Title
Generalization of Wide Neural Networks from the Perspective of Linearization and Kernel Learning

Author
Jin, Hui

Publication Date
2022

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Generalization of Wide Neural Networks from the Perspective of Linearization and Kernel Learning

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mathematics

by

Hui Jin

2022

Abstract of the Dissertation

Generalization of Wide Neural Networks from the
Perspective of Linearization and Kernel Learning

by

Hui Jin
Doctor of Philosophy in Mathematics
University of California, Los Angeles, 2022
Professor Guido Francisco Montúfar Cuartas, Chair

Recently people showed that wide neural networks can be approximated by linear models under gradient descent [JGH18a, LXS19a]. In this dissertation we study generalization of wide neural networks by the linearization of the network, thus some result from kernel learning can directly apply [SH02, CD07]. In Chapter 2, we investigate gradient descent training of wide neural networks and the corresponding implicit bias in function space. We approximate the wide neural networks by corresponding linearized models and show that the implicit bias can be characterized by certain interpolating splines, thus we can use the approximation theory of splines to study the generalization of wide neural networks. In Chapter 3, we show that the decay rate of generalization error of Gaussian Process Regression is determined by the decay rate of the eigenspectrum of the prior and the eigenexpansion coefficients of the target function. This result can be applied to study the generalization error of infinitely wide neural networks with ReLU activations. Since the asymptotic generalization error is closely related to the asymptotic spectrum of the kernel, in Chapter 4 we study the asymptotic spectrum of the Neural Tangent Kernel (NTK) by its power series expansion. We first show that under certain assumptions, the NTK of deep feedforward networks in the infinite width limit can be expressed as a power series. Later on we show that the eigenvalues of the NTK can be expressed the coefficients of the power series. From this expression we show that the decay rate of the eigenvalues is determined by the decay rate of the power series coefficients.

The dissertation of Hui Jin is approved.

Stanley J. Osher

Deanna M. Hunter

Quanquan Gu

Guido Francisco Montúfar Cuartas, Committee Chair

University of California, Los Angeles

2022

Table of Contents

1 Introduction **1**

2 Implicit Bias of Gradient Descent for Mean Squared Error Regression with Two-Layer Wide Neural Network **4**

2.1 Introduction 4

2.2 Notations and Problem Setup 6

2.3 Main Results 8

2.3.1 Univariate Regression 8

2.3.2 Multivariate Regression 10

2.3.3 Discussion of the Main Results 13

2.4 Wide Networks and Parameter Space 19

2.4.1 Implicit Bias in Parameter Space for a Linearized Model 19

2.4.2 Training Only the Output Layer Approximates Training All Parameters 20

2.5 Infinite Width Limit of Shallow Networks 22

2.6 Implicit Bias for Univariate Regression 23

2.7 Implicit Bias for Multivariate Regression 25

2.8 Conclusion 30

Appendices 31

2.A Numerical Illustration of the Theoretical Results 32

2.B Additional Background on the NTK, Initialization, and Parametrization . . . 39

2.B.1 NTK Convergence and Positive-definiteness 39

2.B.2 Anti-Symmetrical Initialization (ASI) 41

2.B.3 Standard vs NTK Parametrization 41

2.B.4 Weight Norm Minimization 42
2.B.5 Basis Parameter for Linearization of the Model 44
2.C Proof of Theorem 1 and Theorem 6 44
2.D Implicit Bias in Parameter Space for a Linearized Model 47
2.E Proof of Theorem 10 52
2.F Training Only the Output Layer Approximates Training a Wide Shallow Network 60
2.G Proof of Theorem 12 62
2.H Proofs of Results for Univariate Regression 67
2.H.1 Proof of Theorem 13 67
2.H.2 Proof of Proposition 14 and Remarks to Proposition 15 71
2.H.3 Proof of Theorem 2 73
2.I Proofs of Results for Multivariate Regression 74
2.I.1 Proof of Theorem 16 74
2.I.2 Proof of Theorem 7 84
2.I.3 Proof of Theorem 8 86
2.I.4 Explicit Form of the Curvature Penalty Function 95
2.J Other Activation Functions for Univariate Regression 97
2.K Effect of Linear Adjustment of the Training Data 99
2.L Neural Networks with Skip Connections 104
2.M Equivalence of Our Characterization and NTK Norm Minimization for Univariate Regression 112
2.N Gradient Descent Trajectory and Trajectory of Smoothing Splines for Univariate Regression 119
2.N.1 Regularized Regression and Early Stopping 119

2.N.2 Trajectory of Smoothing Splines with Uniform Curvature Penalty . . 125
2.N.3 Trajectory of Spatially Adaptive Smoothing Splines 126
2.O Solution to the Variational Problems for Univariate Regression after Training 127
2.O.1 Interpolating Splines with Uniform Curvature Penalty 127
2.O.2 Spatially Adaptive Interpolating Splines 128
2.P Possible Generalizations 130
2.P.1 Deep Networks and Other Architectures 130
2.P.2 Other Loss Functions 130
2.P.3 Other Optimization Procedures 131

3 Learning Curves for Gaussian Process Regression with Power-law Priors and Targets 132

3.1 Introduction 132
3.2 Bayesian Learning and Generalization Error for GPs 136
3.3 Asymptotic Analysis of GP Regression with Power-law Priors 139
3.3.1 Notations and Assumptions 139
3.3.2 Asymptotics of the Normalized Stochastic Complexity 141
3.3.3 Asymptotics of the Bayesian Generalization Error 144
3.3.4 Asymptotics of the Excess Mean Squared Error 145
3.4 Experiments 147
3.5 Conclusion 149
Appendices 150
3.A Experiments for Arc-Cosine Kernels of Different Orders 150
3.B Proofs Related to the Marginal Likelihood 156
3.C Helper Lemmas 158

3.D Proof of the Main Results . . . 179
3.D.1 Proofs Related to the Asymptotics of the Normalized Stochastic Complexity . . . 179
3.D.2 Proofs Related to the Asymptotics of the Generalization Error . . . 201
3.D.3 Proofs Related to the Excess Mean Squared Generalization Error . . 219

4 Asymptotic Spectrum of the NTK via a Power Series Expansion . . . 221
4.1 Introduction . . . 221
4.2 Notations and Preliminaries . . . 223
4.2.1 Hermite Expansion . . . 224
4.2.2 NTK Parametrization . . . 224
4.3 Expressing the NTK as a Power Series . . . 226
4.4 Analyzing the Asymptotic Spectrum of the NTK via its Power Series . . . 229
Appendices . . . 230
4.A Background Material . . . 231
4.A.1 Gaussian Kernel . . . 231
4.A.2 Neural Tangent Kernel (NTK) . . . 232
4.A.3 Unit Variance Initialization . . . 235
4.A.4 Hermite Expansions . . . 238
4.B Expressing the NTK as a Power Series . . . 240
4.B.1 Deriving a Power Series for the NTK . . . 240
4.B.2 Analyzing the Coefficients of the NTK Power Series . . . 247
4.C Analyzing the Spectrum of the NTK via its power series . . . 252
4.C.1 Experimental validation of results on the NTK spectrum . . . 252
4.C.2 Analysis of the Asymptotic Spectrum: Uniform Data . . . 253

5 Conclusion **266**

References **268**

List of Figures

2.1 Illustration of implicit bias for univariate regression 11

2.2 Illustration of implicit bias for multivariate regression 18

2.3 Illustration of implicit bias for univariate regression with Gaussian initialization 35

2.4 Illustration of implicit bias for univariate regression with Gaussian initialization and small bias variance 35

2.5 Illustration of implicit bias for univariate regression with larger data set and larger networks 36

2.6 Training only output layer vs training all parameters of the network 37

2.7 Effect of not adjusting the data 38

2.8 Illustration of Theorem 6. Similar to Figure 2.2, with the same initialization, but with a larger data set. 38

2.9 Illustration of Theorem 6. Similar to Figure 2.2, but with the Gaussian initialization $\boldsymbol{\mathcal{W}} \sim \mathcal{N}(0, I_d)$ and $\mathcal{B} \sim \mathcal{N}(0, 1)$. 39

2.10 Illustration of Theorem 6 with the Gaussian initialization 40

2.11 Network outputs of different scale of initialization 45

2.12 Scatter plots of D_1, D and R^2 104

2.13 Plot of functions $h_1(x)$ and $h_2(x)$. The left panel plots the two function when $\lambda_j = 1$. The right panel plots the two function when $\lambda_j = 5$. 122

2.14 Trajectories of functions obtained by gradient descent training a neural network and by smoothing splines of the training data with decreasing regularization strength 124

3.1 Normalized SC (top) and Bayesian generalization error (bottom) for GPR with the kernel $k^{(1)}_{\text{w/o bias}}$ and the target functions in Table 3.1 148

3.2 Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(1)}_{\text{w/ bias}}$ and the target functions in Table 3.3 152

