambda,\eta} = 16\beta(3C^{\lambda} + C_0^{\lambda})$ and then it's obtained that

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \leq C'_{\beta,\lambda,\eta}\left(\exp\left(-\frac{\lambda n^{p^*}}{C_1}\right) + nM^{-\lambda/(2n)}\right).$$

Similarly, we need to balance the two terms on the RHS. Here we choose n to be the smallest integer not less than $N_{p^*} \in \mathbb{N}$ such that

$$C_{\lambda,d} n^{p^*+1} \leq \log M < C_{\lambda,d}(n+1)^{p^*+1}$$

where $C_{\lambda,d} = 2\left(\frac{1}{C_1} + \frac{1}{\lambda}\right) > 2$ and N_{p^*} is determined by p^* in the form of $N_{p^*} = \min\{n \in \mathbb{N} : n^{p^*} \geq \log n\}$. It follows that

$$\frac{2}{C_1} n^{p^*+1} + \frac{2}{\lambda} n \log n \leq C_{\lambda,d} n^{p^*+1} \leq \log M$$

and that $\exp\left(-\frac{\lambda n^{p^*}}{C_1}\right) \geq nM^{-\lambda/(2n)}$. Then we have

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \leq 2C'_{\beta,\lambda,\eta} \exp\left(-\frac{\lambda n^{p^*}}{C_1}\right).$$

Similarly, we also have $\log M < C_{\lambda,d}(n+1)^{p^*+1} \leq C_{\lambda,d} 2^{p^*+1} n^{p^*+1}$, and then we can obtain the final result

$$\sup_{f\in B_{d,\alpha,\gamma}} |F(f) - \Phi(f)| \leq 2C'_{\beta,\lambda,\eta} \exp\left(-\frac{\lambda}{C_{\lambda,d,p^*}} (\log M)^{p^*/(p^*+1)}\right)$$

with $C_{\lambda,d,p^*} = C_1(C_{\lambda,d}2^{p^*+1})^{p^*/(p^*+1)}$, when M is bigger than N'_{λ,d,p^*} where $N'_{\lambda,d,p^*} = \min\{m \in \mathbb{N} : \log m \geq C_{\lambda,d}N_{p^*}^{p^*+1}\}$. This proves Theorem 3.7. ■

Appendix B

Finite-Dimensional Projection

LEMMA 9. *[64] $S : H \to G$ be a bounded linear operator between a Hilbert space H with $\dim(H) \geq n$ and another Hilbert space G. Let $\sigma_i = \sqrt{\lambda_i}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the eigenvalues of $W = S^*S : H \to H$ with $W(e_i) = \lambda_i e_i$ and orthonormal $\{e_i\}$. Then the linear algorithm*

$$A_n(f) = \sum_{i=1}^{n} \langle f, e_i \rangle S(e_i)$$

is the n-th optimal approximation.

Proof of Lemma 6

Proof. We adopt a Divide and Conquer method here by noticing that $\min(x) = \min\{g(x)\}$ where

$$g(x) = \begin{cases} (\min(x_1, x_2), \min(x_3, x_4), \ldots, \min(x_{d-1}, x_d)), & \text{if } d \text{ is even} \\ (\min(x_1, x_2), \min(x_3, x_4), \ldots, \min(x_{d-2}, x_{d-1}), x_d), & \text{if } d \text{ is odd.} \end{cases}$$

Note that $\min(x_i, x_j) = -\max(-x_i, -x_j) = -\left(\frac{-x_i - x_j}{2} + \frac{|-x_i + x_j|}{2}\right)$ and that $|-x_j + x_i| = \sigma(x_j - x_i) + \sigma(x_i - x_j)$. Then it can be easily obtained that $\max(-x_j, -x_i)$ can be represented as a convolution operation with size 2 kernels and 4 channels. Let $\omega = \{w^{(1)}, w^{(2)}, w^{(3)}, w^{(4)}\}$ with $\omega^{(1)} = \left(-\frac{1}{2}, 0\right), \omega^{(2)} = \left(0, -\frac{1}{2}\right), \omega^{(3)} = \left(-\frac{1}{2}, \frac{1}{2}\right)$ and $\omega^{(4)} = \left(\frac{1}{2}, -\frac{1}{2}\right)$ and bias vectors

$b_1 = \mathbf{1}, b_2 = \mathbf{1}, b_3 = \mathbf{0}$ and $b_4 = \mathbf{0}$. Then

$$\sigma(\omega^{(1)} * [x] + b_1) = \left(1, 1 - \frac{x_1}{2}, 1 - \frac{x_1}{2}, 1 - \frac{x_2}{2}, 1 - \frac{x_3}{2}, \dots, \right)$$
$$\sigma(\omega^{(2)} * [x] + b_2) = \left(1 - \frac{x_1}{2}, 1 - \frac{x_1}{2}, 1 - \frac{x_2}{2}, 1 - \frac{x_3}{2}, \dots, \right)$$
$$\sigma(\omega^{(3)} * [x] + b_3) = \left(\sigma\left(\frac{x_1}{2}\right), 0, \sigma\left(\frac{x_2 - x_1}{2}\right), \sigma\left(\frac{x_3 - x_2}{2}\right), \dots, \right)$$
$$\sigma(\omega^{(4)} * [x] + b_4) = \left(\sigma\left(-\frac{x_1}{2}\right), 0, \sigma\left(\frac{x_1 - x_2}{2}\right), \sigma\left(\frac{x_2 - x_3}{2}\right), \dots, \right).$$

It is easy to notice that starting from the third element, the sums of the elements at odd positions make up the negatives of desired minimums. These values can be retrieved by the downsampling operator with scaling parameter $\nu = 2$. With the replication padding, the last minimum sampled is always $\min(x_d, x_d)$ or $\min(x_{d-1}, x_d)$, instead of the value $\min(x_d, 0)$ at the end of vectors. No matter whether d is even or odd, the minimums of every two neighbor elements defined above are the elements of the output vector of $-\mathcal{D}_2(\omega * x + b)$ with $b = -\mathbf{2}$. Next, we show that the above CNN with $\lceil \log_2(d) \rceil$ layers can represent the minimum function.

The input dimension can be always written as $d = 2^p + q$ where $p, q \in \mathbb{N}$ and $q < 2^p$. Recall that the expression of the downsampling operator is $\mathcal{D}_2(v) = (v_{2k+1})_{1 \le k \le \mathbf{k}}$ where $\mathbf{k} = \lfloor \frac{D+1}{2} \rfloor$ for $v \in \mathbb{R}^{D+3}$.

Given $p', q' \in \mathbb{N}$ such that $q' < 2^{p'}$. Assume that for any $p \le p', q \le q'$ with $q < 2^p$, the minimum of $2^p + q$ elements can be obtained by the above CNN with $\lceil \log_2(2^p + q) \rceil$, i.e., $p + 1$ layers.

For $p = p' + 1, q = q'$, we know that after one layer of the CNN with downsampling, the dimension is reduced to $\left\lfloor \frac{2^{p'+1} + q' + 1}{2} \right\rfloor = 2^{p'} + \lfloor \frac{q'+1}{2} \rfloor$. Then the minimum can be obtained with another p' layers due to $\lfloor \frac{q'+1}{2} \rfloor \le q' < 2^{p'}$. Thus for $D = 2^{p'+1} + q'$, we can get the minimum by the CNN with $p' + 1$ layers.

For $p = p', q = q' + 1$, if $q' + 1 = 2^{p'}$, it corresponds to the case where $p = p' + 1$ and $q = 0$, under which the minimum can be retrieved with the CNN of $p' + 1$ layers. Otherwise, after one layer of the CNN, the dimension is reduced to $2^{p'-1} + \lfloor \frac{q'}{2} + 1 \rfloor$. Then the minimum can

be obtained with another $p' - 1$ layers, since $q' < 2^{p'} - 1$. Thus, the CNN of p' layers can get the minimum with the input dimension $2^{p'} + q' + 1$.

By induction, we show that for any input vector with dimension d, the minimum of elements can be obtained with the CNN of $\lceil \log_2(d) \rceil$ layers. ■

Chapter 4

In-Context Learning of Efficient Transformers

4.1 Introduction

Transformer-based neural networks have become the foundation of modern deep learning frameworks for natural language processing [14, 6], computer vision [24, 68] and scientific discovery applications [35]. Especially when trained with large and diverse corpora, transformers exhibit remarkable few/zero-shot generalization capabilities across various downstream tasks [6]. An underlying mechanism for this behavior is in-context learning, in which pretrained large language models (LLMs) condition on instructions or a few input-output pairs (both referred to as *prompts*) and make predictions on test examples without parameter updates. Many theoretical and empirical studies [105, 1, 19, 115] have demonstrated that the emergence of in-context learning capability is closely related with the context-aware structure of the attention mechanism [101] which enables each token in a sequence to adaptively weight information from all other tokens and to produce a representation conditioned on the sequence context.

However, the original softmax attention mechanism in Vaswani et al. [101] is a double-edged sword: while it is beneficial for context-based representation learning, it suffers from quadratic computational complexity with respect to the context length [112]. As context length grows dramatically, the quadratic computational and memory costs of the standard attention increasingly hinder autoregressive training and inference, undermining LLM performance in long-context modeling scenarios such as processing entire codebases, preserving coherence in long conversations and performing in-depth reasoning across several documents. Therefore,

it's crucial for the design of LLMs to alleviate the curse of quadratic complexity and improve long-context processing capabilities. Numerous works have been recently proposed with this motivation and show performance comparable to the standard attention, including RetNet [97], Mamba [22], and Gated Linear Attention [107]. One branch of these works is known as linear attention [38, 7, 107, 114], which replaces the exponential similarity function with a dot product of key/query functions and yields a linear computational complexity with respect to the context length. This reduction in time and memory cost enables much longer contexts and lower latency during the inference. Meanwhile, the introduction of the hidden-state memory matrix and the forgetting gate improves algorithm stability over utra-long contexts [107]. Although linear attention models have outperformed the softmax attention on some long-context modeling tasks, the theoretical understanding and design principles of these models, especially for in-context learning, remain limited and unexplored.

In this chapter, we investigate the approximation and generalization abilities of linear transformers with context-augmented inputs to reveal the advantage of the linear attention mechanism for the in-context learning scenario. We establish a connection between in-context learning and domain generalization frameworks and show that transformer-based neural networks perform in-context learning as domain generalization [4, 5], and this connection demonstrates the essence of LLMs' remarkable few/zero-shot generalization capabilities without parameter updates during testing. We work with the formulation in Liu and Zhou [49] by representing the context information as a kernel embedding from context probability distributions to vector-valued functions, and exhibit how each word token interacts with the context embedding through an inner product of a tensor product Hilbert space. Based on this framework, we construct a linear transformer to perform in-context learning via a two-staged sampling process, which shows the internal mechanism of the robust generalization capabilities of LLMs. Our main contributions are as follows.

- We present a theoretical analysis framework for the family of linear transformers, one of the most compelling alternatives to the softmax attention in practice. By connecting domain generalization framework with in-context learning, we rigorously prove that linear transformers perform a robust generalization ability under

distribution shifts, which builds the theoretical foundation for understanding the generalization abilities of linear transformers in long-context modeling.

- We observe a fast eigendecay phenomenon in the softmax attention weight matrix products of LLMs. We prove that this phenomenon helps linear transformers alleviate the negative effects of distribution shifts, achieve dimension-independent convergence rates in approximation and generalization analysis and efficiently mimic the behavior of the standard attention.
- We investigate the application of likelihood ratio moments to control distribution shifts in domain generalization. We apply a relaxed condition for unbounded likelihood ratios of probability distributions defined on the noncompact space $\mathbb{R}^d$ and obtain a distribution-dependent concentration inequality for a second stage estimation by the connection between subgaussian norm and finite Rényi divergence.
- We propose a new algorithm for linear conversion of LLMs with the softmax attentions based on our theoretical analysis framework. This conversion scheme captures information from data distributions and parameter matrices in pretrained softmax LLMs. It provides a new perspective for designing new activation functions and training loss for linear conversion of softmax LLMs.

In the following part of this chapter, we first introduce the motivation and definition of linear transformers and a two-staged sampling process as our learning framework. Section 4.3 presents the main results on approximation and generalization, with a proof sketch in Subsection 4.3.3. Section 4.4 provides further discussion, and Appendix 4.5 contains the full proofs.

4.2 Linear Transformers and Formulations for In-Context Learning

In this section, we first define the structure of linear Transformers whose inputs are pairs $(\hat{\rho}, x)$, where $\hat{\rho}$ the empirical version of distribution ρ from which x is sampled. We refer to learning with samples $(\hat{\rho}, x)$ as *in-context learning*, and these samples are generated from the two-staged sampling process defined in Subsection 4.2.2.

4.2.1 Linear Transformers

The standard Transformer [101] consists of blocks of attention mechanisms and shallow networks to process sequential inputs. Let the input sequence $Q = [x_1, \cdots, x_n]^T$ with token vectors $x_i \in \mathbb{R}^d$ for $1 \le i \le n$. Then Q is an input sequence of length n with feature dimension d. The softmax attention is defined as, for $1 \le i \le n$,

$$\mathrm{SoftmaxAttn}(x_i|Q) = \frac{\sum_{j=1}^n \mathrm{sim}(x_i, x_j)(W_v x_j)}{\sum_{j=1}^n \mathrm{sim}(x_i, x_j)} \in \mathbb{R}^d \text{ with } \mathrm{sim}(x_i, x_j) = \exp\left(\frac{\langle W_q x_i, W_k x_j\rangle}{\sqrt{d'}}\right) \tag{4.1}$$

where $W_v \in \mathbb{R}^{d\times d}, W_q \in \mathbb{R}^{d'\times d}, W_k \in \mathbb{R}^{d'\times d}$ are parameter matrices for *value*, *query*, and *key* token vectors respectively. Intuitively, the attention mechanism $\mathrm{SoftmaxAttn}$ takes the input sequence Q as context and produces a refined context-aware representation $\mathrm{SoftmaxAttn}(x_i|Q)$ for each query token x_i in Q.

To establish connections between each token and their context and to demonstrate the benefits of context-aware representation produced by the attention mechanism, we follow Liu and Zhou [49] to assume that Q is a realization with n samples drawn i.i.d. from a Borel probability measure ρ on $\mathbb{R}^d$, with ρ regarded as ***the ground truth context***. Then the RHS of the expression (4.1) can be written as

$$\mathrm{SoftmaxAttn}(\hat{\rho}, x_i) = \frac{\int \mathrm{sim}(x_i, x)(W_v x)\, d\hat{\rho}(x)}{\int \mathrm{sim}(x_i, x)\, d\hat{\rho}(x)} \in \mathbb{R}^d \tag{4.2}$$

where $(\hat{\rho}, x_i)$ is a context-augmented input and $\hat{\rho} = \delta([x_1, \cdots, x_n])$ is ***the accessible context*** with $\delta(\mathcal{S})$ defined as the empirical distribution generated by the dataset $\mathcal{S}$. With a richer accessible context $\hat{\rho}$ by more and more samplings from ρ, the empirical distribution $\hat{\rho}$ can recover the information of the population distribution ρ and produce more refined context-aware representation for each token x_i, which is consistent with the empirical practice of increasing the context window length n.

However, it's easy to observe that for each query token x_i ($1 \le i \le n$), we must evaluate $\mathrm{sim}(x_i, x_j)$ for all $1 \le j \le n$ and then normalize by $\sum_{j=1}^n \mathrm{sim}(x_i, x_j)$, which creates an $n \times n$ attention score matrix and causes both time and memory cost to scale quadratically in

the context length n [112]. Such quadratic growth poses significant challenges for efficient algorithm designs in long-context modeling scenarios. The key idea of linear Transformers is to reduce the quadratic time and memory cost to linear dependence on the context length n by decoupling queries and keys in $\mathrm{sim}(x_i, x_j)$. By replacing the similarity function $\mathrm{sim}(x_i, x_j)$ with $\phi(x_i)^T\phi(x_j)$ where $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a feature mapping, a simple linear attention module can be written as

$$\mathrm{LinearAttn}(\hat{\rho}, x_i) = \frac{\phi(x_i)^T \int \phi(x)(W_v x)\, d\hat{\rho}(x)}{\phi(x_i)^T \int \phi(x)\, d\hat{\rho}(x)} = \frac{\big[\int (W_v x)\phi(x)^T\, d\hat{\rho}(x)\big]\phi(x_i)}{\big[\int \phi(x)^T\, d\hat{\rho}(x)\big]\phi(x_i)}. \quad (4.3)$$

It's easy to observe that a universal memory matrix $\int (W_v x)\phi(x)^T d\hat{\rho}(x) \in \mathbb{R}^{d\times d'}$ can be shared across all query tokens, thus eliminating the need to compute and store the quadratic attention score matrix in (4.1). Beyond this computational efficiency, recent empirical studies [73, 107, 114] have achieved performances comparable to the softmax attention using linear attentions. Motivated by normalization-free linear attentions in Qin et al. [73] and the design of shallow neural network feature mapping ϕ in Zhang et al. [114], we define *Linear Transformers* as follows. Let $\sigma : \mathbb{R} \to \mathbb{R}$ denote the ReLU activation function $\sigma(u) = \max\{u, 0\}$, and denote $\sigma_{\tanh} : \mathbb{R} \to \mathbb{R}$ the tanh activation function $\sigma_{\tanh}(u) = \frac{\exp(u)-\exp(-u)}{\exp(u)+\exp(-u)}$.

DEFINITION 4.1. *A Linear Transformer* T_n *with the structure of* $\mathrm{T}_{n,m,\tilde{m}}$ *and* $m = m(n)$, $\tilde{m} = \tilde{m}(n)$ *is defined as*

$$\mathrm{T}_n(\rho, x) = \sum_{j=1}^{n} \alpha_j \sigma \left(\sum_{q=1}^{m(n)} \phi_{q,\tilde{m}(n)}(x) \left(\sum_{p=1}^{m(n)} \mathcal{T}_v \left[\int \phi_{p,\tilde{m}(n)}(y) A_{p,q}^{(j)} y\, d\rho(y) \right] \right) + b_j \right) + b_0 \quad (4.4)$$

for context-augmented input (ρ, x) *with* $A_{p,q}^{(j)} \in \mathbb{R}^{d\times d}$, $b_j, b_0 \in \mathbb{R}^d$ *and* $\alpha_j \in \mathbb{R}$ *for* $1 \le j \le n$, *and two-hidden-layer tanh neural networks* $\{\phi_{q,\tilde{m}(n)}\}_{q=1}^{m(n)}$ *with the product gate, in the form of*

$$\phi_{q,\tilde{m}(n)} = \mathcal{T}_{1,\odot}\Big(\mathcal{NN}_{q,\tilde{m}(n)}\Big) \text{ and } \mathcal{NN}_{q,\tilde{m}(n)}(x) = W_{q,2}\sigma_{\tanh}(W_{q,1}\sigma_{\tanh}(W_{q,0}x + b_{q,0}) + b_{q,1})$$

with $W_{q,0} \in \mathbb{R}^{8d\tilde{m}(n)\times d}$, $b_{q,0} \in \mathbb{R}^{8d\tilde{m}(n)}$, $W_{q,1} \in \mathbb{R}^{8d\tilde{m}(n)\times 8d\tilde{m}(n)}$, $b_{q,1} \in \mathbb{R}^{8d\tilde{m}(n)}$ *and* $W_{q,2} \in \mathbb{R}^{d\times 8d\tilde{m}(n)}$, *where* $\mathcal{T}_{1,\odot}(z) = \prod_l (\mathcal{T}_1(z))^{(l)}$ *denotes the product of all entries* $(\mathcal{T}_1(z))^{(l)}$ *of a*

truncated vector $\mathcal{T}_1(z)$ and truncation operators $\mathcal{T}_1$ and $\mathcal{T}_v$ apply element-wise on input vectors such that $(\mathcal{T}_v(z))^{(l)} = \operatorname{sgn}(z^{(l)}) \min(v, |z^{(l)}|)$ with truncation level $v > 0$.

REMARK 4.1. *Linear transformers in (4.4) show that for each query x, there's a universal memory unit compressing information from the context on each attention head (p, q), in the form of $\mathcal{T}_v \left[\int \phi_{p,\tilde{m}(n)}(y) A_{p,q}^{(j)} y \, d\rho(y) \right]$. We note that the truncation operator $\mathcal{T}_v$ is actually not required to derive the final generalization bound, but is necessary to obtain an oracle inequality for each linear transformers in hypothesis space, since we consider Borel probability measure ρ defined on the entire $\mathbb{R}^d$ in this chapter. To construct the accessible context $\hat{\rho}$, we collect unbounded samples from $\mathbb{R}^d$ and create the memory unit $\int \phi_{p,\tilde{m}(n)}(y) A_{p,q}^{(j)} y \, d\hat{\rho}(y)$ that is unbounded and will effect sampling estimation if the outer shallow neural network does not have a special structure. Thus, for the algorithm stability, we introduce the truncation operator $\mathcal{T}_v$ to keep the memory unit stable, similar as the idea in Yang et al. [107].*

Linear transformers defined by (4.4) *do not have a normalization factor as equation* (4.3) *does in the form of $\phi(x_i)^T \int \phi(x) \, d\hat{\rho}(x)$. One benefit of normalization-free linear transformer is that it allows a more flexible design of the feature mapping ϕ. In practice, to ensure the normalization factor in* (4.3) *is nonzero, ϕ is often constrained to be nonnegative. It is not required any more with normalization-free linear transformers. Qin et al. [73] also shows that RMSNorm [113] can play a better role than normalization factor for stabilizing the optimization of linear transformers. For more details, we refer readers to Subsection 4.4.1.*

We apply two activation functions (ReLU and tanh) and a product gate in the construction of linear Transformer architecture, which is actually components of the modern activation function SwiGLU [86] for LLMs. The tanh activation is chosen in linear attention to approximate functions in a reproducing kernel Hilbert space (RKHS) induced by a smooth kernel. We include more discussion on this point in Subsection 4.4.2.

4.2.2 Two-Staged Sampling Framework for In-Context Learning

Our first novelty is to formulate the sequential modeling of transformers in in-context learning as the processing context-augmented inputs $(\hat{\rho}, x)$ as in (4.2)(4.3)(4.4) where samples like $(\hat{\rho}, x)$ are generated by a ***two-staged sampling framework*** from domain generalization [4, 5].

Let $\mathcal{X} = \mathbb{R}^d$ denote the input space and $\mathcal{Y} = \big\{y \in \mathbb{R}^d : \|y\|_2 \leq M\big\}$ a closed ball with radius $M > 0$ be the output space. Let $\mathcal{B}_2(\mathcal{X})$ and $\mathcal{B}_2(\mathcal{X} \times \mathcal{Y})$ denote the set of all Borel probability measures with finite second moments on $\mathcal{X}$ and $\mathcal{X} \times \mathcal{Y}$, respectively. We equip $\mathcal{B}_2(\mathcal{X})$ and $\mathcal{B}_2(\mathcal{X} \times \mathcal{Y})$ with Wasserstein-2 distances denoted by W_2 which metrize the weak topology on $\mathcal{B}_2(\mathcal{X})$ and $\mathcal{B}_2(\mathcal{X} \times \mathcal{Y})$. Then for context-augment inputs (ρ, x), we define a complete separable metric space $\Omega = \mathcal{B}_2(\mathcal{X}) \times \mathcal{X}$ equipped with the metric d_Ω as

$$d_\Omega\big((\rho, x), (\rho', x')\big) = \sqrt{W_2(\rho, \rho')^2 + \|x - x\|_2^2}.$$

DEFINITION 4.2. *A **two-staged sampling process** by meta probability measure $\mathcal{P}_\mathcal{G}$ is defined as follows: in the first stage sampling with $N \in \mathbb{N}$, $(\rho_{XY}^{(i)})_{i=1}^N$ are independently sampled from a meta Borel probability measure $\mathcal{P}_\mathcal{G}$ on $\mathcal{B}_2(\mathcal{X} \times \mathcal{Y})$; in the second stage sampling with $n_i \in \mathbb{N}$ for $1 \leq i \leq N$, a dataset $\mathbb{S} = \{(\hat{\rho}_X^{(i)}, X_{ij}, Y_{ij})_{j=1}^{n_i}\}_{i=1}^N$ is created by (X_{ij}, Y_{ij}) sampled independently from $\rho_{XY}^{(i)}$ and $\hat{\rho}_X^{(i)} = \delta([X_{i1}, ..., X_{in_i}])$.*

With the two-staged sampling process induced by $\mathcal{P}_\mathcal{G}$, for a prediction function $\Phi : \Omega \to \mathbb{R}^d$, we define the population risk for in-context learning as

$$\mathcal{E}(\Phi) = \mathbb{E}_{\rho_{XY} \sim \mathcal{P}_\mathcal{G}} \mathbb{E}_{(X,Y) \sim \rho_{XY}} \|\Phi(\rho_X, X) - Y\|_2^2 \tag{4.5}$$

and the empirical risk with a dataset $\mathbb{S} = \{(\hat{\rho}_X^{(i)}, X_{ij}, Y_{ij})_{j=1}^{n_i}\}_{i=1}^N$ as

$$\mathcal{E}_\mathbb{S}(\Phi) = \frac{1}{N} \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \|\Phi(\hat{\rho}_X^{(i)}, X_{ij}) - Y_{ij}\|_2^2.$$

Let $\mathcal{H}_{\mathrm{T}_n}$ be the hypothesis space of a collection of linear transformers in Definition 4.1 and $\mathrm{T}_{\mathbb{S},n} \in \mathcal{H}_{\mathrm{T}_n}$ be the function learned from the empirical risk minimization (ERM) algorithm

by

$$\mathrm{T}_{\mathbb{S},n} = \arg\min_{\mathrm{T}_n \in \mathcal{H}_{\mathrm{T}_n}} \mathcal{E}_{\mathbb{S}}(\mathrm{T}_n). \tag{4.6}$$

Because of the boundedness of the output space $\mathcal{Y}$, our estimator is given by $\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})$ where $\mathscr{T}_M$ is a truncation operator defined on $\mathbb{R}^d$ such that $\mathscr{T}_M(y) = y$ if $\|y\|_2 \leq M$ otherwise $M\frac{y}{\|y\|_2}$.

REMARK 4.2. *For an in-context learning dataset $\mathbb{S}$, each sample consists of a context-augmented input $(\hat{\rho}_X, X)$ and a label Y. Removing the accessible context $\hat{\rho}_X$ from an input reduces the learning problem to classical regression. On the other side, if we drop the query token X, the problem will degenerate to distribution regression as considered in our previous work [49].*

4.2.3 Latent Feature Space for Context-Augmented Inputs

Our second purpose is to investigate how context-augmented samples interact within the attention mechanism and to formulate this interaction into an inner product of a tensor-product Hilbert space (***the latent feature space***).

To mimic normalization-free attention for (ρ, x) inspired by (4.2) as $\int_{\mathcal{X}} \mathrm{sim}(x, x')x'd\rho(x')$, we first introduce an anisotropic Gaussian kernel $k_{\boldsymbol{\lambda}}$ on $\mathcal{X} \times \mathcal{X}$ as

$$k_{\boldsymbol{\lambda}}(x, x') = \exp(-(x - x')^T \Sigma_{\boldsymbol{\lambda}}(x - x'))$$

with shape parameter vector $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_d]^T \in \mathbb{R}^d$ and $\Sigma_{\boldsymbol{\lambda}} = \mathrm{diag}(\lambda_1^2, \cdots, \lambda_d^2)$. (For more details about the choice of the anisotropic Gaussian kernel, see Subsection 4.4.3.) We use kernel $k_{\boldsymbol{\lambda}}$ to measure the similarity between different tokens and then the attention mechanism induced by $k_{\boldsymbol{\lambda}}$ outputs $\int_{\mathcal{X}} k_{\boldsymbol{\lambda}}(x, x')x'd\rho(x')$ for the context-augmented input (ρ, x).

To better understand how the attention mechanism processes context information for context-augmented inputs, we introduce the following definition for context embedding.

DEFINITION 4.3. *Let $\mathcal{H}_{k_{\boldsymbol{\lambda}}}$ be the reproducing kernel Hilbert space (RKHS) induced by $k_{\boldsymbol{\lambda}}$. For each context $\rho \in \mathcal{B}_2(\mathcal{X})$, we define $K_{\boldsymbol{\lambda}}(\rho) : \mathcal{X} \to \mathbb{R}^d$ as*

$$K_{\boldsymbol{\lambda}}(\rho) = \int_{\mathcal{X}} k_{\boldsymbol{\lambda}}(\,\cdot\,, x)x \, d\rho(x) \in \mathcal{H}_{k_{\boldsymbol{\lambda}}} \otimes \mathbb{R}^d.$$

The above definition is an abstract perspective for *key-value cache* [72], where $K_{\boldsymbol{\lambda}}$ embeds context information ρ into a dictionary function mapping each query x_{query} to a context-aware representation

$$K_{\boldsymbol{\lambda}}(\rho)(x_{\text{query}}) = \int_{\mathcal{X}} k_{\boldsymbol{\lambda}}(x_{\text{query}}, x)x \, d\rho(x) \in \mathbb{R}^d.$$

Next, we define a feature mapping $\mathrm{I}_{\boldsymbol{\lambda}}$ and a latent feature space $\mathcal{H}_{\mathcal{F}}$ for context-augemented inputs $(\rho, x) \in \Omega$. In the latent feature space $\mathcal{H}_{\mathcal{F}}$, the similarity measure between context inputs not only depends on context information, but also has the properties of the attention mechanism.

DEFINITION 4.4. *Define a latent feature space $\mathcal{H}_{\mathcal{F}} = (\mathcal{H}_{k_{\boldsymbol{\lambda}}} \otimes \mathbb{R}^d) \otimes \mathcal{H}_{k_{\boldsymbol{\lambda}}}$. Define $\mathrm{I}_{\boldsymbol{\lambda}} : \mathcal{B}_2(\mathcal{X}) \times \mathcal{X} \to \mathcal{H}_{\mathcal{F}}$ by*

$$\mathrm{I}_{\boldsymbol{\lambda}}(\rho, x) = K_{\boldsymbol{\lambda}}(\rho) \otimes k_{\boldsymbol{\lambda}}(x, \cdot) \in \mathcal{H}_{\mathcal{F}}, \tag{4.7}$$

which introduces an inner product for similarity measure between context-augmented inputs $(\rho, x), (\rho', x') \in \Omega$ as

$$\langle \mathrm{I}_{\boldsymbol{\lambda}}(\rho, x), \mathrm{I}_{\boldsymbol{\lambda}}(\rho', x') \rangle_{\mathcal{H}_{\mathcal{F}}} = \langle K_{\boldsymbol{\lambda}}(\rho), K_{\boldsymbol{\lambda}}(\rho') \rangle_{\mathcal{H}_{k_{\boldsymbol{\lambda}}} \otimes \mathbb{R}^d} k_{\boldsymbol{\lambda}}(x, x'). \tag{4.8}$$

The similarity between context-augmented inputs (ρ, x) and (ρ, x') defined in (4.8), depends not only on the token values but also on the contexts ρ and ρ'. This is consistent with a basic observation in natural language processing: a word may have different meanings in different contexts.

REMARK 4.3. *The definition of $\mathrm{I}_{\boldsymbol{\lambda}}$ is inspired by the similarity measure in the domain generalization literature [4, 5] where the mapping $(\rho, x) \mapsto \Phi_{k_{\mathcal{B}}}(\rho) \otimes \Phi_{k_{\mathcal{X}}}(x)$ is considered with $\Phi_{k_{\mathcal{B}}}$,$\Phi_{k_{\mathcal{X}}}$ the canonical feature maps of kernels $k_{\mathcal{B}}$ [8], $k_{\mathcal{X}}$ respectively. In Appendix 4.5.3, we show that both $K_{\boldsymbol{\lambda}}$ and $\mathrm{I}_{\boldsymbol{\lambda}}$ are injective and continuous mappings. The injection of $K_{\boldsymbol{\lambda}}$*

is one of the key features of attention modules: compress context distributions into elements in $\mathcal{H}_{k_{\boldsymbol{\lambda}}} \otimes \mathbb{R}^d$ and fight against the negative effects of distribution shifts, while preserving the ability to distinguish different context distributions. It would be interesting to extend the results in Sriperumbudur et al. [96] to a quantitative analysis on dissimilar distribution with a small distance in $\mathcal{H}_{k_{\boldsymbol{\lambda}}} \otimes \mathbb{R}^d$ to investigate this tradeoff created by $K_{\boldsymbol{\lambda}}$.

4.3 Main Results on Linear Transformers for In-Context Learning

This section states the main results of approximation and generalization analysis for in-context learning with linear transformers. We first introduce some assumptions on $\mathcal{P}_{\mathcal{G}}$ for *the two-staged sampling process* and $k_{\boldsymbol{\lambda}}$ for *the latent feature space*.

The two-staged sampling process induced by $\mathcal{P}_{\mathcal{G}}$ in Definition 4.2 introduces a probability measure $\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}$ on $\mathcal{B}_2(\mathcal{X})$ for context information by

$$\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(E) = \mathcal{P}_{\mathcal{G}}\big(\{\mu \in \mathcal{B}_2(\mathcal{X} \times \mathcal{Y}) : \mu \circ \pi_{\mathcal{X}}^{-1} \in E\}\big) \tag{4.9}$$

for any Borel set $E \in \mathcal{B}_2(\mathcal{X})$ with the coordinate map $\pi_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$, and also a probability measure $\nu_{\mathcal{G}}$ for context-augmented inputs on the product σ-algebra of Ω by $\nu_{\mathcal{G}}(B \times A) = \int_B \rho(A) d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho)$ for any Borel set $B \in \mathcal{B}_2(\mathcal{X}), A \in \mathcal{X}$.

ASSUMPTION 4.5. *$\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}$ is supported on a subset $\mathcal{B}_{2,b}(\mathcal{X})$ of $\mathcal{B}_2(\mathcal{X})$ defined as*

$$\mathcal{B}_{2,b}(\mathcal{X}) = \big\{\rho \in \mathcal{B}_2(\mathcal{X}) : \mathbb{E}_{X\sim\rho}\|X\|_2^4 \le C_{\mathcal{B}}^2\big\}$$

with a constant $C_{\mathcal{B}} > 1$. We also denote $\Omega_{\mathcal{B}} = \big\{(\rho, x) \in \Omega : \rho \in \mathcal{B}_{2,b}(\mathcal{X})\big\}$.