3.3 Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(2)}_{\text{w/o bias}}$ and the target functions in Table 3.4. 153

3.4 Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(2)}_{\text{w/ bias}}$ and the target functions in Table 3.5. 154

3.5 Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(0)}_{\text{w/o bias}}$ and the target functions in Table 3.6. 155

3.6 Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(0)}_{\text{w/ bias}}$ and the target functions in Table 3.7. 156

4.1 Asymptotic NTK Spectrum 253

List of Tables

2.1 Experimental settings. 103

3.1 Target functions used in the experiments for the first order arc-cosine kernel without bias $k^{(1)}_{\text{w/o bias}}$ 148

3.2 The different kernel functions used in our experiments, their values of α, the corresponding neural network activation function 151

3.3 Target functions used in the experiments for the first order arc-cosine kernel with bias $k^{(1)}_{\text{w/ bias}}$ 152

3.4 Target functions used in the experiments for the second order arc-cosine kernel without bias $k^{(2)}_{\text{w/o bias}}$ 153

3.5 Target functions used in the experiments for the second order arc-cosine kernel with bias $k^{(2)}_{\text{w/ bias}}$ 154

3.6 Target functions used in the experiments for the zero order arc-cosine kernel without bias $k^{(0)}_{\text{w/o bias}}$ 155

3.7 Target functions used in the experiments for the zero order arc-cosine kernel with bias $k^{(0)}_{\text{w/ bias}}$ 156

4.1 Percentage of $\sum_{p=0}^{\infty} \kappa_{p,2}$ accounted for by the first $T+1$ NTK coefficients assuming $\gamma_w^2 = 1$, $\gamma_b^2 = 0$, $\sigma_w^2 = 1$ and $\sigma_b^2 = 1 - \mathbb{E}[\phi(Z)^2]$. 228

Acknowledgments

Approaching the completion of my dissertation, I would like to take this opportunity to thank everyone who has helped me during my five and a half years PhD study.

I would like thank my advisor Guido for his great help and support. Guido is a very kind, patient and knowledgeable advisor. He has introduced me to the area of machine learning theory and showed me how to do researches. Thanks to his help, I am able to finish this dissertation.

I would like to thank my other committee members, Stanley Osher, Deanna Needell and Quanquan Gu for their valuable suggestions, time and feedbacks.

I would like to thank all my collaborators. I would like to thank Pradeep Kr. Banerjee for his ideas, suggestions and help during the writing of our learning curve paper. I would like to thank Benjamin Bowman, Michael Murray and Yoni Dukler for their insightful discussion about research.

I would like to thank all my friends. I would like to thank Zhangji Zhao, Weiyi Liu, Hanshen Huang, Yang Shen and many other PhD students at UCLA for their help. At the end, I would like to thank my parent Minghai Jin and Guizhen Li for their support, encouragement and endless love.

This thesis contains content from three papers. Chapter 2 is a version of [JM20]. This work was done under the supervision of Guido Montufar. Chapter 3 is a version of [JBM22]. This work was done in collaboration with Pradeep Banerjee under the supervision of Guido Montufar. Chapter 4 is a version of [MJB23]. This work was done in collaboration with Michael Murray and Benjamin Bowman under the supervision of Guido Montufar. I thank Michael Murray for proposing the idea of NTK power series and Benjamin Bowman for proposing the idea of studying the top eigenvalues of NTK by the power series.

The work in this dissertation was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 757983) and NSF CAREER Grant DMS-2145630.

Vita

2013–2017	B.S. Information and Computing Science, Peking University.
2017–2022	Teaching Assistant, Mathematics Department, UCLA.
2022–Present	Graduate Student Researcher, Mathematics Department, UCLA.
2020	Visiting Researcher, Max Planck Institute for Mathematics in the Sciences, Germany.
2021	Visiting Researcher, Max Planck Institute for Mathematics in the Sciences, Germany.
06/13-09/16	Research Intern, Liangpai Investment Management, Shanghai, China.

Publications and Preprints

Jin, Hui, and Guido Montúfar. "Implicit bias of gradient descent for mean squared error regression with wide neural networks." arXiv preprint arXiv:2006.07356 (2020).

Jin, Hui, Pradeep Kr Banerjee, and Guido Montúfar. "Learning curves for Gaussian process regression with power-law priors and targets." International Conference on Learning Representations (2022). arxiv.org/pdf/2110.12231

Jin, Hui, Xie He, Yanghui Wang, Hao Li, and Andrea L. Bertozzi. "Noisy subgraph isomorphisms on multiplex networks." 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019.

Murray, Michael, Hui Jin, Benjamin Bowman and Guido Montúfar. "Characterizing the spectrum of the NTK via a power series expansion." Submitted to The Eleventh International Conference on Learning Representations (2023). https://arxiv.org/pdf/2211.07844

CHAPTER 1

Introduction

Deep learning and neural networks have achieved significant success in artificial intelligence and have been widely applied to many areas, such computer vision, natural language processing, recommendation system and reinforcement learning. The deep neural networks used in practice are highly overparameterized, which means that the number of trainable parameters are way larger than the number of training samples. Overparametrized networks are highly expressive [Bar93, MPC14, PLR16a] and easy to optimize [NH17, KHK19]. Moreover, in many applications overparameterized networks could achieve small generalization error.

However, traditional statistical learning theory such as Probably Approximately Correct (PAC) learning theory [Val84] fails to explain why overparameterized networks could generalize well. This is because overparameterized networks could even fit random training data easily [ZBH21], which means that the hypothesis class of overparameterized networks cannot satisfies uniform convergence. [ZBH21] also shows that explicit regularization is not enough to explain the small generalization error and implicit bias of the optimization method such as stochastic gradient descent (SGD) is important to generalization.

Recently [JGH18a] showed that gradient descent on a wide neural network can be characterized by kernel gradient descent in function space with respect to the Neural Tangent Kernel (NTK). The NTK is fixed during training in the infinite-width limit. [LXS19a] showed that the training dynamics of wide neural networks is approximated by the linearization of the networks at initialization. These results allow us to use linearized models and kernel learning [CD07, CBP21] to analyze the generalization of overparametrized networks. It is noted that similar to overparametrized networks, kernel learning can also easily fit random labels while

performs well on test data in certain tasks [BMM18]. By adopting the tool of linearized models and kernel learning with respect to the NTK, a bunch of optimization and generalization results of overparametrized networks can be obtained [DLL19, DZP19, ADH19a, ALS19a, ZCZ20]. In this dissertation, we are going to follow the approach of linearized models and kernel learning and analyze several problems regarding the generalization of wide neural networks.

In Chapter 2, we study the implicit bias of gradient descent on wide neural networks, which is important to explaining the generalization of overparametrized networks. We approximate the wide neural networks by corresponding linearized models and compute the implicit bias in function space. For univariate regression, we show that the solution of training a width-n shallow ReLU network is within $n^{-1/2}$ of the function which fits the training data and whose difference from the initial function has the smallest 2-norm of the second derivative weighted by a curvature penalty that depends on the probability distribution that is used to initialize the network parameters. We compute the curvature penalty function explicitly for various common initialization procedures. For instance, asymmetric initialization with a uniform distribution yields a constant curvature penalty, and thence the solution function is the natural cubic spline interpolation of the training data. For stochastic gradient descent we obtain the same implicit bias result. We obtain a similar result for different activation functions. For multivariate regression we show an analogous result, whereby the second derivative is replaced by the Radon transform of a fractional Laplacian. For initialization schemes that yield a constant penalty function, the solutions are polyharmonic splines. Moreover, we show that the training trajectories are captured by trajectories of smoothing splines with decreasing regularization strength.

In Chapter 3, we study the generalization error of kernel learning such as Gaussian Process Regression (GPR) and Kernel Ridge Regression (KRR), which are closely related to infinitely wide neural networks. We characterize the power-law asymptotics of learning curves for Gaussian process regression (GPR) under the assumption that the eigenspectrum of the prior and the eigenexpansion coefficients of the target function follow a power law. Under similar assumptions, we leverage the equivalence between GPR and kernel ridge regression (KRR) to

show the generalization error of KRR. Gaussian process kernel and the neural tangent kernel (NTK) in several cases (e.g. with ReLU activations) is known to have a power-law spectrum. Hence our methods can be applied to study the generalization error of infinitely wide neural networks with ReLU activations.

In Chapter 4, we study the asymptotic spectrum of the Neural Tangent Kernel (NTK) via a power series expansion of the NTK. The asymptotic spectrum of the NTK can be used to study the asymptotics of learning curves as we show in Chapter 3. Under mild conditions on the network initialization we show that the NTK of arbitrarily deep feedforward networks in the infinite width limit can be expressed in the form of a power series. The power series expansion of the NTK facilitate us to study the spectrum of the NTK. For data drawn uniformly on the sphere we derive an explicit formula for the eigenvalues of the NTK, given the coefficient of the NTK power series. This result shows that faster decay in the NTK coefficients implies a faster decay in its spectrum. From this we recover existing results on eigenvalue asymptotics for ReLU networks and comment on how the activation function influences the RKHS.

CHAPTER 2

Implicit Bias of Gradient Descent for Mean Squared Error Regression with Two-Layer Wide Neural Network *

2.1 Introduction

Understanding why artificial neural networks trained in the overparametrized regime and without explicit regularization generalize well in practice is one of the key challenges in contemporary deep learning [ZBH17]. A series of works have observed that this phenomenon must involve some form of capacity control beyond the network size [NTS15] and, specifically, an implicit bias resulting from the parameter optimization procedures [NTS17]. By implicit bias we mean that among the many candidate hypotheses that fit the training data, the optimization procedure selects one which satisfies additional properties benefitting its performance on new data. In this chapter we investigate the implicit bias of gradient descent parameter optimization for mean squared error regression with wide shallow ReLU networks. Our theory shows that gradient descent is biased towards smooth functions. More precisely, the trained functions are well captured by interpolating splines depending on the initial function and the probability distribution that is used to initialize the network parameters.

Under appropriate conditions, we intuitively expect that gradient descent will be biased towards solutions close to the initial parameter. Indeed, considering overparametrized neural networks, [OS19] showed that gradient descent finds a global minimizer of the training objective which is close to the initialization. This intuition is spot-on for least squares regression with linearized models. In this case, [ZXL20] showed that gradient flow optimization converges

*This chapter is adapted from [JM20].

to the global minimum which is closest to the initialization in parameter space. Although neural networks have a non-linear parametrization, [JGH18a] and [LXS19b] showed that the training dynamics of wide neural networks is well approximated by the dynamics of the linearization at a suitable initialization. This is referred to as the kernel regime, in contrast to the adaptive regime where the models are not well approximated by their linearization. Also, [COB19] showed that, under appropriate scaling of the output weights, a model can converge to zero training loss while hardly varying its parameters. This phenomenon is referred to as "lazy training". On the other hand, it is also possible to relate properties of the parameters to properties of the represented functions. [SES19] studied infinite-width univariate (single input) neural networks and showed that, under a standard parametrization, the complexity of the represented functions, as measured by the 1-norm of the second derivative, can be controlled by the 2-norm of the parameters. [OWS20] extended these results to the multivariate setting. Using these results, one can show that gradient descent with ℓ_2 weight penalty leads to simple functions. We will pursue an approach following these ideas, where we first approximate the gradient dynamics of a wide network in terms of a linear model and then establish a function space description of the implicit bias in parameter space.

The implicit bias of parameter optimization has also been investigated in terms of the properties of the loss function at the points reached by different optimization procedures [KMN17, WZW17, DPB17]. [GLS18a] analyze the implicit bias of different optimization methods (natural gradient, steepest and mirror descent) for linear regression and separable linear classification problems, and obtain characterizations in terms of minimum norm or max-margin solutions. Several works have studied the implicit bias of optimization for classification tasks in terms of margins. [SHN18] showed that in classification problems with separable data, gradient descent with linear networks converges to a max-margin solution. [GLS18b] presented a result on implicit bias for deep linear convolutional networks, and [JT19] studied non-separable data. [CB20] showed that gradient flow for logistic regression with infinitely wide two-layer networks yields a max-margin classifier in a certain space. In the adaptive regime, [MBG18] showed that gradient flow for shallow ReLU networks initialized

close to zero quantizes features depending on the training data but not on the network size. [BGL21] showed the evolution of the tangent features during training which can be interpreted as feature selection and compression. [WTP19] obtained results for univariate regression contrasting the kernel regime and the adaptive regime. We will obtain a related result for univariate regression in the kernel regime and a corresponding result for the multivariate case.

This chapter is organized as follows. In Section 2.2 we provide settings and notation. We present our main results in Section 2.3, along with a discussion. The main techniques pertaining wide networks and the infinite width limit are presented in Sections 2.4 and 2.5. In Sections 2.6 and 2.7, we present the main derivations for the implicit bias in function space for univariate and multivariate regression. In the interest of a concise presentation, technical proofs and extended discussions are deferred to appendices.

2.2 Notations and Problem Setup

Consider a fully connected network with d inputs, one hidden layer of width n, and a single output. For any given input $\mathbf{x} \in \mathbb{R}^d$, the output of the network is

$$f(\mathbf{x}, \theta) = \sum_{i=1}^{n} W_i^{(2)} \phi(\langle \mathbf{W}_i^{(1)}, \mathbf{x} \rangle + b_i^{(1)}) + b^{(2)}, \tag{2.1}$$

where ϕ is an entry-wise activation function, $\mathbf{W}^{(1)} = (\mathbf{W}_1^{(1)}, \ldots, \mathbf{W}_n^{(1)})^T \in \mathbb{R}^{n \times d}$, $\mathbf{W}_i^{(1)} = (W_{i,1}^{(1)}, \ldots, W_{i,d}^{(1)})^T \in \mathbb{R}^d$, $\mathbf{W}^{(2)} = (W_1^{(2)}, \ldots, W_n^{(2)})^T \in \mathbb{R}^n$, $\mathbf{b}^{(1)} = (b_1^{(1)}, \ldots, b_n^{(1)})^T \in \mathbb{R}^n$ and $b^{(2)} \in \mathbb{R}$ are the weights and biases of the first and second layer. We write $\theta = \mathrm{vec}(\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{W}^{(2)}, b^{(2)})$ for the vector of all network parameters. These parameters are initialized by independent samples of pre-specified random variables $\mathcal{W}$ and $\mathcal{B}$ as follows:

$$\begin{aligned} W_{i,j}^{(1)} &\overset{d}{=} \sqrt{1/d}\, \mathcal{W}, \quad b_i^{(1)} \overset{d}{=} \sqrt{1/d}\, \mathcal{B}, \\ W_i^{(2)} &\overset{d}{=} \sqrt{1/n}\, \mathcal{W}, \quad b^{(2)} \overset{d}{=} \sqrt{1/n}\, \mathcal{B}. \end{aligned} \tag{2.2}$$

In the analysis of [JGH18a, LXS19b], $\mathcal{W}$ and $\mathcal{B}$ are Gaussian $\mathcal{N}(0, \sigma^2)$. In the default initialization of PyTorch [PGM19], $\mathcal{W}$ and $\mathcal{B}$ have uniform distribution $\mathrm{Unif}(-\sigma, \sigma)$. More generally, we will also allow weight-bias pairs $(\mathbf{W}_i^{(1)}, b_i^{(1)})$ of units in the hidden layer to be sampled from the joint distribution of a sub-Gaussian $(\boldsymbol{\mathcal{W}}, \mathcal{B})$, where $\boldsymbol{\mathcal{W}}$ is a d-dimensional random vector and $\mathcal{B}$ is a random variable. The parameters of the second layer are still sampled from random variables $\mathcal{W}^{(2)}$ and $\mathcal{B}^{(2)}$. Then the parameters of the network are initialized as follows:

$$\begin{aligned} (\mathbf{W}_i^{(1)}, b_i^{(1)}) &\overset{d}{=} (\boldsymbol{\mathcal{W}}, \mathcal{B}) \\ W_i^{(2)} &\overset{d}{=} \sqrt{1/n}\, \mathcal{W}^{(2)}, \quad b^{(2)} \overset{d}{=} \sqrt{1/n}\, \mathcal{B}^{(2)}. \end{aligned} \tag{2.3}$$

The setting (2.1) is known as the standard parametrization. Some works [JGH18a,LXS19b] use the so-called NTK parametrization, where the factor $\sqrt{1/n}$ is carried outside of the trainable parameter (for details see Appendix 2.B.3). If we fix the learning rate for all parameters, gradient descent leads to different trajectories under these two parametrizations (for details see Appendix 2.B.3). Our results are presented for the standard parametrization.