ASSUMPTION 4.6. *There exists $\gamma > 1$ and $\kappa, C_{\mathcal{G}} > 0$ such that*

$$\int_{\mathcal{B}_2(\mathcal{X})} \|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}^2 \, d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho) \le C_{\mathcal{G}}$$

where $\omega_\kappa(\rho)$ is the likelihood ratio defined as $\frac{d\rho}{d\rho_\kappa}$ and ρ_κ is Gaussian probability measure with zero mean and covariance matrix $\kappa^2 \mathrm{I}_d$.

We provide two examples of meta probability measure $\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}$ satisfying Assumptions 4.5 and 4.6 in Appendix 4.5.3.

ASSUMPTION 4.7. *For the anisotropic Gaussian kernel $k_{\boldsymbol{\lambda}}$, we assume $\boldsymbol{\lambda} = (\lambda_l)_{l=1}^d$ is a sequence of shape parameters such that $\lambda_{(l)} \leq C_\theta l^{-\theta}$ for $1 \leq l \leq d$ with the order $\lambda_{(1)} \geq \lambda_{(2)} \geq \cdots \geq \lambda_{(d)} > 0$ where $C_\theta, \theta > 0$ are two constants independent of d.*

REMARK 4.4. *By the Portmanteau theorem, $\mathcal{B}_{2,b}(\mathcal{X})$ is closed in the W_2-topology, which allows probability measures supported on $\mathcal{B}_{2,b}(\mathcal{X})$ can be extended to probability measures on the entire $\mathcal{B}_2(\mathcal{X})$, matching the framework of the two-staged sampling process.*

The likelihood ratio is a popular tool to control distribution shifts in both theoretical and empirical studies [54, 85] and actually the quantity $\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}$ in Assumption 4.6 is closely related to the Rényi divergence [82, 100] of distribution ρ with respect to the reference distribution ρ_κ.

The fast decay of shape parameters in $\boldsymbol{\lambda}$ is often observed in practice. More details about Assumption 4.7 can be found in Subsection 4.4.3.

4.3.1 Approximation of Variation Normed Functions

Our third contribution is to propose an explicit hypothesis space under which linear Transformers achieve dimension-independent convergence rates for operator approximation without assuming access to the structure of a latent feature space $\mathcal{H}_{\mathcal{F}}$, enabling further study on generalization error analysis.

First we impose a regularity condition on the true predictor for (4.5) using the latent feature mapping $\mathrm{I}_{\boldsymbol{\lambda}}$ in (4.7) and a *variation normed space*. Let $\mathcal{H}_{\mathrm{para}} = \mathcal{H}_{\mathcal{F}} \oplus \mathbb{R}$ (the extra dimension is left for bias weights) and

$$\mathrm{B}(\mathcal{H}_{\mathrm{para}}; \mathbb{R}^d) = \left\{ W \in \mathcal{L}(\mathcal{H}_{\mathrm{para}}; \mathbb{R}^d) \,\middle|\, \|W\|_{\mathrm{op}} \leq 1 \right\}$$

the closed unit ball with the operator norm in the space $\mathcal{L}(\mathcal{H}_{\mathrm{para}}; \mathbb{R}^d)$ of all bounded linear operators from $\mathcal{H}_{\mathrm{para}}$ to $\mathbb{R}^d$. Note that the finite dimension of the output space makes $\mathcal{L}(\mathcal{H}_{\mathrm{para}}; \mathbb{R}^d)$

identical with the Hilbert space of Hilbert-Schmidt operators, and hence $\mathrm{B}(\mathcal{H}_{\mathrm{para}};\mathbb{R}^d)$ is weakly compact in $\mathcal{L}(\mathcal{H}_{\mathrm{para}};R^d)$. Let $\mathcal{M}(\mathrm{B}(\mathcal{H}_{\mathrm{para}};\mathbb{R}^d))$ denote the space of all signed Radon measures on $\mathrm{B}(\mathcal{H}_{\mathrm{para}};\mathbb{R}^d)$.

DEFINITION 4.8. *For $\mu \in \mathcal{M}(\mathrm{B}(\mathcal{H}_{\mathrm{para}};\mathbb{R}^d))$, let $F_\mu(h) = \int_{\mathrm{B}(\mathcal{H}_{\mathrm{para}};\mathbb{R}^d)} \sigma(Wh)\, d\mu(W)$ for $h \in \mathcal{H}_{\mathrm{para}}$. The variation normed space $\mathcal{F}_1$ of $\mathbb{R}^d$-valued functions on $\mathcal{H}_{\mathrm{para}}$ is defined as*

$$\mathcal{F}_1(\mathcal{H}_{\mathrm{para}};\mathbb{R}^d) = \Big\{ F : \mathcal{H}_{\mathrm{para}} \to \mathbb{R}^d \;\Big|\; \|F\|_{\mathcal{F}_1} := \inf_{\mu: F=F_\mu} \|\mu\|_{\mathcal{M}} < \infty \Big\}.$$

Then we give a dimension-independent approximation result with the hypothesis space

$$\mathcal{H}_{\mathrm{T}_n} = \Bigg\{ \mathrm{T}_n : \|\alpha\|_1 \le 2C_F, \sum_{p,q=1}^{m(n)} \big\|A_{p,q}^{(j)}\big\|_F^2 \le d, \|b_j\|_2 \le \sqrt{2dC_{\mathcal{B}}} \text{ for each } 1 \le j \le n,$$
$$\|b_0\|_2 \le C_F\sqrt{2dC_{\mathcal{B}}},\ \|\Theta_{\tanh}\|_\infty \le c_1(c_2\log(n))^{c_3(\log n)^2} \text{ and truncation level } v = C_{\mathcal{B}} \Bigg\}. \tag{4.10}$$

where $C_F > 0$ is a constant, c_1, c_2, c_3 are constants depending on θ, γ and $\Theta_{\tanh}$ denotes the parameters in two-layered tanh neural networks satisfying a sparse structure such that

$$W_{q,j} = \mathrm{diag}(W_{q,j}^{(1)}, ..., W_{q,j}^{(d)}) \text{ for } j = 0, 1, 2$$

where for $1 \le l \le d$, $W_{q,0}^{(l)} \in \mathbb{R}^{8\tilde{m}(n)\times 1}$, $W_{q,1}^{(l)} \in \mathbb{R}^{8\tilde{m}(n)\times 8\tilde{m}(n)}$ and $W_{q,2}^{(l)} \in \mathbb{R}^{1\times 8\tilde{m}(n)}$.

Let $\tilde{X} = (\rho_X, X)$ be the context-augmented input with the ground truth context, and $\tilde{X}_{ij} = (\hat{\rho}_X^{(i)}, X_{ij})$ with the accessible context. We rewrite $\mathcal{E}(\Phi) = \mathbb{E}_{(\tilde{X},Y)\sim\mathbb{P}_{\mathcal{G}}}\|\Phi(\tilde{X}) - Y\|_2^2$ where $\mathbb{P}_{\mathcal{G}}$ is a probability measure induced by the two-staged sampling process with $\mathcal{P}_{\mathcal{G}}$ [5]. Then the regression function for the population risk $\mathcal{E}$ is defined as

$$\Phi_{\mathcal{G}}(\tilde{X}) = \int_{\mathcal{Y}} y\, d\,\mathbb{P}_{\mathcal{G}}(\cdot|\tilde{X}). \tag{4.11}$$

THEOREM 4.9. *Let $\Phi_{\mathcal{G}} = F(\mathrm{I}_{\boldsymbol{\lambda}}(\cdot), 1)$ with $F \in \mathcal{F}_1(\mathcal{H}_{\mathrm{para}};\mathbb{R}^d)$ and $\|F\|_{\mathcal{F}_1} \le C_F$. For $0 < \xi < \theta$ and $n > C'_{\kappa,\theta,\gamma}$, there exists a $\mathrm{T} \in \mathcal{H}_{\mathrm{T}_{2n}}$ such that*

$$\|\mathrm{T} - \Phi_{\mathcal{G}}\|_{L^2(\nu_{\mathcal{G}})}^2 \le C_*^2 n^{-1} \text{ and } \|\mathrm{T}\|_{C(\Omega)} \le 2C_F\sqrt{d(1+C_{\mathcal{B}})}$$

with $m = \left\lceil n^{\frac{\gamma}{2(\gamma-1)\xi}} \right\rceil$, $\tilde{m} = \left\lceil \left(\frac{1}{2} + \frac{\gamma}{4(\gamma-1)\xi} \right) \log n \right\rceil$ *in (4.4), where* C_* *is a constant depending on* $\kappa, \xi, \gamma, \boldsymbol{\lambda}, C_{\mathcal{G}}, C_{\mathcal{B}}, C_F$ *and* $\mathrm{poly}(d)$.

REMARK 4.5. *Variation normed spaces for neural network approximation have been well studied in Barron [3], Bach [2], Korolev [40], Siegel and Xu [90], Yang and Zhou [108]. Roughly speaking, a variation normed space can be viewed as the collection of shallow neural networks with infinity width and thus can mimic the function class of target functions arising in practice. Definition 4.8 is a special case of Korolev [40] where the authors consider neural networks with values in a Banach space, extending the original result in Barron [3, Theorem 4].*

4.3.2 Generalization Analysis of In-Context Learning

Our final contribution is to address unbounded sampling (without domain restrictions) in the two-staged sampling process and establish an oracle inequality in Appendix 4.5.2. Combining Theorem 4.9 with this oracle inequality, we obtain a dimension-independent generalization rate.

THEOREM 4.10. *Let* $d \geq 2$ *and* $n \geq \max\{3, C'_{\kappa,\theta,\gamma}, \frac{C'^2_*}{64MC_{d,F,\mathcal{B}}}\}$ *and* $\Phi_{\mathcal{G}} = F(\mathrm{I}_{\boldsymbol{\lambda}}(\cdot), 1)$ *for some* $F \in \mathcal{F}_1(\mathcal{H}_{\mathrm{para}}; \mathbb{R}^d)$. *If the parameter* n *of the hypothesis space* $\mathcal{H}_{\mathrm{T}_n}$ *and the second-stage sample size* ϑ *are chosen as*

$$n = \lfloor \mathcal{K}_1 N^{\frac{1}{2+\gamma/[(\gamma-1)\xi]}} \rfloor \text{ and } \vartheta = N^3$$

and $m, \tilde{m}$ *chosen as in Theorem 4.9, then for the estimator generated by the ERM framework for the two-staged sampling process, we have*

$$\mathbb{E}\{\|\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}) - \Phi_{\mathcal{G}}\|^2_{L^2(\nu_{\mathcal{G}})}\} \leq \mathcal{K}_3 N^{-\frac{1}{2+\gamma/[(\gamma-1)\xi]}} (\log N)^3 \tag{4.12}$$

where $\mathcal{K}_3$ *is a constant depending on* $\kappa, \xi, \gamma, \boldsymbol{\lambda}, C_{\mathcal{G}}, C_{\mathcal{B}}, C_F$ *and* $\mathrm{poly}(d)$.

REMARK 4.6. *Although controlling distribution shift via likelihood ratio moments with divergence order* $\gamma > 1$ *provides* $L^2(\rho)$ *error bounds for every target distributions* ρ *satisfying Assumptions 4.5 and 4.6, the factor* $\frac{\gamma-1}{\gamma}$ *in RHS of* (4.12) *significantly slows the convergence*

as γ approaches 1. This makes a tradeoff between convergence rate and the capacity of admissible target distributions: smaller γ allows more distributions satisfying Assumption 4.6 as shown in Example 3 but greatly slows convergence. However, we observe the fast spectral decay phenomenon in LLMs (shown in Figure 4.1 of Subsection 4.4.3) that mitigates this slowdown for in-context learning. Indeed, under Assumption 4.7, the exponent becomes $\frac{(\gamma-1)\xi}{\gamma}$ that slows the rate of tending to infinity as $\gamma \to 1$ when ξ is large.

4.3.3 Proof Sketch

The proof of the approximation result in Theorem 4.9 relies on the following error decomposition

$$\begin{aligned} &\|\Phi_{\mathcal{G}} - \mathrm{T}_{2n,m,\tilde{m}}\|_{L^2(\nu_{\mathcal{G}})} \\ \leq &\|\Phi_{\mathcal{G}} - \mathcal{N}_{2n}\|_{L^2(\nu_{\mathcal{G}})} + \|\mathcal{N}_{2n} - \Psi_{2n,m}\|_{L^2(\nu_{\mathcal{G}})} + \|\Psi_{2n,m} - \mathrm{T}_{2n,m,\tilde{m}}\|_{L^2(\nu_{\mathcal{G}})} \end{aligned} \tag{4.13}$$

for $\mathrm{T}_{2n,m,\tilde{m}} \in \mathcal{H}_{\mathrm{T}_{2n}}$, where $\mathcal{N}_{2n}$ is a shallow neural network with operator-valued parameters in $\mathcal{L}(\mathcal{H}_{\mathrm{para}}; \mathbb{R}^d)$ and $\Psi_{2n,m}$ is a neural network with latent polynomial features depending on the parameterization of $k_{\boldsymbol{\lambda}}$, both explicitly constructed in Section 4.5.1.

The approximation error $\|\Phi_{\mathcal{G}} - \mathcal{N}_{2n}\|_{L^2(\nu_{\mathcal{G}})}$ is estimated by random approximation results from [40] in Appendix 4.5.1. The estimation of the last two terms on the RHS of (4.13) relies on Assumption 4.6 to bound L^2 errors under first-stage samples ρ_X with L^2 error under the reference probability distribution ρ_κ, as described in Appendix 4.5.1.1 and 4.5.1.2. More specifically, for the second term $\|\mathcal{N}_{2n} - \Psi_{2n,m}\|_{L^2(\nu_{\mathcal{G}})}$, $\mathcal{N}_{2n}$ can be regarded as a neural network with countably infinite feature-function parameters in an RKHS, and $\Psi_{2n,m}$ is a neural network constructed by the optimal choice of selecting only m feature-function parameters based on the parameterization of $k_{\boldsymbol{\lambda}}$. For the last term $\|\Psi_{2n,m} - \mathrm{T}_{2n,m,\tilde{m}}\|_{L^2(\nu_{\mathcal{G}})}$, the idea is to show the linear attentions in $\mathrm{T}_{2n,m,\tilde{m}}$ are fast universal approximators for any feature-function parameters in RKHS $\mathcal{H}_{k_{\boldsymbol{\lambda}}}$ without any prior knowledge on the parameterization of $k_{\boldsymbol{\lambda}}$. The approximation is considered with the supremum norm on the bounded domain $[-B, B]^d$, and the error outside this domain is controlled by a Gaussian tail decay (see Appendix 4.5.3).

For the generalization analysis, we define the first-stage sampling error $\mathcal{E}_N$ as

$$\mathcal{E}_N(\Phi) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big[\|\Phi(\tilde{X}) - Y\|_2^2|\rho_{XY}^{(i)}\Big]$$

and the second-stage sampling error $\mathcal{E}_{N,\mathcal{X}}$ with ground truth context as

$$\mathcal{E}_{N,\mathcal{X}}(\Phi) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}\|\Phi(\rho_X^{(i)}, X_{ij}) - Y_{ij}\|_2^2.$$

Then for any $\Phi \in \mathcal{H}_{\mathrm{T}_n}$, we have the following error decomposition

$$\begin{aligned}\mathcal{E}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}(\Phi_{\mathcal{G}}) = {}& \mathcal{E}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}_N(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) + \mathcal{E}_N(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}_{N,\mathcal{X}}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}))\\ &+ \mathcal{E}_{N,\mathcal{X}}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}_{\mathbb{S}}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) + \mathcal{E}_{\mathbb{S}}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}_{\mathbb{S}}(\Phi)\\ &+ \mathcal{E}_{\mathbb{S}}(\Phi) - \mathcal{E}_{N,\mathcal{X}}(\Phi) + \mathcal{E}_{N,\mathcal{X}}(\Phi) - \mathcal{E}_N(\Phi) + \mathcal{E}_N(\Phi) - \mathcal{E}(\Phi)\\ &+ \mathcal{E}(\Phi) - \mathcal{E}(\Phi_{\mathcal{G}})\end{aligned}$$

with $\mathcal{E}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}(\Phi_{\mathcal{G}}) = \|\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}) - \Phi_{\mathcal{G}}\|_{L^2(\nu_{\mathcal{G}})}^2$ and $\mathcal{E}_{\mathbb{S}}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}_{\mathbb{S}}(\Phi) \leq 0$.

Let

$$\begin{aligned}&\mathcal{E}_1(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) = \Big(\mathcal{E}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}(\Phi_{\mathcal{G}})\Big) - \Big(\mathcal{E}_N(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}_N(\Phi_{\mathcal{G}})\Big),\\ &\mathcal{E}_1'(\Phi) = \Big(\mathcal{E}_N(\Phi) - \mathcal{E}_N(\Phi_{\mathcal{G}})\Big) - \Big(\mathcal{E}(\Phi) - \mathcal{E}(\Phi_{\mathcal{G}})\Big),\\ &\mathcal{E}_2(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) = \mathcal{E}_N(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}_{N,\mathcal{X}}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})),\\ &\mathcal{E}_2'(\Phi) = \mathcal{E}_{N,\mathcal{X}}(\Phi) - \mathcal{E}_N(\Phi),\\ &\mathcal{E}_3(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) = \mathcal{E}_{N,\mathcal{X}}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}_{\mathbb{S}}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})),\\ &\mathcal{E}_3'(\Phi) = \mathcal{E}_{\mathbb{S}}(\Phi) - \mathcal{E}_{N,\mathcal{X}}(\Phi) \text{ and } \mathcal{E}_4(\Phi) = \mathcal{E}(\Phi) - \mathcal{E}(\Phi_{\mathcal{G}}).\end{aligned}$$

Then we have

$$\begin{aligned}\mathcal{E}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}(\Phi_{\mathcal{G}}) \leq {}& \mathcal{E}_1(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) + \mathcal{E}_1'(\Phi) + \mathcal{E}_2(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) + \mathcal{E}_2'(\Phi)\\ &+ \mathcal{E}_3(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) + \mathcal{E}_3'(\Phi) + \mathcal{E}_4(\Phi).\end{aligned} \tag{4.14}$$

In the first-stage sampling, we control the effect of unbounded sampling on $\mathcal{E}_1(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}))$ by using the covering number under a pseudo-metric defined by the supremum of distribution expectations over $\mathcal{B}_{2,b}(\mathcal{X})$ (Appendix 4.5.2.3). In the second-stage sampling, the situation becomes more complicated with the structure of linear transformers, so we decompose the second-stage sampling error into two parts: $\mathcal{E}_2(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})), \mathcal{E}_2'(\Phi)$ with *ground truth contexts* (Appendix 4.5.2.4) and $\mathcal{E}_3(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})), \mathcal{E}_3'(\Phi)$ only with *accessible contexts* (Appendix 4.5.2.5).

For the case with *ground truth contexts*, the similar idea with the first-stage sampling applies: after introducing Rademacher complexity for sampling estimation, we bound the empirical process by the Dudley integral. We then use concavity to move the expectations over second stage samples into the covering number expression, thereby eliminating the effect of unbounded samples. For the last sampling estimation with *accessible contexts*, the problem becomes even more challenging in the presence of memory units in linear Transformers, since both the input samples and the memory-unit outputs are unbounded. Here, we apply an extension [59] of Azuma-McDiarmind's inequality under subgaussian conditions to obtain a distribution-dependent probability concentration inequality where an observation on the relation between $\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}$ and subgaussian norm plays an important role. To obtain the final generalization error, we apply the expectation identity for non-negative random variables to bring $\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}$ out of the denominator and the exponential, so that we can take an expectation of $\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}$ with respect to $\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}$ by Assumption 4.6.

4.4 Related Works and Discussions

4.4.1 Normalization Factor and RMSNorm

Early linear attention models [38, 7] have some softmax-inspired design features (4.1), such as inserting a normalization denominator as in (4.3). However, Qin et al. [73] demonstrates both theoretically and empirically that linear Transformers with (4.3) make the gradients for attention matrices unbounded and lead to a less stable optimization and worse convergence. To alleviate this negative effect, Qin et al. [73] removes the normalization factor in (4.3) and

applies RMSNorm [113] to the linear attention output:

$$O_{\text{norm}} = \text{RMSNorm}(Q(K^T V)) \tag{4.15}$$

where

$$Q = [\phi(Q_1), \cdots, \phi(Q_n)]^T \in \mathbb{R}^{n \times d'},$$
$$K = [\phi(K_1), \cdots, \phi(K_n)]^T \in \mathbb{R}^{n \times d'},$$
$$V = [V_1, \cdots, V_n]^T \in \mathbb{R}^{n \times d}$$

and for input $A = (a_{ij})_{i,j} \in \mathbb{R}^{n \times d}$, $A' = \text{RMSNorm}(A) \in \mathbb{R}^{n \times d}$ is defined as

$$A' = (a'_{ij})_{i,j} \text{ such that } a'_{ij} = \frac{a_{ij}}{\sqrt{\frac{1}{d}\sum_{j=1}^{d} a_{ij}^2 + \epsilon}} \cdot \beta_j$$

with learnable scaling factor $\boldsymbol{\beta} = (\beta_1, ..., \beta_d)^T \in \mathbb{R}^d$. RMSNorm reduces the amount of computation and increases efficiency over LayerNorm, and it is widely used in the open-weight LLMs like Qwen3 [106].

4.4.2 Activation Functions in LLM

The design of activation functions has evolved with the development of LLMs. While the original transformer used ReLU activation by default, early LLMs such as BERT and GPT-2/3 employed the Gaussian Error Linear Unit (GeLU) activation [25], and it then became the standard choice. For $x \in \mathbb{R}^d$, GeLU is defined as

$$\text{GeLU}(x) = x \odot \text{F}(x)$$

where $\odot$ denotes the Hadamard product and F denotes the cumulative distribution function for the standard gaussian and applies element-wise on d-dimensional vectors. In practice, GeLU activation function is often implemented via a tanh approximation [1] as

$$\text{GeLU}(x) \approx 0.5x \odot \left[1 + \sigma_{\tanh}\left(\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^{\odot 3}\right)\right)\right].$$

[1] Refer to GeLU implementation in Pytorch 2.8 documentation: https://docs.pytorch.org/docs/stable/generated/torch.nn.GELU.html

where $x^{\odot 3}$ denotes the Hadamard power of order 3. Similar activation functions such as Swish activation [80] were introduced later. It's worth noting that Swish activation is defined as

$$\mathrm{Swish}_\beta(x) = x \odot \mathrm{Sigmoid}(\beta\, x)$$

where $\beta \in \mathbb{R}$ is a constant or learnable parameter and $\mathrm{Sigmoid}$ applies element-wise on x with

$$\mathrm{Sigmoid}(a) = \frac{1}{1+\exp(-a)} = \frac{1}{2}(1 + \sigma_{\tanh}(a/2)) \text{ for } a \in \mathbb{R}. \tag{4.16}$$

SiLU is a special case of Swish with $\beta = 1$.

Finally, combining all tricks above, Shazeer [86] proposed SwiGLU which is widely used in modern LLMs and defined as

$$\mathrm{SwiGLU}_\beta(x) = W_o((W_1 x + b_1) \odot \mathrm{Swish}_\beta(W_2 x + b_2)).$$

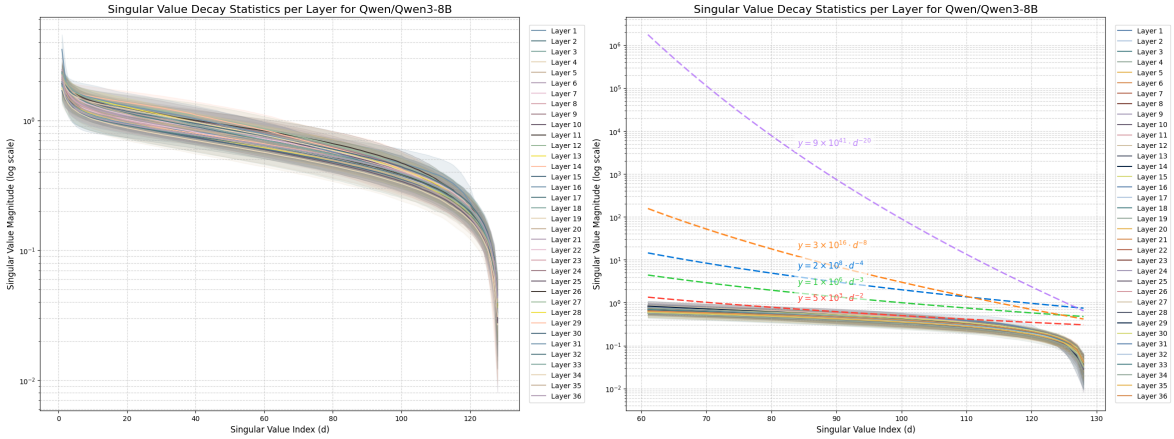
In our definition of linear Transformers (4.1), we use three nonlinear components: Tanh activation, product gate, and ReLU activation. It is easy to observe the connection between tanh activation and SwiGLU by Equation (4.16). For ReLU activation, when β is large enough, $\mathrm{Sigmoid}(\beta\, a)$ approximates the indicator function (except at zero) and Swish_β behaves like a ReLU activation function. For product gate, when $\beta = 0$, $\mathrm{Swish}_\beta(x) = x/2$ and then we can obtain $x \odot x$ by SwiGLU activation with a suitable choice of parameters [80].

4.4.3 Linear Conversion of Softmax LLMs

Distilling knowledge from pretrained softmax LLMs into subquadratic models [114] has recently attracted interest in the research community. Here, we present a perspective on linearizing pretrained softmax LLMs, derived from our theoretical analysis framework.

Recall the softmax attention module (4.1). Following Tsai et al. [99], Liu and Zhou [49], we take a kernelized viewpoint of attention modules by letting similarity function $\mathrm{sim}(x_i, x_j) = \exp(\langle W_q x_i, W_k x_j \rangle)$. Then (4.1) can be written as

$$\mathrm{SoftmaxAttn}(x_i|Q) = \frac{1}{\mathrm{Z}(x_i)} \sum_{j=1}^{n} \mathrm{sim}(x_i, x_j)(W_v x_j) \tag{4.17}$$



(A) Fast Decay of Singular Values (B) Polynomial Fit of Decay Trend

FIGURE 4.1. Fast Eigendecay of Qwen3-8B (Ghost in the Kernel)

where $\mathrm{Z}(x_i) = \sum_{j=1}^{n} \mathrm{sim}(x_i, x_j)$ is a normalization factor. For each pair of query and key weight matrices (W_q, W_k), we perform singular value decomposition $W_q^T W_k = W_1^T \Sigma_{\boldsymbol{\lambda}} W_2$ where W_1, W_2 are $d \times d$ orthogonal matrices and $\Sigma_{\boldsymbol{\lambda}}$ is a positive diagonal matrix with rank d_{hidden}. It's obtained that

$$\begin{aligned} \mathrm{sim}(x_i, x_j) &= \exp(x_i^T W_1^T \Sigma_{\boldsymbol{\lambda}} W_2 x_j) = \exp(\tilde{x}_i^T \Sigma_{\boldsymbol{\lambda}} \hat{x}_j) \\ &= \exp\left(\frac{1}{2}\tilde{x}_i^T \Sigma_{\boldsymbol{\lambda}} \tilde{x}_i\right) \exp\left(\frac{1}{2}\hat{x}_j^T \Sigma_{\boldsymbol{\lambda}} \hat{x}_j\right) \exp\left(-\frac{1}{2}(\tilde{x}_i - \hat{x}_j)^T \Sigma_{\boldsymbol{\lambda}} (\tilde{x}_i - \hat{x}_j)\right) \end{aligned} \tag{4.18}$$

with query $\tilde{x}_i = W_1 x_i$ and key $\hat{x}_j = W_2 x_j$. From the above expression, we observe that the roles of the key and query matrices can be decomposed into kernel asymmetry between queries and keys, represented by orthogonal matrices W_1, W_2, and geometric information, represented by a diagonal matrix $\Sigma_{\boldsymbol{\lambda}}$. In Figure 4.1, we show a numerical demonstration of singular values in diagonal matrices $\Sigma_{\boldsymbol{\lambda}}$ in the attention modules of large language model Qwen3 [106] , which exhibits a rapid decay of singular values across layers, nearly exponential for large d. The rapid decay of shape parameters enables us to design linear transformers that efficiently mimic softmax attention and alleviate the negative effects of distribution shifts in our analysis.

For the construction of a linear attention, the key is to decouple the interaction in $\mathrm{sim}(x_i, x_j)$ between queries and keys as shown in (4.3). For the similarity function in Equation 4.18, the

target is to find a decoupling of query-key interaction between $\tilde{x}$ and $\hat{x}$ for the anisotropic gaussian kernel $k_{\boldsymbol{\lambda}}(\tilde{x},\hat{x}) = \exp(-\frac{1}{2}(\tilde{x}-\hat{x})^T\Sigma_{\boldsymbol{\lambda}}(\tilde{x}-\hat{x}))$ to mimic context modeling in softmax attention. In Appendix 4.5.3, we know that there's an optimal linear approximation scheme for context embedding, denoted as $(\psi_q^{\boldsymbol{\lambda}})_{q=1}^m$ with explicit expressions, which only depends on the geometric information of $\Sigma_{\boldsymbol{\lambda}}$ and underlying context distributions on $\mathbb{R}^d$. With a suitable choice of m, we have

$$\mathrm{sim}(x_i,x_j) \approx \sum_{q=1}^{m} \underbrace{\exp\left(\frac{1}{2}\tilde{x}_i^T\Sigma_{\boldsymbol{\lambda}}\tilde{x}_i\right)\psi_q^{\boldsymbol{\lambda}}(\tilde{x}_i)}_{query} \cdot \underbrace{\exp\left(\frac{1}{2}\hat{x}_i^T\Sigma_{\boldsymbol{\lambda}}\hat{x}_i\right)\psi_q^{\boldsymbol{\lambda}}(\hat{x}_i)}_{key}. \tag{4.19}$$

Compared with previous methods [38, 7] for linear attention, the choice of $(\psi_q^{\boldsymbol{\lambda}})$ captures the latent data structures learned by the pretrained softmax LLM and makes use of its parameters. However, in practice, it may be difficult to directly apply the approximation (4.19), because the semantic distributions of tokens are highly anisotropic and hard to estimate, while we use an isotropic Gaussian to roughly control the class of distributions in the two-staged sampling process. But it still provides us an idea to design new activation functions and regularity for feature mappings for linear conversion of pretrained softmax LLMs, just as in Zhang et al. [114].

4.5 Proof of Main Results on Linear Transformers for In-Context Learning

4.5.1 Theorem 4.9: Approximation Scheme by Linear Transformers

By Proposition 3, we know that $\mathrm{I}_{\boldsymbol{\lambda}} : \Omega \to \mathcal{H}_{\mathcal{F}}$ is continuous, which introduces a pushforward probability measure $\mu_{\mathcal{G}}$ on $\mathcal{H}_{\mathcal{F}}$ by $\mu_{\mathcal{G}} = \nu_{\mathcal{G}} \circ \mathrm{I}_{\boldsymbol{\lambda}}^{-1}$. Then we define a probability measure $\tilde{\mu}_{\mathcal{G}} = \mu_{\mathcal{G}} \times \delta_1$ on $\mathcal{H}_{\mathrm{para}}$. For $h \in \mathcal{H}_{\mathrm{para}}$, let $h_{\mathcal{F}} = \mathrm{Proj}_{\mathcal{H}_{\mathcal{F}}}(h)$ and we have

$$\begin{aligned}\int_{\mathcal{H}_{\mathrm{para}}} \|h\|^2_{\mathcal{H}_{\mathrm{para}}}\, d\tilde{\mu}_{\mathcal{G}}(h) &= 1 + \int_{\mathcal{H}_{\mathcal{F}}} \|h_{\mathcal{F}}\|^2_{\mathcal{H}_{\mathcal{F}}}\, d\mu_{\mathcal{G}}(h_{\mathcal{F}}) = 1 + \int_{\Omega_{\mathcal{B}}} \|K_{\boldsymbol{\lambda}}(\rho) \otimes k_{\boldsymbol{\lambda}}(x,\cdot)\|^2_{\mathcal{H}_{\mathcal{F}}}\, d\nu_{\mathcal{G}}(\rho, x) \\ &= 1 + \int_{\Omega_{\mathcal{B}}} \|K_{\boldsymbol{\lambda}}(\rho)\|^2_{\mathcal{H}_{k_{\boldsymbol{\lambda}}} \otimes \mathbb{R}^d} \|k_{\boldsymbol{\lambda}}(x,\cdot)\|^2_{\mathcal{H}_{k_{\boldsymbol{\lambda}}}}\, d\nu_{\mathcal{G}}(\rho, x) \leq 1 + C_{\mathcal{B}}.\end{aligned}$$

We also note that for $W \in \mathcal{L}(\mathcal{H}_{\mathrm{para}}; \mathbb{R}^d)$, $\|W\|_{\mathrm{op}} \leq \|W\|_{\mathrm{HS}} \leq \sqrt{d}\|W\|_{\mathrm{op}}$ where $\|\cdot\|_{\mathrm{HS}}$ denotes Hilbert-Schmidt norm of $\mathcal{HS}(\mathcal{H}_{\mathrm{para}}; \mathbb{R}^d)$. let $\{e_l\}_{l=1}^d$ denote an orthonormal basis of $\mathbb{R}^d$ and for any $h \in \mathcal{H}_{\mathrm{para}}$, we have $\langle Wh, e_l\rangle_{\mathbb{R}^d} = \langle h, W^* e_l\rangle_{\mathcal{H}_{\mathrm{para}}}$. Denote $w_l = W^* e_l \in \mathcal{H}_{\mathrm{para}}$ and then

$$Wh = \sum_{l=1}^d \langle h, w_l\rangle_{\mathcal{H}_{\mathrm{para}}} e_l = \sum_{l=1}^d (e_l \otimes w_l) h. \tag{4.20}$$

It implies that the elements in $\mathrm{B}(\mathcal{H}_{\mathrm{para}}; \mathbb{R}^d)$ share the form $W = \sum_{l=1}^d e_l \otimes w_l$ where $w_l \in \mathcal{H}_{\mathrm{para}}$ such that

$$\|W\|_{\mathrm{op}} = \sup_{\|h\|_{\mathcal{H}_{\mathrm{para}}}=1} \left(\sum_{l=1}^d \langle w_l, h\rangle^2_{\mathcal{H}_{\mathrm{para}}}\right)^{\frac{1}{2}} \leq 1.$$

For the probability measure $\tilde{\mu}_{\mathcal{G}}$ and $F \in \mathcal{F}_1(\mathcal{H}_{\mathrm{para}}; \mathbb{R}^d)$, we have the following lemma from Theorem 3.24 in [40] for the approximation by shallow nets with operator-valued parameters.