We consider a regression problem for data $\{(\mathbf{x}_j, y_j)\}_{j=1}^M$ with inputs $\mathcal{X} = \{\mathbf{x}_j\}_{j=1}^M$ and outputs $\mathcal{Y} = \{y_j\}_{j=1}^M$. For a loss function $\ell\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the empirical risk (also called training error) is $L(\theta) = \frac{1}{M}\sum_{j=1}^M \ell(f(\mathbf{x}_j, \theta), y_j)$. We will mainly focus on the square loss $\ell(y, \hat{y}) = \frac{1}{2}\|y - \hat{y}\|^2$, in which case L is the mean squared error. We use full batch gradient descent with a fixed learning rate η to minimize $L(\theta)$. Writing θ_t for the parameter at time t, and θ_0 for the initialization, this defines an iteration

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta) = \theta_t - \eta \nabla_\theta f(\mathcal{X}, \theta_t)^T \nabla_{f(\mathcal{X}, \theta_t)} L, \tag{2.4}$$

where $f(\mathcal{X}, \theta_t) = [f(\mathbf{x}_1, \theta_t), \ldots, f(\mathbf{x}_M, \theta_t)]^T$ is the vector of network outputs for all training inputs, and $\nabla_{f(\mathcal{X}, \theta_t)} L$ is the gradient of L as a function of the network outputs $f(\mathcal{X}, \theta_t)$. We will use subscript i to index neurons and subscript t to index time. Furthermore, we denote by $\hat{\Theta}_n$ the empirical neural tangent kernel (NTK) of the standard parametrization (2.1) at time 0, which is the matrix $\hat{\Theta}_n = \frac{1}{n}\nabla_\theta f(\mathcal{X}, \theta_0) \nabla_\theta f(\mathcal{X}, \theta_0)^T$. We write C^k for the space of

real valued functions with continuous kth derivatives and Lip for the space of Lipschitz continuous functions. We use the notations O_p to denote the standard mathematical orders in probability.[1]

2.3 Main Results

In this section we describe our main results for univariate and multivariate regression, followed by an interpretation and overview of the proof steps developed in the next sections.

2.3.1 Univariate Regression

We have the following description of the implicit bias in function space when applying gradient descent to univariate least squares regression with wide ReLU neural networks.

Theorem 1 (Implicit bias of gradient descent for univariate regression)**.** *Consider a feed-forward network with a single input unit, a hidden layer of n rectified linear units, and a single linear output unit. Assume standard parametrization* (2.1) *and parameter initialization* (2.3)*, which means for each hidden unit the input weight and bias are initialized from a sub-Gaussian $(\mathcal{W}, \mathcal{B})$ with joint density $p_{\mathcal{W},\mathcal{B}}$. Then, for any finite data set $\{(x_j, y_j)\}_{j=1}^M$ and sufficiently large n there exist constants $u, v \in \mathbb{R}$ so that optimization of the mean squared error on the adjusted training data $\{(x_j, y_j - ux_j - v)\}_{j=1}^M$ by full-batch gradient descent with sufficiently small step size converges to a parameter θ^* for which the output function $f(x, \theta^*)$ attains zero training error. Furthermore, letting $\zeta(x) = \int_{\mathbb{R}} |W|^3 p_{\mathcal{W},\mathcal{B}}(W, -Wx)\ \mathrm{d}W$ and $S = \mathrm{supp}(\zeta) \cap [\min_j x_j, \max_j x_j]$, we have $\sup_{x \in S} \|f(x, \theta^*) - g^*(x)\|_2 = O_p(n^{-\frac{1}{2}})$ over the random initialization θ_0, where g^* solves following variational problem:*

$$\begin{aligned} \min_{g \in C^2(S)} \quad & \int_S \frac{1}{\zeta(x)} (g''(x) - f''(x, \theta_0))^2\ \mathrm{d}x \\ \text{subject to} \quad & g(x_j) = y_j - ux_j - v, \quad j = 1, \ldots, M. \end{aligned} \tag{2.5}$$

[1] $X_n = O_p(a_n)$ as $n \to \infty$ means that for any $\epsilon > 0$, there exists a finite $M_\epsilon > 0$ and a finite $N_\epsilon > 0$ such that $\mathbb{P}(|X_n / a_n| > M_\epsilon) < \epsilon, \forall n > N_\epsilon$.

The proof is provided in Appendix 2.C. Our main theorem also holds when the network parameters are trained by stochastic gradient descent. For details, see Theorem 24 and Remark 25 in Appendix 2.D. We will give an interpretation of the result in Section 2.3.3. We first give the explicit form of ζ for several common parameter initialization procedures.

Theorem 2 (Explicit form of the curvature penalty for common initializations)**.**

(a) Gaussian initialization. Assume that $\mathcal{W}$ *and* $\mathcal{B}$ *are independent,* $\mathcal{W} \sim \mathcal{N}(0, \sigma_w^2)$ *and* $\mathcal{B} \sim \mathcal{N}(0, \sigma_b^2)$*. Then* $\zeta(x) = \frac{2\sigma_w^3 \sigma_b^3}{\pi(\sigma_b^2 + x^2\sigma_w^2)^2}$*.*

(b) Binary-uniform initialization. Assume that $\mathcal{W}$ *and* $\mathcal{B}$ *are independent,* $\mathcal{W} \in \{-1, 1\}$ *and* $\mathcal{B} \sim \mathrm{Unif}(-a_b, a_b)$ *with* $a_b \geq I$*. Then* ζ *is constant on* $[-I, I]$*.*

(c) Uniform initialization. Assume that $\mathcal{W}$ *and* $\mathcal{B}$ *are independent,* $\mathcal{W} \sim \mathrm{Unif}(-a_w, a_w)$ *and* $\mathcal{B} \sim \mathrm{Unif}(-a_b, a_b)$ *with* $\frac{a_b}{a_w} \geq I$*. Then* ζ *is constant on* $[-I, I]$*.*

The proof is provided in Appendix 2.H.3.

Remark 3. *Theorem 2 (b) and (c) show that for certain parameter initialization distributions, the function* ζ *is constant on an interval. In this case, the solution* $(g(x) - f(x, \theta_0))$ *to the variational problem* (2.5) *in Theorem 1 corresponds to cubic spline interpolation with natural boundary conditions (see, e.g., [ANW67]). For general* ζ*, the solution corresponds to a spatially adaptive natural cubic spline, which can be computed numerically by solving a linear system and theoretically in an RKHS formalism (see Appendix 2.O for details).*

For different activation functions, we have the following corollary, proved in Appendix 2.J.

Corollary 4 (Different activation functions)**.** *Use the same settings as in Theorem 1 except with activation function* ϕ *instead of ReLU. Suppose that* ϕ *is a Green's function of a linear operator* L*, i.e.,* $\mathrm{L}\phi = \delta$*, where* δ *denotes the Dirac delta function. Assume that* ϕ *is homogeneous of degree* k*, i.e.,* $\phi(ax) = a^k\phi(x)$ *for all* $a > 0$*. Then we can find a function* p *satisfying* $\mathrm{L}p \equiv 0$ *and adjust the training data* $\{(x_j, y_j)\}_{j=1}^M$ *to* $\{(x_j, y_j - p(x_j)\}_{j=1}^M$*. After*

that, the statement in Theorem 1 holds with the variational problem (2.5) *changed to*

$$\begin{aligned} \min_{g\in C^2(S)} \quad & \int_S \frac{1}{\zeta(x)}[\mathrm{L}(g(x)-f(x,\theta_0))]^2 \,\mathrm{d}x \\ \text{subject to} \quad & g(x_j)=y_j-p(x_j), \quad j=1,\ldots,M, \end{aligned} \tag{2.6}$$

where $\zeta(x)=p_{\mathcal{C}}(x)\mathbb{E}(\mathcal{W}^{2k}|\mathcal{C}=x)$ *and* $S=\mathrm{supp}(\zeta)\cap[\min_j x_j, \max_j x_j]$.

Based on Theorem 1, we can also give an approximate description of the optimization trajectory in function space. If we substitute the constraints $g(x_j)=y_j$ in (2.5) by a quadratic penalty $\frac{1}{\lambda}\frac{1}{M}\sum_{j=1}^{M}(g(x_j)-y_j)^2$, then we obtain the variational problem for a so-called spatially adaptive smoothing spline (see [AS96a, PSH06]). This problem can be solved explicitly and can be shown to approximate early stopping. In Appendix 2.N we provide details for the following observation.

Remark 5 (Training trajectory)**.** *The output function of the network after gradient descent training for* t *steps with learning rate* $\bar{\eta}/n$ *is approximated by the solution to following optimization problem:*

$$\min_{g\in C^2(S)} \quad \sum_{j=1}^{M}[g(x_j)-y_j]^2+\frac{1}{\bar{\eta}t}\int_S \frac{1}{\zeta(x)}(g''(x)-f''(x,\theta_0))^2 \,\mathrm{d}x. \tag{2.7}$$

2.3.2 Multivariate Regression

For multivariate regression, we have the following generalization of Theorem 1.