LEMMA 10. *For any probability measure μ on $\mathcal{H}_{\mathrm{para}}$ with a finite second moment and $F \in \mathcal{F}_1(\mathcal{H}_{\mathrm{para}}; \mathbb{R}^d)$, there exists a shallow neural network $\mathcal{N}_n$ such that*

$$\|F - \mathcal{N}_n\|^2_{L^2_{\mu}(\mathcal{H}_{\mathrm{para}}; \mathbb{R}^d)} \leq \frac{\|F\|^2_{\mathcal{F}_1}\, \mathbb{E}_{h\sim\mu}\|h\|^2_{\mathcal{H}_{\mathrm{para}}}}{n},$$

and the shallow neural network $\mathcal{N}_n$ has the form

$$\mathcal{N}_n(h) = \sum_{j=1}^{n} \alpha_j \sigma(W_j h) \text{ with } \alpha_j \in \mathbb{R}, W_j \in \mathrm{B}(\mathcal{H}_{\mathrm{para}}; \mathbb{R}^d) \text{ for } 1 \le j \le n \text{ and } \|\alpha\|_1 \le \|F\|_{\mathcal{F}_1}.$$

We apply the above lemma with $\mu = \tilde{\mu}_{\mathcal{G}}$. Then for $h = (h_{\mathcal{F}}, 1)$, $\mathcal{N}_n(h)$ can be further written by (4.20) into

$$\begin{aligned}
\mathcal{N}_n(h) &= \sum_{j=1}^{n} \alpha_j \sigma(W_j h) = \sum_{j=1}^{n} \alpha_j \sigma\left(\sum_{l=1}^{d} (e_l \otimes w_{j,l}) h \right) \\
&= \sum_{j=1}^{n} \alpha_j \sigma\left(\sum_{l=1}^{d} (\langle v_{j,l}, h_{\mathcal{F}} \rangle_{\mathcal{H}_{\mathcal{F}}} + b_{j,l}) e_l \right) \\
&= \sum_{j=1}^{n} \alpha_j \sigma\left(\sum_{l=1}^{d} \langle v_{j,l}, h_{\mathcal{F}} \rangle_{\mathcal{H}_{\mathcal{F}}} e_l + b_j \right)
\end{aligned}$$

where $b_j = \sum_{l=1}^{d} b_{j,l} e_l \in \mathbb{R}^d$, $w_{j,l} = (v_{j,l}, b_{j,l})$ with $v_{j,l} \in \mathcal{H}_{\mathcal{F}}$ and $b_{j,l} \in \mathbb{R}$ such that

$$\|W_j\|_{\mathrm{op}} = \sup_{\|h\|_{\mathcal{H}_{\mathrm{para}}}=1} \left(\sum_{l=1}^{d} \langle w_{j,l}, h \rangle_{\mathcal{H}_{\mathrm{para}}}^2 \right)^{\frac{1}{2}} \le 1.$$

We also know that for $h = (h_{\mathcal{F}}, 1)$ in the support of $\tilde{\mu}_{\mathcal{G}}$, $h_{\mathcal{F}} = K_{\boldsymbol{\lambda}}(\rho) \otimes k_{\boldsymbol{\lambda}}(x, \cdot)$ for some $(\rho, x) \in \Omega_{\mathcal{B}}$. It follows that $\|h\|_{\mathcal{H}_{\mathrm{para}}} \le \sqrt{1 + C_{\mathcal{B}}}$ and $\|W_j h\|_2 \le \sqrt{1 + C_{\mathcal{B}}}$ for $h \in \mathrm{supp}(\tilde{\mu}_{\mathcal{G}})$. Then we construct a double-width shallow neural network $\mathcal{N}_{2n}$ as

$$\begin{aligned}
\mathcal{N}_{2n}(h) &= \sum_{j=1}^{n} \alpha_j \Big(\sigma(W_j h + \sqrt{1 + C_{\mathcal{B}}} \mathbf{1}_d) - \sigma(W_j h - \sqrt{1 + C_{\mathcal{B}}} \mathbf{1}_d) - \sqrt{1 + C_{\mathcal{B}}} \mathbf{1}_d \Big) \\
&= \sum_{j=1}^{2n} \alpha'_j \sigma\left(W'_j h + (-1)^{\lfloor \frac{j-1}{n} \rfloor} \sqrt{1 + C_{\mathcal{B}}} \mathbf{1}_d \right) - \sum_{j=1}^{n} \alpha_j \sqrt{1 + C_{\mathcal{B}}} \mathbf{1}_d \\
&= \sum_{j=1}^{2n} \alpha'_j \sigma\left(\sum_{l=1}^{d} \langle v'_{j,l}, h_{\mathcal{F}} \rangle_{\mathcal{H}_{\mathcal{F}}} e_l + b'_j \right) + b'_0
\end{aligned}$$

to realize the same approximant in Lemma 10 by adding additional bias vectors and letting $\alpha'_j = -\alpha'_{j+n} = \alpha_j$, $v'_{j,l} = v'_{j+n,l} = v_{j,l}$, $b'_j = b_j + \sqrt{1 + C_{\mathcal{B}}} \mathbf{1}_d$, $b'_{j+n} = b_j - \sqrt{1 + C_{\mathcal{B}}} \mathbf{1}_d$ for $1 \le j \le n$, and $b_0 = -\sum_{j=1}^{n} \alpha_j \sqrt{1 + C_{\mathcal{B}}} \mathbf{1}_d$. In the following content, we still use $(\alpha_j, v_{j,l}, b_j, b_0)$ as notations for parameters in $\mathcal{N}_{2n}$ with no confusion.

4.5.1.1 Neural Network with Latent Polynomial Features

Recall that for (ρ, x) sampled from the probability measure $\nu_{\mathcal{G}}$ on $\Omega_{\mathcal{B}}$, we take the feature $\mathrm{I}_{\boldsymbol{\lambda}}(\rho, x) = K_{\boldsymbol{\lambda}}(\rho) \otimes k_{\boldsymbol{\lambda}}(x, \cdot) \in \mathcal{H}_{\mathcal{F}}$ as the model input. Let $(\psi_q^{\boldsymbol{\lambda}})_{q\in\mathbb{N}}$ be the orthonormal basis of $\mathcal{H}_{k_{\boldsymbol{\lambda}}}$ and $(r_q^{\boldsymbol{\lambda}})_{q\in\mathbb{N}}$ the eigenvalue sequence as defined in Appendix 4.5.3. Then for each pair (j, l), $v_{j,l}$ can be written as

$$v_{j,l} = \sum_{p,q\in\mathbb{N}} \sum_{1\le s\le d} a_{p,q,s}^{(j,l)} (\psi_p^{\boldsymbol{\lambda}} \otimes e_s) \otimes \psi_q^{\boldsymbol{\lambda}} \text{ with } \sum_{p,q\in\mathbb{N}} \sum_{1\le s\le d} (a_{p,q,s}^{(j,l)})^2 \le 1.$$

It follows that

$$\begin{aligned}\langle v_{j,l}, \mathrm{I}_{\boldsymbol{\lambda}}(\rho, x)\rangle_{\mathcal{H}_{\mathcal{F}}} &= \sum_{p,q\in\mathbb{N}} \sum_{1\le s\le d} a_{p,q,s}^{(j,l)} \langle \psi_p^{\boldsymbol{\lambda}} \otimes e_s \otimes \psi_q^{\boldsymbol{\lambda}}, K_{\boldsymbol{\lambda}}(\rho) \otimes k_{\boldsymbol{\lambda}}(x, \cdot)\rangle_{\mathcal{H}_{\mathcal{F}}} \\ &= \sum_{p,q\in\mathbb{N}} \sum_{1\le s\le d} a_{p,q,s}^{(j,l)} \psi_q^{\boldsymbol{\lambda}}(x) \int \psi_p^{\boldsymbol{\lambda}}(y) e_s^T y \, d\rho(y).\end{aligned}$$

The basic idea for constructing a neural network with latent polynomial features is to estimate the truncation error for the query index q and the context memory index p.

First we consider to estimate the truncation error on index q. We make one step further by writing

$$\langle v_{j,l}, \mathrm{I}_{\boldsymbol{\lambda}}(\rho, x)\rangle_{\mathcal{H}_{\mathcal{F}}} = \sum_{q\in\mathbb{N}} \left(\sum_{p\in\mathbb{N}} \sum_{1\le s\le d} a_{p,q,s}^{(j,l)} \int \psi_p^{\boldsymbol{\lambda}}(y) e_s^T y \, d\rho(y) \right) \psi_q^{\boldsymbol{\lambda}}(x) =: \sum_{q\in\mathbb{N}} b_q^{(j,l)}(\rho) \psi_q^{\boldsymbol{\lambda}}(x).$$

Define

$$A^{(j,l)} : l^2(\mathbb{N} \times [d]) \to l^2(\mathbb{N}) \text{ such that } (A^{(j,l)} z)_q = \sum_{p\in\mathbb{N}} \sum_{s\in[d]} a_{p,q,s}^{(j,l)} z_{p,s}$$

where $[d] := \{1, \ldots, d\}$ and $q \in \mathbb{N}$. It is easy to see that $A^{(j,l)}$ is a Hilbert-Schmidt operator with $\|A^{(j,l)}\|_{\mathrm{HS}}^2 = \sum_{p,q\in\mathbb{N},s\in[d]} (a_{p,q,s}^{(j,l)})^2 \le 1$. Also note that by dominated convergence theorem,

$$\sum_{p\in\mathbb{N}} \sum_{s\in[d]} \left(\int \psi_p^{\boldsymbol{\lambda}}(y) e_s^T y \, d\rho(y) \right)^2 \le \int \sum_{p\in\mathbb{N}} (\psi_p^{\boldsymbol{\lambda}}(y))^2 \|y\|_2^2 \, d\rho(y) = \int k_{\boldsymbol{\lambda}}(y, y) \|y\|_2^2 \, d\rho(y) \le C_{\mathcal{B}},$$

which shows that $\int \psi^{\boldsymbol{\lambda}}(y) y \, d\rho(y) := \left(\int \psi_p^{\boldsymbol{\lambda}}(y) e_s^T y \, d\rho(y)\right)_{p\in\mathbb{N}, s\in[d]} \in l^2(\mathbb{N}\times[d])$. It follows that

$$
\begin{aligned}
C_{\mathcal{B}}^{\frac{1}{2}} &\geq \|A^{(j,l)}\|_{\mathrm{HS}} \left\| \int \psi^{\boldsymbol{\lambda}}(y) y \, d\rho(y) \right\|_{l^2(\mathbb{N}\times[d])} \\
&\geq \left\| A^{(j,l)} \left(\int \psi^{\boldsymbol{\lambda}}(y) y \, d\rho(y) \right) \right\|_{l^2(\mathbb{N})} = \left(\sum_{q\in\mathbb{N}} (b_q^{(j,l)}(\rho))^2 \right)^{\frac{1}{2}}
\end{aligned}
$$

and $\sum_{q\in\mathbb{N}} b_q^{(j,l)}(\rho)\psi_q^{\boldsymbol{\lambda}} \in \mathcal{H}_{k_{\boldsymbol{\lambda}}}$ for each (j,l). Let

$$
\Psi_{2n,m_1}(\rho, x) := \sum_{j=1}^{2n} \alpha_j \sigma \left(\sum_{l=1}^{d} \sum_{q=1}^{m_1} b_q^{(j,l)}(\rho) \psi_q^{\boldsymbol{\lambda}}(x) e_l + b_j \right) + b_0.
$$

Then we have

$$
\begin{aligned}
&\|\mathcal{N}_{2n}(\mathrm{I}_{\boldsymbol{\lambda}}(\cdot), 1) - \Psi_{2n,m_1}\|_{L^2(\nu_{\mathcal{G}})}^2 \\
&= \int_{\Omega_{\mathcal{B}}} \|\mathcal{N}_{2n}(\mathrm{I}_{\boldsymbol{\lambda}}(\rho, x), 1) - \Psi_{2n,m_1}(\rho, x)\|_2^2 \, d\nu_{\mathcal{G}}(\rho, x) \\
&\leq \int_{\Omega_{\mathcal{B}}} \left\| \sum_{j=1}^{2n} |\alpha_j| \sum_{l=1}^{d} \left| \sum_{q\geq m_1+1} b_q^{(j,l)}(\rho) \psi_q^{\boldsymbol{\lambda}}(x) \right| e_l \right\|_2^2 d\nu_{\mathcal{G}}(\rho, x) \\
&= \int_{\Omega_{\mathcal{B}}} \sum_{l=1}^{d} \left(\sum_{j=1}^{2n} |\alpha_j| \left| \sum_{q\geq m_1+1} b_q^{(j,l)}(\rho) \psi_q^{\boldsymbol{\lambda}}(x) \right| \right)^2 d\nu_{\mathcal{G}}(\rho, x) \\
&\leq \int_{\Omega_{\mathcal{B}}} \sum_{l=1}^{d} \|\alpha\|_1 \left[\sum_{j=1}^{2n} |\alpha_j| \left(\sum_{q\geq m_1+1} b_q^{(j,l)}(\rho) \psi_q^{\boldsymbol{\lambda}}(x) \right)^2 \right] d\nu_{\mathcal{G}}(\rho, x) \\
&= \int_{\mathcal{B}_{2,b}(\mathcal{X})} \|\alpha\|_1 \sum_{l=1}^{d} \left[\sum_{j=1}^{2n} |\alpha_j| \int_{\mathcal{X}} \left(\sum_{q\geq m_1+1} b_q^{(j,l)}(\rho) \psi_q^{\boldsymbol{\lambda}}(x) \right)^2 d\rho(x) \right] d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho) \\
&\leq \int_{\mathcal{B}_{2,b}(\mathcal{X})} \|\alpha\|_1^2 \sum_{l=1}^{d} \left[\max_{1\leq j\leq 2n} \int_{\mathcal{X}} \left(\sum_{q\geq m_1+1} b_q^{(j,l)}(\rho) \psi_q^{\boldsymbol{\lambda}}(x) \right)^2 d\rho(x) \right] d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho) \\
&=: \int_{\mathcal{B}_{2,b}(\mathcal{X})} \|\alpha\|_1^2 \sum_{l=1}^{d} \left(\max_{1\leq j\leq 2n} \mathcal{E}_{j,l}(\rho) \right) d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho). \qquad (4.21)
\end{aligned}
$$

If $1<\gamma<\infty$, for each pair $(j,l)\in\{1,\ldots,n\}\times\{1,\ldots,d\}$, we obtain that

$$\begin{aligned}
\mathcal{E}_{j,l}(\rho)&=\int_{\mathcal{X}}\left(\sum_{q\geq m_1+1}b_q^{(j,l)}(\rho)\psi_q^{\boldsymbol{\lambda}}(x)\right)^2 d\rho(x)\\
&=\int_{\mathcal{X}}\left(\sum_{q\geq m_1+1}b_q^{(j,l)}(\rho)\psi_q^{\boldsymbol{\lambda}}(x)\right)^2 (\omega_\kappa(\rho)(x))d\rho_\kappa(x)\\
&\leq\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}\left[\int_{\mathcal{X}}\left(\sum_{q\geq m_1+1}b_q^{(j,l)}(\rho)\psi_q^{\boldsymbol{\lambda}}(x)\right)^{\frac{2\gamma}{\gamma-1}} d\rho_\kappa(x)\right]^{\frac{\gamma-1}{\gamma}}\\
&\leq\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}\left[\int_{\mathcal{X}}\left(\sum_{q\geq m_1+1}b_q^{(j,l)}(\rho)\psi_q^{\boldsymbol{\lambda}}(x)\right)^{2} d\rho_\kappa(x)\right]^{\frac{\gamma-1}{\gamma}}\left\|\sum_{q\geq m_1+1}b_q^{(j,l)}(\rho)\psi_q^{\boldsymbol{\lambda}}\right\|_\infty^{\frac{2}{\gamma}}\\
&\leq\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}\left\|\sum_{q\geq m_1+1}b_q^{(j,l)}(\rho)\psi_q^{\boldsymbol{\lambda}}\right\|_{L^2(\rho_\kappa)}^{\frac{2(\gamma-1)}{\gamma}}\left\|\sum_{q\geq m_1+1}b_q^{(j,l)}(\rho)\psi_q^{\boldsymbol{\lambda}}\right\|_{\mathcal{H}_{k_{\boldsymbol{\lambda}}}}^{\frac{2}{\gamma}}\\
&=\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}\left\|\sum_{q\geq m_1+1}b_q^{(j,l)}(\rho)\psi_q^{\boldsymbol{\lambda}}\right\|_{L^2(\rho_\kappa)}^{\frac{2(\gamma-1)}{\gamma}}\left(\sum_{q\geq m_1+1}(b_q^{(j,l)}(\rho))^2\right)^{\frac{1}{\gamma}}.
\end{aligned}$$

Insert the above estimation back into (4.21). It can be derived by the approximation result in Appendix 4.5.3 that

$$\begin{aligned}
&\|\mathcal{N}_{2n}(\mathrm{I}_{\boldsymbol{\lambda}}(\cdot),1)-\Psi_{2n,m_1}\|_{L^2(\nu_{\mathcal{G}})}^2\\
\leq&\|\alpha\|_1^2\int_{\mathcal{B}_{2,b}(\mathcal{X})}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}\sum_{l=1}^d\left[\max_{1\leq j\leq 2n}\left\|\sum_{q\geq m_1+1}b_q^{(j,l)}(\rho)\psi_q^{\boldsymbol{\lambda}}\right\|_{L^2(\rho_\kappa)}^{\frac{2(\gamma-1)}{\gamma}}\left(\sum_{q\geq m_1+1}(b_q^{(j,l)}(\rho))^2\right)^{\frac{1}{\gamma}}\right]d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho)\\
\leq&4C_F^2\,\mathbb{E}_{\rho\sim\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}}\ \|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}\sum_{l=1}^d\left[\max_{1\leq j\leq 2n}\left(\left(\sum_{q\in\mathbb{N}}(b_q^{(j,l)}(\rho))^2\right)^{\frac{1}{2}}C_{\kappa,\xi}m_1^{-\xi}\right)^{\frac{2(\gamma-1)}{\gamma}}\left(\sum_{q\in\mathbb{N}}(b_q^{(j,l)}(\rho))^2\right)^{\frac{1}{\gamma}}\right]\\
=&4C_F^2\int_{\mathcal{B}_{2,b}(\mathcal{X})}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}\sum_{l=1}^d\left[\max_{1\leq j\leq 2n}\left(\sum_{q\in\mathbb{N}}(b_q^{(j,l)}(\rho))^2\right)\right]C_{\kappa,\xi,\gamma}m_1^{-\xi\cdot\frac{2(\gamma-1)}{\gamma}}\,d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho)\\
\leq&4C_F^2\int_{\mathcal{B}_{2,b}(\mathcal{X})}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}dC_{\mathcal{B}}C_{\kappa,\xi,\gamma}m_1^{-\xi\cdot\frac{2(\gamma-1)}{\gamma}}\,d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho)\leq C_1m_1^{-\xi\cdot\frac{2(\gamma-1)}{\gamma}}
\end{aligned}$$

where $C_F = \|F\|_{\mathcal{F}_1}$ and $C_1 = 4dC_F^2 C_{\mathcal{B}} C_{\kappa,\xi,\gamma} C_{\mathcal{G}}^{\frac{1}{2}}$.

For $\gamma = \infty$, it's easy to obtain that

$$\mathcal{E}_{j,l}(\rho) \le \|\omega_\kappa(\rho)\|_{L^\infty(\rho_\kappa)}\| \sum_{q \ge m_1+1} b_q^{(j,l)}(\rho)\psi_q^{\boldsymbol{\lambda}}\|_{L^2(\rho_\kappa)}^2$$

and then

$$\|\mathcal{N}_{2n} - \Psi_{2n,m_1}\|_{L^2(\pi)}^2 \le C_1 m_1^{-2\xi}.$$

Next we perform a truncation on the index p: we let

$$\Psi_{2n,m_1,m_2}(\rho,x) := \sum_{j=1}^{2n} \alpha_j \sigma \left(\sum_{l=1}^{d}\sum_{q=1}^{m_1}\sum_{p=1}^{m_2}\sum_{s=1}^{d} a_{p,q,s}^{(j,l)} \psi_q^{\boldsymbol{\lambda}}(x) \int \psi_p^{\boldsymbol{\lambda}}(y)(e_l e_s^T) y d\rho(y) + b_j \right) + b_0.$$

and $a_{p,s}^{(j,l)}(x) := \sum_{q=1}^{m_1} a_{p,q,s}^{(j,l)} \psi_q^{\boldsymbol{\lambda}}(x)$. Then we have

$$\begin{aligned}
&\|\Psi_{2n,m_1} - \Psi_{2n,m_1 m_2}\|_{L^2(\nu_{\mathcal{G}})}^2 = \int_{\Omega_{\mathcal{B}}} \|\Psi_{2n,m_1}(\rho,x) - \Psi_{2n,m_1,m_2}(\rho,x)\|_2^2 \, d\nu_{\mathcal{G}}(\rho,x) \\
&\le \int_{\Omega_{\mathcal{B}}} \left\| \sum_{j=1}^{2n} |\alpha_j| \sum_{l=1}^{d} \left| \sum_{q=1}^{m_1} \sum_{p \ge m_2+1} \sum_{s=1}^{d} a_{p,q,s}^{(j,l)} \psi_q^{\boldsymbol{\lambda}}(x) \int \psi_p^{\boldsymbol{\lambda}}(y) e_s^T y \, d\rho(y) \right| e_l \right\|_2^2 d\nu_{\mathcal{G}}(\rho,x) \\
&= \int_{\Omega_{\mathcal{B}}} \sum_{l=1}^{d} \left(\sum_{j=1}^{2n} |\alpha_j| \left| \sum_{q=1}^{m_1} \sum_{p \ge m_2+1} \sum_{s=1}^{d} a_{p,q,s}^{(j,l)} \psi_q^{\boldsymbol{\lambda}}(x) \int \psi_p^{\boldsymbol{\lambda}}(y) e_s^T y \, d\rho(y) \right| \right)^2 d\nu_{\mathcal{G}}(\rho,x) \\
&=: \int_{\Omega_{\mathcal{B}}} \sum_{l=1}^{d} \left(\sum_{j=1}^{2n} |\alpha_j| \left| \sum_{p \ge m_2+1} \sum_{s=1}^{d} a_{p,s}^{(j,l)}(x) \int \psi_p^{\boldsymbol{\lambda}}(y) e_s^T y \, d\rho(y) \right| \right)^2 d\nu_{\mathcal{G}}(\rho,x) \\
&\le \int_{\Omega_{\mathcal{B}}} \|\alpha\|_1 \sum_{l=1}^{d} \sum_{j=1}^{2n} |\alpha_j| \left(\sum_{p \ge m_2+1} \sum_{s=1}^{d} a_{p,s}^{(j,l)}(x) \int \psi_p^{\boldsymbol{\lambda}}(y) e_s^T y \, d\rho(y) \right)^2 d\nu_{\mathcal{G}}(\rho,x) \\
&=: \int_{\Omega_{\mathcal{B}}} \|\alpha\|_1 \sum_{l=1}^{d} \sum_{j=1}^{2n} |\alpha_j| \, \mathcal{E}'_{j,l}(\rho,x) \, d\nu_{\mathcal{G}}(\rho,x). \qquad (4.22)
\end{aligned}$$

Similarly, by dominated convergence theorem and the integral shift to the gaussian distribution ρ_κ, we obtain

$$\begin{aligned}
\mathcal{E}'_{j,l}(\rho,x) &= \left(\int \sum_{s=1}^{d}(e_s^T y)\left(\sum_{p\geq m_2+1} a_{p,s}^{(j,l)}(x)\psi_p^{\boldsymbol{\lambda}}(y)\right) d\rho(y)\right)^2 \\
&= \left(\int \sum_{s=1}^{d}(e_s^T y)a_s^{(j,l)}(x,y)\, d\rho(y)\right)^2 \leq \left(\int \|y\|_2 \left(\sum_{s=1}^{d} a_s^{(j,l)}(x,y)^2\right)^{\frac{1}{2}} d\rho(y)\right)^2 \\
&\leq \mathbb{E}_\rho\|Y\|_2^2 \int \sum_{s=1}^{d} a_s^{(j,l)}(x,y)^2\, d\rho(y) \leq C_{\mathcal{B}} \sum_{s=1}^{d}\int a_s^{(j,l)}(x,y)^2 \omega_\kappa(\rho)(y)\, d\rho_\kappa(y) \\
&\leq C_{\mathcal{B}}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)} \sum_{s=1}^{d}\left[\int \left(a_s^{(j,l)}(x,y)\right)^{\frac{2\gamma}{\gamma-1}} d\rho_\kappa(y)\right]^{\frac{\gamma-1}{\gamma}} \\
&\leq C_{\mathcal{B}}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)} \sum_{s=1}^{d} \underbrace{\|a_s^{(j,l)}(x,\cdot)\|_{L^2(\rho_\kappa)}^{\frac{2(\gamma-1)}{\gamma}}}_{\mathrm{I}_1^{(j,l,s)}(x)} \underbrace{\|a_s^{(j,l)}(x,\cdot)\|_\infty^{\frac{2}{\gamma}}}_{\mathrm{I}_2^{(j,l,s)}(x)}.
\end{aligned}$$

Then (4.22) can be further bounded by

$$\begin{aligned}
&C_{\mathcal{B}}\int_{\mathcal{B}_{2,b}(\mathcal{X})} \|\alpha\|_1\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)} \sum_{j=1}^{2n}|\alpha_j| \sum_{s,l=1}^{d}\left(\int_{\mathcal{X}} \mathrm{I}_1^{(j,l,s)}(x)\mathrm{I}_2^{(j,l,s)}(x)\, d\rho(x)\right) d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho) \\
\leq\, &C_{\mathcal{B}}\int_{\mathcal{B}_{2,b}(\mathcal{X})} \|\alpha\|_1^2\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)} \max_{1\leq j\leq 2n}\sum_{s,l=1}^{d}\left(\int_{\mathcal{X}} \mathrm{I}_1^{(j,l,s)}(x)\mathrm{I}_2^{(j,l,s)}(x)\, d\rho(x)\right) d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho) =: \Delta_0
\end{aligned}$$

where the integral of $\mathrm{I}_1^{(j,l,s)}(x)\mathrm{I}_2^{(j,l,s)}(x)$ with respect to ρ can be bounded by

$$\int_{\mathcal{X}} \mathrm{I}_1^{(j,l,s)}(x)\mathrm{I}_2^{(j,l,s)}(x)\, d\rho(x) \leq \left\|\mathrm{I}_1^{(j,l,s)}\right\|_{L^{\frac{\gamma}{\gamma-1}}(\rho)} \left\|\mathrm{I}_2^{(j,l,s)}\right\|_{L^\gamma(\rho)}.$$

For the first term $\mathrm{I}_1^{(j,l,s)}$,

$$\begin{aligned}
\left\|\mathrm{I}_1^{(j,l,s)}\right\|_{L^{\frac{\gamma}{\gamma-1}}(\rho)} &= \left(\int_{\mathcal{X}} \left\|a_s^{(j,l)}(x,\cdot)\right\|_{L^2(\rho_\kappa)}^2 \, d\rho(x)\right)^{\frac{\gamma-1}{\gamma}} \\
&= \left(\int\int \left(\sum_{p\geq m_2+1} a_{p,s}^{(j,l)}(x)\psi_p^{\boldsymbol{\lambda}}(y)\right)^2 d\rho_\kappa(y)\, d\rho(x)\right)^{\frac{\gamma-1}{\gamma}} \\
&= \left(\int\int \sum_{p\geq m_2+1} (a_{p,s}^{(j,l)}(x))^2(\psi_p^{\boldsymbol{\lambda}}(y))^2 \, d\rho_\kappa(y)\, d\rho(x)\right)^{\frac{\gamma-1}{\gamma}} \\
&= \left(\int \sum_{p\geq m_2+1} \left(\int (a_{p,s}^{(j,l)}(x))^2 \, d\rho(x)\right) (\psi_p^{\boldsymbol{\lambda}}(y))^2 \, d\rho_\kappa(y)\right)^{\frac{\gamma-1}{\gamma}} \\
&= \left\| \sum_{p\geq m_2+1} \left\|a_{p,s}^{(j,l)}\right\|_{L^2(\rho)} \psi_p^{\boldsymbol{\lambda}} \right\|_{L^2(\rho_\kappa)}^{\frac{2(\gamma-1)}{\gamma}}.
\end{aligned}$$

Perform the domain shift to each coefficient of $\psi_p^{\boldsymbol{\lambda}}$ again and we can obtain

$$\begin{aligned}
\left\|a_{p,s}^{(j,l)}\right\|_{L^2(\rho)} &= \int \left(\sum_{q=1}^{m_1} a_{p,q,s}^{(j,l)} \psi_q^{\boldsymbol{\lambda}}(x)\right)^2 \omega_\kappa(\rho)(x)\, d\rho_\kappa(x) \\
&\leq \|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)} \left\|\sum_{q=1}^{m_1} a_{p,q,s}^{(j,l)} \psi_q^{\boldsymbol{\lambda}}\right\|_{L^2(\rho_\kappa)}^{\frac{2(\gamma-1)}{\gamma}} \left\|\sum_{q=1}^{m_1} a_{p,q,s}^{(j,l)} \psi_q^{\boldsymbol{\lambda}}\right\|_{\mathcal{H}_{k_{\boldsymbol{\lambda}}}}^{\frac{2}{\gamma}} \\
&= \|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)} \left(\sum_{q=1}^{m_1} (a_{p,q,s}^{(j,l)})^2 r_q^{\boldsymbol{\lambda}}\right)^{\frac{\gamma-1}{\gamma}} \left(\sum_{q=1}^{m_1} (a_{p,q,s}^{(j,l)})^2\right)^{\frac{1}{\gamma}} \\
&\leq \|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)} (r_1^{\boldsymbol{\lambda}})^{\frac{\gamma-1}{\gamma}} \sum_{q=1}^{m_1} (a_{p,q,s}^{(j,l)})^2,
\end{aligned}$$

which implies that $\sum_{p\geq m_2+1} \left\|a_{p,s}^{(j,l)}\right\|_{L^2(\rho)} \psi_p^{\boldsymbol{\lambda}} \in \mathcal{H}_{k_{\boldsymbol{\lambda}}}$ with the RKHS norm

$$\left(\sum_{p\geq m_2+1} \left\|a_{p,s}^{(j,l)}\right\|_{L^2(\rho)}^2\right)^{\frac{1}{2}} \leq \left(\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)} (r_1^{\boldsymbol{\lambda}})^{\frac{\gamma-1}{\gamma}} \sum_{p\geq m_2+1} \sum_{q=1}^{m_1} (a_{p,q,s}^{(j,l)})^2\right)^{\frac{1}{2}}.$$

For the second term $\mathrm{I}_2^{(j,l,s)}$,

$$\begin{aligned}
\left\|\mathrm{I}_2^{(j,l,s)}\right\|_{L^\gamma(\rho)} &= \left(\int \left\|a_{\boldsymbol{\lambda}}^{(j,l,s)}(x,\cdot)\right\|_\infty^2 d\rho(x)\right)^{\frac{1}{\gamma}} \leq \left(\int \left\|a_{\boldsymbol{\lambda}}^{(j,l,s)}(x,\cdot)\right\|_{\mathcal{H}_{k_{\boldsymbol{\lambda}}}}^2 d\rho(x)\right)^{\frac{1}{\gamma}} \\
&\leq \left(\int \left\|\sum_{p\geq m_2+1} a_{p,s}^{(j,l)}(x)\psi_p^{\boldsymbol{\lambda}}\right\|_{\mathcal{H}_{k_{\boldsymbol{\lambda}}}}^2 d\rho(x)\right)^{\frac{1}{\gamma}} = \left(\int \sum_{p\geq m_2+1} (a_{p,s}^{(j,l)}(x))^2\, d\rho(x)\right)^{\frac{1}{\gamma}} \\
&\leq \left(\int \sum_{p\geq m_2+1} \left(\sum_{q=1}^{m_1} a_{p,q,s}^{(j,l)}\psi_q^{\boldsymbol{\lambda}}(x)\right)^2 d\rho(x)\right)^{\frac{1}{\gamma}} \\
&\leq \left(\int \left(\sum_{p\geq m_2+1}\sum_{q=1}^{m_1}(a_{p,q,s}^{(j,l)})^2\right)\left(\sum_{q=1}^{m_1}\psi_q^{\boldsymbol{\lambda}}(x)^2\right) d\rho(x)\right)^{\frac{1}{\gamma}} \\
&\leq \left(\sum_{p\geq m_2+1}\sum_{q=1}^{m_1}(a_{p,q,s}^{(j,l)})^2\right)^{\frac{1}{\gamma}} \left(\int \sum_{q=1}^{\infty}\psi_q^{\boldsymbol{\lambda}}(x)^2\, d\rho(x)\right)^{\frac{1}{\gamma}} \\
&= \left(\sum_{p\geq m_2+1}\sum_{q=1}^{m_1}(a_{p,q,s}^{(j,l)})^2\right)^{\frac{1}{\gamma}}.
\end{aligned}$$

Combine the estimations for $\mathrm{I}_1^{(j,l,s)}$ and $\mathrm{I}_2^{(j,l,s)}$ and we get an upper bound that

$$\begin{aligned}
\Delta_0 &\leq 4C_{\mathcal{B}}C_F^2 \int_{\mathcal{B}_{2,b}(\mathcal{X})} \|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)} \left(\max_{1\leq j\leq 2n}\sum_{s,l=1}^{d} \left\|\mathrm{I}_1^{(j,l)}\right\|_{L^{\frac{\gamma}{\gamma-1}}(\rho)} \left\|\mathrm{I}_2^{(j,l)}\right\|_{L^\gamma(\rho)}\right) d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho) \\
&\leq 4C_{\mathcal{B}}C_F^2 \int_{\mathcal{B}_{2,b}(\mathcal{X})} \|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)} \Bigg\{ \max_{1\leq j\leq 2n}\sum_{s,l=1}^{d} \\
&\quad \left[\left(\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}(r_1^{\boldsymbol{\lambda}})^{\frac{\gamma-1}{\gamma}}\sum_{p\in\mathbb{N}}\sum_{q=1}^{m_1}(a_{p,q,s}^{(j,l)})^2\right)^{\frac{1}{2}} C_{\kappa,\xi} m_2^{-\xi}\right]^{\frac{2(\gamma-1)}{\gamma}} \cdot \left(\sum_{p\in\mathbb{N}}\sum_{q=1}^{m_1}(a_{p,q,s}^{(j,l)})^2\right)^{\frac{1}{\gamma}} \Bigg\} d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho) \\
&\leq 4dC_{\mathcal{B}}C_F^2 (r_1^{\boldsymbol{\lambda}})^{\frac{(\gamma-1)^2}{\gamma^2}} \int_{\mathcal{B}_{2,b}(\mathcal{X})} \|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}^{\frac{2\gamma-1}{\gamma}} C_{\kappa,\xi,\gamma} m_2^{-\xi\cdot\frac{2(\gamma-1)}{\gamma}} d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho)
\end{aligned}$$

$$\leq 4dC_{\mathcal{B}}C_F^2C_{\kappa,\xi,\gamma}(r_1^{\boldsymbol{\lambda}})^{\frac{(\gamma-1)^2}{\gamma^2}}m_2^{-\xi\cdot\frac{2(\gamma-1)}{\gamma}}\int_{\mathcal{B}_{2,b}(\mathcal{X})}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}^{\frac{2\gamma-1}{\gamma}}\,d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho)$$

$$\leq C_2 m_2^{-\xi\cdot\frac{2(\gamma-1)}{\gamma}}$$

where $C_2 = 4dC_{\mathcal{B}}C_F^2C_{\kappa,\xi,\gamma}(r_1^{\boldsymbol{\lambda}})^{\frac{(\gamma-1)^2}{\gamma^2}}C_{\mathcal{G}}^{\frac{2\gamma-1}{2\gamma}}$.