Theorem 6 (Implicit bias of gradient descent for multivariate regression)**.** *Consider the same network settings as in Theorem 1 except with* d *input units instead of a single input unit. Assume that* $\boldsymbol{\mathcal{W}}$ *is a random vector with* $\mathbb{P}(\|\boldsymbol{\mathcal{W}}\|=0)=0$ *and* $\mathcal{B}$ *is a random variable; the distribution of* $(\boldsymbol{\mathcal{W}},\mathcal{B})$ *is symmetric, i.e.,* $(\boldsymbol{\mathcal{W}},\mathcal{B})$ *and* $(-\boldsymbol{\mathcal{W}},-\mathcal{B})$ *have the same distribution; and* $\|\boldsymbol{\mathcal{W}}\|_2$ *and* $\mathcal{B}$ *are both sub-Gaussian. Then, for any finite data set* $\{(\mathbf{x}_j,y_j)\}_{i=1}^{M}$ *and sufficiently large* n *there exist a constant vector* $\mathbf{u}$ *and a constant* v *so that optimization of*

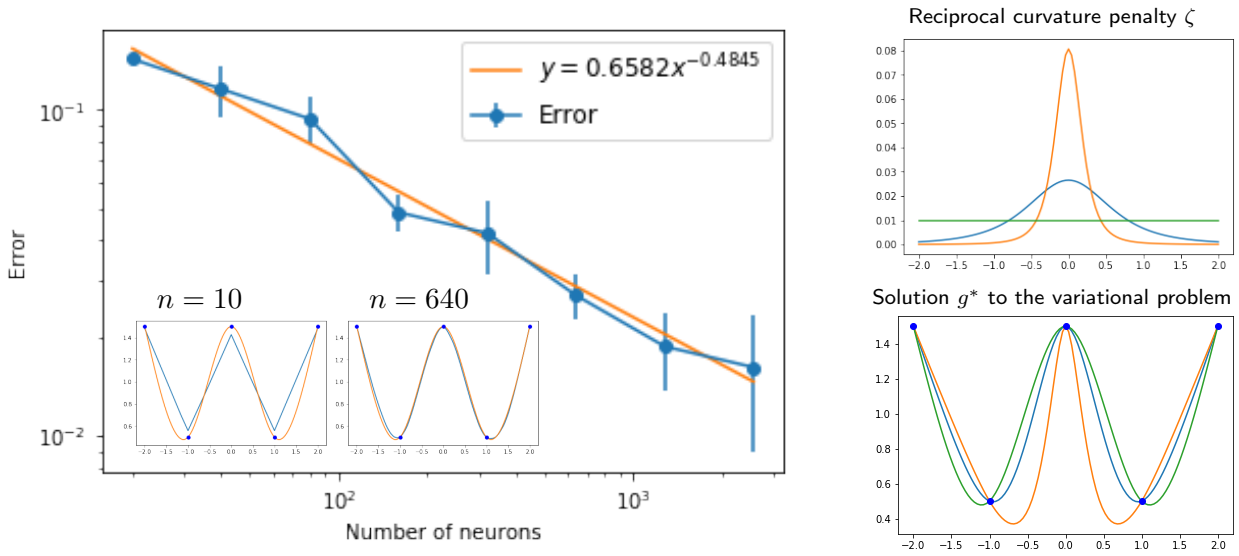

Figure 2.1: Illustration of Theorem 1. Left: Uniform error between the solution g^* to the variational problem and the functions $f(\cdot, \theta^*)$ obtained by gradient descent training with uniform initialization $\mathcal{W} \sim \text{Unif}(-1, 1)$, $\mathcal{B} \sim \text{Unif}(-2, 2)$, against the number of neurons n. The inset shows the training data (dots), g^* (orange), and $f(\cdot, \theta^*)$ (blue) for two values of n. Right: Effect of the curvature penalty function on the shape of the solution function. The bottom shows g^* for various ζ shown at the top. The green curve is for ζ constant on $[-2, 2]$, derived from $\mathcal{W} \sim \text{Unif}(-1, 1)$, $\mathcal{B} \sim \text{Unif}(-2, 2)$; blue is for $\zeta(x) = 1/(1 + x^2)^2$, derived from $\mathcal{W} \sim \mathcal{N}(0, 1)$, $\mathcal{B} \sim \mathcal{N}(0, 1)$; and orange for $\zeta(x) = 1/(0.1 + x^2)^2$, derived from $\mathcal{W} \sim \mathcal{N}(0, 1)$, $\mathcal{B} \sim \mathcal{N}(0, 0.1)$. Theorem 2 shows how to compute ζ for these distributions.

the mean squared error on the adjusted training data $\{(\mathbf{x}_j, y_j - \langle \mathbf{u}, \mathbf{x}_j \rangle - v)\}_{j=1}^M$ *by full-batch gradient descent with sufficiently small step size converges to a parameter* θ^* *for which* $f(\mathbf{x}, \theta^*)$ *attains zero training error. Furthermore, let* $\mathcal{U} = \|\boldsymbol{\mathcal{W}}\|_2$, $\boldsymbol{\mathcal{V}} = \boldsymbol{\mathcal{W}}/\|\boldsymbol{\mathcal{W}}\|_2$, $\mathcal{C} = -\mathcal{B}/\|\boldsymbol{\mathcal{W}}\|_2$ *and* $\zeta(\boldsymbol{V}, c) = p_{\boldsymbol{\mathcal{V}},\mathcal{C}}(\boldsymbol{V}, c)\mathbb{E}(\mathcal{U}^2 | \boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c)$, *where* $p_{\boldsymbol{\mathcal{V}},\mathcal{C}}$ *is the joint density of* $(\boldsymbol{\mathcal{V}}, \mathcal{C})$. *Then, for any compact set* $D \subset \mathbb{R}^d$, *we have* $\sup_{\mathbf{x} \in D} \|f(\mathbf{x}, \theta^*) - g^*(\mathbf{x})\|_2 = O_p(n^{-\frac{1}{2}})$ *over the random initialization* θ_0, *where* g^* *solves following variational problem:*

$$\begin{aligned} \min_{g \in \text{Lip}(\mathbb{R}^d)} \quad & \int_{\text{supp}(\zeta)} \frac{\left(\mathcal{R}\{(-\Delta)^{(d+1)/2}(g - f(\cdot, \theta_0))\}(\boldsymbol{V}, c)\right)^2}{\zeta(\boldsymbol{V}, c)} \, \mathrm{d}\boldsymbol{V}\mathrm{d}c \\ \text{subject to} \quad & g(\mathbf{x}_j) = y_j, \quad j = 1, \ldots, M \\ & \mathcal{R}\{(-\Delta)^{(d+1)/2}(g - f(\cdot, \theta_0))\}(\boldsymbol{V}, c) = 0, \quad (\boldsymbol{V}, c) \notin \text{supp}(\zeta) \\ & (-\Delta)^{(d+1)/2}(g - f(\cdot, \theta_0)) \in L^p(\mathbb{R}^d), \ 1 \le p < d/(d-1). \end{aligned} \tag{2.8}$$

Here $\mathcal{R}$ is the Radon transform defined by $\mathcal{R}\{f\}(\boldsymbol{\omega}, b) \coloneqq \int_{\langle \boldsymbol{\omega}, \mathbf{x} \rangle = b} f(\mathbf{x}) \mathrm{d}s(\mathbf{x})$, the fractional power of the negative Laplacian $(-\Delta)^{(d+1)/2}$ is defined in Fourier domain by $\widehat{(-\Delta)^{(d+1)/2} f}(\boldsymbol{\xi}) = \|\boldsymbol{\xi}\|^{d+1} \widehat{f}(\boldsymbol{\xi})$, and $\mathrm{Lip}(\mathbb{R}^d)$ is the space of Lipschitz continuous functions on $\mathbb{R}^d$.

The proof is given in Appendix 2.C. In Proposition 17 we will show that for specific distributions of $(\boldsymbol{\mathcal{W}}, \mathcal{B})$, the function $\zeta(\boldsymbol{V}, c)$ is constant on $\mathrm{supp}(\zeta)$, which greatly simplifies the variational problem (2.8). We prove the following theorem in Appendix 2.I.2.

Theorem 7 (Variational problem for constant ζ)**.** *Suppose $\boldsymbol{\mathcal{W}}$ is uniformly distributed on $\mathbb{S}^{d-1}$ and $\mathcal{B}$ is uniformly distributed on $[-a_b, a_b]$. Assume that $a_b \geq \max_i \|\mathbf{x}_i\|_2$. Then the variational problem* (2.8) *is equivalent to*

$$\begin{aligned} \min_{h \in \mathrm{Lip}(\mathbb{R}^d) \cap C(\mathbb{R}^d)} \quad & \int_{\mathbb{R}^d} \left((-\Delta)^{(d+3)/4} (h(\mathbf{x}) - f(\mathbf{x}, \theta_0)) \right)^2 \mathrm{d}\mathbf{x} \\ \text{subject to} \quad & h(\mathbf{x}_j) = y_j, \quad j = 1, \dots, M \\ & (-\Delta)^{(d+1)/2} (h(\mathbf{x}) - f(\mathbf{x}, \theta_0)) \in L^p(\mathbb{R}^d), \ 1 \leq p < d/(d-1). \end{aligned} \tag{2.9}$$

We can solve the simplified variational problem (2.9) explicitly. We prove the following theorem in Appendix 2.I.3.