4.5.1.2 Linear Transformer with Adaptive Attention Heads

Recall that

$$\Psi_{2n,m}(\rho,x) := \sum_{j=1}^{2n}\alpha_j\sigma\left(\sum_{l=1}^{d}\sum_{p,q=1}^{m}\sum_{s=1}^{d}a_{p,q,s}^{(j,l)}\psi_q^{\boldsymbol{\lambda}}(x)\int\psi_p^{\boldsymbol{\lambda}}(y)(e_le_s^T)y d\rho(y)+b_j\right)+b_0$$
$$=\sum_{j=1}^{2n}\alpha_j\sigma\left(\sum_{p,q=1}^{m}\psi_q^{\boldsymbol{\lambda}}(x)\int\psi_p^{\boldsymbol{\lambda}}(y)A_{p,q}^{(j)}y\,d\rho(y)+b_j\right)+b_0$$

with $A_{p,q}^{(j)} = \sum_{l=1}^{d}\sum_{s=1}^{d}a_{p,q,s}^{(j,l)}e_le_s^T = [a_{p,q,s}^{(j,l)}]_{1\leq l\leq d,1\leq s\leq d}$.

Now we approximate each $\psi_q^{\boldsymbol{\lambda}}$ with a neural network ϕ_q, and let

$$\mathrm{T}_{2n,m}(\rho,x) := \sum_{j=1}^{2n}\alpha_j\sigma\left(\sum_{q=1}^{m}\phi_q(x)\left(\sum_{p=1}^{m}\int\phi_p(y)A_{p,q}^{(j)}y\,d\rho(y)\right)+b_j\right)+b_0.$$

Then we have the error decomposition:

$$\|\Psi_{2n,m}-\mathrm{T}_{2n,m}\|_{L^2(\nu_{\mathcal{G}})}\leq\|\Psi_{2n,m}-\widetilde{\mathrm{T}}_{2n,m}\|_{L^2(\nu_{\mathcal{G}})}+\|\widetilde{\mathrm{T}}_{2n,m}-\mathrm{T}_{2n,m}\|_{L^2(\nu_{\mathcal{G}})}$$

where

$$\widetilde{\mathrm{T}}_{2n,m}(\rho,x) = \sum_{j=1}^{2n}\alpha_j\sigma\left(\sum_{q=1}^{m}\phi_q(x)\left(\sum_{p=1}^{m}\int\psi_p^{\boldsymbol{\lambda}}(y)A_{p,q}^{(j)}y\,d\rho(y)\right)+b_j\right)+b_0.$$

Similarly with error estimations for the truncated error, let

$$\widetilde{b}_q^{(j,l)}(\rho) := \sum_{1\leq p\leq m,1\leq s\leq d}a_{p,q,s}^{(j,l)}\int\psi_p^{\boldsymbol{\lambda}}(y)e_s^Ty\,d\rho(y)$$

and we have

$$
\begin{aligned}
&\left\|\Psi_{2n,m}-\widetilde{\mathrm{T}}_{2n,m}\right\|^2_{L^2(\nu_{\mathcal{G}})}\\
&\leq \int_{\Omega_{\mathcal{B}}}\left\|\sum_{j=1}^{2n}|\alpha_j|\sum_{l=1}^{d}\left|\sum_{q=1}^{m}\widetilde{b}_q^{(j,l)}(\rho)\big(\psi_q^{\boldsymbol{\lambda}}(x)-\phi_q(x)\big)\right|e_l\right\|_2^2 d\nu_{\mathcal{G}}(\rho,x)\\
&=\int_{\Omega_{\mathcal{B}}}\sum_{l=1}^{d}\left(\sum_{j=1}^{2n}|\alpha_j|\left|\sum_{q=1}^{m}\widetilde{b}_q^{(j,l)}(\rho)\big(\psi_q^{\boldsymbol{\lambda}}(x)-\phi_q(x)\big)\right|\right)^2 d\nu_{\mathcal{G}}(\rho,x)\\
&\leq \int_{\mathcal{B}_{2,b}(\mathcal{X})}\|\alpha\|_1\sum_{l=1}^{d}\sum_{j=1}^{2n}|\alpha_j|\int_{\mathcal{X}}\left(\sum_{q\in[m]}\widetilde{b}_q^{(j,l)}(\rho)\big(\psi_q^{\boldsymbol{\lambda}}(x)-\phi_q(x)\big)\right)^2 d\rho(x)d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho)\\
&\leq \int_{\mathcal{B}_{2,b}(\mathcal{X})}\|\alpha\|_1^2\sum_{l=1}^{d}\max_{1\leq j\leq 2n}\left(\sum_{q=1}^{m}(\widetilde{b}_q^{(j,l)}(\rho))^2\right)\int_{\mathcal{X}}\sum_{q=1}^{m}\big(\psi_q^{\boldsymbol{\lambda}}(x)-\phi_q(x)\big)^2 d\rho(x)d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho)\\
&\leq d\|\alpha\|_1^2 C_{\mathcal{B}}\int_{\mathcal{B}_{2,b}(\mathcal{X})}\sum_{q=1}^{m}\int_{\mathcal{X}}\big(\psi_q^{\boldsymbol{\lambda}}(x)-\phi_q(x)\big)^2\, d\rho(x)\, d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho)\\
&=: dC_{\mathcal{B}}\|\alpha\|_1^2\sum_{q=1}^{m}\int_{\mathcal{B}_{2,b}(\mathcal{X})}\widetilde{\mathcal{E}}_q(\rho)\, d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho).
\end{aligned}
$$

Take $\phi_q = \phi_{q,\tilde{m}}$ to be the two-hidden-layer tanh neural network with a product gate in Appendix 4.5.3. Then we have

$$\widetilde{\mathcal{E}}_q(\rho)\leq (4d^2+C_{\kappa,\gamma}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)})\exp(-2\tilde{m}\log\tilde{m}).$$

Take the estimation back and it follows that

$$\left\|\Psi_{2n,m}-\widetilde{\mathrm{T}}_{2n,m,\tilde{m}}\right\|^2_{L^2(\nu_{\mathcal{G}})}\leq C_3 m\exp(-2\tilde{m}\log\tilde{m})$$

with $C_3=4dC_{\mathcal{B}}C_F^2\left(4d^2+C_{\kappa,\gamma}C_{\mathcal{G}}^{\frac{1}{2}}\right)$ for $\tilde{m}>C_{\kappa,\theta,\gamma}$.

Then we use the same group of two-hidden-layer neural networks $\{\phi_{q,\tilde{m}}\}_{1\leq q\leq m}$ to approximate $\{\psi_q^{\boldsymbol{\lambda}}\}_{1\leq q\leq m}$ in the context memory part. Now let a linear Transformer

$$\mathrm{T}_{2n,m,\tilde{m}}(\rho,x)=\sum_{j=1}^{2n}\alpha_j\sigma\left(\sum_{q=1}^{m}\phi_{q,\tilde{m}}(x)\left(\sum_{p=1}^{m}\int\phi_{p,\tilde{m}}(y)A_{p,q}^{(j)}y\, d\rho(y)\right)+b_j\right)+b_0.$$

Then the approximation error for context functions can be measured as

$$\begin{aligned}
&\left\|\widetilde{\mathrm{T}}_{2n,m,\tilde{m}} - \mathrm{T}_{2n,m,\tilde{m}}\right\|^2_{L^2(\nu_{\mathcal{G}})} \\
&\leq \mathbb{E}_{\rho\sim\mathcal{P}^{\mathcal{X}}_{\mathcal{G}}}\|\alpha\|_1^2 \sum_{l=1}^{d} \max_{1\leq j\leq 2n} \int_{\mathcal{X}} \Big(\sum_{1\leq p,q\leq m, 1\leq s\leq d} a^{(j,l)}_{p,q,s}\phi_{q,\tilde{m}}(x) \\
&\int (\psi^{\boldsymbol{\lambda}}_p(y) - \phi_{p,\tilde{m}}(y)) e_s^T y \, d\rho(y)\Big)^2 d\rho(x) \\
&\leq \int_{\mathcal{B}_{2,b}(\mathcal{X})} \|\alpha\|_1^2 \sum_{l=1}^{d} \max_{1\leq j\leq 2n} \left(\int_{\mathcal{X}} \sum_{1\leq p\leq m, 1\leq s\leq d} (\tilde{a}^{(j,l)}_{p,s}(x))^2 d\rho(x)\right) \cdot \\
&\left(\sum_{1\leq p\leq m, 1\leq s\leq d} \left(\int (\psi^{\boldsymbol{\lambda}}_p(y) - \phi_{p,\tilde{m}}(y)) e_s^T y \, d\rho(y)\right)^2\right) d\mathcal{P}^{\mathcal{X}}_{\mathcal{G}}(\rho)
\end{aligned}$$

where

$$\tilde{a}^{(j,l)}_{p,s}(x) := \sum_{q=1}^{m} a^{(j,l)}_{p,q,s}\phi_{q,\tilde{m}}(x).$$

For the first factor, let

$$\mathrm{I}^{(j,l)}_3(\rho) = \int_{\mathcal{X}} \sum_{1\leq p\leq m, 1\leq s\leq d} (\tilde{a}^{(j,l)}_{p,s}(x))^2 \, d\rho(x),$$

and we can obtain

$$\begin{aligned}
\sum_{l=1}^{d} \max_{1\leq j\leq 2n} \mathrm{I}^{(j,l)}_3(\rho) &\leq d\sum_{q=1}^{m} \int_{\mathcal{X}} (\phi_{q,\tilde{m}}(x))^2 \, d\rho(x) \leq d\sum_{q=1}^{m} (2\|\psi^{\boldsymbol{\lambda}}_q\|^2_{L^2(\rho)} + 2\|\psi^{\boldsymbol{\lambda}}_q - \phi_{q,\tilde{m}}\|^2_{L^2(\rho)}) \\
&\leq 2d + 2dm(4d^2 + C_{\kappa,\gamma}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}) \exp(-2\tilde{m}\log\tilde{m}).
\end{aligned}$$

For the second factor, it's easy to observe that

$$\begin{aligned}
&\sum_{1\leq p\leq m, 1\leq s\leq d} \left(\int (\psi^{\boldsymbol{\lambda}}_p(y) - \phi_{p,\tilde{m}}(y)) e_s^T y \, d\rho(y)\right)^2 \leq \sum_{p=1}^{m} \|\psi^{\boldsymbol{\lambda}}_p - \phi_{q,\tilde{m}}\|^2_{L^2(\rho)} \mathbb{E}_{X\sim\rho}\|X\|_2^2 \\
&\leq C_{\mathcal{B}} m(4d^2 + C_{\kappa,\gamma}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}) \exp(-2\tilde{m}\log\tilde{m}).
\end{aligned}$$

Choose $\tilde{m}$ such that $m\exp(-2\tilde{m}\log\tilde{m}) < 1$. Combine two estimations. It is obtained that

$$\begin{aligned}
&\left\|\tilde{\mathrm{T}}_{n,m,\tilde{m}} - \mathrm{T}_{n,m,\tilde{m}}\right\|^2_{L^2(\nu_{\mathcal{G}})} \\
&\leq 8C_F^2 C_{\mathcal{B}} d \int_{\mathcal{B}_2(\mathcal{X})} (20d^4 + 9d^2 C_{\kappa,\gamma}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)} + C^2_{\kappa,\gamma}\|\omega_\kappa(\rho)\|^2_{L^\gamma(\rho_\kappa)}) m\exp(-2\tilde{m}\log\tilde{m})\, d\mathcal{P}^{\mathcal{X}}_{\mathcal{G}}(\rho) \\
&\leq C_4 m\exp(-2\tilde{m}\log\tilde{m})
\end{aligned}$$

with $C_4 = 8C_F^2 C_{\mathcal{B}} d(20d^4 + 9d^2(1 + C_{\kappa,\gamma}C_{\mathcal{G}})^2)$.

Combining all the estimations, we can achieve the following convergence rate for $n \geq C'_{\kappa,\theta,\gamma}$ with $C'_{\kappa,\theta,\gamma} = \exp\left(\frac{4(\gamma-1)\xi C_{\kappa,\theta,\gamma}}{2(\gamma-1)\xi+\gamma}\right)$:

$$\begin{aligned}
&\|F(\mathrm{I}_{\boldsymbol{\lambda}}(\cdot),1) - \mathrm{T}_{2n,m,\tilde{m}}\|_{L^2(\nu_{\mathcal{G}})} \\
&\leq \|F(\mathrm{I}_{\boldsymbol{\lambda}}(\cdot),1) - \mathcal{N}_{2n}(\mathrm{I}_{\boldsymbol{\lambda}}(\cdot),1)\|_{L^2(\nu_{\mathcal{G}})} + \|\mathcal{N}_{2n}(\mathrm{I}_{\boldsymbol{\lambda}}(\cdot),1) - \Psi_{2n,m}\|_{L^2(\nu_{\mathcal{G}})} + \|\Psi_{2n,m} - \mathrm{T}_{2n,m,\tilde{m}}\|_{L^2(\nu_{\mathcal{G}})} \\
&\leq ((1+C_{\mathcal{B}})C_F^2)^{\frac{1}{2}} n^{-\frac{1}{2}} + \left(C_1^{\frac{1}{2}} + C_2^{\frac{1}{2}}\right) m^{-\frac{\xi(\gamma-1)}{\gamma}} + \left(C_3^{\frac{1}{2}} + C_4^{\frac{1}{2}}\right) m^{\frac{1}{2}}\exp(-\tilde{m}\log\tilde{m}) \leq C_* n^{-\frac{1}{2}},
\end{aligned}$$

with $C_* = 3\max\left\{(1+C_{\mathcal{B}})^{\frac{1}{2}}C_F, C_1^{\frac{1}{2}} + C_2^{\frac{1}{2}}, C_3^{\frac{1}{2}} + C_4^{\frac{1}{2}}\right\}$,

$$m = \left\lceil n^{\frac{\gamma}{2(\gamma-1)\xi}} \right\rceil \text{ and } \tilde{m} = \left\lceil \left(\frac{1}{2} + \frac{\gamma}{4(\gamma-1)\xi}\right)\log n \right\rceil.$$

■

4.5.2 Oracle Inequality: Sampling Error for Linear Transformers

In this Subsection, we derive an oracle inequality for the two-stage sampling estimation. We first prove the compactness of the hypothesis space, which guarantees the existence of $\mathrm{T}_{\mathbb{S},n}$ in (4.6); we then estimate the covering numbers used to bound the empirical process.

4.5.2.1 Compact subspaces in $C(\Omega)$

Recall that (Ω, d_Ω) is a complete separable metric space. To prove that $\mathcal{H}_{\mathrm{T}_n}$ is compact in $C(\Omega)$, it's sufficient to show that $\mathcal{H}_{\mathrm{T}_n}$ is sequentially compact in $C(\Omega)$. By Arzelà-Ascoli Theorem, it's sufficient to check the equi-boundedness and equi-continuity of $\mathcal{H}_{\mathrm{T}_n}$. For

any $\mathrm{T}_n \in \mathcal{H}_{\mathrm{T}_n}$, we can pick a group of parameters $\left((\alpha_j),(b_j),(A_{p,q}^{(j)}),\Theta_{\tanh}\right)$ satisfying the conditions of the hypothesis space $\mathcal{H}_{\mathrm{T}_n}$. For simplicity, we denote

$$\sigma_j(\rho,x)=\sigma\left(\sum_{q=1}^{m(n)}\phi_{q,\tilde{m}(n)}(x)\left(\sum_{p=1}^{m(n)}\mathcal{T}_{C_{\mathcal{B}}}\left[\int\phi_{p,\tilde{m}(n)}(y)A_{p,q}^{(j)}y\,d\rho(y)\right]\right)+b_j\right).$$

Then we have

$$\begin{aligned}
&\|\mathrm{T}_n(\rho,x)-\mathrm{T}_n(\rho',x')\|_2=\left\|\sum_{j=1}^{n}\alpha_j(\sigma_j(\rho,x)-\sigma_j(\rho',x'))\right\|_2\\
\leq&\|\alpha\|_1\max_{1\leq j\leq n}\|\sigma_j(\rho,x)-\sigma_j(\rho',x')\|_2\\
\leq&\|\alpha\|_1\max_{1\leq j\leq n}\sum_{p,q=1}^{m(n)}\Big\|\phi_{q,\tilde{m}(n)}(x)\mathcal{T}_{C_{\mathcal{B}}}\left[\int\phi_{p,\tilde{m}(n)}(y)A_{p,q}^{(j)}y\,d\rho(y)\right]\\
&-\phi_{q,\tilde{m}(n)}(x')\mathcal{T}_{C_{\mathcal{B}}}\left[\int\phi_{p,\tilde{m}(n)}(y)A_{p,q}^{(j)}y\,d\rho'(y)\right]\Big\|_2\\
\leq&\|\alpha\|_1\max_{1\leq j\leq n}\sum_{p,q=1}^{m(n)}\Bigg\|\phi_{q,\tilde{m}(n)}(x)\Big(\mathcal{T}_{C_{\mathcal{B}}}\left[\int\phi_{p,\tilde{m}(n)}(y)A_{p,q}^{(j)}y\,d\rho(y)\right]\\
&\qquad-\mathcal{T}_{C_{\mathcal{B}}}\left[\int\phi_{p,\tilde{m}(n)}(y)A_{p,q}^{(j)}y\,d\rho'(y)\right]\Big)\\
&\qquad+\mathcal{T}_{C_{\mathcal{B}}}\left[\int\phi_{p,\tilde{m}(n)}(y)A_{p,q}^{(j)}y\,d\rho'(y)\right](\phi_{q,\tilde{m}(n)}(x)-\phi_{q,\tilde{m}(n)}(x'))\Bigg\|_2\\
\leq&\|\alpha\|_1\max_{1\leq j\leq n}\sum_{p,q=1}^{m(n)}\Bigg(\underbrace{\left\|\int(\phi_{p,\tilde{m}(n)}(y)A_{p,q}^{(j)}y-\phi_{p,\tilde{m}(n)}(y')A_{p,q}^{(j)}y')\,d\rho(y)d\rho'(y')\right\|_2}_{\Delta(\rho,\rho')}\\
&\qquad+\sqrt{d}C_{\mathcal{B}}\left|\phi_{q,\tilde{m}(n)}(x)-\phi_{q,\tilde{m}(n)}(x')\right|\Bigg).
\end{aligned}$$

Recall that $\phi_{q,\tilde{m}(n)}(x) = \prod_{l=1}^{d} \mathcal{T}_1(\phi_{q,\tilde{m}(n)}^{(l)}(x))$, where $(\phi_{q,\tilde{m}(n)}^{(l)})_{l=1}^{d}$ is a group of two-hidden-layer tanh neural networks shown in Appendix 4.5.3. It's easy to observe that

$$
\begin{aligned}
&\left|\phi_{q,\tilde{m}(n)}(x) - \phi_{q,\tilde{m}(n)}(x')\right| \\
=&\left|\prod_{l=1}^{d} \mathcal{T}_1(\phi_{q,\tilde{m}(n)}^{(l)}(x)) - \prod_{l=1}^{d} \mathcal{T}_1(\phi_{q,\tilde{m}(n)}^{(l)}(x'))\right| \leq d \max_{1\leq l\leq d} \left|\phi_{q,\tilde{m}(n)}^{(l)}(x) - \phi_{q,\tilde{m}(n)}^{(l)}(x')\right| \\
\leq& d \max_{1\leq l\leq d} \left|c_{q,l}^T[\sigma_{\tanh}(W_{q,1}^{(l)}\sigma_{\tanh}(W_{q,0}^{(l)}x + b_{q,0}^{(l)}) + b_{q,1}^{(l)}) - \sigma_{\tanh}(W_{q,1}^{(l)}\sigma_{\tanh}(W_{q,0}^{(l)}x' + b_{q,0}^{(l)}) + b_{q,1}^{(l)})]\right| \\
\leq& d \max_{1\leq l\leq d} \|c_{q,l}^T\|_\infty \left\|W_{q,1}^{(l)}(\sigma_{\tanh}(W_{q,0}^{(l)}x + b_{q,0}^{(l)}) - \sigma_{\tanh}(W_{q,0}^{(l)}x' + b_{q,0}^{(l)}))\right\|_\infty \\
\leq& d \max_{1\leq l\leq d} \|c_{q,l}^T\|_\infty \left\|W_{q,1}^{(l)}\right\|_\infty \left\|W_{q,0}^{(l)}\right\|_\infty \|x - x'\|_\infty \leq \left(d \max_{1\leq l\leq d} \|c_{q,l}^T\|_\infty \left\|W_{q,1}^{(l)}\right\|_\infty \left\|W_{q,0}^{(l)}\right\|_\infty\right)\|x - x'\|_2 \\
\leq& C_{n,d}\|x - x'\|_2.
\end{aligned}
$$

Since the parameters in tanh neural networks are uniformly bounded by $\|\Theta_{\tanh}\|_\infty$, $C_{n,d}$ just depends on n and d. It also follows that for any $\tau \in \prod(\rho, \rho')$,

$$
\begin{aligned}
\Delta(\rho, \rho') &\leq \int \left\|\phi_{p,\tilde{m}(n)}(y)A_{p,q}^{(j)}y - \phi_{p,\tilde{m}(n)}(y')A_{p,q}^{(j)}y'\right\|_2 d\tau(y, y') \\
&\leq \left\|A_{p,q}^{(j)}\right\|_F \int \|\phi_{p,\tilde{m}(n)}(y)y - \phi_{p,\tilde{m}(n)}(y')y'\|_2 \, d\tau(y, y') \\
&\leq \left\|A_{p,q}^{(j)}\right\|_F \left(\int \left\|\phi_{p,\tilde{m}(n)}(y)(y - y')\right\|_2 d\tau(y, y') + \int \left\|y'(\phi_{p,\tilde{m}(n)}(y) - \phi_{p,\tilde{m}(n)}(y'))\right\|_2 d\tau(y, y')\right) \\
&\leq \left\|A_{p,q}^{(j)}\right\|_F \left(\int \|y - y'\|_2 \, d\tau(y, y') + C_{n,d}\sqrt{\mathbb{E}_{\rho'}\|Y'\|_2^2}\sqrt{\mathbb{E}_\tau\|Y - Y'\|_2^2}\right) \\
&\leq \left\|A_{p,q}^{(j)}\right\|_F (1 + C_{n,d}\sqrt{\mathbb{E}_{\rho'}\|Y'\|_2^2})\sqrt{\mathbb{E}_\tau\|Y - Y'\|_2^2}.
\end{aligned}
$$

Take the above estimation back. It can be obtained that

$$
\begin{aligned}
&\|\mathrm{T}_n(\rho, x) - \mathrm{T}_n(\rho', x')\|_2 \\
\leq& \|\alpha\|_1 \max_{1\leq j\leq n} \sum_{p,q=1}^{m(n)} \left(\left\|A_{p,q}^{(j)}\right\|_F (1 + C_{n,d}\|Y\|_{L^2(\rho')})\|Y - Y'\|_{L^2(\tau)} + \sqrt{d}C_{\mathcal{B}}C_{n,d}\|x - x'\|_2\right) \\
\leq& \|\alpha\|_1 m\sqrt{d}(1 + C_{n,d}\|Y\|_{L^2(\rho')} + \sqrt{d}C_{\mathcal{B}}C_{n,d})(\|Y - Y'\|_{L^2(\tau)} + \|x - x'\|_2)
\end{aligned}
$$

which holds for any $\tau \in \prod(\rho, \rho')$. It follows that

$$
\|\mathrm{T}_n(\rho, x) - \mathrm{T}_n(\rho', x')\|_2 \leq C_F m\sqrt{d}(1 + C_{n,d}\|Y\|_{L^2(\rho')} + \sqrt{d}C_{\mathcal{B}}C_{n,d})d_\Omega((\rho, x), (\rho', x'))
$$

and proves that $\mathcal{H}_{\mathrm{T}_n}$ is equi-continuous at each $(\rho', x') \in \Omega$.

For any $(\rho, x) \in \Omega$, we have

$$\begin{aligned}\|\mathrm{T}_n(\rho,x)\|_2 &\le \|b_0\|_2 + \|\alpha\|_1 \max_{1\le j\le n}\left(\|b_j\|_2 + \left\|\sum_{p,q=1}^{m(n)} \phi_{q,\tilde{m}(n)}(x)\mathcal{T}_{C_{\mathcal{B}}}\left[\int \phi_{p,\tilde{m}(n)}(y)A_{p,q}^{(j)} y\, d\rho(y)\right]\right\|_2\right)\\ &\le C_F\sqrt{dC_{\mathcal{B}}} + C_F\left(\sqrt{2dC_{\mathcal{B}}} + \sum_{p,q=1}^{m(n)}\left\|\mathcal{T}_{C_{\mathcal{B}}}\left[\int \phi_{p,\tilde{m}(n)}(y)A_{p,q}^{(j)} y\, d\rho(y)\right]\right\|_2\right)\\ &\le C_F(2\sqrt{dC_{\mathcal{B}}} + m^2 C_{\mathcal{B}}\sqrt{d}).\end{aligned}$$

which shows that $\{\mathrm{T}_n(\rho,x) : \mathrm{T}_n \in \mathcal{H}_{\mathrm{T}_n}\}$ is bounded for each $(\rho, x) \in \Omega$. Therefore, $\mathcal{H}_{\mathrm{T}_n}$ is compact in $C(\Omega)$ which guarantees the existence of $\mathrm{T}_{\mathbb{S},n}$ and the below covering number.

4.5.2.2 Covering Number Estimations

Note that in the first stage and pseudo second stage sampling, we have access to the true distributions sampled from $\mathcal{P}_{\mathcal{G}}$ on $\mathcal{B}_{2,b}(\mathcal{X})$ such that the second moments are uniformly bounded by $C_{\mathcal{B}}$, which however doesn't hold true for the empirical distributions in the second stage sampling.

Recall that the hypothesis space is defined as

$$\begin{aligned}\mathcal{H}_{\mathrm{T}_n} = \Bigg\{\mathrm{T}_n : \|\alpha\|_1 \le 2C_F, \sum_{p,q=1}^{m(n)} \|A_{p,q}^{(j)}\|_F^2 \le d, \|b_j\|_2 \le \sqrt{2dC_{\mathcal{B}}} \text{ for each } 1\le j\le n,\\ \|b_0\|_2 \le C_F\sqrt{2dC_{\mathcal{B}}}, \text{ and } \|\Theta_{\tanh}\|_\infty \le c_1(c_2'\log(n))^{c_3'(\log n)^2}\Bigg\}.\end{aligned}$$

For $\sigma_j(\rho, x)$ defined above, it's easy to observe that for any $(\rho, x) \in \Omega_{\mathcal{B}}$,

$$\begin{aligned}
\|\sigma_j(\rho,x)\|_2 &\le \left\| b_j + \sum_{p,q=1}^{m(n)} \phi_{q,\tilde{m}(n)}(x) \mathcal{T}_{C_{\mathcal{B}}} \left[\int \phi_{p,\tilde{m}(n)}(y) A_{p,q}^{(j)} y \, d\rho(y) \right] \right\|_2 \\
&\le \|b_j\|_2 + \sum_{p,q=1}^{m(n)} \int \left\| A_{p,q}^{(j)} y \right\|_2 d\rho(y) \\
&\le \|b_j\|_2 + C_{\mathcal{B}}^{\frac{1}{2}} \sum_{p,q=1}^{m(n)} \left\| A_{p,q}^{(j)} \right\|_F \\
&\le \sqrt{2dC_{\mathcal{B}}} + \sqrt{C_{\mathcal{B}}} m \sqrt{d} \le m\sqrt{2dC_{\mathcal{B}}}.
\end{aligned}$$

Choose $\mathrm{T}_n, \bar{\mathrm{T}}_n \in \mathcal{H}_{\mathrm{T}_n}$ with $\|\alpha - \bar{\alpha}\|_1 \le \epsilon$, $\left(\sum_{p,q=1}^{m(n)} \left\| A_{p,q}^{(j)} - \bar{A}_{p,q}^{(j)} \right\|_F^2 \right)^{\frac{1}{2}} \le \epsilon$, $\|b_j - \bar{b}_j\|_2 \le \epsilon$ for $1 \le j \le n$, and $\|b_0 - \bar{b}_0\|_2 \le \epsilon$. Then we have

$$\begin{aligned}
\left\| \mathrm{T}_n(\rho,x) - \bar{\mathrm{T}}_n(\rho,x) \right\|_2 &= \left\| (b_0 - \bar{b}_0) + \sum_{j=1}^{n} (\alpha_j \sigma_j(\rho,x) - \bar{\alpha}_j \bar{\sigma}_j(\rho,x)) \right\|_2 \\
&\le \|b_0 - \bar{b}_0\|_2 + \left\| \sum_{j=1}^{n} (\alpha_j - \bar{\alpha}_j)\sigma_j(\rho,x) + \bar{\alpha}_j(\sigma_j(\rho,x) - \bar{\sigma}_j(\rho,x)) \right\|_2 \\
&\le \epsilon + \|\alpha - \bar{\alpha}\|_1 \max_{1\le j\le n} \|\sigma_j(\rho,x)\|_2 + \|\bar{\alpha}\|_1 \max_{1\le j\le n} \|\sigma_j(\rho,x) - \bar{\sigma}_j(\rho,x)\|_2 \\
&\le 2\sqrt{dC_{\mathcal{B}}} m\epsilon + 2C_F \max_{1\le j\le n} \|\sigma_j(\rho,x) - \bar{\sigma}_j(\rho,x)\|_2.
\end{aligned}$$

For the second term in the above inequality, it follows that

$$\begin{aligned}
&\|\sigma_j(\rho,x) - \bar{\sigma}_j(\rho,x)\|_2 \\
\le &\left\| (b_j - \bar{b}_j) + \sum_{p,q=1}^{m(n)} \left(\phi_{q,\tilde{m}(n)}(x) \mathcal{T}_{C_{\mathcal{B}}} \left[\int \phi_{p,\tilde{m}(n)}(y) A_{p,q}^{(j)} y \, d\rho(y) \right] \right. \right. \\
&\qquad \left. \left. - \bar{\phi}_{q,\tilde{m}(n)}(x) \mathcal{T}_{C_{\mathcal{B}}} \left[\int \bar{\phi}_{p,\tilde{m}(n)}(y) \bar{A}_{p,q}^{(j)} y \, d\rho(y) \right] \right) \right\|_2 \\
\le &\|b_j - \bar{b}_j\|_2 + \sum_{p,q=1}^{m(n)} \left\| (\phi_{q,\tilde{m}(n)}(x) - \bar{\phi}_{q,\tilde{m}(n)}(x)) \mathcal{T}_{C_{\mathcal{B}}} \left[\int \phi_{p,\tilde{m}(n)}(y) A_{p,q}^{(j)} y \, d\rho(y) \right] \right\|_2 \\
&+ \sum_{p,q=1}^{m(n)} \left\| \bar{\phi}_{q,\tilde{m}(n)}(x) \left(\mathcal{T}_{C_{\mathcal{B}}} \left[\int \phi_{p,\tilde{m}(n)}(y) A_{p,q}^{(j)} y \, d\rho(y) \right] - \mathcal{T}_{C_{\mathcal{B}}} \left[\int \bar{\phi}_{p,\tilde{m}(n)}(y) \bar{A}_{p,q}^{(j)} y \, d\rho(y) \right] \right) \right\|_2
\end{aligned}$$

$$
\begin{aligned}
&\leq \|b_j-\bar{b}_j\|_2+\sum_{p,q=1}^{m(n)}\left|\phi_{q,\tilde{m}(n)}(x)-\bar{\phi}_{q,\tilde{m}(n)}(x)\right|\int\left\|A_{p,q}^{(j)}y\right\|_2\,d\rho(y)\\
&\qquad+\sum_{p,q=1}^{m(n)}\left\|\int\left(\phi_{p,\tilde{m}(n)}(y)A_{p,q}^{(j)}y-\bar{\phi}_{p,\tilde{m}(n)}(y)\bar{A}_{p,q}^{(j)}y\right)d\rho(y)\right\|_2\\
&\leq \|b_j-\bar{b}_j\|_2+C_{\mathcal{B}}^{\frac{1}{2}}\sum_{p,q=1}^{m(n)}\left\|A_{p,q}^{(j)}\right\|_F\left|\phi_{q,\tilde{m}(n)}(x)-\bar{\phi}_{q,\tilde{m}(n)}(x)\right|+\sum_{p,q=1}^{m(n)}\left\|\int\phi_{p,\tilde{m}(n)}(y)(A_{p,q}^{(j)}-\bar{A}_{p,q}^{(j)})y\,d\rho(y)\right\|_2\\
&\qquad+\sum_{p,q=1}^{m(n)}\left\|\int(\phi_{p,\tilde{m}(n)}(y)-\bar{\phi}_{p,\tilde{m}(n)}(y))\bar{A}_{p,q}^{(j)}y\,d\rho(y)\right\|_2\\
&\leq \|b_j-\bar{b}_j\|_2+C_{\mathcal{B}}^{\frac{1}{2}}\max_{1\leq q\leq m(n)}\left|\phi_{q,\tilde{m}(n)}(x)-\bar{\phi}_{q,\tilde{m}(n)}(x)\right|\sum_{p,q=1}^{m(n)}\left\|A_{p,q}^{(j)}\right\|_F+C_{\mathcal{B}}^{\frac{1}{2}}\sum_{p,q=1}^{m(n)}\left\|A_{p,q}^{(j)}-\bar{A}_{p,q}^{(j)}\right\|_F\\
&\qquad+\sum_{p,q=1}^{m(n)}\int\left|\phi_{p,\tilde{m}(n)}(y)-\bar{\phi}_{p,\tilde{m}(n)}(y)\right|\left\|\bar{A}_{p,q}^{(j)}y\right\|_2\,d\rho(y)\\
&\leq \epsilon+\sqrt{C_{\mathcal{B}}}m\epsilon+\sqrt{dC_{\mathcal{B}}}m\max_{1\leq q\leq m(n)}\left|\phi_{q,\tilde{m}(n)}(x)-\bar{\phi}_{q,\tilde{m}(n)}(x)\right|\\
&\qquad+\sqrt{dC_{\mathcal{B}}}m\max_{1\leq p\leq m(n)}\|\phi_{p,\tilde{m}(n)}-\bar{\phi}_{p,\tilde{m}(n)}\|_{L^2(\rho)}.
\end{aligned}
$$

Recall that the above two-hidden-layer tanh neural networks have the form $\phi_{p,\tilde{m}(n)}=\mathcal{T}_{1,\odot}\big(\phi_{p,\tilde{m}(n)}^{(1)},\ldots,\phi_{p,\tilde{m}(n)}^{(d)}\big)$ and then it can be obtained that for any $x\in\mathbb{R}^d$,

$$
\begin{aligned}
&\left|\phi_{p,\tilde{m}(n)}(x)-\bar{\phi}_{p,\tilde{m}(n)}(x)\right|\leq d\max_{1\leq l\leq d}\left|\phi_{p,\tilde{m}(n)}^{(l)}(x)-\bar{\phi}_{p,\tilde{m}(n)}^{(l)}(x)\right|\\
&\leq d\max_{1\leq l\leq d}\Big|c_{p,l}^T\underbrace{\left[\sigma_{\tanh}(W_{p,1}^{(l)}\sigma_{\tanh}(W_{p,0}^{(l)}x+b_{p,0}^{(l)})+b_{p,1}^{(l)})\right]}_{=:f_p^{(l)}(x)}\\
&\qquad-\bar{c}_{p,l}^T\underbrace{\left[\sigma_{\tanh}(\overline{W}_{p,1}^{(l)}\sigma_{\tanh}(\overline{W}_{p,0}^{(l)}x+\bar{b}_{p,0}^{(l)})+\bar{b}_{p,1}^{(l)})\right]}_{=:\bar{f}_p^{(l)}(x)}\Big|\\
&\leq d\max_{1\leq l\leq d}\left|c_{p,l}^Tf_p^{(l)}(x)-\bar{c}_{p,l}^Tf_p^{(l)}(x)+\bar{c}_{p,l}^Tf_p^{(l)}(x)-\bar{c}_{p,l}^T\bar{f}_p^{(l)}(x)\right|\\
&\leq d\max_{1\leq l\leq d}\left(\|c_{p,l}-\bar{c}_{p,l}\|_1+\|\bar{c}_{p,l}\|_1\left\|f_p^{(l)}(x)-\bar{f}_p^{(l)}(x)\right\|_\infty\right)\\
&\leq 8d\,\tilde{m}(n)\max_{1\leq l\leq d}\left\{\|c_{p,l}-\bar{c}_{p,l}\|_\infty+\|\bar{c}_{p,l}\|_\infty\left\|f_p^{(l)}(x)-\bar{f}_p^{(l)}(x)\right\|_\infty\right\},
\end{aligned}
$$

where

$$\begin{aligned}
&\left\|f_p^{(l)}(x)-\bar{f}_p^{(l)}(x)\right\|_\infty \\
\leq&\left\|(W_{p,1}^{(l)}\sigma_{\tanh}(W_{p,0}^{(l)}x+b_{p,0}^{(l)})+b_{p,1}^{(l)})-(\overline{W}_{p,1}^{(l)}\sigma_{\tanh}(\overline{W}_{p,0}^{(l)}x+\bar{b}_{p,0}^{(l)})+\bar{b}_{p,1}^{(l)})\right\|_\infty \\
\leq&\left\|b_{p,1}^{(l)}-\bar{b}_{p,1}^{(l)}\right\|_\infty+\left\|W_{p,1}^{(l)}-\overline{W}_{p,1}^{(l)}\right\|_\infty+\left\|W_{p,1}^{(l)}\right\|_\infty\left\|(W_{p,0}^{(l)}-\overline{W}_{p,0}^{(l)})x+b_{p,0}^{(l)}-\bar{b}_{p,0}^{(l)}\right\|_\infty.
\end{aligned}$$

Choose that

$$\begin{aligned}
&\|c_{p,l}-\bar{c}_{p,l}\|_\infty\leq\epsilon,\left\|b_{p,1}^{(l)}-\bar{b}_{p,1}^{(l)}\right\|_\infty\leq\epsilon,\left\|W_{p,1}^{(l)}-\overline{W}_{p,1}^{(l)}\right\|_{\max}\leq\epsilon,\\
&\left\|W_{p,0}^{(l)}-\overline{W}_{p,0}^{(l)}\right\|_{\max}\leq\epsilon,\left\|b_{p,0}^{(l)}-\bar{b}_{p,0}^{(l)}\right\|_\infty\leq\epsilon.
\end{aligned}\tag{4.23}$$

Then it's easy to see that

$$\left\|f_p^{(l)}(x)-\bar{f}_p^{(l)}(x)\right\|_\infty\leq 24(1+\|x\|_\infty)[c_4\tilde{m}(n)]^{c_5(\tilde{m}(n))^2}\epsilon,$$

which implies that

$$\left|\phi_{p,\tilde{m}(n)}(x)-\bar{\phi}_{p,\tilde{m}(n)}(x)\right|\leq 192d\,(2+\|x\|_\infty)[c_4\tilde{m}(n)]^{3c_5(\tilde{m}(n))^2}\epsilon$$

and that

$$\begin{aligned}
&\sqrt{dC_{\mathcal{B}}}m\Big(\max_{1\leq p,q\leq m(n)}\left|\phi_{q,\tilde{m}}(x)-\bar{\phi}_{q,\tilde{m}}(x)\right|+\|\phi_{p,\tilde{m}}-\bar{\phi}_{p,\tilde{m}}\|_{L^2(\rho)}\Big)\\
\leq&\sqrt{dC_{\mathcal{B}}}m\Big(192d\left(4+\|x\|_\infty+\int_{\mathcal{X}}\|x\|_\infty\,d\rho\right)[c_4\tilde{m}(n)]^{3c_5(\tilde{m}(n))^2}\Big)\epsilon\\
\leq&192C_{\mathcal{B}}d^{\frac{3}{2}}(5+\|x\|_\infty)(m(n))[c_4\tilde{m}(n)]^{3c_5(\tilde{m}(n))^2}\epsilon.
\end{aligned}$$

We can conclude that

$$\left\|\mathrm{T}_n(\rho,x)-\bar{\mathrm{T}}_n(\rho,x)\right\|_2\leq 386C_FC_{\mathcal{B}}d^{\frac{3}{2}}(9+\|x\|_\infty)(m(n))[c_4\tilde{m}(n)]^{3c_5(\tilde{m}(n))^2}\epsilon=:\Xi(x,\epsilon).$$

For any pseudometric space $(\mathcal{H},d_{\mathcal{H}})$ and $\epsilon>0$, the covering number of $(\mathcal{H},d_{\mathcal{H}})$ with radius ϵ is defined as $\inf\{|\mathcal{D}|:\mathcal{D}\subset\mathcal{H},\text{for any }\Psi\in\mathcal{H}\text{ there exists }\Phi\in\mathcal{D}\text{ with }d_{\mathcal{H}}(\Psi,\Phi)\leq\epsilon\}$.

We denote by $\mathrm{N}(\mathcal{H}_{\mathrm{T}_n}, \epsilon, d_{\mathcal{B}_{2,b}(\mathcal{X})})$ the uniform ϵ-covering number for the first stage sampling with the pseudometric

$$d_{\mathcal{B}_{2,b}(\mathcal{X})}(\Phi_1, \Phi_2) = \sup_{\rho \in \mathcal{B}_{2,b}(\mathcal{X})} \mathbb{E}_\rho \|\Phi_1(\tilde{X}) - \Phi_2(\tilde{X})\|_2.$$

It follows that for the above $\mathrm{T}_n, \bar{\mathrm{T}}_n$ and the parameter selections,

$$d_{\mathcal{B}_{2,b}(\mathcal{X})}(\mathrm{T}_n, \bar{\mathrm{T}}_n) \leq \sup_{\rho \in \mathcal{B}_{2,b}(\mathcal{X})} \mathbb{E}_\rho \Xi(X, \epsilon) \leq c_6 C_F C_{\mathcal{B}}^2 d^{\frac{3}{2}}(m(n))[c_4 \tilde{m}(n)]^{3c_5(\tilde{m}(n))^2} \epsilon.$$

Then the $\tilde{\epsilon}$-covering number of $\mathcal{H}_{\mathrm{T}_n}$ with respect to $d_{\mathcal{B}_{2,b}(\mathcal{X})}$ can bounded as

$$\begin{aligned}
&\mathrm{N}(\mathcal{H}_{\mathrm{T}_n}, \tilde{\epsilon}, d_{\mathcal{B}_{2,b}(\mathcal{X})}) \\
&\leq \left(1 + \frac{4C_F}{\epsilon}\right)^n \left(1 + \frac{2C_F\sqrt{2dC_{\mathcal{B}}}}{\epsilon}\right)^{n\left(m^2d^2+d\right)+1} \left(1 + \frac{(c_2'\tilde{m})^{c_3'\tilde{m}^2}}{\epsilon}\right)^{[8\tilde{m}d+8\tilde{m}+(8\tilde{m})^2+8\tilde{m}+8\tilde{m}]dm} \\
&\leq \left(1 + \frac{4C_F\sqrt{dC_{\mathcal{B}}}}{\epsilon}\right)^{3nm^2d^2} \left(1 + \frac{(c_2'\tilde{m})^{c_3'\tilde{m}^2}}{\epsilon}\right)^{96d^2\tilde{m}^2m} \\
&\leq \left(1 + \frac{4c_6C_FC_{\mathcal{B}}^3d^2m(c_4\tilde{m})^{3c_5\tilde{m}^2}}{\tilde{\epsilon}}\right)^{3d^2nm^2} \left(1 + \frac{c_6C_FC_{\mathcal{B}}^2d^{\frac{3}{2}}m(c_4\tilde{m})^{6c_5\tilde{m}^2}}{\tilde{\epsilon}}\right)^{96d^2\tilde{m}^2m} \\
&\leq \left(1 + \frac{8C_{d,F,\mathcal{B}}}{\tilde{\epsilon}}\right)^{100d^2nm^2} ((c_4m\tilde{m})^{600c_5d^2nm^2\tilde{m}^2}),
\end{aligned}$$

with $C_{d,F,\mathcal{B}} = c_6C_FC_{\mathcal{B}}^3d^2$, which is followed by

$$\log \mathrm{N}(\mathcal{H}_{\mathrm{T}_n}, \tilde{\epsilon}, d_{\mathcal{B}_{2,b}(\mathcal{X})}) \leq 100nd^2m^2 \log\left(1 + \frac{8C_{d,F,\mathcal{B}}}{\tilde{\epsilon}}\right) + 600c_5d^2nm^2\tilde{m}^2 \log(c_4m\tilde{m}).$$

Conditioned on the pseudo second-stage samples $(\rho_X^{(i)}, X_{ij})_{1\leq i\leq N, 1\leq j\leq n_i}$, we can define the empirical $L^2_{\hat{P}}$ ϵ-covering number $\mathrm{N}(\mathcal{H}_{\mathrm{T}_n}, \epsilon, d_{\hat{P}})$ with the pseudometric

$$d_{\hat{P}}(\Phi_1, \Phi_2) = \left(\frac{1}{N}\sum_{i=1}^N \frac{1}{n_i}\sum_{j=1}^{n_i} \|\Phi_1(\rho_X^{(i)}, X_{ij}) - \Phi_2(\rho_X^{(i)}, X_{ij})\|_2^2\right)^{\frac{1}{2}}.$$

Then for the above $\mathrm{T}_n, \bar{\mathrm{T}}_n$ and the parameter selections, we have

$$\begin{aligned} d_{\hat{P}}(\mathrm{T}_n, \bar{\mathrm{T}}_n) &\leq \left(\frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \Xi(X_{ij}, \epsilon)^2 \right)^{\frac{1}{2}} \\ &\leq 386\sqrt{2} C_F C_{\mathcal{B}} d^{\frac{3}{2}} \left(9 + \sqrt{\frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \|X_{ij}\|_\infty^2} \right) m[c_4 \tilde{m}]^{3c_5 \tilde{m}^2} \epsilon \\ &\leq c_6 C_F C_{\mathcal{B}} d^{\frac{3}{2}} \underbrace{\left(9 + \sqrt{\frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \|X_{ij}\|_2^2} \right)}_{=:\|\hat{X}\|_2} m(c_4 \tilde{m})^{3c_5 \tilde{m}^2} \epsilon. \end{aligned}$$

Similarly, it's easy to see that

$$\log \mathrm{N}(\mathcal{H}_{\mathrm{T}_n}, \tilde{\epsilon}, d_{\hat{P}}) \leq 100 n d^2 m^2 \log \left(1 + \frac{8 c_6 C_F C_{\mathcal{B}}^2 d^2 \|\hat{X}\|_2}{\tilde{\epsilon}} \right) + 600 c_5 n d^2 m^2 \tilde{m}^2 \log(c_4 m \tilde{m}).$$

4.5.2.3 First-Stage Sampling Error Estimation

The following lemma is from Lemma 3.19 [10].

LEMMA 11. *Let $\mathscr{J}$ be a set of functions on Z and $\mathscr{B}, c > 0$ such that for each $\mathcal{J} \in \mathscr{J}$, $|\mathcal{J} - \mathbb{E}(\mathcal{J})| \leq \mathscr{B}$ and $\mathbb{E}(\Gamma^2) \leq c\mathbb{E}(\Gamma)$ almost surely. Then for every $\epsilon > 0$ and $0 < r \leq 1$,*

$$\mathbb{P}_{\mathbf{z} \in Z^m} \left\{ \sup_{\mathcal{J} \in \mathscr{J}} \frac{\mathbb{E}(\mathcal{J}) - \mathbb{E}_{\mathbf{z}}(\mathcal{J})}{\sqrt{\mathbb{E}(\mathcal{J}) + \epsilon}} > 4r\sqrt{\epsilon} \right\} \leq \mathrm{N}(\mathscr{J}, r\epsilon, \|\cdot\|_{L^\infty(Z)}) \exp\left\{ -\frac{r^2 m \epsilon}{2c + \frac{2}{3}\mathscr{B}} \right\}.$$

We consider the class of functions $\mathcal{J}_\Phi : \mathcal{B}_{2,b}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$, denoted by

$$\mathscr{J}(\mathcal{H}_{\mathrm{T}_n}) := \left\{ \mathcal{J}_\Phi : \mathcal{J}_\Phi(\rho_{XY}) = \mathbb{E}[\|\mathscr{T}_M \Phi(\tilde{X}) - Y\|_2^2 | \rho_{XY}] - \mathbb{E}[\|\Phi_{\mathcal{G}}(\tilde{X}) - Y\|_2^2 | \rho_{XY}], \Phi \in \mathcal{H}_{\mathrm{T}_n} \right\}.$$

Then for each $\mathcal{J}_\Phi \in \mathscr{J}(\mathcal{H}_{\mathrm{T}_n})$, we have

$$\mathbb{E}(\mathcal{J}_\Phi) = \mathcal{E}(\mathscr{T}_M \Phi) - \mathcal{E}(\Phi_{\mathcal{G}}) \text{ and } \frac{1}{N} \sum_{i=1}^{N} \mathcal{J}_\Phi(\rho_{XY}^{(i)}) = \mathcal{E}_N(\mathscr{T}_M \Phi) - \mathcal{E}_N(\Phi_{\mathcal{G}}),$$

where $\mathcal{E}(\mathscr{T}_M\Phi) - \mathcal{E}(\Phi_{\mathcal{G}}) = \|\mathscr{T}_M\Phi - \Phi_{\mathcal{G}}\|^2_{L^2_{\nu_{\mathcal{G}}}}$. Also notice that

$$\begin{aligned}
|\mathcal{J}_\Phi(\rho_{XY})| &= \left|\mathbb{E}\big[\|\mathscr{T}_M\Phi(\tilde{X}) - Y\|_2^2 - \|\Phi_{\mathcal{G}}(\tilde{X}) - Y\|_2^2\big|\rho_{XY}\big]\right| \\
&\le \left|\mathbb{E}\big[(\|\mathscr{T}_M\Phi(\tilde{X}) - Y\|_2 + \|\Phi_{\mathcal{G}}(\tilde{X}) - Y\|_2)(\|\mathscr{T}_M\Phi(\tilde{X}) - \Phi_{\mathcal{G}}\|_2)\big|\rho_{XY}\big]\right| \\
&\le 8M^2,
\end{aligned}$$

which implies that $|\mathcal{J}_\Phi(\rho_{XY}) - \mathbb{E}(\mathcal{J}_\Phi)| \le 16M^2$ and

$$\begin{aligned}
\mathbb{E}(\mathcal{J}_\Phi^2) &\le \mathbb{E}_{\rho_{XY}\sim\mathcal{P}_{\mathcal{G}}}\Big[\mathbb{E}\big[(\|\mathscr{T}_M\Phi(\tilde{X}) - Y\|_2^2 - \|\Phi_{\mathcal{G}}(\tilde{X}) - Y\|_2^2)^2|\rho_{XY}\big]\Big] \\
&\le 16M^2\mathbb{E}_{\rho_X\sim\mathcal{P}^X_{\mathcal{G}}}(\mathbb{E}[\|\mathscr{T}_M\Phi(\tilde{X}) - \Phi_{\mathcal{G}}\|_2|\rho_X])^2 \le 16M^2\|\mathscr{T}_M\Phi - \Phi_{\mathcal{G}}\|^2_{L^2_\nu} = 16M^2\mathbb{E}(\mathcal{J}_\Phi).
\end{aligned}$$

Then for any $\Phi_1, \Phi_2 \in \mathcal{H}_{\mathrm{T}_n}$, it follows that

$$\begin{aligned}
|\mathcal{J}_{\Phi_1}(\rho_{XY}) - \mathcal{J}_{\Phi_2}(\rho_{XY})| &= \left|\mathbb{E}[\|\mathscr{T}_M\Phi_1(\tilde{X}) - Y\|_2^2|\rho_{XY}] - \mathbb{E}[\|\mathscr{T}_M\Phi_2(\tilde{X}) - Y\|_2^2|\rho_{XY}]\right| \\
&\le 4M\left|\mathbb{E}\big[\|\mathscr{T}_M\Phi_1(\tilde{X}) - \mathscr{T}_M\Phi_2(\tilde{X})\|_2|\rho_X\big]\right| \\
&\le 4M\left|\mathbb{E}\big[\|\Phi_1(\tilde{X}) - \Phi_2(\tilde{X})\|_2|\rho_X\big]\right| \le 4Md_{\mathcal{B}_{2,b}(\mathcal{X})}(\Phi_1, \Phi_2),
\end{aligned}$$

which shows that

$$\mathrm{N}(\mathscr{J}(\mathcal{H}_{\mathrm{T}_n}), \epsilon, \|\cdot\|_{L^\infty(\mathcal{B}_{2,b}(\mathcal{X}\times\mathcal{Y}))}) \le \mathrm{N}\left(\mathcal{H}_{\mathrm{T}_n}, \frac{\epsilon}{4M}, d_{\mathcal{B}_{2,b}(\mathcal{X})}\right).$$

Then with the uniform ratio inequality, we have that

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\Phi\in\mathcal{H}_{\mathrm{T}_n}} \frac{(\mathcal{E}(\mathscr{T}_M\Phi) - \mathcal{E}(\Phi_{\mathcal{G}})) - (\mathcal{E}_N(\mathscr{T}_M\Phi) - \mathcal{E}_N(\Phi_{\mathcal{G}}))}{\sqrt{\mathcal{E}(\mathscr{T}_M\Phi) - \mathcal{E}(\Phi_{\mathcal{G}}) + \epsilon}} > \sqrt{\epsilon}\right\} \\
&\le \mathrm{N}\left(\mathscr{J}(\mathcal{H}_{\mathrm{T}_n}), \frac{\epsilon}{4}, \|\cdot\|_{L^\infty(\mathcal{B}_{2,b}(\mathcal{X}\times\mathcal{Y}))}\right)\exp\left\{-\frac{3N\epsilon}{2048M^2}\right\} \\
&\le \mathrm{N}\left(\mathcal{H}_{\mathrm{T}_n}, \frac{\epsilon}{16M}, d_{\mathcal{B}_{2,b}(\mathcal{X})}\right)\exp\left\{-\frac{3N\epsilon}{2048M^2}\right\}.
\end{aligned}$$

It's easy to see that $\sqrt{(\mathcal{E}(\mathscr{T}_M\Phi) - \mathcal{E}(\Phi_{\mathcal{G}}) + \epsilon)\epsilon} \le \frac{1}{2}(\mathcal{E}(\mathscr{T}_M\Phi) - \mathcal{E}(\Phi_{\mathcal{G}})) + \epsilon$, which follows by taking $\Phi = \mathrm{T}_{\mathbb{S},n}$ that

$$\begin{aligned}
&\mathbb{P}\left\{\mathcal{E}_1(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) > \frac{1}{2}\big(\mathcal{E}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}(\Phi_{\mathcal{G}})\big) + \epsilon\right\} \\
&\le \mathrm{N}\left(\mathcal{H}_{\mathrm{T}_n}, \frac{\epsilon}{16M}, d_{\mathcal{B}_{2,b}(\mathcal{X})}\right)\exp\left\{-\frac{3N\epsilon}{2048M^2}\right\}.
\end{aligned}$$

Similarly, For each $\Phi \in \mathcal{H}_{\mathrm{T}_n}$, note that $\|\Phi\|_{C(\Omega_{\mathcal{B}})}$ is uniformly bounded. We define

$$\widetilde{\mathcal{J}}_\Phi(\rho_{XY}) = \mathbb{E}[\|\Phi(\tilde{X}) - Y\|_2^2 | \rho_{XY}] - \mathbb{E}[\|\Phi_{\mathcal{G}}(\tilde{X}) - Y\|_2^2 | \rho_{XY}].$$

$\widetilde{\mathcal{J}}_\Phi$ can be considered as a random variable with $\left|\widetilde{\mathcal{J}}_\Phi(\rho_{XY})\right| \leq (3M + \|\Phi\|_{C(\Omega_{\mathcal{B}})})^2$ and it follows that

$$\left|\widetilde{\mathcal{J}}_\Phi(\rho_{XY}) - \mathbb{E}(\widetilde{\mathcal{J}}_\Phi)\right| \leq 2(3M + \|\Phi\|_{C(\Omega_{\mathcal{B}})})^2$$

and

$$\mathbb{E}(\widetilde{\mathcal{J}}_\Phi^2) \leq (3M + \|\Phi\|_{C(\Omega_{\mathcal{B}})})^2 \mathcal{E}_4(\Phi).$$

Then with the Bernstein inequality, we have

$$\mathbb{P}\{\mathcal{E}_1'(\Phi) > \epsilon\} \leq \exp\left\{-\frac{N\epsilon^2}{2(3M + \|\Phi\|_{C(\Omega_{\mathcal{B}})})^2 \left(\mathcal{E}_4(\Phi) + \frac{2}{3}\epsilon\right)}\right\}.$$

4.5.2.4 Second-Stage Sampling Error with Ground Truth Context

Assume that $n_1 = \cdots = n_N = \vartheta$. Conditioned on the given first-stage samples $(\rho_{XY}^{(i)})_{1\leq i\leq N}$, the random variables $(X_{ij}, Y_{ij})_{1\leq i\leq N, 1\leq j\leq \vartheta}$ are independent but not identically distributed: for each $1 \leq i \leq N$, $(X_{ij}, Y_{ij}) \sim \rho_{XY}^{(i)}$. Let $\widetilde{S} = (\rho_X^{(i)}, X_{ij}, Y_{ij})_{1\leq i\leq N, 1\leq j\leq \vartheta}$. We introduce a random variable

$$\Lambda(\widetilde{S}) = \sup_{\bar{\Phi}\in\mathscr{T}_M(\mathcal{H}_{\mathrm{T}_n})} \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}\Big(\mathbb{E}\big[\|\bar{\Phi}(\tilde{X}) - Y\|_2^2 \big| \rho_{XY}^{(i)}\big] - \|\bar{\Phi}(\rho_X^{(i)}, X_{ij}) - Y_{ij}\|_2^2\Big),$$

which follows that

$$\left|\Lambda(\widetilde{S}) - \Lambda(\widetilde{S}^{\setminus(i,j)})\right| \leq \sup_{\bar{\Phi}\in\mathscr{T}_M(\mathcal{H}_{\mathrm{T}_n})} \frac{1}{N\vartheta}\left|\|\bar{\Phi}(\rho_X^{(i)}, X_{ij}) - Y_{ij}\|_2^2 - \|\bar{\Phi}(\rho_X^{(i)}, X_{ij}') - Y_{ij}'\|_2^2\right| \leq \frac{8M^2}{N\vartheta}$$

where $\widetilde{S}^{\setminus(i,j)}$ denotes the sample $\widetilde{S}$ with a change on (i,j)-th variable with $(X_{ij}, Y_{ij}) \overset{i.i.d}{\sim} (X_{ij}', Y_{ij}')$ while all others fixed. Then by Azuma-McDiarmind's inequality, it can be derived that

$$\mathbb{P}\left\{\left|\Lambda(\widetilde{S}) - \mathbb{E}\big[\Lambda(\widetilde{S})|(\rho_{XY}^{(i)})_i\big]\right| > \epsilon\right\} \leq 2\exp\left\{-\frac{(N\vartheta)\,\epsilon^2}{32M^4}\right\}, \text{ for any } \epsilon > 0.$$

Let $(\zeta_{ij})_{1\le i\le N,1\le j\le \vartheta}$ be i.i.d Rademacher random variables. Then by the symmetrization [44], we have

$$\mathbb{E}\big[\Lambda(\widetilde{S})\big|(\rho_{XY}^{(i)})_i\big] \le \mathbb{E}_{\left[(X_{ij},Y_{ij})\sim\rho_{XY}^{(i)}\right]_i}\mathbb{E}_{(\zeta_{ij})}\sup_{\bar{\Phi}\in\mathscr{T}_M(\mathcal{H}_{\mathrm{T}_n})}\frac{2}{N}\sum_{i=1}^{N}\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}\zeta_{ij}\|\bar{\Phi}(\rho_X^{(i)},X_{ij})-Y_{ij}\|_2^2.$$

Since $|l_y(u)-l_y(u')|\le 4M\|u-u'\|_2$ where $l_y(u):=\|u-y\|_2^2$ with $\|u\|_2$ and $\|y\|_2$ less than M, it follows by the vector-contraction inequality [58] that

$$\mathbb{E}[\Lambda(\widetilde{S})|(\rho_{XY}^{(i)})_i] \le 8\sqrt{2}M\mathbb{E}_{\left[X_{ij}\sim\rho_X^{(i)}\right]_i}\mathbb{E}_{(\boldsymbol{\zeta}_{ij})}\sup_{\bar{\Phi}\in\mathscr{T}_M(\mathcal{H}_{\mathrm{T}_n})}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}\langle\boldsymbol{\zeta}_{ij},\bar{\Phi}(\rho_X^{(i)},X_{ij})\rangle$$

where $\boldsymbol{\zeta}_{ij}$ is a random vector in $\mathbb{R}^d$ with each component being i.i.d Rademacher random variable.

We define a family of zero-mean random variables indexed by $\bar{\Phi}\in\mathscr{T}_M(\mathcal{H}_{\mathrm{T}_n})$ as

$$Z_{\bar{\Phi}} := \frac{1}{\sqrt{N\vartheta}}\sum_{i=1}^{N}\sum_{j=1}^{\vartheta}\langle\boldsymbol{\zeta}_{ij},\bar{\Phi}(\rho_X^{(i)},X_{ij})\rangle,$$

which implies that

$$\mathbb{E}[\Lambda(\widetilde{S})|(\rho_{XY}^{(i)})_i] \le 8\sqrt{2}M\mathbb{E}_{\left[X_{ij}\sim\rho_X^{(i)}\right]_i}\mathbb{E}\left[\sup_{\bar{\Phi}\in\mathscr{T}_M(\mathcal{H}_{\mathrm{T}_n})}\frac{1}{\sqrt{N\vartheta}}Z_{\bar{\Phi}}\middle|(X_{ij})_{i,j}\right].$$

For any $\bar{\Phi},\bar{\Phi}'\in\mathscr{T}_M(\mathcal{H}_{\mathrm{T}_n})$,

$$\mathbb{E}[\exp(v(Z_{\bar{\Phi}}-Z_{\bar{\Phi}'}))|(X_{ij})_{i,j}] \le \exp(v^2 d_\vartheta(\bar{\Phi},\bar{\Phi}')^2/2),\quad \forall v\in\mathbb{R}$$

where

$$d_\vartheta(\bar{\Phi},\bar{\Phi}') := \|\bar{\Phi}-\bar{\Phi}'\|_{L^2(P_\vartheta)} = \left(\frac{1}{N\vartheta}\sum_{i=1}^{N}\sum_{j=1}^{\vartheta}\|\bar{\Phi}(\rho_X^{(i)},X_{ij})-\bar{\Phi}'(\rho_X^{(i)},X_{ij})\|_2^2\right)^{\frac{1}{2}}.$$

Then, conditioned on $(X_{ij})_{i,j}$, $Z_{\bar{\Phi}}$ is a subgaussian process indexed by $\bar{\Phi}\in\mathscr{T}_M(\mathcal{H}_{\mathrm{T}_n})$ with respect to d_ϑ. It follows by the Dudley Integral [103] that

$$\mathbb{E}\left[\sup_{\bar{\Phi}\in\mathscr{T}_M(\mathcal{H}_{\mathrm{T}_n})}Z_{\bar{\Phi}}\middle|(X_{ij})_{i,j}\right] \le 32\int_0^{2M}\sqrt{\log \mathrm{N}(u,\mathscr{T}_M(\mathcal{H}_{\mathrm{T}_n}),d_\vartheta)}\,du.$$

It follows that

$$\begin{aligned}
&\mathbb{E}[\Lambda(\widetilde{S})|(\rho_{XY}^{(i)})_i] \leq \frac{256\sqrt{2}M}{\sqrt{N\vartheta}}\mathbb{E}_{[X_{ij}\sim\rho_X^{(i)}]_i}\left(\int_0^{2M}\sqrt{\log \mathrm{N}(u,\mathcal{H}_{\mathrm{T}_n},d_{\hat{\rho}_\vartheta})}\,du\right)\\
\leq&\frac{256\sqrt{2}M}{\sqrt{N\vartheta}}\mathbb{E}_{[X_{ij}\sim\rho_X^{(i)}]_i}\int_0^{2M}\sqrt{\underbrace{100nd^2m^2}_{\varrho_1}\log\left(1+\frac{8c_6C_FC_{\mathcal{B}}^2d^2\|\hat{X}\|_2}{u}\right)+\underbrace{600c_5nd^2m^2\tilde{m}^2\log(c_4m\tilde{m})}_{\varrho_2}}\,du\\
&\leq \frac{256\sqrt{2}M}{\sqrt{N\vartheta}}\mathbb{E}_{[X_{ij}\sim\rho_X^{(i)}]_i}\left(2M\sqrt{\varrho_2}+\sqrt{\varrho_1}\int_0^{2M}\sqrt{\log\left(1+\frac{8c_6C_FC_{\mathcal{B}}^2d^2\|\hat{X}\|_2}{u}\right)}\,du\right)\\
&\leq \frac{256\sqrt{2}M}{\sqrt{N\vartheta}}\mathbb{E}_{[X_{ij}\sim\rho_X^{(i)}]_i}\left(2M\sqrt{\varrho_2}+\sqrt{\varrho_1}\sqrt{8c_6C_FC_{\mathcal{B}}^2d^2\|\hat{X}\|_2}\int_0^{2M}u^{-\frac{1}{2}}\,du\right)\\
&= \frac{256\sqrt{2}M}{\sqrt{N\vartheta}}\left(2M\sqrt{\varrho_2}+8\sqrt{\varrho_1c_6C_FC_{\mathcal{B}}^2d^2M}\;\mathbb{E}_{[X_{ij}\sim\rho_X^{(i)}]_i}\sqrt{\|\hat{X}\|_2}\right)
\end{aligned}$$

where

$$\mathbb{E}_{[X_{ij}\sim\rho_X^{(i)}]_i}\sqrt{\|\hat{X}\|_2}\leq\sqrt{\mathbb{E}_{[X_{ij}\sim\rho_X^{(i)}]_i}\|\hat{X}\|_2}=\sqrt{9+\mathbb{E}\sqrt{\frac{1}{N\vartheta}\sum_{i=1}^{N}\sum_{j=1}^{\vartheta}\|X_{ij}\|_2^2}}\leq 9+C_{\mathcal{B}}^{\frac{1}{2}}.$$

Then we can obtain that

$$\mathbb{E}[\Lambda(\widetilde{S})|(\rho_{XY}^{(i)})_i]\leq\frac{C_{M,\mathcal{B},d}}{\sqrt{N\vartheta}}\sqrt{n}m\tilde{m}^2$$

with $C_{M,\mathcal{B},d}=256\sqrt{2}M\max\left\{20dM\sqrt{6c_5},80d^2(9+C_{\mathcal{B}}^{1/2})\sqrt{c_6C_FC_{\mathcal{B}}^2M}\right\}$.