Theorem 8 (Closed form solution)**.** *Suppose $h(\mathbf{x})$ solves the variational problem* (2.9)*. Then $h(\mathbf{x})$ is given by*

$$h(\mathbf{x}) - f(\mathbf{x}, \theta_0) = \sum_{j=1}^{M} \lambda_j \|\mathbf{x} - \mathbf{x}_j\|^3 + \langle \mathbf{u}, \mathbf{x}_i \rangle + v, \tag{2.10}$$

where the coefficients λ_j, $\mathbf{u}$ and v are determined by

$$\begin{cases} \sum_{j=1}^{M} \lambda_j \|\mathbf{x}_i - \mathbf{x}_j\|^3 + \langle \mathbf{u}, \mathbf{x}_i \rangle + v = y_i - f(\mathbf{x}_i, \theta_0), \quad i = 1, \dots, M \\ \sum_{j=1}^{M} \lambda_j = 0 \\ \sum_{j=1}^{M} \lambda_j \mathbf{x}_j = \mathbf{0}\,. \end{cases} \tag{2.11}$$

Remark 9. *A function of the form* (2.10)–(2.11) *is referred to as a polyharmonic spline (see [Pot81]), which is a special type of radial basis function interpolation [Du 08]. When*

$d = 1$ (i.e., the univariate case), this corresponds to the natural cubic spline interpolation described in Remark 3. Finally, we observe that the training trajectory of gradient descent for multivariate regression can be approximately described by a sequence of so-called polyharmonic smoothing splines [Seg19] with decreasing regularization parameter, similar to the description (2.7) *for the univariate case.*

2.3.3 Discussion of the Main Results

Interpretation An intuitive interpretation of Theorem 1 is that gradient descent optimization is biased towards smooth functions. At those regions of the input space where ζ is smaller, we can expect the difference between the functions after and before training to have a small curvature. We call $\rho = 1/\zeta$ a curvature penalty function. The theorem gives an explicit description of the bias in function space depending on the initialization. In Theorem 2 we obtain the explicit form of ζ for various common parameter initialization procedures. In particular, when the parameters are initialized independently from a uniform distribution on a finite interval, ζ is constant and the problem is solved by the natural cubic spline interpolation of the data.

We illustrate Theorem 1 numerically in Figure 2.1 and more extensively in Appendix 2.A. In close agreement with the theory, the solution to the variational problem captures the solution of gradient descent training uniformly with error of order $n^{-1/2}$. To illustrate the effect of the curvature penalty function, Figure 2.1 also shows the solutions to the variational problem for different values of ζ corresponding to different initialization distributions. We see that indeed at input points where ζ is small resp. peaks strongly, the solution function tends to have a lower curvature resp. use a higher curvature in order to fit the training data. This description could be used to formulate heuristics for parameter initialization either to ease optimization or to induce specific smoothness priors on the solutions. In particular, in Proposition 15 we will show that any curvature penalty $1/\zeta$ can be implemented by an appropriate choice of the parameter initialization distribution.

Similar to the univariate case, in the multivariate case gradient descent implicitly controls the complexity of the solution functions obtained upon training. In this case the complexity is measured by the weighted 2-norm of the Radon transform of the $(d+1)/2$ power of the negative Laplacian. The weight function ζ is again determined by the distribution used to initialize the parameters. Although the precise interpretation of these expressions is no longer as straightforward, intuitively the implicit bias corresponds to penalizing a global notion of overall curvature across hyperplanes in the input space. For certain parameter initialization distributions, Theorem 8 shows that the network output after training is a polyharmonic spline. We illustrate Theorem 6 numerically in Figure 2.2 and more extensively in Appendix 2.A. Again in close agreement with the theory, the solution to the variational problem captures the solution returned by gradient descent training with a uniform error of order $n^{-1/2}$.

These results show that the effective capacity of the network, understood as the set of possible output functions after training, is well captured by a space of cubic splines (polyharmonic splines for multivariate regression) relative to the initial function. This is a space with dimension of order M (the number of training examples) independently of the number of parameters of the network.

We note that under suitable asymmetric parameter initialization (see Appendix 2.B.2), it is possible to achieve $f(\cdot, \theta_0) \equiv 0$. Then in Theorem 1 and Theorem 6, the regularization is on the curvature of the output function itself (rather than its difference to the initial function). Further, we note that although Theorem 1 and Theorem 6 describe gradient descent training with linearly adjusted data, they also approximately describe training with the original training data (see Appendix 2.K for more details). The adjustment of the training data simply accounts for the fact that the second derivative and the Laplace operator are invariant to addition of linear terms. In practice we can use the coefficients $\mathbf{u}$ and v of linear regression $y_j = \langle \mathbf{u}, \mathbf{x}_j \rangle + v + \epsilon_j$, $j = 1, \ldots, M$, and set the adjusted data as $\{(\mathbf{x}_j, \epsilon_j)\}_{j=1}^M$. Furthermore, if we change the network architecture by adding skip connections from the inputs to the outputs, our result holds for the original training data without any adjustments. Details are

provided in Appendix 2.L.

Generalization results Theorem 1 allows us to show how gradient descent on wide neural networks learns a target function. In following paragraphs, we show how the solution of variational problem (2.5), (2.7) and (2.8) converges to a target function as the amount of data increases.

In the so-called univariate noiseless model, the training outputs are given by $y_j = g_0(x_j)$, where $g_0 \colon [a, b] \mapsto \mathbb{R}$ is the target function. Let $a = x_0 < x_1 < \cdots < x_M < x_{M+1} = b$ and $h = \max_i x_{i+1} - x_i$. If ζ is constant on $[a, b]$, the solution g^* of (2.5) is the cubic interpolation spline of training data. [HM76] showed in the context of splines that for a target function $g_0 \in C^4([a, b])$ one has $\|g^* - g_0\|_\infty \leq C\|g_0^{(4)}\|_\infty h^4$, where $g_0^{(4)}$ is the fourth derivative of g_0.

For univariate noisy models, the training outputs are given by $y_j = g_0(x_j) + \epsilon_j$, where ϵ_j are zero-mean independent random variables with a common variance σ^2. In this case we use early stopping to smooth out the noise and the training result is characterized by the solution of (2.7). If ζ is constant on $[a, b]$, the solution g^* of (2.7) is the cubic smoothing spline of training data. [Rag83, Theorem 5.8] showed that if $g_0 \in C^2([a, b])$ and $\{x_j\}_{j=1}^M$ are the uniform partition of $[a, b]$, then $\mathbb{E}\|g^* - g_0\|_2^2 \leq C\big((1/t + (1/M)^4)\|g_0''\|^2 + t^{1/4}/M\big)$, where t is the number of training steps. If we choose t to be $\Theta(M^{4/5})$, then $\mathbb{E}\|g^* - g_0\|_2^2 = O(M^{-4/5})$. This gives us some hints about how to choose the stopping time depending on the number of training samples. Similar observations can be obtained for more general settings. [Rag83] also gives out the error bound of g^* for non-uniform training inputs. [EL06] shows a similar result if $\{x_j\}_{j=1}^M$ are sampled independently from a distribution.

If ζ is non-constant on $[a, b]$, the solution g^* of (2.7) is called the spatially adaptive smoothing spline of the training data. [WDS13, Corollary 1] showed that if $g_0 \in C^4([a, b])$, $\zeta \in C^3([a, b])$, $t = \Theta(M^{4/9})$ and $\{x_j\}_{j=1}^M$ are sampled from a distribution on $[a, b]$ with bounded positive density function $q \in C^3([a, b])$, $|g^*(x) - g_0(x)| = O_p(M^{-4/9})$. If the curvature of the target function changes a lot on its domain, spatially adaptive smoothing splines with properly chosen ζ perform better than cubic smoothing splines. [WDS13, Corollary 1] showed that optimal ζ is the solution of a variational problem if the target function is known. They

approximate the optimal ζ by a piecewise constant function and estimate the target function from training data by interpolating splines. Then they numerically solve the variational problem and get the suitable ζ for the training data. [AS96b] and [SBR10] proposed to choose ζ based on an estimation of the second derivative of g_0. [LG10] used a piecewise constant ζ and proposed a search algorithm to find such ζ. Proposition 15 showed the way to choose the joint distribution of weight and bias parameters in order that ζ is proportional to a given function. Once we find out a proper ζ according to the training data using the methods in above literature, we can initialize the weight and bias parameters by the corresponding joint distribution and train the wide neural network by gradient descent. According to the theory, such special parameter initialization should perform better than uniform or Gaussian initialization.

For multivariate noiseless models, if ζ is constant over its support, the solution g^* of variational problem (2.8) is the polyharmonic spline. Then the error bound between g^* and target function g_0 is shown in [Pot81, Theorem 3.2].