Then we have

$$\mathbb{P}\left\{\sup_{\Phi\in\mathcal{H}_{\mathrm{T}_n}}\mathcal{E}_2(\mathscr{T}_M(\Phi))>\epsilon+C_{M,\mathcal{B},d}\frac{\sqrt{n}m\tilde{m}^2}{\sqrt{N\vartheta}}\middle|\rho_{XY}^{(1)},\ldots,\rho_{XY}^{(N)}\right\}\leq\exp\left(-\frac{(N\vartheta)\epsilon^2}{32M^4}\right).$$

Similarly, for each $\Phi\in\mathcal{H}_{\mathrm{T}_n}$, we define

$$\tilde{\Lambda}_\Phi(\widetilde{S})=\frac{1}{N\vartheta}\sum_{i=1}^{N}\sum_{j=1}^{\vartheta}\left(\|\Phi(\rho_X^{(i)},X_{ij})-Y_{ij}\|_2^2-\mathbb{E}\big[\|\Phi(\tilde{X})-Y\|_2^2|\rho_{XY}^{(i)}\big]\right).$$

Similarly, we have

$$\left|\tilde{\Lambda}_\Phi(\widetilde{S})-\tilde{\Lambda}_\Phi(\widetilde{S}^{\setminus(i,j)})\right|\leq\frac{1}{N\vartheta}\left|\|\Phi(\rho_X^{(i)},X_{ij})-Y_{ij}\|_2^2-\|\Phi(\rho_X^{(i)},X'_{ij})-Y'_{ij}\|_2^2\right|\leq\frac{4(M+\|\Phi\|_{C(\Omega_{\mathcal{B}})})^2}{N\vartheta}.$$

Then by Azuma-McDiarmind's inequality, we can derive that

$$\mathbb{P}\{\mathcal{E}_2'(\Phi) > \epsilon | \rho_{XY}^{(1)}, \ldots, \rho_{XY}^{(N)}\} \leq \exp\left\{-\frac{(N\vartheta)\epsilon^2}{8(M + \|\Phi\|_{C(\Omega_{\mathcal{B}})})^4}\right\}.$$

4.5.2.5 Second-Stage Sampling Error with Accessible Context

Now, we estimate the sampling error with accessible context information, i.e., the empirical distributions. We have

$$\begin{aligned}
&\sup_{\Phi\in\mathcal{H}_{\mathrm{T}_n}} |\mathcal{E}_3(\mathscr{T}_M(\Phi))| \\
&= \sup_{\Phi\in\mathcal{H}_{\mathrm{T}_n}} \left| \frac{1}{N\vartheta} \sum_{i=1}^{N} \sum_{j=1}^{\vartheta} \left(\|\mathscr{T}_M(\Phi)(\rho_X^{(i)}, X_{ij}) - Y_{ij}\|_2^2 - \|\mathscr{T}_M(\Phi)(\hat{\rho}_X^{(i)}, X_{ij}) - Y_{ij}\|_2^2 \right) \right| \\
&\leq \sup_{\Phi\in\mathcal{H}_{\mathrm{T}_n}} \frac{1}{N\vartheta} \sum_{i=1}^{N} \sum_{j=1}^{\vartheta} 4M \left\| \Phi(\rho_X^{(i)}, X_{ij}) - \Phi(\hat{\rho}_X^{(i)}, X_{ij}) \right\|_2 \\
&\leq \sup_{\Phi\in\mathcal{H}_{\mathrm{T}_n}} \frac{4M}{N} \sum_{i=1}^{N} \left(\|\alpha\|_1 \max_{1\leq j'\leq n} \sum_{p,q=1}^{m(n)} \left\| \int_{\mathcal{X}} \phi_{p,\tilde{m}(n)}(x) A_{p,q}^{(j')} x \, d\rho_X^{(i)} - \int_{\mathcal{X}} \phi_{p,\tilde{m}(n)}(x) A_{p,q}^{(j')} x \, d\hat{\rho}_X^{(i)} \right\|_2 \right) \\
&\leq \sup_{\phi\in\mathcal{NN}(\Theta_{\tilde{m}})} \frac{8MC_F}{N} \sum_{i=1}^{N} m\sqrt{d} \max_{1\leq p\leq m} \left\| \int_{\mathcal{X}} \phi_{p,\tilde{m}(n)}(x) x \, d\rho_X^{(i)} - \int_{\mathcal{X}} \phi_{p,\tilde{m}(n)}(x) x \, d\hat{\rho}_X^{(i)} \right\|_2 \\
&\leq \frac{8MC_F\sqrt{d}m}{N} \sum_{i=1}^{N} \underbrace{\sup_{\phi\in\mathcal{NN}(\Theta_{\tilde{m}})} \left\| \int_{\mathcal{X}} \phi(x) x \, d\rho_X^{(i)} - \int_{\mathcal{X}} \phi(x) x \, d\hat{\rho}_X^{(i)} \right\|_2}_{=:\mathcal{V}^{(i)}(\hat{\rho}_X^{(i)})} \\
&\leq 8MC_F\sqrt{d}m \max_{1\leq i\leq N} \mathcal{V}^{(i)}(\hat{\rho}_X^{(i)}).
\end{aligned}$$

Then for $x = (x_1, \ldots, x_\vartheta) \in \mathcal{X}^\vartheta$ and the samples $(X_{i1}, \ldots, X_{i\vartheta})$ in $\hat{\rho}_X^{(i)}$, we define

$$\begin{aligned}
\mathcal{V}_j^{(i)}(\hat{\rho}_X^{(i)})(x) = \mathcal{V}^{(i)}(&\delta[x_1, \ldots, x_{j-1}, X_{ij}, x_{j+1}, \ldots, x_\vartheta]) \\
&- \mathbb{E}_{X_{ij}'\sim\rho_X^{(i)}}(\mathcal{V}^{(i)}(\delta[x_1, \ldots, x_{j-1}, X_{ij}', x_{j+1}, \ldots, x_\vartheta]))
\end{aligned}$$

for $1 \leq j \leq \vartheta$ where $\delta[S]$ is defined as the empirical distribution generated by the dataset S.

We easily obtain that

$$\left| \mathcal{V}_j^{(i)}(\hat{\rho}_X^{(i)})(x) \right|$$

$$
\begin{aligned}
&= \left|\mathbb{E}_{X'_{ij}\sim\rho_X^{(i)}}\Big(\mathcal{V}^{(i)}(\delta[x_1,\ldots,x_{j-1},X_{ij},x_{j+1},\ldots,x_\vartheta]) - \mathcal{V}^{(i)}(\delta[x_1,\ldots,x_{j-1},X'_{ij},x_{j+1},\ldots,x_\vartheta])\Big)\right| \\
&\le \mathbb{E}_{X'_{ij}\sim\rho_X^{(i)}}\left|\sup_{\phi\in\mathcal{NN}(\Theta_{\tilde{m}})}\left\|\frac{1}{\vartheta}\left(\sum_{j'\neq j}\phi(x_{j'})x_{j'}+\phi(X_{ij})X_{ij}\right)-\mathbb{E}_{\rho_X^{(i)}}(\phi(X)X)\right\|_2\right. \\
&\qquad\left.-\sup_{\phi\in\mathcal{NN}(\Theta_{\tilde{m}})}\left\|\frac{1}{\vartheta}\left(\sum_{j'\neq j}\phi(x_{j'})x_{j'}+\phi(X'_{ij})X'_{ij}\right)-\mathbb{E}_{\rho_X^{(i)}}(\phi(X)X)\right\|_2\right| \\
&\le \mathbb{E}_{X'_{ij}\sim\rho_X^{(i)}}\left(\sup_{\phi\in\mathcal{NN}(\Theta_{\tilde{m}})}\frac{1}{\vartheta}\left\|\phi(X_{ij})X_{ij}-\phi(X'_{ij})X'_{ij}\right\|_2\right) \\
&\le \mathbb{E}_{X'_{ij}\sim\rho_X^{(i)}}\left(\sup_{\phi\in\mathcal{NN}(\Theta_{\tilde{m}})}\frac{1}{\vartheta}(\|\phi(X_{ij})X_{ij}\|_2+\|\phi(X'_{ij})X'_{ij}\|_2)\right) \\
&\le \frac{1}{\vartheta}(\|X_{ij}\|_2+\mathbb{E}_{X'_{ij}\sim\rho_X^{(i)}}\|X'_{ij}\|_2) \le \frac{1}{\vartheta}(\|X_{ij}\|_2+\sqrt{C_{\mathcal{B}}}).
\end{aligned}
$$

For each $\rho_X^{(i)}\in\mathcal{B}_{2,b}(\mathcal{X})$, we have the ratio condition that $\|\omega_\kappa(\rho_X^{(i)})\|_{L^\gamma(\rho_\kappa)}<\infty$. Then for $X_{ij}\sim\rho_X^{(i)}$ and any $p\ge 1$, we have

$$
\begin{aligned}
\left(\mathbb{E}_{\rho_X^{(i)}}\|X_{ij}\|_2^p\right)^{\frac{1}{p}} &= \left(\int_{\mathcal{X}}\|x\|_2^p\cdot\omega_\kappa(\rho_X^{(i)})(x)\,d\rho_\kappa\right)^{\frac{1}{p}} \le \|\omega_\kappa(\rho_X^{(i)})\|_{L^\gamma(\rho_\kappa)}^{\frac{1}{p}}(\mathbb{E}_{\rho_\kappa}\|X\|_2^{p\gamma'})^{\frac{1}{p\gamma'}} \\
&\le (1+\|\omega_\kappa(\rho_X^{(i)})\|_{L^\gamma(\rho_\kappa)})(\mathbb{E}_{\rho_\kappa}\|X\|_2^{p\gamma'})^{\frac{1}{p\gamma'}}
\end{aligned}
$$

with $\gamma'=\frac{\gamma}{\gamma-1}$. For $X\sim\rho_\kappa$, $\|X\|_2$ is a norm subgaussian random variable [34] such that $(\mathbb{E}_{\rho_\kappa}\|X\|_2^p)^{\frac{1}{p}}\le c\kappa\sqrt{dp}$ for any $p\ge 1$ where c is an absolute constant. Then we have

$$
\left(\mathbb{E}_{\rho_X^{(i)}}\|X_{ij}\|_2^p\right)^{\frac{1}{p}} \le c\kappa\sqrt{\gamma'}\big(1+\left\|\omega_\kappa(\rho_X^{(i)})\right\|_{L^\gamma(\rho_\kappa)}\big)\sqrt{p}
$$

for any $p\ge 1$. We define the subgaussian norm [103] for a random variable Z as $\|Z\|_{\psi_2}=\sup_{p\ge1}\frac{(\mathbb{E}|Z|^p)^{\frac{1}{p}}}{\sqrt{p}}$. Then $\|X_{ij}\|_2$ is a subgaussian random variable with $\|\|X_{ij}\|_2\|_{\psi_2}\le c\kappa\sqrt{\gamma'}(1+\|\omega_\kappa(\rho_X^{(i)})\|_{L^\gamma(\rho_\kappa)})$. It also implies that for any $x\in X^\vartheta$, $\mathcal{V}_j^{(i)}(\hat{\rho}_X^{(i)})(x)$ is also a subgaussian random variable with

$$
\left\|\mathcal{V}_j^{(i)}(\hat{\rho}_X^{(i)})(x)\right\|_{\psi_2}\le\frac{1}{\vartheta}\Big(\|\|X_{ij}\|_2\|_{\psi_2}+\sqrt{C_{\mathcal{B}}}\Big),
$$

because for any $p \geq 1$,

$$\left(\mathbb{E}_{X_{ij}\sim\rho_X^{(i)}}\left|\mathcal{V}_j^{(i)}(\hat{\rho}_X^{(i)})(x)\right|^p\right)^{\frac{1}{p}} \leq \frac{1}{\vartheta}\left[(\mathbb{E}_{\rho_X^{(i)}}\|X_{ij}\|_2^p)^{\frac{1}{p}} + \sqrt{C_{\mathcal{B}}}\right]$$

by Minkowski inequality. Then by Theorem 3 in Maurer and Pontil [59], we have for any $\epsilon > 0$,

$$\mathbb{P}\{\mathcal{V}^{(i)}(\hat{\rho}_X^{(i)}) - \mathbb{E}\mathcal{V}^{(i)}(\hat{\rho}_X^{(i)}) > \epsilon\} \leq \exp\left(-\frac{\vartheta\epsilon^2}{\psi_2(\rho_X^{(i)})}\right)$$

where

$$\psi_2(\rho_X^{(i)}) = 32e\Big(c\kappa\sqrt{\gamma'}(1 + \|\omega_\kappa(\rho_X^{(i)})\|_{L^\gamma(\rho_\kappa)}) + \sqrt{C_{\mathcal{B}}}\Big)^2.$$

We also have

$$\begin{aligned}
\mathbb{E}\mathcal{V}^{(i)}(\hat{\rho}_X^{(i)}) &= \mathbb{E}_{X_{ij}\sim P_X^{(i)}} \sup_{\phi\in\mathcal{NN}(\Theta_{\tilde{m}})} \left\|\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}\phi(X_{ij})X_{ij} - \mathbb{E}_{P_X^{(i)}}(\phi(X)X)\right\|_2 \\
&\leq \mathbb{E}_{X_{ij},X'_{ij}\overset{i.i.d}{\sim}P_X^{(i)}} \sup_{\phi\in\mathcal{NN}(\Theta_{\tilde{m}})} \left\|\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}[\phi(X_{ij})X_{ij} - \phi(X'_{ij})X'_{ij}]\right\|_2 \\
&= \frac{1}{\vartheta}\mathbb{E}_{X_{ij},X'_{ij}\overset{i.i.d}{\sim}P_X^{(i)}}\mathbb{E}_{\zeta_{ij}} \sup_{\phi\in\mathcal{NN}(\Theta_{\tilde{m}})} \left\|\sum_{j=1}^{\vartheta}\zeta_{ij}[\phi(X_{ij})X_{ij} - \phi(X'_{ij})X'_{ij}]\right\|_2 \\
&\leq \frac{2}{\vartheta}\mathbb{E}_{X_{ij}}\mathbb{E}_{\zeta_{ij}} \sup_{\phi\in\mathcal{NN}(\Theta_{\tilde{m}})} \left\|\sum_{j=1}^{\vartheta}\zeta_{ij}\phi(X_{ij})X_{ij}\right\|_2 \\
&= \frac{2}{\vartheta}\mathbb{E}_{X_{ij}}\mathbb{E}_{\zeta_{ij}} \sup_{\phi\in\mathcal{NN}(\Theta_{\tilde{m}})} \sup_{\mathrm{u}\in S^{d-1}} \left(\sum_{j=1}^{\vartheta}\zeta_{ij}\phi(X_{ij})\mathrm{u}^T X_{ij}\right) \\
&= \frac{2}{\vartheta}\mathbb{E}_{X_{ij}}\mathbb{E}_{\zeta_{ij}} \sup_{f\in\mathrm{U}} \left(\sum_{j=1}^{\vartheta}\zeta_{ij}f(X_{ij})\right)
\end{aligned}$$

where $(\zeta_{ij})_{j=1}^{\vartheta}$ are independent Rademacher random variables and

$$\mathrm{U} := \{f : f(x) = \phi(x)\mathrm{u}^T x \text{ with } \phi \in \mathcal{NN}(\Theta_{\tilde{m}}), \mathrm{u} \in S^{d-1}\}.$$

We define a family of zero-mean random variables index by $f \in \mathrm{U}$ as

$$Z_f^{(i)} := \frac{1}{\sqrt{\vartheta}}\sum_{j=1}^{\vartheta}\zeta_{ij}f(X_{ij}),$$

which implies that

$$\mathbb{E}\mathcal{V}^{(i)}(\hat{\rho}_X^{(i)}) \leq \mathbb{E}_{X_{ij}}\mathbb{E}\left[\sup_{f\in \mathrm{U}} \frac{1}{\sqrt{\vartheta}} Z_f^{(i)} \middle| (X_{ij})_j\right].$$

For any $f, f' \in \mathrm{U}$,

$$\mathbb{E}[\exp(v(Z_f^{(i)} - Z_{f'}^{(i)}))|(X_{ij})_{i,j}] \leq \exp(v^2 d_\vartheta^{(i)}(f,f')^2/2),\ \forall v \in \mathbb{R}$$

where

$$d_\vartheta^{(i)}(f,f') := \left(\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}(f(X_{ij}) - f'(X_{ij}))^2\right)^{\frac{1}{2}}.$$

Then, conditioned on $(X_{ij})_j$, $Z_f^{(i)}$ is a subgaussian process indexed by $f \in \mathrm{U}$ with respect to $d_\vartheta^{(i)}$. It's easy to see that for any $f, f' \in \mathrm{U}$,

$$d_\vartheta^{(i)}(f,f') = \left(\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}\left(\phi_f(X_{ij})\mathrm{u}_f^T X_{ij} - \phi_{f'}(X_{ij})\mathrm{u}_{f'}^T X_{ij}\right)^2\right)^{\frac{1}{2}} \leq 2\sqrt{\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}\|X_{ij}\|_2^2} =: 2M_i.$$

If $\|\mathrm{u}_f - \mathrm{u}_{f'}\|_2 \leq \epsilon$ and parameters in ϕ_f and $\phi_{f'}$ satisfy (4.23),

$$\begin{aligned}
d_\vartheta^{(i)}(f,f') &\leq \left(\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}\left(\phi_f(X_{ij})\mathrm{u}_f^T X_{ij} - \phi_{f'}(X_{ij})\mathrm{u}_f^T X_{ij}\right)^2\right)^{\frac{1}{2}} \\
&\quad + \left(\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}\left(\phi_{f'}(X_{ij})\mathrm{u}_f^T X_{ij} - \phi_{f'}(X_{ij})\mathrm{u}_{f'}^T X_{ij}\right)^2\right)^{\frac{1}{2}} \\
&\leq \left(\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}\left(\phi_f(X_{ij}) - \phi_{f'}(X_{ij})\right)^2\|X_{ij}\|_2^2\right)^{\frac{1}{2}} + M_i\|\mathrm{u}_f - \mathrm{u}_{f'}\|_2 \\
&\leq \left(\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}\left(192d(2+\|X_{ij}\|_\infty)[c_4\tilde{m}]^{3c_5\tilde{m}^2}\epsilon\right)^2\|X_{ij}\|_2^2\right)^{\frac{1}{2}} + M_i\epsilon \\
&\leq \left(M_i + 192d(c_4\tilde{m})^{3c_5\tilde{m}^2}\sqrt{\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}(8\|X_{ij}\|_2^2 + 2\|X_{ij}\|_2^4)}\right)\epsilon.
\end{aligned}$$

Then the covering number

$$\mathrm{N}(\tilde{\epsilon}, \mathrm{U}, d_\vartheta^{(i)}) \le \left(1 + \frac{(c_2'\tilde{m})^{c_3'\tilde{m}^2}}{\epsilon}\right)^{96d^2\tilde{m}^2} \left(1 + \frac{2}{\epsilon}\right)^d \le \left(1 + \frac{(c_2'\tilde{m})^{c_3'\tilde{m}^2}}{\epsilon}\right)^{100d^2\tilde{m}^2}$$
$$\le \left(1 + \frac{\Delta^{(i)}}{\tilde{\epsilon}}\right)^{100d^2\tilde{m}^2} (c_6\tilde{m})^{c_7\tilde{m}^4},$$

where $\Delta^{(i)} := M_i + \sqrt{\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}(8\|X_{ij}\|_2^2 + 2\|X_{ij}\|_2^4)}$, $c_6 = 192d(c_2' + c_4)$ and $c_7 = 100(c_3' + 3c_5)d^2$.

Then by Dudley integral, we have

$$\mathbb{E}\left[\sup_{f\in\mathrm{U}} Z_f^{(i)} \middle| (X_{ij})_j\right] \le 32\int_0^{2M_i} \sqrt{\log \mathrm{N}(v, \mathrm{U}, \mathrm{d}_\vartheta^{(\mathrm{i})})}\, dv$$
$$\le 32\int_0^{2M_i} \sqrt{100d^2\tilde{m}^2 \log\left(1 + \frac{\Delta^{(i)}}{v}\right) + c_7\tilde{m}^4\log(c_6\tilde{m})}\, dv$$
$$\le 32\left(2M_i\tilde{m}^2\sqrt{c_7\log(c_6\tilde{m})} + 10d\tilde{m}\int_0^{2M_i}\sqrt{\log\left(1 + \frac{\Delta^{(i)}}{v}\right)}\, dv\right)$$
$$\le 32\left(2M_i\tilde{m}^2\sqrt{c_7\log(c_6\tilde{m})} + 10d\tilde{m}\int_0^{2M_i}\sqrt{\Delta^{(i)}}v^{-\frac{1}{2}}\, dv\right)$$
$$\le 32\Big(2M_i\tilde{m}^2\sqrt{c_7\log(c_6\tilde{m})} + 20d\tilde{m}\sqrt{2M_i\Delta^{(i)}}\Big).$$

Then we have

$$\mathbb{E}\mathcal{V}^{(i)}(\hat{\rho}_X^{(i)}) \le \frac{1}{\sqrt{\vartheta}}\mathbb{E}_{X_{ij}}32\Big(2M_i\tilde{m}^2\sqrt{c_7\log(c_6\tilde{m})} + 20d\tilde{m}\sqrt{2M_i\Delta^{(i)}}\Big)$$
$$= \frac{1}{\sqrt{\vartheta}}\left(64\tilde{m}^2\sqrt{c_7\log(c_6\tilde{m})}\mathbb{E}_{X_{ij}}M_i + 640d\tilde{m}\mathbb{E}_{X_{ij}}\sqrt{2M_i\Delta^{(i)}}\right)$$

where $\mathbb{E}_{X_{ij}}M_i = \mathbb{E}_{X_{ij}}\sqrt{\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}\|X_{ij}\|_2^2} \le \sqrt{\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}\mathbb{E}_{X_{ij}}\|X_{ij}\|_2^2} = \sqrt{C_{\mathcal{B}}}$ and

$$\mathbb{E}_{X_{ij}}\sqrt{2M_i\Delta^{(i)}} = \mathbb{E}_{X_{ij}}\sqrt{2M_i\left(M_i + \sqrt{\frac{1}{\vartheta}\sum_{j=1}^{\vartheta}(8\|X_{ij}\|_2^2 + 2\|X_{ij}\|_2^4)}\right)}$$

$$
\begin{aligned}
&\leq \mathbb{E}_{X_{ij}} \sqrt{2M_i \left(M_i + \sqrt{\frac{1}{\vartheta}\sum_{j=1}^{\vartheta} 8\|X_{ij}\|_2^2} + \sqrt{\frac{1}{\vartheta}\sum_{j=1}^{\vartheta} 2\|X_{ij}\|_2^4} \right)} \\
&= \mathbb{E}_{X_{ij}} \sqrt{(2+4\sqrt{2})M_i^2 + 2\sqrt{2}M_i \sqrt{\frac{1}{\vartheta}\sum_{j=1}^{\vartheta} \|X_{ij}\|_2^4}} \\
&\leq \sqrt{(2+4\sqrt{2})\mathbb{E}_{X_{ij}} M_i^2 + 2\sqrt{2}\sqrt{\mathbb{E}_{X_{ij}} M_i^2} \sqrt{\mathbb{E}_{X_{ij}} \left(\frac{1}{\vartheta}\sum_{j=1}^{\vartheta} \|X_{ij}\|_2^4 \right)}} \\
&\leq \sqrt{(2+4\sqrt{2})C_{\mathcal{B}} + 2\sqrt{2}\sqrt{C_{\mathcal{B}}}C_{\mathcal{B}}} \leq \sqrt{2+6\sqrt{2}} C_{\mathcal{B}}^{\frac{3}{4}}.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\mathbb{E}\mathcal{V}^{(i)}(\hat{\rho}_X^{(i)}) \leq \frac{1}{\sqrt{\vartheta}} \left(64\sqrt{C_{\mathcal{B}}}\tilde{m}^2 \sqrt{c_7 \log(c_6 \tilde{m})} + 640d\sqrt{2+6\sqrt{2}} C_{\mathcal{B}}^{\frac{3}{4}} \tilde{m} \right) \\
&\leq \frac{c_8}{\sqrt{\vartheta}} \tilde{m}^2 \sqrt{\log(\tilde{m})}
\end{aligned}
$$

where $c_8 = 64\sqrt{C_{\mathcal{B}} c_7 (\log c_6 + 1)} + 640d\sqrt{2+6\sqrt{2}} C_{\mathcal{B}}^{\frac{3}{4}}$.

Then we have

$$
\mathbb{P}\left\{ \mathcal{V}^{(i)}(\hat{\rho}_X^{(i)}) > \epsilon + \frac{c_8 \tilde{m}^2 \sqrt{\log(\tilde{m})}}{\sqrt{\vartheta}} \right\} \leq \exp\left(-\frac{\vartheta \epsilon^2}{\Psi_2(\rho_X^{(i)})} \right),
$$

and it follows that

$$
\begin{aligned}
\mathbb{P}\left\{ \sup_{\Phi \in \mathcal{H}_{\mathrm{T}_n}} |\mathcal{E}_3(\mathscr{T}_M(\Phi))| > 8MC_F\sqrt{d}m \left(\epsilon + \frac{c_8 \tilde{m}^2 \sqrt{\log(\tilde{m})}}{\sqrt{\vartheta}} \right) \middle| \rho_X^{(1)}, \ldots, \rho_X^{(N)} \right\} \\
\leq N \exp\left(-\frac{\vartheta \epsilon^2}{\max_{1 \leq i \leq N} \Psi_2(\rho_X^{(i)})} \right)
\end{aligned}
$$

which is equivalent to the inequality that

$$\mathbb{P}\left\{\sup_{\Phi\in\mathcal{H}_{\mathrm{T}_n}}|\mathcal{E}_3(\mathscr{T}_M(\Phi))|>\epsilon+\frac{8c_8\sqrt{d}MC_F\tilde{m}^2m\sqrt{\log(\tilde{m})}}{\sqrt{\vartheta}}\middle|\rho_X^{(1)},\ldots,\rho_X^{(N)}\right\}$$
$$\leq N\exp\left(-\frac{\vartheta\epsilon^2}{64M^2C_F^2dm^2\max_{1\leq i\leq N}\psi_2(\rho_X^{(i)})}\right).$$

Similarly, for any $\Phi\in\mathcal{H}_{\mathrm{T}_n}$, we have both $\|\Phi\|_{C(\Omega_{\mathcal{B}})}$ and $\|\Phi\|_{C(\Omega)}$ uniformly bounded respectively. Then we have

$$\begin{aligned}\mathcal{E}_3(\Phi)&=\frac{1}{N\vartheta}\sum_{i=1}^{N}\sum_{j=1}^{\vartheta}\left(\|\Phi(\hat{\rho}_X^{(i)},X_{ij})-Y_{ij}\|_2^2-\|\Phi(\rho_X^{(i)},X_{ij})-Y_{ij}\|_2^2\right)\\&\leq\frac{1}{N\vartheta}\sum_{i=1}^{N}\sum_{j=1}^{\vartheta}(2M+\|\Phi\|_{C(\Omega_{\mathcal{B}})}+\|\Phi\|_{C(\Omega)})\left\|\Phi(\hat{\rho}_X^{(i)},X_{ij})-\Phi(\rho_X^{(i)},X_{ij})\right\|_2\\&\leq 4(M+\|\Phi\|_{C(\Omega)})C_F\sqrt{d}m\max_{1\leq i\leq N}\left\|\int_{\mathcal{X}}\phi_\Phi(x)x\,d\hat{\rho}_X^{(i)}-\int_{\mathcal{X}}\phi_\Phi(x)x\,d\rho_X^{(i)}\right\|_2,\end{aligned}$$

which follows that

$$\begin{aligned}&\mathbb{P}\left\{\mathcal{E}_3'(\Phi)>\epsilon+\frac{4c_8C_F\sqrt{d}(M+\|\Phi\|_{C(\Omega)})\tilde{m}^2m\sqrt{\log(\tilde{m})}}{\sqrt{\vartheta}}\middle|\rho_X^{(1)},\ldots,\rho_X^{(N)}\right\}\\\leq&N\exp\left(-\frac{\vartheta\epsilon^2}{16(M+\|\Phi\|_{C(\Omega)})^2C_F^2dm^2\max_{1\leq i\leq N}\psi_2(\rho_X^{(i)})}\right)\end{aligned}$$

Recall that for any $\Phi\in\mathcal{H}_{\mathrm{T}_n}$,

$$\begin{aligned}&\mathcal{E}((\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})))-\mathcal{E}(\Phi_{\mathcal{G}})=\|\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})-\Phi_{\mathcal{G}}\|_{L^2(\nu_{\mathcal{G}})}^2\\\leq&\mathcal{E}_1(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}))+\mathcal{E}_1'(\Phi)+\mathcal{E}_2(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}))+\mathcal{E}_2'(\Phi)+\mathcal{E}_3(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}))+\mathcal{E}_3'(\Phi)+\mathcal{E}_4(\Phi).\end{aligned}$$

Then combine all inequalities, and by the union bound, for any $\Phi \in \mathcal{H}_{\mathrm{T}_n}$ we have

$$
\begin{aligned}
&\mathbb{P}\Big\{ \|\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}) - \Phi_{\mathcal{G}}\|_{L^2(\pi)}^2 > 2\mathcal{E}_4(\Phi) + 2C_{M,\mathcal{B},d}\frac{\sqrt{n}m\tilde{m}^2}{\sqrt{N\vartheta}} \\
&\quad + c_9(3M + \|\Phi\|_{C(\Omega)})\frac{\tilde{m}^2 m\sqrt{\log(\tilde{m})}}{\sqrt{\vartheta}} + 12\epsilon \Big\} \\
&\leq \mathrm{N}\left(\mathcal{H}_{\mathrm{T}_n}, \frac{\epsilon}{16M}, d_{\mathcal{B}_{2,b}(\mathcal{X})}\right) \exp\left\{-\frac{3N\epsilon}{2048M^2}\right\} + \exp\left\{-\frac{N\epsilon^2}{2(3M + \|\Phi\|_{C(\Omega_{\mathcal{B}})})^2\left(\mathcal{E}_4(\Phi) + \frac{2}{3}\epsilon\right)}\right\} \\
&\quad + 2\exp\left\{-\frac{(N\vartheta)\epsilon^2}{32(M + \|\Phi\|_{C(\Omega_{\mathcal{B}})})^4}\right\} \\
&\quad + 2\mathbb{E}_{P_X^{(i)}\sim\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}} N\exp\left\{-\frac{\vartheta\epsilon^2}{64(M + \|\Phi\|_{C(\Omega)})^2 C_F^2 dm^2 \max_{1\leq i\leq N}\psi_2(\rho_X^{(i)})}\right\}
\end{aligned}
$$

where $\mathcal{E}_4(\Phi) = \|\Phi - \Phi_{\mathcal{G}}\|_{L^2(\nu_{\mathcal{G}})}^2$ and $c_9 = 8c_8 C_F\sqrt{d}$. ■

4.5.3 Theorem 4.10: Generalization Bound for Linear Transformers

From the approximation results, for $n \geq 3$, there exists a transformer $\mathrm{T} \in \mathcal{H}_{\mathrm{T}_n}$ such that

$$
\|\mathrm{T} - \Phi_{\mathcal{G}}\|_{L^2(\nu_{\mathcal{G}})} \leq C_*\left(\left\lfloor\frac{n}{2}\right\rfloor\right)^{-\frac{1}{2}} \leq C'_* n^{-\frac{1}{2}} \text{ with } C'_* = 2C_*
$$

and by the approximant construction, we have $\|\mathrm{T}\|_{C(\Omega_{\mathcal{B}})} \leq \|\mathrm{T}\|_{C(\Omega)} \leq 2C_F\sqrt{d(1 + C_{\mathcal{B}})} = \mathcal{A}_1$. Apply the above oracle inequality with letting $\Phi = \mathrm{T}$ and $\overline{\psi_2(\rho_X)} := \frac{1}{N}\sum_{i=1}^N \psi_2(\rho_X^{(i)})$, and we obtain that

$$
\begin{aligned}
&\mathbb{P}\Big\{ \|\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}) - \Phi_{\mathcal{G}}\|_{L^2(\pi)}^2 > 2C_*'^2 n^{-1} + 2C_{M,\mathcal{B},d}\frac{\sqrt{n}m\tilde{m}^2}{\sqrt{N\vartheta}} \\
&\quad + 8c_8 C_F\sqrt{d}(3M + \mathcal{A}_1)\frac{\tilde{m}^2 m\sqrt{\log(\tilde{m})}}{\sqrt{\vartheta}} + 12\epsilon \Big\} \\
&\leq \exp\left\{100d^2 nm^2\log\left(1 + \frac{128MC_{d,F,\mathcal{B}}}{\epsilon}\right) + 600c_5 d^2 nm^2\tilde{m}^2\log(c_4 m\tilde{m}) - \frac{3N\epsilon}{2048M^2}\right\} \\
&\quad + \exp\left\{-\frac{N\epsilon^2}{2(3M + \mathcal{A}_1)^2\left(C_*'^2 n^{-1} + \frac{2}{3}\epsilon\right)}\right\} + 2\exp\left\{-\frac{(N\vartheta)\epsilon^2}{32(M + \mathcal{A}_1)^4}\right\}
\end{aligned}
$$

$$+\ 2\mathbb{E}_{P_X^{(i)}\sim\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}} N\exp\left\{-\frac{\vartheta\epsilon^2}{64(M+\mathcal{A}_1)^2 dC_F^2 m^2 N\overline{\Psi_2(\rho_X)}}\right\}.$$

We follow the parameter selections in the approximation by letting $m = \lceil n^{\frac{\gamma}{2(\gamma-1)\xi}}\rceil$ and $\tilde{m} = \lceil(\frac{1}{2}+\frac{\gamma}{4(\gamma-1)\xi})\log n\rceil$. It's obtained that

$$\begin{aligned}
&\mathbb{P}\left\{\|\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}) - \Phi_{\mathcal{G}}\|_{L^2(\pi)}^2 > 2C_*'^2 n^{-1} + \mathcal{A}_2\frac{n^{\frac{1}{2}+\frac{\gamma}{2(\gamma-1)\xi}}(\log n)^2}{\sqrt{N\vartheta}} + \mathcal{A}_3\frac{n^{\frac{\gamma}{2(\gamma-1)\xi}}(\log n)^3}{\sqrt{\vartheta}} + 12\epsilon\right\}\\
&\le \exp\left\{200d^2 n^{1+\frac{\gamma}{(\gamma-1)\xi}}\log\left(1+\frac{128MC_{d,F,\mathcal{B}}}{\epsilon}\right) + \mathcal{A}_4 n^{1+\frac{\gamma}{(\gamma-1)\xi}}(\log n)^3 - \frac{3N\epsilon}{2048M^2}\right\}\\
&\qquad + \exp\left\{-\frac{N\epsilon^2}{2(3M+\mathcal{A}_1)^2\left(C_*'^2 n^{-1}+\frac{2}{3}\epsilon\right)}\right\} + 2\exp\left\{-\frac{(N\vartheta)\epsilon^2}{32(M+\mathcal{A}_1)^4}\right\}\\
&\qquad + 2\mathbb{E}_{P_X^{(i)}\sim\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}} N\exp\left\{-\frac{\vartheta\epsilon^2}{128d(M+\mathcal{A}_1)^2 C_F^2 n^{\frac{\gamma}{(\gamma-1)\xi}} N\overline{\Psi_2(\rho_X)}}\right\}
\end{aligned}$$

where $\mathcal{A}_2 = \left(1+\frac{\gamma}{2(\gamma-1)\xi}\right)^2 C_{M,\mathcal{B},d}$, $\mathcal{A}_3 = 32(3M+\mathcal{A}_1)C_F\sqrt{dC_{\mathcal{B}}}$ and

$$\mathcal{A}_4 = \left(1+\frac{\gamma}{(\gamma-1)\xi}\right)\left[\left(1200c_5 d^2\left(\frac{1}{2}+\frac{\gamma}{4(\gamma-1)\xi}\right)^2 + \log c_4\left(\frac{1}{2}+\frac{\gamma}{4(\gamma-1)\xi}\right)^2\right)\right].$$

If we take $\epsilon \ge 2C_*'^2 n^{-1}(\log n)^3$, it follows that

$$\begin{aligned}
&\mathbb{P}\left\{\|\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}) - \Phi_{\mathcal{G}}\|_{L^2(\pi)}^2 > 13\epsilon + \mathcal{A}_2\frac{n^{\frac{1}{2}+\frac{\gamma}{2(\gamma-1)\xi}}(\log n)^2}{\sqrt{N\vartheta}} + \mathcal{A}_3\frac{n^{\frac{\gamma}{2(\gamma-1)\xi}}(\log n)^3}{\sqrt{\vartheta}}\right\}\\
&\le \exp\left\{200d^2 n^{1+\frac{\gamma}{(\gamma-1)\xi}}\log\left(1+\frac{64MC_{d,F,\mathcal{B}}}{C_*'^2}\,n\right) + \mathcal{A}_4 n^{1+\frac{\gamma}{(\gamma-1)\xi}}(\log n)^3 - \frac{3N\epsilon}{2048M^2}\right\}\\
&\quad + \exp\left\{-\frac{3N\epsilon}{8(3M+\mathcal{A}_1)^2}\right\} + 2\exp\left\{-\frac{(N\vartheta)\epsilon^2}{32(M+\mathcal{A}_1)^4}\right\}\\
&\quad + 2\mathbb{E}_{P_X^{(i)}\sim\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}} N\exp\left\{-\frac{\vartheta\epsilon^2}{128d(M+\mathcal{A}_1)^2 C_F^2 n^{\frac{\gamma}{(\gamma-1)\xi}} N\overline{\Psi_2(\rho_X)}}\right\}.
\end{aligned}$$

Take $n = \lfloor \mathcal{K}_1 N^{\frac{1}{2+\gamma/[(\gamma-1)\xi]}} \rfloor$ with $\mathcal{K}_1 = \left(\min\left\{ \frac{3C_*'^2}{819200M^2d^2\left(1+\log\left(\frac{64MC_{d,F,\mathcal{B}}}{C_*}\right)\right)}, \frac{3C_*'^2}{4096M^2\mathcal{A}_4} \right\} \right)^{\frac{1}{2+\frac{\gamma\xi}{(\gamma-1)}}}$ and the second stage data size $\vartheta \geq N$, then we have

$$
\begin{aligned}
&\mathbb{P}\{\|\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}) - \Phi_{\mathcal{G}}\|_{L^2(\pi)}^2 > \mathcal{A}_5\epsilon\} \\
\leq& \exp\left\{\frac{3N\epsilon}{8192M^2} + \frac{3N\epsilon}{8192M^2} - \frac{3N\epsilon}{2048M^2}\right\} + \exp\left\{-\frac{3N\epsilon}{8(3M+\mathcal{A}_1)^2}\right\} + 2\exp\left\{-\frac{C_*'^2N\epsilon}{16\mathcal{K}_1(M+\mathcal{A}_1)^4}\right\} \\
&+ 2\mathbb{E}_{P_X^{(i)}\sim\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}} N\exp\left\{-\frac{\vartheta\epsilon}{\mathcal{A}_6 N^{\frac{3(\gamma-1)\xi+2\gamma}{2(\gamma-1)\xi+\gamma}}\overline{\psi_2(\rho_X)}}\right\} \\
\leq& 4\exp\left\{-\frac{N\epsilon}{\mathcal{A}_7}\right\} + 2\mathbb{E}_{P_X^{(i)}\sim\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}} N\exp\left\{-\frac{\vartheta\epsilon}{\mathcal{A}_6 N^{\frac{3(\gamma-1)\xi+2\gamma}{2(\gamma-1)\xi+\gamma}}\overline{\psi_2(\rho_X)}}\right\},
\end{aligned}
$$

where

$$
\mathcal{A}_5 = 13 + \frac{\mathcal{A}_2\mathcal{K}_1^{\frac{3}{2}+\frac{\gamma}{2(\gamma-1)\xi}} + \mathcal{A}_3\mathcal{K}_1^{1+\frac{\gamma}{(\gamma-1)\xi}}}{2C_*'^2},\ \mathcal{A}_6 = 128d(M+\mathcal{A}_1)^2C_F^2\mathcal{K}_1^{\frac{\gamma}{(\gamma-1)\xi}},
$$

and $\mathcal{A}_7 = \min\left\{\frac{3}{3096M^2}, \frac{3}{8(3M+\mathcal{A}_1)^2}, \frac{C_*'^2}{16\mathcal{K}_1(M+\mathcal{A}_1)^4}\right\}$.

Take $t = \mathcal{A}_5\epsilon$. Then when $t \geq 2\mathcal{A}_5C_*'^2n^{-1}(\log n)^3$, we have

$$
\mathbb{P}\{\|\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}) - \Phi_{\mathcal{G}}\|_{L^2(\nu_{\mathcal{G}})}^2 > t\} \leq 4\exp\left\{-\frac{Nt}{\mathcal{A}_5\mathcal{A}_7}\right\} + 2\mathbb{E}_{\rho_X^{(i)}\sim\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}} N\exp\left\{-\frac{\vartheta t}{\mathcal{A}_5\mathcal{A}_6 N^{\frac{3(\gamma-1)\xi+2\gamma}{2(\gamma-1)\xi+\gamma}}\overline{\psi_2(\rho_X)}}\right\}.
$$

It implies that

$$
\begin{aligned}
&\mathbb{E}\{\mathcal{E}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}(\Phi_{\mathcal{G}})\} \\
=&\mathbb{E}\|\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}) - \Phi_{\mathcal{G}}\|_{L^2(\nu_{\mathcal{G}})}^2 = \int_0^\infty \mathbb{P}\{\|\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}) - \Phi_{\mathcal{G}}\|_{L^2(\nu_{\mathcal{G}})}^2 > t\}\,dt \\
=& \left(\int_0^{2\mathcal{A}_5C_*'^2n^{-1}(\log n)^3} + \int_{2\mathcal{A}_5C_*'^2n^{-1}(\log n)^3}^\infty\right) \mathbb{P}\{\|\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}) - \Phi_{\mathcal{G}}\|_{L^2(\nu_{\mathcal{G}})}^2 > t\}\,dt \\
\leq& 2\mathcal{A}_5C_*'^2n^{-1}(\log n)^3 + \int_0^\infty \mathbb{P}\{\|\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n}) - \Phi_{\mathcal{G}}\|_{L^2(\nu_{\mathcal{G}})}^2 > t\}\,dt
\end{aligned}
$$

$$\leq 2\mathcal{K}_2 N^{-\frac{1}{2+\gamma/[(\gamma-1)\xi]}} (\log N)^3 + \int_0^\infty 4\exp\left\{-\frac{N^{\frac{1}{2+\gamma/[(\gamma-1)\xi]}} t}{\mathcal{A}_5\mathcal{A}_7}\right\} dt$$

$$+ 2\mathbb{E}_{\rho_X^{(i)}\sim\mathcal{P}_\mathcal{G}^\mathcal{X}} N \int_0^\infty \exp\left\{-\frac{\vartheta t}{\mathcal{A}_5\mathcal{A}_6 N^{\frac{3(\gamma-1)\xi+2\gamma}{2(\gamma-1)\xi+\gamma}} \overline{\psi_2(\rho_X)}}\right\} dt$$

$$= 2\mathcal{K}_2 N^{-\frac{1}{2+\gamma/[(\gamma-1)\xi]}} (\log N)^3 + 4\mathcal{A}_5\mathcal{A}_7 N^{-\frac{1}{2+\gamma/[(\gamma-1)\xi]}} + 2\mathcal{A}_5\mathcal{A}_6 \frac{N^{\frac{5(\gamma-1)\xi+3\gamma}{2(\gamma-1)\xi+\gamma}}}{\vartheta} \mathbb{E}_{\rho_X^{(i)}\sim\mathcal{P}_\mathcal{G}^\mathcal{X}} \left(\overline{\psi_2(\rho_X)}\right)$$

where $\mathcal{K}_2 = 4\mathcal{A}_5 C_*'^2 \mathcal{K}_1^{-1} \left(\log \mathcal{K}_1 + \frac{1}{2+\gamma/[(\gamma-1)\xi]}\right)^3$. Take $\vartheta = N^3$ and we obtain that

$$\mathbb{E}\{\mathcal{E}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}(\Phi_\mathcal{G})\} \leq \left(2\mathcal{K}_2 + 4\mathcal{A}_5\mathcal{A}_7 + 2\mathcal{A}_5\mathcal{A}_6 \mathbb{E}_{\rho_X^{(i)}\sim\mathcal{P}_\mathcal{G}^\mathcal{X}} \left(\overline{\psi_2(\rho_X)}\right)\right) N^{-\frac{1}{2+\gamma/[(\gamma-1)\xi]}} (\log N)^3$$

where

$$\begin{aligned}
\mathbb{E}_{\rho_X^{(i)}\sim\mathcal{P}_\mathcal{G}^\mathcal{X}} \left(\overline{\psi_2(\rho_X)}\right) &= \mathbb{E}_{\rho_X^{(i)}\sim\mathcal{P}_\mathcal{G}^\mathcal{X}} \left(\frac{1}{N}\sum_{i=1}^N \psi_2(\rho_X^{(i)})\right) \\
&= 32e\, \mathbb{E}_{\rho_X^{(i)}\sim\mathcal{P}_\mathcal{G}^\mathcal{X}} \frac{1}{N}\sum_{i=1}^N \left(c\kappa\sqrt{\gamma'}(1 + \|\omega_\kappa(\rho_X^{(i)})\|_{L^\gamma(\rho_\kappa)}) + \sqrt{C_\mathcal{B}}\right)^2 \\
&\leq 64e\, \mathbb{E}_{\rho_X\sim\mathcal{P}_\mathcal{G}^\mathcal{X}} \left(c^2\kappa^2\gamma'(1 + \|\omega_\kappa(\rho_X)\|_{L^\gamma(\rho_\kappa)})^2 + C_\mathcal{B}\right) \\
&\leq\ 64e(C_\mathcal{B} + 2c^2\kappa^2\gamma') + 128e\, \mathbb{E}_{\rho_X\sim\mathcal{P}_\mathcal{G}^\mathcal{X}} \|\omega_\kappa(\rho_X)\|^2_{L^\gamma(\rho_\kappa)} \\
&\leq\ 64e(C_\mathcal{B} + 2c^2\kappa^2\gamma' + 2C_\mathcal{G}).
\end{aligned}$$

It follows that

$$\mathbb{E}\{\mathcal{E}(\mathscr{T}_M(\mathrm{T}_{\mathbb{S},n})) - \mathcal{E}(\Phi_\mathcal{G})\} \leq \mathcal{K}_3 N^{-\frac{1}{2+\gamma/[(\gamma-1)\xi]}} (\log N)^3$$

with $\mathcal{K}_3 = 2\mathcal{K}_2 + 4\mathcal{A}_5\mathcal{A}_7 + 128e\mathcal{A}_5\mathcal{A}_6(C_\mathcal{B} + 2c^2\kappa^2\gamma' + 2C_\mathcal{G})$. ■

Appendix C

Context Embedding and Feature Mapping

PROPOSITION 2. *$K_{\boldsymbol{\lambda}} : \mathcal{B}_2(\mathcal{X}) \to \mathcal{H}_{k_{\boldsymbol{\lambda}}} \otimes \mathbb{R}^d$ is an injective and continuous mapping.*

PROPOSITION 3. *The embedding operator* $\mathrm{I}_{\boldsymbol{\lambda}} : \Omega \to \mathcal{H}_{\mathcal{F}}$ *is injective and continuous.*

Proof. Continuity: Recall that $\Omega = \mathcal{B}_2(\mathcal{X}) \times \mathcal{X}$ is a metric space equipped with d_Ω. Also observe that

$$\begin{aligned}\|\mathrm{I}_{\boldsymbol{\lambda}}(\rho, x) - \mathrm{I}_{\boldsymbol{\lambda}}(\rho', x')\|_{\mathcal{H}_{\mathcal{F}}} &\leq \|\mathrm{I}_{\boldsymbol{\lambda}}(\rho, x) - \mathrm{I}_{\boldsymbol{\lambda}}(\rho', x)\|_{\mathcal{H}_{\mathcal{F}}} + \|\mathrm{I}_{\boldsymbol{\lambda}}(\rho', x) - \mathrm{I}_{\boldsymbol{\lambda}}(\rho', x')\|_{\mathcal{H}_{\mathcal{F}}} \\ &= \|K_{\boldsymbol{\lambda}}(\rho - \rho') \otimes k_{\boldsymbol{\lambda}}(x, \cdot)\|_{\mathcal{H}_{\mathcal{F}}} + \|K_{\boldsymbol{\lambda}}(\rho') \otimes (k_{\boldsymbol{\lambda}}(x, \cdot) - k_{\boldsymbol{\lambda}}(x', \cdot))\|_{\mathcal{H}_{\mathcal{F}}},\end{aligned}$$

in which

$$\begin{aligned}&\|k_{\boldsymbol{\lambda}}(x, \cdot) - k_{\boldsymbol{\lambda}}(x', \cdot)\|^2_{\mathcal{H}_{k_{\boldsymbol{\lambda}}}} = 2(1 - \exp\{-(x - x')^T \Sigma_{\boldsymbol{\lambda}}(x - x')\}) \\ &\leq 2(x - x')^T \Sigma_{\boldsymbol{\lambda}}(x - x') \leq 2\|\Sigma_{\boldsymbol{\lambda}}\|_2 \|x - x\|_2^2,\end{aligned}$$

and for any $\tau \in \prod(\rho, \rho')$,

$$\begin{aligned}&\|K_{\boldsymbol{\lambda}}(\rho - \rho')\|_{\mathcal{H}_{k_{\boldsymbol{\lambda}}} \otimes \mathbb{R}^d} = \left\| \int_{\mathcal{X} \times \mathcal{X}} (k_{\boldsymbol{\lambda}}(\cdot, y)y - k_{\boldsymbol{\lambda}}(\cdot, y')y') \, d\rho(y) d\rho'(y') \right\|_{\mathcal{H}_{k_{\boldsymbol{\lambda}}} \otimes \mathbb{R}^d} \\ &\leq \int_{\mathcal{X} \times \mathcal{X}} \|k_{\boldsymbol{\lambda}}(\cdot, y)y - k_{\boldsymbol{\lambda}}(\cdot, y')y'\|_{\mathcal{H}_{k_{\boldsymbol{\lambda}}} \otimes \mathbb{R}^d} \, d\tau(y, y') \\ &\leq \int_{\mathcal{X} \times \mathcal{X}} \|k_{\boldsymbol{\lambda}}(\cdot, y)(y - y')\|_{\mathcal{H}_{k_{\boldsymbol{\lambda}}} \otimes \mathbb{R}^d} + \|(k_{\boldsymbol{\lambda}}(\cdot, y) - k_{\boldsymbol{\lambda}}(\cdot, y'))y'\|_{\mathcal{H}_{k_{\boldsymbol{\lambda}}} \otimes \mathbb{R}^d} \, d\tau(y, y') \\ &\leq \int_{\mathcal{X} \times \mathcal{X}} \|y - y'\|_2 \, d\tau(y, y') + \int_{\mathcal{X} \times \mathcal{X}} \sqrt{2}\|\Sigma_{\boldsymbol{\lambda}}\|_2^{\frac{1}{2}} \|y - y'\|_2 \|y'\|_2 \, d\tau(y, y') \\ &\leq \left(\int_{\mathcal{X} \times \mathcal{X}} \|y - y'\|_2^2 \, d\tau(y, y') \right)^{\frac{1}{2}} + \sqrt{2}\|\Sigma_{\boldsymbol{\lambda}}\|_2^{\frac{1}{2}} \left(\int_{\mathcal{X} \times \mathcal{X}} \|y - y'\|_2^2 \, d\tau(y, y') \right)^{\frac{1}{2}} (\mathbb{E}_{\rho'} \|Y'\|_2^2)^{\frac{1}{2}} \\ &\leq \left(1 + \sqrt{2}\|\Sigma_{\boldsymbol{\lambda}}\|_2^{\frac{1}{2}} (\mathbb{E}_{\rho'} \|Y'\|_2^2)^{\frac{1}{2}}\right) \left(\int_{\mathcal{X} \times \mathcal{X}} \|y - y'\|_2^2 \, d\tau(y, y') \right)^{\frac{1}{2}}.\end{aligned}$$

Since the above inequality holds for any $\tau \in \prod(\rho, \rho')$, it follows that

$$\|K_{\boldsymbol{\lambda}}(\rho - \rho')\|_{\mathcal{H}_{k_{\boldsymbol{\lambda}}} \otimes \mathbb{R}^d} \leq (1 + \sqrt{2}\|\Sigma_{\boldsymbol{\lambda}}\|_2^{\frac{1}{2}} (\mathbb{E}_{\rho'} \|Y'\|_2^2)^{\frac{1}{2}}) W_2(\rho, \rho')$$

and that

$$\begin{aligned}\|\mathrm{I}_{\boldsymbol{\lambda}}(\rho, x) - \mathrm{I}_{\boldsymbol{\lambda}}(\rho', x')\|_{\mathcal{H}_{\mathcal{F}}} &\leq \left(1 + \sqrt{2}\|\Sigma_{\boldsymbol{\lambda}}\|_2^{\frac{1}{2}} (\mathbb{E}_{\rho'} \|Y'\|_2^2)^{\frac{1}{2}}\right) W_2(\rho, \rho') + \sqrt{2}\|\Sigma_{\boldsymbol{\lambda}}\|_2^{\frac{1}{2}} (\mathbb{E}_{\rho'} \|Y'\|_2^2) \|x - x'\|_2 \\ &\leq (1 + 2\sqrt{2}\|\Sigma_{\boldsymbol{\lambda}}\|_2^{\frac{1}{2}} (\mathbb{E}_{\rho'} \|Y'\|_2^2)) d_\Omega((\rho, x), (\rho', x')).\end{aligned}$$

Injection: $k_{\boldsymbol{\lambda}}(x,y) = g_{\boldsymbol{\lambda}}(x-y) = \exp\left\{-\frac{1}{2}(x-y)^T\Sigma_{\boldsymbol{\lambda}}^{-1}(x-y)\right\}$ with $\Sigma_{\boldsymbol{\lambda}} = \mathrm{diag}(2\lambda_1^{-2},\ldots,2\lambda_d^{-2}) \succ 0$. For each $1 \le j \le d$, let

$$\widehat{K_{\boldsymbol{\lambda}}^{(j)}(\rho)}(\omega) = \int_{\mathbb{R}^d} e^{-i\omega^T y} K_{\boldsymbol{\lambda}}^{(j)}(\rho)(y)\, dy = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\omega^T y} g_{\boldsymbol{\lambda}}(y-x)x^{(j)}\, d\rho(x)\, dy.$$

By Fubini Theorem, we have

$$\begin{aligned}
\widehat{K_{\boldsymbol{\lambda}}^{(j)}(\rho)}(\omega) &= \int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d} e^{-i\omega^T y} g_{\boldsymbol{\lambda}}(y-x)\, dy\right] x^{(j)} d\rho(x) \\
&= (2\pi)^{\frac{d}{2}}\det(\Sigma_{\boldsymbol{\lambda}})^{\frac{1}{2}} e^{-\frac{1}{2}\omega^T\Sigma_{\boldsymbol{\lambda}}\omega}\int_{\mathbb{R}^d} x^{(j)} e^{-i\omega^T x}\, d\rho(x) \\
&= (2\pi)^{\frac{d}{2}}\det(\Sigma_{\boldsymbol{\lambda}})^{\frac{1}{2}} e^{-\frac{1}{2}\omega^T\Sigma_{\boldsymbol{\lambda}}\omega}\; i\partial_j \mathrm{F}(\rho)(\omega),
\end{aligned}$$

where $\mathrm{F}(\rho)$ is Fourier transform of probability measure ρ.

It follows that $\widehat{K_{\boldsymbol{\lambda}}(\rho)}(\omega) = (2\pi)^{\frac{d}{2}}\det(\Sigma_{\boldsymbol{\lambda}})^{\frac{1}{2}} e^{-\frac{1}{2}\omega^T\Sigma_{\boldsymbol{\lambda}}\omega}\; i\nabla\mathrm{F}(\rho)(\omega)$. Take $\mu = \rho_1 - \rho_2$ with $\rho_1, \rho_2 \in \mathcal{B}_2(\mathcal{X})$. If $K_{\boldsymbol{\lambda}}(\mu) = 0$, then $K_{\boldsymbol{\lambda}}(\mu)(y) = 0$ for any $y \in \mathbb{R}^d$. It implies that $\widehat{K_{\boldsymbol{\lambda}}(\mu)} \equiv 0$ and $\nabla\mathrm{F}(\mu) \equiv 0$. Note that $\mathrm{F}(\mu)(0) = \mu(\mathcal{X}) = \rho_1(\mathcal{X}) - \rho_2(\mathcal{X}) = 0$. It can be obtained that $\mathrm{F}(\mu) \equiv 0$. Then by the inversion of Fourier transform of measures, we have $\rho_1 = \rho_2$, which shows that $K_{\boldsymbol{\lambda}}$ is an injective mapping on $\mathcal{B}_2(\mathcal{X})$. It also follows that $\mathrm{I}_{\boldsymbol{\lambda}}$ is injective since $x \mapsto k_{\boldsymbol{\lambda}}(x,\cdot)$ is also an injective mapping. ■

Examples for Marginal Meta Probability Measure

EXAMPLE 2. *(Distributions with Compact Support and Bounded Density).*

For $\mathrm{B}, \mathrm{C} > 0$*, we define the probability class* $\mathcal{G}(\mathrm{B},\mathrm{C})$ *of all probability measures with a Lebesgue density bounded by* C *almost surely and supported on the the closed ball of radius* B *centered at zero. Then* $\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}$ *is a probability measure supported on* $\mathcal{G}(\mathrm{B},\mathrm{C})$.

Proof. Take (μ_n) a sequence in $\mathcal{G}(\mathrm{B},\mathrm{C})$ with $\mu_n \to \mu$ in $(\mathcal{B}_2(\mathcal{X}), W_2)$. Denote K the closed ball of radius B centered at zero. Then by Portmanteau Theorem, the weak convergence of measures implies that $1 = \limsup_{n\to\infty}\mu_n(\mathrm{K}) \le \mu(\mathrm{K}) \le 1$. Therefore, $\mu(\mathrm{K}) = 1$.

Also by the weak convergence, we have $\int \varphi d\mu = \lim_{n\to\infty} \int \varphi d\mu_n \leq \mathrm{C} \int \varphi d\nu$ for any $\varphi \in C_c^+(\mathrm{K})$, where ν is Lebesgue measure. Then by the density of function class $C_c^+(\mathrm{K})$, we have $\mu(E) \leq \mathrm{C}\nu(E)$ for any Borel set $E \subset \mathrm{K}$, which follows that μ is absolute continuous with respect to ν and $d\mu/d\nu \leq \mathrm{C}$ almost everywhere on K. It implies that $\mu \in \mathcal{G}(\mathrm{B}, \mathrm{C})$ and then $\mathcal{G}(\mathrm{B}, \mathrm{C})$ is closed in $(\mathcal{B}_2(\mathcal{X}), W_2)$.

It's easy to obtain that

$$\begin{aligned}\mathbb{E}_{\rho\sim\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}} \|\omega_\kappa(\rho)\|^2_{L^\gamma(\rho_\kappa)} &= \int_{\mathcal{G}(\mathrm{B},\mathrm{C})} \left(\int_{\mathrm{K}} \left(\frac{d\rho}{d\rho_\kappa} \right)^\gamma d\rho_\kappa \right)^{\frac{2}{\gamma}} d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho) \\ &\leq \mathrm{C}^2 (2\pi\kappa^2)^{\frac{(\gamma-1)d}{\gamma}} \exp\left(\frac{(\gamma-1)\mathrm{B}^2}{\gamma\kappa^2} \right).\end{aligned}$$

■

EXAMPLE 3. *(Distributions in Diffusion Generative Modeling [94]).*

For $\mathrm{B} > 0$ *and* $0 < t_0 < T < \sqrt{\frac{\gamma}{\gamma-1}}\kappa$*, we define the probability class* $\mathcal{G}_{[t_0,T]}(\mathrm{B})$ *as the collection of the convolutions between two probability distributions*

$$\Big\{ \mu * \rho_{\tilde{\kappa}} : \mu \textit{ supported on the closed ball with radius } \mathrm{B} \textit{ in } \mathbb{R}^d, \\ \textit{Gaussian measure } \rho_{\tilde{\kappa}} \textit{ with } t_0 \leq \tilde{\kappa} \leq T \Big\}$$

with

$$(\mu * \rho_{\tilde{\kappa}})(x) = (2\pi\tilde{\kappa}^2)^{-\frac{d}{2}} \int_{\|y\|_2 \leq \mathrm{B}} \exp\left(-\frac{\|x-y\|^2}{2\tilde{\kappa}^2} \right) d\mu(y).$$

The marginal meta probability measure $\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}$ *is defined as a joint probability measure on* $\mathcal{B}_{2,b}(\mathcal{X}) \times [t_0, T]$ *as* $\tilde{\mathcal{P}}_{\mathcal{G}}^{\mathcal{X}} \times \mathrm{Uniform}[t_0, T]$ *where* $\tilde{\mathcal{P}}_{\mathcal{G}}^{\mathcal{X}}$ *is a probability measure defined on* $\mathcal{B}_{2,b}(\mathcal{X})$.

Proof. Since the convolution between μ and $\rho_{\tilde{\kappa}}$ can be considered as the probability distribution of random variable $X + Z_{\tilde{\kappa}}$ with $X \sim \mu$ and $Z_{\tilde{\kappa}} \sim$ gaussian distribution $\rho_{\tilde{\kappa}}$ independently, $W_2(\mu_1 * \rho_{\tilde{\kappa}}, \mu_2 * \rho_{\tilde{\kappa}}) \leq W_2(\mu_1, \mu_2)$ by the coupling argument. Similarly for $(\mu_1, t_1), (\mu_2, t_2) \in$

$\mathcal{B}_{2,b}(\mathcal{X}) \times [t_0, T]$, we have

$$W_2(\mu_1 * \rho_{t_1}, \mu_2 * \rho_{t_2}) \leq W_2(\mu_1, \mu_2) + W_2(\rho_{t_1}, \rho_{t_2}) \leq W_2(\mu_1, \mu_2) + \sqrt{d}\,|t_1 - t_2|\,,$$

which shows that $\tilde{\mathrm{I}} : (\mu, \tilde{\kappa}) \mapsto \mu * \rho_{\tilde{\kappa}}$ is continuous. Then the meta probability in domain generalization framework can be defined with $\mathcal{P}_{\mathcal{G}}^{\mathcal{X}} = (\tilde{\mathcal{P}}_{\mathcal{G}}^{\mathcal{X}} \times \mathrm{Uniform}[t_0, T]) \circ \tilde{\mathrm{I}}^{-1}$. It's easy to see that for any $\mu * \rho_{\tilde{\kappa}} \in \mathcal{G}_{[t_0,T]}(\mathrm{B})$, we have

$$\begin{aligned}\mathbb{E}\|X + Z\|_2^4 &\leq \mathbb{E}(2\|X\|_2^2 + 2\|Z\|_2^2)^2 = 4(\mathbb{E}\|X\|_2^4 + \mathbb{E}\|Z\|_2^4) + 8(\mathbb{E}\|X\|_2^2)(\mathbb{E}\|Z\|_2^2) \\ &\leq 4(\mathrm{B}^4 + d(d+2)T^4) + 8d\mathrm{B}^2T^2,\end{aligned}$$

and

$$\|\omega_\kappa(\mu * \rho_{\tilde{\kappa}})\|_{L^\gamma(\rho_\kappa)}^\gamma = (2\pi\kappa^2)^{\frac{d(\gamma-1)}{2}} \int_{\mathbb{R}^d} (\mu * \rho_{\tilde{\kappa}}(x))^\gamma \exp\left(\frac{\gamma-1}{2\kappa^2}\|x\|_2^2\right) dx$$

$$\begin{aligned}&= (2\pi\kappa^2)^{\frac{d(\gamma-1)}{2}} (2\pi\tilde{\kappa}^2)^{-\frac{d\gamma}{2}} \\ &\quad \int_{\mathbb{R}^d} \left(\int_{\|y\|_2 \leq \mathrm{B}} \exp\left(-\frac{\|x-y\|_2^2}{2\tilde{\kappa}^2}\right) d\mu(y)\right)^\gamma \exp\left(\frac{\gamma-1}{2\kappa^2}\|x\|_2^2\right) dx \\ &= (2\pi\kappa^2)^{\frac{d(\gamma-1)}{2}} (2\pi\tilde{\kappa}^2)^{-\frac{d\gamma}{2}} \\ &\quad \int_{\mathbb{R}^d} \left(\int_{\|y\|_2 \leq \mathrm{B}} \exp\left(\frac{2x^Ty - \|y\|_2^2}{2\tilde{\kappa}^2}\right) d\mu(y)\right)^\gamma \exp\left(-\left(\frac{\gamma}{2\tilde{\kappa}^2} - \frac{\gamma-1}{2\kappa^2}\right)\|x\|_2^2\right) dx.\end{aligned}$$

It's easy to see that $\|\omega_\kappa(\mu * \rho_{\tilde{\kappa}})\|_{L^\gamma(\rho_\kappa)} < \infty$ if and only if $v_{\tilde{\kappa}} := \frac{\gamma}{2\tilde{\kappa}^2} - \frac{\gamma-1}{2\kappa^2}$ which is equivalent to the condition $\tilde{\kappa} < \sqrt{\frac{\gamma}{\gamma-1}}\kappa$. Moreover, Let $b_{\tilde{\kappa}} = \frac{\gamma}{\tilde{\kappa}^2}$. By Jensen's inequality, we have

$$\begin{aligned}&\|\omega_\kappa(\mu * \rho_{\tilde{\kappa}})\|_{L^\gamma(\rho_\kappa)}^\gamma \\ &\leq \mathcal{A}_{d,\kappa,\gamma}\tilde{\kappa}^{-d\gamma} \int_{\mathbb{R}^d} \int_{\|y\|_2 \leq \mathrm{B}} \exp\left(-\frac{\gamma\|x-y\|_2^2}{2\tilde{\kappa}^2}\right) d\mu(y) \exp\left(\frac{\gamma-1}{2\kappa^2}\|x\|_2^2\right) dx \\ &\leq \mathcal{A}_{d,\kappa,\gamma}\tilde{\kappa}^{-d\gamma} \int_{\|y\|_2 \leq \mathrm{B}} \int_{\mathbb{R}^d} \exp\left(-\Big(\frac{\gamma}{2\tilde{\kappa}^2} - \frac{\gamma-1}{2\kappa^2}\Big)\|x\|_2^2\right) \exp\Big(\frac{2\gamma x^Ty - \gamma\|y\|_2^2}{2\tilde{\kappa}^2}\Big) dx\, d\mu(y) \\ &\leq \mathcal{A}_{d,\kappa,\gamma}\tilde{\kappa}^{-d\gamma} \int_{\|y\|_2 \leq \mathrm{B}} \int_{\mathbb{R}^d} \exp(-v_{\tilde{\kappa}}\|x\|_2^2 + b_{\tilde{\kappa}}y^Tx)\, dx \exp\Big(-\frac{\gamma\|y\|_2^2}{2\tilde{\kappa}^2}\Big) d\mu(y)\end{aligned}$$

$$
\begin{aligned}
&= \mathcal{A}_{d,\kappa,\gamma}\tilde{\kappa}^{-d\gamma}\int_{\|y\|_2\le \mathrm{B}}\int_{\mathbb{R}^d}\exp\Big(-v_{\tilde{\kappa}}\Big\|x-\frac{b_{\tilde{\kappa}}y}{2v_{\tilde{\kappa}}}\Big\|_2^2\Big)\,dx\exp\Big(\frac{b_{\tilde{\kappa}}^2}{4v_{\tilde{\kappa}}}\|y\|_2^2-\frac{\gamma}{2\tilde{\kappa}^2}\|y\|_2^2\Big)\,d\mu(y)\\
&= \mathcal{A}_{d,\kappa,\gamma}\Big(\frac{\pi}{v_{\tilde{\kappa}}}\Big)^{\frac{d}{2}}\tilde{\kappa}^{-d\gamma}\int_{\|y\|_2\le \mathrm{B}}\exp\Big((\frac{b_{\tilde{\kappa}}^2}{4v_{\tilde{\kappa}}}-\frac{\gamma}{2\tilde{\kappa}^2})\|y\|_2^2\Big)d\mu(y)\\
&\le \mathcal{A}_{d,\kappa,\gamma}\Big(\frac{\pi}{v_{\tilde{\kappa}}}\Big)^{\frac{d}{2}}\tilde{\kappa}^{-d\gamma}\exp\left(\frac{\gamma(\gamma-1)\tilde{\kappa}^2}{2(\gamma\kappa^2\tilde{\kappa}^2-(\gamma-1)\tilde{\kappa}^4)}\mathrm{B}^2\right)=:f(\tilde{\kappa})^\gamma,
\end{aligned}
$$

We can observe that $f(\tilde{\kappa})$ is a continuous function on $[t_0,T]$ and $f(\tilde{\kappa})\sim\tilde{\kappa}^{-d\left(1-\frac{1}{\gamma}\right)},\tilde{\kappa}\to 0$. For the domain generalization assumption, we have

$$
\begin{aligned}
\int_{\mathcal{B}_2(\mathcal{X})}\|\omega_\kappa(\rho)\|^2_{L^\gamma(\rho_\kappa)}\,d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho) &= \int_{\mathcal{G}_{[t_0,T]}(\mathrm{B})}\|\omega_\kappa(\rho)\|^2_{L^\gamma(\rho_\kappa)}\,d\mathcal{P}_{\mathcal{G}}^{\mathcal{X}}(\rho)\\
&= \frac{1}{T-t_0}\int_{t_0}^{T}\int_{\mathcal{B}_{2,c}(\mathcal{X})}\|\omega_\kappa(\mu*\rho_{\tilde{\kappa}})\|^2_{L^\gamma(\rho_\kappa)}\,d\tilde{\mathcal{P}}_{\mathcal{G}}^{\mathcal{X}}(\mu)\,d\tilde{\kappa}\\
&\le \frac{1}{T-t_0}\int_{t_0}^{T}f(\tilde{\kappa})^2\,d\tilde{\kappa}\le C_{t_0,T,\gamma,\mathrm{B},d}. \qquad \blacksquare
\end{aligned}
$$

Approximation in Gaussian Space

Optimal Linear Approximation

Note that the space $\mathcal{H}_{k_{\boldsymbol{\lambda}}}$ is actually the tensor product of unvariate RKHS with the kernels $\exp\{-\lambda_l^2(a-b)^2\}$ for $a,b\in\mathbb{R}$, so we first consider the univariate case with $k_{\lambda_1}(a,b)=\exp\{-\lambda_1^2(a-b)^2\}$ where $a,b\in\mathbb{R}$ and the gaussian measure $\rho_{1,\kappa}$ with density function $(2\pi\kappa^2)^{-\frac{1}{2}}\exp\left\{-\frac{a^2}{2\kappa^2}\right\}$.