Strategy of the proof In Section 2.4 we observe that for a linearized model, gradient descent with sufficiently small step size finds the minimizer of the training objective which is closest to the initial parameter (similar to a result by [ZXL20]). Then Theorem 10 shows that the training dynamics of a linearized wide network is well approximated in parameter and in function space by that of a lower dimensional linear model which trains only the output weights. This property has appeared in different contexts [Dan17] and is sometimes taken for granted in the literature. We show that it holds for the standard parametrization, although it does not hold for the NTK parametrization, which leads to the adaptive regime. Under these settings, the implicit bias of gradient descent amounts to minimizing distance from the initial parameter, subject to fitting the training data. In Section 2.5, we relate this description of the implicit bias in parameter space to an alternative optimization problem. In Theorem 12 we show that the solution to this alternative problem has a well defined limit as the width of the network tends to infinity, which allows us to obtain a variational description. In Section 2.6, we focus on the case of univariate regression. In Theorem 13 we translate the

description of the bias from parameter space to function space. In Section 2.7, we turn to the case of multivariate regression and use the inversion formula of the dual Radon transform to analyze the optimization objective. Finally, we exploit recent results (Theorem H.1 in [LXS19b]) bounding the difference in function space of the solutions obtained from training a wide network and its linearization to conclude the proof.

Related works [ZXL20] described the implicit bias of gradient descent in the kernel regime as minimizing a kernel norm from initialization, subject to fitting the training data. Our result can be regarded as making the kernel norm explicit, thus providing an interpretable description of the bias in function space and further illuminating the role of the parameter initialization procedure. We prove the equivalence in Appendix 2.M. [CG19] derived the generalization bounds for overparametrized deep neural networks under stochastic gradient descent training. They also approximated the neural network by a linearized model, which is called a neural tangent random feature (NTRF) model in their work.

[SES19] showed that infinitely wide networks with 2-norm weight regularization represent functions with smallest 1-norm of the second derivative, an example of which are linear splines (see Appendix 2.B.4 for more details). A recent work by [PN19] further develops this direction for two-layer networks with certain activation functions that interpolate data while minimizing a weight norm. In contrast, our result characterizes the solutions of training from a given initialization without explicit regularization, which turn out to minimize a weighted 2-norm of the second derivative and hence correspond to cubic splines. Another recent work [HTW19] discusses ridge weight penalty, adaptive splines, and early stopping for one-input ReLU networks training only the output layer. The spline perspective for univariate shallow ReLU networks has recently been also discussed by [SPD20]. [WTP19] showed a similar result in the kernel regime for shallow ReLU networks training only the output layer from zero initialization. In contrast, we consider the initialization of the second layer and show that the difference from the initial output function is implicitly regularized by gradient descent. We show the result of training both layers can be approximated by training only the second layer in Theorem 10. In addition, we give the explicit form of ζ in Theorem 2, while the

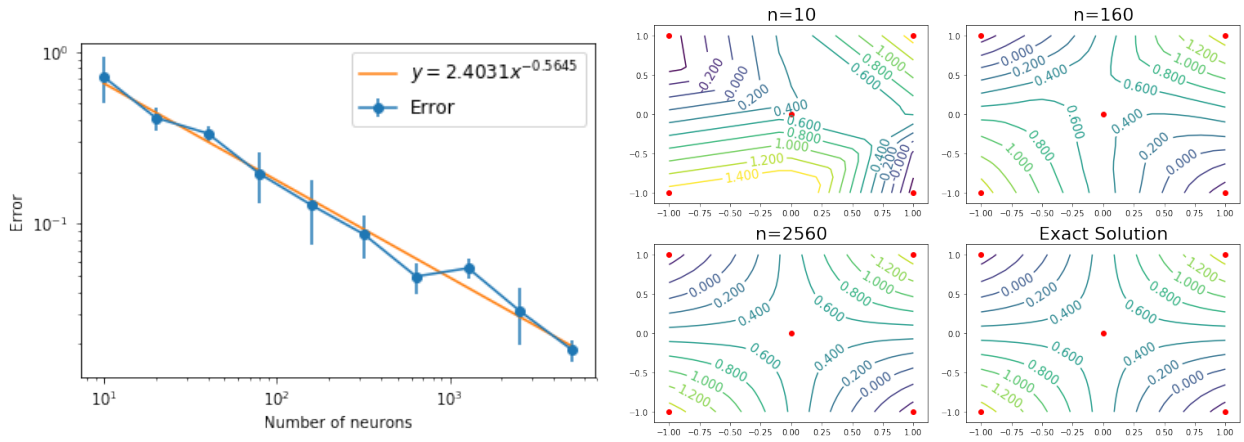

Figure 2.2: Illustration of Theorem 6. Left: Uniform error between the solution g^* to the variational problem and the functions $f(\cdot, \theta^*)$ obtained by gradient descent training of a neural network (in this case with initialization $\boldsymbol{\mathcal{W}} \sim \text{Unif}(\mathbb{S}^1)$, $\mathcal{B} \sim \text{Unif}(-2, 2)$), against the number of neurons. Right: The input training data (dots), the contour plots of trained network functions with 10, 160, 2560 neurons, and the exact solution to the variational problem.

description given by [WTP19] has a minor error because of a typo in their computation. Significantly, our results also cover multivariate regression, different activation functions, and training trajectories.

In the multivariate case, [OWS20] studied infinite-width neural networks with parameters having bounded norm. They showed that the complexity of the functions represented by the network, as measured by the 1-norm of the Radon transform of the $(d+1)/2$-power of the negative Laplacian of the function, can be controlled by the 2-norm of the parameters. Rather than bounding the 2-norm of the parameters, our result describes the implicit bias of gradient descent and in turn we obtain a weighted 2-norm. A recent work by [PN21] considers adding an explicit regularization of 1-norm of the Radon tranform in function space for multivariate regression, and uses the representer theorem to obtain the solution to the variational problem. In contrast, we consider gradient descent without explicit regularization and the implicit bias turns out to be a weighted 2-norm.

2.4 Wide Networks and Parameter Space

In this section, we characterize the implicit bias in parameter space and show that, under our initialization and parametrization scheme, training only the output layer approximates training all parameters.

2.4.1 Implicit Bias in Parameter Space for a Linearized Model

In this section we describe how training a linearized network or a wide network by gradient descent leads to solutions having parameter values close to the initial parameter values. First, we consider the following linearized model:

$$f^{\text{lin}}(\mathbf{x}, \omega) = f(\mathbf{x}, \theta_0) + \nabla_\theta f(\mathbf{x}, \theta_0)(\omega - \theta_0). \tag{2.12}$$

We write ω for the parameter of the linearized model, in order to distinguish it from the parameter θ of the nonlinearized model. The empirical loss of the linearized model is defined by

$$L^{\text{lin}}(\omega) = \sum_{j=1}^{M} \ell(f^{\text{lin}}(\mathbf{x}_j, \omega), y_j). \tag{2.13}$$

The gradient descent iteration for the linearized model is given by

$$\omega_0 = \theta_0, \quad \omega_{t+1} = \omega_t - \eta \nabla_\theta f(\mathcal{X}, \theta_0)^T \nabla_{f^{\text{lin}}(\mathcal{X}, \omega_t)} L^{\text{lin}}. \tag{2.14}$$

Next, we consider wide neural networks. According to Theorem H.1 in [LXS19b],

$$\sup_t \|f^{\text{lin}}(\mathbf{x}, \omega_t) - f(\mathbf{x}, \theta_t)\|_2 = O_p(n^{-\frac{1}{2}})$$

This means that gradient descent training of a wide network or of the linearization of the network results in similar trajectories and solutions in function space. Both solution functions fit the training data perfectly, meaning $f^{\text{lin}}(\mathcal{X}, \omega_\infty) = f(\mathcal{X}, \theta_\infty) = \mathcal{Y}$, and they are also

approximately equal outside of the training data.

Under the assumption that $\text{rank}(\nabla_\theta f(\mathcal{X}, \theta_0)) = M$, the gradient descent iterations (2.14) of the linearized network converge to the unique global minimum that is closest to initialization [GLS18a, ZXL20]. More precisely, ω_∞ is the solution to following constrained optimization problem (further details are provided in Appendix 2.D):

$$\min_{\omega} \|\omega - \theta_0\|_2 \quad \text{s.t. } f^{\text{lin}}(\mathcal{X}, \omega) = \mathcal{Y}. \tag{2.15}$$

2.4.2 Training Only the Output Layer Approximates Training All Parameters

In the following we consider networks with a single hidden layer of n ReLUs and a linear output, $f(\mathbf{x}, \theta) = \sum_{i=1}^n W_i^{(2)}[\langle \mathbf{W}_i^{(1)}, \mathbf{x}\rangle + b_i^{(1)}]_+ + b^{(2)}$. We show that the functions and parameter vectors obtained by training the linearized model are close to those obtained by training only the output layer. In view of the previous subsection, this implies that training all parameters of a wide network or training only the output layer results in similar functions.