For $j\ge 1$, the eigenvalues and eigenfunctions are given [18, 81] by

$$
r_{\lambda_1,j}=(\sqrt{2}\kappa)^{-1}\lambda_1^{2j-2}/\mathcal{C}_1^{j-\frac{1}{2}}\text{ where }\mathcal{C}_1=\lambda_1^2+\frac{1}{4\kappa^2}+\frac{1}{2\kappa}\sqrt{\frac{1}{4\kappa^2}+2\lambda_1^2},
$$

and

$$
\tilde{\varphi}_{\lambda_1,j}(a)=\exp\left(-\left(\frac{1}{2\kappa}\sqrt{\frac{1}{4\kappa^2}+2\lambda_1^2}-\frac{1}{4\kappa^2}\right)a^2\right)H_{j-1}\left(\frac{1}{\kappa^{\frac{1}{2}}}\left(\frac{1}{4\kappa^2}+2\lambda_1^2\right)^{\frac{1}{4}}a\right)
$$

where H_{j-1} is the Hermite polynomial of degree $j-1$, given by

$$H_{j-1}(a) = (-1)^{j-1} e^{a^2} \frac{d^{j-1}}{da^{j-1}} e^{-a^2} \text{ for } a \in \mathbb{R}$$

such that

$$\int_{\mathbb{R}} H_{j-1}^2(a) \exp(-a^2)\, da = \sqrt{\pi} 2^{j-1} (j-1)! \text{ for } j \in \mathbb{N}.$$

Then we can take a orthonormal basis of $L^2(\rho_{1,\kappa})$ to be $\{\varphi_{\lambda_1,j}\}_{j\in\mathbb{N}}$ by

$$\varphi_{\lambda_1,j}(a) = \sqrt{\frac{(1+8\kappa^2\lambda_1^2)^{\frac{1}{4}}}{2^{j-1}(j-1)!}} \exp\left(-\frac{2\lambda_1^2 a^2}{\sqrt{1+8\kappa^2\lambda_1^2}+1}\right) H_{j-1}\left(\frac{1}{\sqrt{2}\kappa}(1+8\kappa^2\lambda_1^2)^{\frac{1}{4}} a\right),$$

and observe that $(\sqrt{2}\kappa)^{-1}\mathcal{C}_1^{-\frac{1}{2}} = 1 - \frac{\lambda_1^2}{\mathcal{C}_1}$, which allows us to rewrite $r_{\lambda_1,j}$ as $r_{\lambda_1,j} = (1-\eta_{\lambda_1})\eta_{\lambda_1}^{j-1}$ with

$$\eta_{\lambda_1} = \frac{\lambda_1^2}{\mathcal{C}_1} = \frac{4\kappa^2\lambda_1^2}{4\kappa^2\lambda_1^2 + 1 + \sqrt{1+8\kappa^2\lambda_1^2}} \in (0,1).$$

For the multivariate case with $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$, let $\boldsymbol{j}$ be a multi-index with $\boldsymbol{j} = (j_1, \ldots, j_d) \in \mathbb{N}^d$. Then the pairs $(r_{\boldsymbol{j}}^{\boldsymbol{\lambda}}, \varphi_{\boldsymbol{j}}^{\boldsymbol{\lambda}})$ of eigenvalues and eigenfunctions are given by

$$r_{\boldsymbol{j}}^{\boldsymbol{\lambda}} := \prod_{l=1}^{d} r_{\lambda_l,j_l} = \prod_{l=1}^{d} (1-\eta_{\lambda_l})\eta_{\lambda_l}^{j_l-1} \text{ and } \varphi_{\boldsymbol{j}}^{\boldsymbol{\lambda}}(x) := \prod_{l=1}^{d} \varphi_{\lambda_l,j_l}(x^{(l)}) \text{ for } x = [x^{(1)}, ..., x^{(d)}] \in \mathbb{R}^d.$$

We can also define an orthonormal basis $(\psi_{\boldsymbol{j}}^{\boldsymbol{\lambda}})$ on $\mathcal{H}_{k_{\boldsymbol{\lambda}}}$ by

$$\psi_{\boldsymbol{j}}^{\boldsymbol{\lambda}}(x) := \prod_{l=1}^{d} \psi_{\lambda_l,j_l}(x^{(l)}) \text{ where } \psi_{\lambda_l,j_l} := \sqrt{r_{\lambda_l,j_l}}\varphi_{\lambda_l,j_l}.$$

For the simplicity of notation, we rearrange the sequence of eigenpairs $(r_{\boldsymbol{j}}^{\boldsymbol{\lambda}}, \psi_{\boldsymbol{j}}^{\boldsymbol{\lambda}})_{\boldsymbol{j}\in\mathbb{N}^d}$ to the sequence $(r_q^{\boldsymbol{\lambda}}, \psi_q^{\boldsymbol{\lambda}})_{q\in\mathbb{N}}$ with the order of a non-increasing sequence of eigenvalues, i.e., $r_1^{\boldsymbol{\lambda}} \geq r_2^{\boldsymbol{\lambda}} \geq \cdots > 0$.

By Corollary 4.12 in [64], the optimal linear approximation error is

$$\mathcal{E}(n, \mathcal{H}_{k_{\boldsymbol{\lambda}}}) := \inf_{\Lambda_n \subset \mathcal{H}_{k_{\boldsymbol{\lambda}}}} \sup_{\|f\|_{\mathcal{H}_{k_{\boldsymbol{\lambda}}}} \leq 1} \left\| f - \mathrm{Proj}_{\Lambda_n}(f) \right\|_{L^2(\rho_\kappa)} = \sqrt{r_{n+1}^{\boldsymbol{\lambda}}}$$

where Λ_n is an n-dimensional subspace of $\mathcal{H}_{k_{\boldsymbol{\lambda}}}$. By Theorem 5.2 in [18], $\mathcal{E}(n, \mathcal{H}_{k_{\boldsymbol{\lambda}}}) \leq C_{\delta,\kappa,\theta} n^{-\max\left(\theta, \frac{1}{2}\right)+\delta}$ for any $\delta > 0$ where $C_{\delta,\kappa,\theta}$ only depends on δ, κ and θ. ■

Approximation of Eigenfunctions by Two-Hidden-Layer Tanh Neural Networks

In this part, we show $L^2(\rho)$-approximation of orthonormal basis defined in Appendix 4.5.3 by neural networks where ρ is a probability distribution with $\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)} < \infty$.

Recall that $\psi_j^{\boldsymbol{\lambda}}$ has a product form of factors being elements with a unit norm in $\mathcal{H}_{k_{\lambda_l}}$. To approximate analytic functions with this form, we apply the shallow neural network with and tanh activation functions and a product-gated output defined in (4.1).

By scaling and translating the variable, for each pair (λ_l, j_l), we define $g_{\lambda_l,j_l}(t) := \psi_{\lambda_l,j_l}\left(2B\left(t - \frac{1}{2}\right)\right)$ for $t \in [0,1]$ with some number $B > 0$. Here we introduce the class of (Q,R)-analytic functions with $Q, R > 0$ in which an analytic function f satisfies the smoothness condition that $\|D^\beta f\|_{L^\infty([0,1]^d)} \leq QR^{-\beta}\beta!$ for all $\beta \in \mathbb{N}$.

By Theorem 1 in [117], for each $\psi_{\lambda_l,j_l} \in \mathcal{H}_{k_{\lambda_l}}$, we have that

$$\begin{aligned}\left|D^\beta g_{\lambda_l,j_l}(t)\right| &= \left|(2B)^\beta D^\beta \psi_{\lambda_l,j_l}\left(2B\left(t - 1/2\right)\right)\right| = \left|(2B)^\beta \left\langle (D^\beta k_{\lambda_j})_{2B\left(t-\frac{1}{2}\right)}, \psi_{\lambda_l,j_l} \right\rangle_{\mathcal{H}_{k_{\lambda_l}}}\right| \\ &\leq (2B)^\beta \sqrt{D^{(\beta,\beta)} k_{\lambda_l}(x,x)} \leq (4B\lambda_l)^\beta \beta! \leq (4BC_\theta)^\beta \beta!.\end{aligned}$$

It implies that g_{λ_l,j_l} is a $(1, (4BC_\theta)^{-1})$-analytic function for each pair (λ_l, j_l).

Indeed, for $k_{\lambda_l}(a,b) = \exp\{-\lambda_l^2(a-b)^2\}$,

$$D^{(\beta,\beta)} k_{\lambda_l} = \partial_a^\beta \partial_b^\beta k_{\lambda_l} = (-\lambda_l^2)^\beta \partial_c^{2\beta} \exp\{-c^2\} \text{ with } c = \lambda_l(a-b).$$

By the definition of Hermite polynomials, $\partial_c^{2\beta} \exp(-c^2) = H_{2\beta}(c) \exp(-c^2)$. It follows that

$$\partial_c^{2\beta} \exp(-c^2)|_{c=0} = H_{2\beta}(0) = (-1)^\beta \frac{(2\beta)!}{\beta!}$$

which is called Hermite numbers of the even order. Then we have

$$\sqrt{D^{(\beta,\beta)} k_{\lambda_l}(x,x)} \leq \lambda_l^\beta \sqrt{\frac{(2\beta)!}{\beta!}} = \lambda_l^\beta \sqrt{\binom{2\beta}{\beta} \beta!} \leq (2\lambda_l)^\beta \beta!,$$

which proves the claim with $Q = 1, R = (4BC_\theta)^{-1}$.

The following lemma follows from an application of Theorem B.7 in [13] and Corollary 5.5 in [12] by taking $s = 4\tilde{m}, N = \tilde{m}$.

LEMMA 12. *For $B > 1$, each $g_{\lambda_l,j_l}(t) = \psi_{\lambda_l,j_l}\left(2B\left(t - \frac{1}{2}\right)\right)$ on $[0,1]$. For $\tilde{m} > 3$, There exists a tanh neural network $\hat{g}^{\tilde{m}}_{\lambda_l,j_l}$ with two hidden layers of width at most $8\tilde{m}$ such that*

$$\|g_{\lambda_l,j_l} - \hat{g}^{\tilde{m}}_{\lambda_l,j_l}\|_{L^\infty([0,1])} \leq 2 \exp\left(-4\tilde{m} \log\left(\frac{\tilde{m}}{6BC_\theta}\right)\right)$$

with the parameters bounded by $c_1'(c_2'\tilde{m})^{160\tilde{m}^2}$ where c_1', c_2' are two absolute constants.

We define $\hat{\psi}^{\tilde{m}}_{\lambda_l,j_l}(t) = \hat{g}^{\tilde{m}}_{\lambda_l,j_l}\left(\frac{t}{2B} + \frac{1}{2}\right)$ and recall the product gate $\mathcal{T}_{1,\odot}$ defined as

$$\mathcal{T}_{1,\odot}(x) = \prod_{l=1}^d \mathcal{T}_1(x_l) \text{ with } \mathcal{T}_1(x_l) = x_l \text{ if } |x_l| < 1 \text{ otherwise } \frac{x_l}{|x_l|}$$

($\mathcal{T}_1$ can be also implemented by a fixed ReLU neural network as $\sigma(x_l+1)-\sigma(x_l-1)-1$). Then we can construct an approximant for $\psi^{\boldsymbol{\lambda}}_{\boldsymbol{j}} = \prod_{l=1}^d \psi_{\lambda_l,j_l}$ by $\hat{\psi}^{\boldsymbol{\lambda}}_{\boldsymbol{j},\tilde{m}} := \mathcal{T}_{1,\odot}\big((\hat{\psi}^{\tilde{m}}_{\lambda_1,j_1}, \ldots, \hat{\psi}^{\tilde{m}}_{\lambda_d,j_d})\big)$ with $L^2(\rho)$ approximation error

$$\begin{aligned}\left\|\psi^{\boldsymbol{\lambda}}_{\boldsymbol{j}} - \hat{\psi}^{\boldsymbol{\lambda}}_{\boldsymbol{j},\tilde{m}}\right\|^2_{L^2(\rho)} &= \left(\int_{\|x\|_\infty \leq B} + \int_{\|x\|_\infty > B}\right) (\psi^{\boldsymbol{\lambda}}_{\boldsymbol{j}}(x) - \hat{\psi}^{\boldsymbol{\lambda}}_{\boldsymbol{j},\tilde{m}}(x))^2 \, d\rho(x) \\ &\leq \sup_{\|x\|_\infty \leq B} (\psi^{\boldsymbol{\lambda}}_{\boldsymbol{j}}(x) - \hat{\psi}^{\boldsymbol{\lambda}}_{\boldsymbol{j},\tilde{m}}(x))^2 + 2\rho(\{x : \|x\|_\infty > B\}).\end{aligned}$$

For the first term, we bound it with Lemma 12 by introducing intermediate terms as follows:

$$\begin{aligned}
&\sup_{\|x\|_\infty\le B}\left|\psi^{\boldsymbol{\lambda}}_{\boldsymbol{j}}(x)-\hat{\psi}^{\boldsymbol{\lambda}}_{\boldsymbol{j},\tilde{m}}(x)\right| = \Big\|\prod_{l=1}^{d}\psi_{\lambda_l,j_l}-\prod_{l=1}^{d}\mathcal{T}_1(\hat{\psi}^{\tilde{m}}_{\lambda_l,j_l})\Big\|_{L^\infty([-B,B]^d)}\\
\le&\Big\|\prod_{l=1}^{d}\psi_{\lambda_l,j_l}-\mathcal{T}_1(\hat{\psi}^{\tilde{m}}_{\lambda_1,j_1})\prod_{l=2}^{d}\psi_{\lambda_l,j_l}+\cdots+\prod_{l'=1}^{h}\mathcal{T}_1(\hat{\psi}^{\tilde{m}}_{\lambda_{l'},j_{l'}})\prod_{l=h+1}^{d}\psi_{\lambda_l,j_l}-\prod_{l'=1}^{h+1}\mathcal{T}_1(\hat{\psi}^{\tilde{m}}_{\lambda_{l'},j_{l'}})\prod_{l=h+2}^{d}\psi_{\lambda_l,j_l}\\
&+\cdots+\prod_{l'=1}^{d-1}\mathcal{T}_1(\hat{\psi}^{\tilde{m}}_{\lambda_l,j_l})\psi_{\lambda_d,j_d}-\prod_{l=1}^{d}\mathcal{T}_1(\hat{\psi}^{\tilde{m}}_{\lambda_l,j_l})\Big\|_{L^\infty([-B,B]^d)}\\
\le& d\max_{0\le h\le d-1}\left\|\prod_{l'=1}^{h}\mathcal{T}_1(\hat{\psi}^{\tilde{m}}_{\lambda_{l'},j_{l'}})\prod_{l=h+1}^{d}\psi_{\lambda_l,j_l}-\prod_{l'=1}^{h+1}\mathcal{T}_1(\hat{\psi}^{\tilde{m}}_{\lambda_{l'},j_{l'}})\prod_{l=h+2}^{d}\psi_{\lambda_l,j_l}\right\|_{L^\infty([-B,B]^d)}\\
\le& d\max_{0\le h\le d-1}\|\psi_{\lambda_{h+1},j_{h+1}}-\mathcal{T}_1(\hat{\psi}^{\tilde{m}}_{\lambda_{h+1},j_{h+1}})\|_{L^\infty([-B,B])}\le d\max_{0\le h\le d-1}\|\psi_{\lambda_{h+1},j_{h+1}}-\hat{\psi}^{\tilde{m}}_{\lambda_{h+1},j_{h+1}}\|_{L^\infty([-B,B])}\\
\le& 2d\exp\left(-4\tilde{m}\log\left(\frac{\tilde{m}}{6BC_\theta}\right)\right).
\end{aligned}$$

For the second term, we bound it by the subgaussian tail decay of probability measures:

$$\begin{aligned}
\rho(\{x:\|x\|_\infty>B\}) &= \int_{\|x\|_\infty>B}\omega_\kappa(\rho)(x)\,d\rho_\kappa(x)\le\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}(\rho_\kappa(\{x:\|x\|_\infty>B\}))^{\frac{\gamma-1}{\gamma}}\\
&\le\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}\left(d\cdot\frac{\kappa}{\sqrt{2\pi}}B^{-1}\exp\left(-\frac{B^2}{2\kappa^2}\right)\right)^{\frac{\gamma-1}{\gamma}}\\
&\le C_{\kappa,\gamma}d\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}\exp\left(-\frac{\gamma-1}{\gamma}\Big(\frac{B^2}{2\kappa^2}+\log B\Big)\right)
\end{aligned}$$

with $C_{\kappa,\gamma}=\left(\frac{\kappa}{\sqrt{2\pi}}\right)^{\frac{\gamma-1}{\gamma}}$. Let $B=\frac{\tilde{m}^{\frac{3}{4}}}{6C_\theta}$ and the above upper bound can written as

$$\begin{aligned}
\rho(\{x:\|x\|_\infty>B\}) &\le dC_{\kappa,\gamma}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}\exp\left(-\frac{\gamma-1}{\gamma}\Big(\frac{\tilde{m}^{\frac{3}{2}}}{72\kappa^2C_\theta^2}+\frac{3}{4}\log\tilde{m}-\log 6C_\theta\Big)\right)\\
&\le dC_{\kappa,\gamma}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}\exp(-2\tilde{m}\log\tilde{m})
\end{aligned}$$

when $\tilde{m}>C_{\kappa,\theta,\gamma}$ with $C_{\kappa,\theta,\gamma}$ a constant only depending on κ,γ and C_θ.

Then combine two estimations and we can the final bound as

$$\begin{aligned}\|\psi_{\boldsymbol{j}}^{\boldsymbol{\lambda}} - \hat{\psi}_{\boldsymbol{j},\tilde{m}}^{\boldsymbol{\lambda}}\|_{L^2(\rho)}^2 &\leq (2d)^2 \exp(-2\tilde{m}\log\tilde{m}) + 2dC_{\kappa,\gamma}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)} \exp(-2\tilde{m}\log\tilde{m}) \\ &\leq (4d^2 + C_{\kappa,\gamma}\|\omega_\kappa(\rho)\|_{L^\gamma(\rho_\kappa)}) \exp(-2\tilde{m}\log\tilde{m})\end{aligned}$$

for $\tilde{m} > C_{\kappa,\theta,\gamma}$. ■

Chapter 5

Conclusion

Through this thesis, we study the approximation and generalization properties of transformer models in a systematic way. One of the main starting points is to model context information as the discretization of an underlying probability distribution. Under this viewpoint, the storage of key–value cache at the computational level can be related to the kernel embedding of probability distributions, which helps connect the computational mechanism of the model with an interpretable mathematical framework. Based on this perspective, Chapter 2 suggests that when pretraining samples are more diverse, the model is in a better position to learn the underlying target functional. At the same time, the analysis remains affected by the curse of dimensionality, which appears to be somewhat inconsistent with the empirical scaling behavior often observed in practice. This issue is particularly relevant because the token embedding dimension d is itself a tunable parameter.

Chapter 3 then examines this issue further within the kernel-embedding framework. Our analysis shows that low-rank and sparsified network architectures can capture the rapidly decaying singular-value structure that often appears in weight matrices, and in this way substantially alleviate the curse of dimensionality. This phenomenon is especially important for Fourier functional networks, where the decay structure plays a fundamental role in determining effective approximation classes. It is also through this perspective that the connection between transformers and neural operators becomes considerably more transparent: both can be understood as architectures designed to approximate mappings between structured function spaces, with compression, spectral decay, and operator structure playing essential roles in their efficiency.

Finally, in Chapter 4, we turn to Transformer encoders of the BERT type and abstract them as mappings from classes of probability distributions to classes of response functions. This abstraction provides a natural mathematical framework for understanding the essence of in-context learning. In this formulation, in-context learning is no longer viewed merely as an empirical phenomenon tied to prompting, but rather as a structured functional mapping from distributional representations of context to task-dependent responses. Taken together, the thesis presents a unified perspective in which kernel embedding, operator learning, and transformer architectures are linked within a common mathematical framework, offering a systematic explanation of how transformers approximate, generalize, and adapt to contextual information.

Prior studies [19, 1, 115, 88] often formulate in-context learning as predicting the label of a given sample conditioned an input prompt containing other samples and labels. As noted in Zhang et al. [115], a model Φ performs in-context learning as

$$\Phi : \mathcal{S} \times \mathcal{X} \to \mathcal{Y}, \quad \mathcal{S} = \cup_{n\in\mathbb{N}}\big\{(x_1, y_1, ..., x_n, y_n) : x_i \in \mathcal{X}, y_i \in \mathcal{Y}\big\}$$

where $\mathcal{X}$ is the input space and $\mathcal{Y}$ the output space, and Φ is trained on prompts of the form $\mathscr{P} = (x_1, h(x_1), ..., x_\vartheta, h(x_\vartheta), x_{\text{query}})$ with $h \sim \mathcal{P}$ a distribution defined on a function space H to minimize the error $\mathbb{E}_{\mathscr{P}} l(\Phi(\mathscr{P}), h(x_{\text{query}}))$ with a loss function l. Previous theoretical work has focused on linear function spaces [115] and Hölder spaces [88]. These studies have demonstrated that transformers can perform well on structured prompts of input-output pairs, but this formulation has two limitations to bridge the gap between theory and application [61]. First, in-context learning emerges as a property of LLMs after pretraining on tasks like autoregression or diffusion-based generation. A pretrained LLM can perform in-context learning without any parameter updates [105], which is not consistent with theoretical settings that require training on structured prompts. Second, prompts for in-context learning are often unstructured and may lack labels. For example, in machine translation from English to French, the input prompt may contain only instructions in English. Empirical studies [61] also show that the correct mapping between inputs and true labels in prompts has little performance gains for in-context learning: model performance with random labels closely matches that with true labels.

We address these problems by the domain generalization framework [4, 5] and formulate in-context learning as operator learning with the two-staged sampling process:

$$\Phi : \hat{\rho}_X^{(i)} \mapsto (h : \mathcal{X} \to \mathcal{Y}), \quad \hat{\rho}_X^{(i)} = \delta([x_{i1}, ..., x_{in_i}]) \text{ with } x_{ij} \in \mathcal{X}. \tag{5.1}$$

This formulation suggests that the operator Φ maps the context distribution $\hat{\rho}_X^{(i)}$ to a response function $h_{\hat{\rho}_X^{(i)}}$ that takes queries from $\mathcal{X}$ and outputs $h_{\hat{\rho}_X^{(i)}}(x_{\text{query}})$ for any $x_{\text{query}} \in \mathcal{X}$, which aligns with both the nature of Transformers as context-based representation learning and also the parameter-freezing setting after pretraining for in-context learning. This operator-learning viewpoint enables us connect the pretrainng stage with in-context learning capacity. With a richer unstructured prompt $[x_{i1}, ..., x_{in}]$ by more and more samplings ($X_{ij} \sim \rho_X^{(i)}$) from the ground truth context distribution $\rho_X^{(i)}$, the empirical context distribution $\hat{\rho}_X^{(i)}$ can recover $\rho_X^{(i)}$ and then $\hat{\Phi}(\hat{\rho}_X^{(i)})$ can well approximate $\Phi(\rho_X^{(i)})$ without parameter updates, where $\hat{\Phi}$ is a pretrained Transformer model to approximate the operator Φ.

Bibliography

[1] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

[2] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

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

[4] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[5] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of machine learning research*, 22(2):1–55, 2021.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

[8] Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[9] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.

[10] Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*, volume 24. Cambridge University Press, 2007.

[11] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[12] Tim De Ryck, Samuel Lanthaler, and Siddhartha Mishra. On the approximation of functions by tanh neural networks. *Neural Networks*, 143:732–750, 2021.

[13] Tim De Ryck, Ameya D Jagtap, and Siddhartha Mishra. Error estimates for physics-informed neural networks approximating the navier–stokes equations. *IMA Journal of Numerical Analysis*, page drac085, 2023.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[15] Josef Dick, Peter Kritzer, Friedrich Pillichshammer, and Henryk Woźniakowski. Approximation of analytic functions in korobov spaces. *Journal of Complexity*, 30(2): 2–28, 2014.

[16] David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. In *International Conference on Learning Representations (Workshop)*, 2014.

[17] Zhiying Fang, Zheng-Chu Guo, and Ding-Xuan Zhou. Optimal learning rates for distribution regression. *Journal of Complexity*, 56:101426, 2020.

[18] Gregory E. Fasshauer, Fred J. Hickernell, and Henryk Woźniakowski. On dimension-independent rates of convergence for function approximation with gaussian kernels. *SIAM Journal on Numerical Analysis*, 50(1):247–271, 2012.

[19] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598, 2022.

[20] Paul Geuchen and Felix Voigtlaender. Optimal approximation using complex-valued neural networks. *Advances in Neural Information Processing Systems*, 36:1681–1737, 2023.

[21] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[22] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

[23] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2022.

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[25] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[26] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[27] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10, 2022.

[28] Markus Holzleitner, Sergei V. Pereverzyev, and Werner Zellinger. Domain generalization by functional regression. *Numerical Functional Analysis and Optimization*, 45(3): 259–281, 2024.

[29] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[30] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2022.

[31] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[32] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Kumar Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.

[33] Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation. *ACM Transactions on Software Engineering and Methodology*, 35(2):1–72, 2026.

[34] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm, 2019.

[35] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[36] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws

for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[37] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6): 422–440, 2021.

[38] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020.

[39] Jason M. Klusowski and Andrew R. Barron. Approximation by combinations of relu and squared relu ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Transactions on Information Theory*, 64(12):7649–7656, 2018.

[40] Yury Korolev. Two-layer neural networks with values in a banach space. *SIAM Journal on Mathematical Analysis*, 54(6):6358–6389, 2022.

[41] Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for fourier neural operators. page 76.

[42] Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 1998.

[43] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

[44] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991. ISBN 978-3-642-20211-7 978-3-642-20212-4.

[45] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021.

[46] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023.

[47] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel

network for partial differential equations. 2020.

[48] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations, 2021.

[49] Peilin Liu and Ding-Xuan Zhou. Generalization analysis of transformers in distribution regression. *Neural Computation*, 37(2):260–293, 2025.

[50] Peilin Liu and Ding-Xuan Zhou. Approximation of functionals on korobov spaces with fourier functional networks. *Under Review at Journal of Machine Learning Research*, 2025.

[51] Peilin Liu, Yuqing Liu, Xiang Zhou, and Ding-Xuan Zhou. Approximation of functionals on korobov spaces with fourier functional networks. *Neural Networks*, 182:106922, 2025.

[52] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[53] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[54] Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in rkhs-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.

[55] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. 1999.

[56] Tong Mao and Ding-Xuan Zhou. Approximation of functions from korobov spaces by deep convolutional neural networks. *Advances in Computational Mathematics*, 48(6): 84, 2022.

[57] Tong Mao, Zhongjie Shi, and Ding-Xuan Zhou. Approximating functions with multi-features by deep convolutional neural networks. *Analysis and Applications*, 21(01): 93–125, 2023.

[58] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

[59] Andreas Maurer and Massimiliano Pontil. Concentration inequalities under sub-gaussian and sub-exponential conditions. In *Advances in Neural Information Processing Systems*, volume 34, pages 7588–7597. Curran Associates, Inc., 2021.

[60] H. N. Mhaskar. Approximation properties of a multilayered feedforward artificial neural network. *Advances in Computational Mathematics*, 1(1):61–80, 1993.

[61] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, 2022.

[62] Hadrien Montanelli and Qiang Du. New error bounds for deep relu networks using sparse grids. *SIAM Journal on Mathematics of Data Science*, 1(1):78–92, 2019.

[63] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[64] Erich Novak and Henryk Woźniakowski. *Tractability of Multivariate Problems. 1: Linear Information*. Number 6. European Mathematical Soc, Zürich, 2008. ISBN 978-3-03719-026-5.

[65] Erich Novak, Ian H. Sloan, and Henryk Wo'zniakowski. Tractability of approximation for weighted korobov spaceson classical and quantum computers. *Foundations of Computational Mathematics*, 4(2):121–156, 2004.

[66] Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

[67] OpenAI. GPT-4 technical report, 2023.

[68] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[69] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[70] Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks, 2018.

[71] Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195, 1999.

[72] Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *Proceedings of machine learning and systems*, 5:606–624, 2023.

[73] Zhen Qin, Xiaodong Han, Weixuan Sun, Dongxu Li, Lingpeng Kong, Nick Barnes, and Yiran Zhong. The devil in linear transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7025–7041, 2022.

[74] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohao Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosFormer: Rethinking softmax in attention. In *International Conference on Learning Representations*, 2022.

[75] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. .

[76] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. .

[77] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

[78] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67, 2020.

[79] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[80] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[81] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Mass, 2006. ISBN 978-0-262-18253-9.

[82] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.

[83] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4), 2020.

[84] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792): 706–710, 2020.

[85] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[86] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[87] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017.

[88] Zhaiming Shen, Alexander Hsu, Rongjie Lai, and Wenjing Liao. Understanding in-context learning on structured manifolds: Bridging attention to kernel methods. *arXiv preprint arXiv:2506.10959*, 2025.

[89] Zhongjie Shi, Zhan Yu, and Ding-Xuan Zhou. Learning theory of distribution regression with neural networks. *Constructive Approximation*, 62(1):61–104, 2025.

[90] Jonathan W Siegel and Jinchao Xu. Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *Foundations of Computational Mathematics*, 24(2):481–537, 2024.

[91] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

[92] Linhao Song, Jun Fan, Di-Rong Chen, and Ding-Xuan Zhou. Approximation of nonlinear functionals using deep relu networks. *Journal of Fourier Analysis and Applications*, 29(4):50, 2023.

[93] Linhao Song, Ying Liu, Jun Fan, and Ding-Xuan Zhou. Approximation of smooth functionals using deep relu networks. *Neural Networks*, 166:424–436, 2023.

[94] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[95] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert R G Lanckriet. Universality, characteristic kernels and rkhs embedding of measures.

[96] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

[97] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

[98] Zoltan Szabo, Bharath K Sriperumbudur, Barnabas Poczos, and Arthur Gretton. Learning theory for distribution regression.

[99] Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4344–4353, 2019.

[100] Tim van Erven and Peter Harremoës. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[101] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[102] Felix Voigtlaender. The universal approximation theorem for complex-valued neural networks. *Applied and Computational Harmonic Analysis*, 64:33–61, 2023.

[103] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 1 edition, February 2019. ISBN 978-1-108-62777-1 978-1-108-49802-9.

[104] Terry Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.

[105] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

[106] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

[107] Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Forty-first International Conference on Machine Learning*, 2024.

[108] Yunfei Yang and Ding-Xuan Zhou. Optimal rates of approximation by shallow relu k neural networks and applications to nonparametric regression. *Constructive Approximation*, 62(2):329–360, 2025.

[109] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94, 2016.

[110] Zhan Yu and Ding-Xuan Zhou. Deep learning theory of distribution regression with cnns. *Advances in Computational Mathematics*, 49(4):51, 2023.

[111] Zhan Yu, Daniel W. C. Ho, Zhongjie Shi, and Ding-Xuan Zhou. Robust kernel-based distribution regression. *Inverse Problems*, 37(10):105014, 2021.

[112] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc., 2020.

[113] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

[114] Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Re. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. In *The Twelfth International Conference on Learning Representations*, 2024.

[115] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

[116] Ding-Xuan Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49(7):1743–1752, 2003.

[117] Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1):456–463, 2008.

[118] Ding-Xuan Zhou. Deep distributed convolutional neural networks: Universality. *Analysis and Applications*, 16(06):895–919, 2018.

[119] Ding-Xuan Zhou. Theory of deep convolutional neural networks: Downsampling. *Neural Networks*, 124:319–327, 2020.

[120] Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020.

[121] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *arXiv:2012.07436 [cs]*, 2020.

[122] Tian-Yi Zhou, Namjoon Suh, Guang Cheng, and Xiaoming Huo. Approximation of RKHS functionals by neural networks. *arXiv preprint arXiv:2403.12187*, 2024.