Let $\theta_0 = \text{vec}(\overline{\mathbf{W}}^{(1)}, \overline{\mathbf{b}}^{(1)}, \overline{\mathbf{W}}^{(2)}, \overline{b}^{(2)})$ be the parameter at initialization so that $f^{\text{lin}}(\cdot, \theta_0) = f(\cdot, \theta_0)$. Denote the trained parameter of the linearized network by $\omega_\infty = \text{vec}(\widehat{\mathbf{W}}^{(1)}, \widehat{\mathbf{b}}^{(1)}, \widehat{\mathbf{W}}^{(2)}, \widehat{b}^{(2)})$. Using initialization (2.3), given $1 \le i \le n$, we have that $\|\overline{\mathbf{W}}_i^{(1)}\|, \overline{b}_i^{(1)} = O_p(1)$ and $\overline{W}_i^{(2)}, \overline{b}^{(2)} = O_p(n^{-\frac{1}{2}})$.[2] Therefore, writing H for the Heaviside function, we have

$$\begin{aligned}
\nabla_{\mathbf{W}_i^{(1)}, b_i^{(1)}} f(\mathbf{x}, \theta_0) &= \Big[\overline{W}_i^{(2)} H(\langle \overline{\mathbf{W}}_i^{(1)}, \mathbf{x}\rangle + \overline{b}_i^{(1)}) \cdot \mathbf{x}\ ,\ \overline{W}_i^{(2)} H(\langle \overline{\mathbf{W}}_i^{(1)}, \mathbf{x}\rangle + \overline{b}_i^{(1)})\Big] = O_p(n^{-\frac{1}{2}}), \\
\nabla_{W_i^{(2)}, b^{(2)}} f(\mathbf{x}, \theta_0) &= \Big[[\langle \overline{\mathbf{W}}_i^{(1)}, \mathbf{x}\rangle + \overline{b}_i^{(1)}]_+\ ,\ 1\Big] = O_p(1).
\end{aligned} \tag{2.16}$$

This implies that when n is large, if we use gradient descent with a constant learning rate for all parameters, then the changes of $\mathbf{W}^{(1)}$, $\mathbf{b}^{(1)}$, $b^{(2)}$ are negligible compared with the changes of $\mathbf{W}^{(2)}$. In turn, approximately we can train just the output weights, $W_i^{(2)}, i = 1, \ldots, n$,

[2] More precisely, given $1 \le i \le n$, $\exists C$, for any $\delta > 0$, s.t. with prob. $1 - \delta$, $|\overline{W}_i^{(2)}|, |\overline{b}^{(2)}| \le Cn^{-1/2}\sqrt{\log \frac{1}{\delta}}$ and $\|\overline{\mathbf{W}}_i^{(1)}\|, |\overline{b}_i^{(1)}| \le C\sqrt{\log \frac{1}{\delta}}$ since the random variables are sub-Gaussian.

and fix all other parameters, which corresponds to training a smaller linear model. Let $\widetilde{\omega}_t = \mathrm{vec}(\overline{\mathbf{W}}^{(1)}, \overline{\mathbf{b}}^{(1)}, \widetilde{\mathbf{W}}_t^{(2)}, \overline{b}^{(2)})$ be the parameter at time t under the update rule where $\overline{\mathbf{W}}^{(1)}, \overline{\mathbf{b}}^{(1)}, \overline{b}^{(2)}$ are kept fixed at their initial values, and

$$\widetilde{\mathbf{W}}_0^{(2)} = \overline{\mathbf{W}}^{(2)}, \quad \widetilde{\mathbf{W}}_{t+1}^{(2)} = \widetilde{\mathbf{W}}_t^{(2)} - \eta \nabla_{\mathbf{W}^{(2)}} L^{\mathrm{lin}}(\widetilde{\omega}_t). \tag{2.17}$$

Let $\widetilde{\omega}_\infty = \lim_{t\to\infty} \widetilde{\omega}_t$. By the above discussion, we expect that $f^{\mathrm{lin}}(\mathbf{x}, \widetilde{\omega}_\infty)$ will be close to $f^{\mathrm{lin}}(\mathbf{x}, \omega_\infty)$. We have the following formal result for mean squared error regression.

Theorem 10 (Training only output weights vs linearized network)**.** *Consider a finite data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^M$. Assume that we use the square loss $\ell(\widehat{y}, y) = \frac{1}{2}\|\widehat{y} - y\|_2^2$; $\inf_n \lambda_{\min}(\hat{\Theta}_n) > 0$. Let ω_t denote the parameters of the linearized model at time t when we train all parameters using (2.14), and let $\widetilde{\omega}_t$ denote the parameters at time t when we only train weights of the output layer using (2.17). If we use the same learning rate η in these two training processes and $\eta < \frac{2}{n\lambda_{\max}(\hat{\Theta}_n)}$, then for any $\mathbf{x} \in \mathbb{R}^d$,*

$$\sup_t |f^{\mathrm{lin}}(\mathbf{x}, \widetilde{\omega}_t) - f^{\mathrm{lin}}(\mathbf{x}, \omega_t)| = O_p(n^{-1}), \ \text{as } n \to \infty.$$

Moreover, in terms of the parameter trajectories we have $\sup_t \|\overline{\mathbf{W}}^{(1)} - \widehat{\mathbf{W}}_t^{(1)}\|_2 = O_p(n^{-1})$, $\sup_t \|\overline{\mathbf{b}}^{(1)} - \widehat{\mathbf{b}}_t^{(1)}\|_2 = O_p(n^{-1})$, $\sup_t \|\widetilde{\mathbf{W}}_t^{(2)} - \widehat{\mathbf{W}}_t^{(2)}\|_2 = O_p(n^{-3/2})$, $\sup_t \|\overline{b}^{(2)} - \widehat{b}_t^{(2)}\| = O_p(n^{-1})$.

The proof is provided in Appendix 2.E. By combining Theorem 10 and the fact that training a linearized model approximates training a wide network (Theorem H.1 in [LXS19b]), we obtain the following.

Corollary 11 (Training only output weights vs training all weights)**.** *Consider the settings of Theorem 10, and assume that the joint distribution of $(\mathcal{W}, \mathcal{B})$ is sub-Gaussian. Given any compact set $D \subset \mathbb{R}^d$, for every $\mathbf{x} \in D$, $\sup_t \|f^{\mathrm{lin}}(\mathbf{x}, \tilde{\omega}_t) - f(\mathbf{x}, \theta_t)\|_2 = O_p(n^{-\frac{1}{2}})$.*

The proof is given in Appendix 2.F. In view of the arguments in this section, in the next sections we will focus on training only the output weights and understanding the corresponding solution functions.

2.5 Infinite Width Limit of Shallow Networks

According to (2.15), gradient descent training of the output weights (2.17) achieves zero loss, $f^{\text{lin}}(\mathbf{x}_j, \widetilde{\omega}_\infty) - f^{\text{lin}}(\mathbf{x}_j, \theta_0) = \sum_{i=1}^n (\widetilde{W}_i^{(2)} - \overline{W}_i^{(2)})[\langle \overline{\mathbf{W}}_i^{(1)}, \mathbf{x}_j \rangle + \overline{b}_i^{(1)}]_+ = y_j - f(\mathbf{x}_j, \theta_0)$, $j = 1, \dots, M$, with minimum $\|\widetilde{\mathbf{W}}^{(2)} - \overline{\mathbf{W}}^{(2)}\|_2^2$. Hence gradient descent is actually solving

$$\min_{\mathbf{W}^{(2)}} \|\mathbf{W}^{(2)} - \overline{\mathbf{W}}^{(2)}\|_2^2 \quad \text{s.t.} \quad \sum_{i=1}^n (W_i^{(2)} - \overline{W}_i^{(2)})[\langle \overline{\mathbf{W}}_i^{(1)}, \mathbf{x}_j \rangle + \overline{b}_i^{(1)}]_+ = y_j - f(\mathbf{x}_j, \theta_0), \ j = 1, \dots, M. \tag{2.18}$$

To simplify the presentation, in the following we let $f^{\text{lin}}(\mathbf{x}, \theta_0) \equiv 0$ by using the Anti-Symmetrical Initialization (ASI) trick (see Appendix 2.B.2). The analysis still goes through without this simplification (see Appendix 2.H).

We reformulate problem (2.18) in a way that allows us to consider the limit of infinitely wide networks, with $n \to \infty$, and obtain a deterministic counterpart, analogous to the convergence of the NTK. Let μ_n denote the empirical distribution of the samples $(\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)})_{i=1}^n$, i.e., $\mu_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A\Big((\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)})\Big)$, where $\mathbb{1}_A$ denotes the indicator function for measurable subsets A in $\mathbb{R}^2$. We further consider a function $\alpha_n \colon \mathbb{R}^2 \to \mathbb{R}$ whose value encodes the difference of the output weight from its initialization for a hidden unit with input weight and bias given by the argument, i.e., $\alpha_n(\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)}) = n(W_i^{(2)} - \overline{W}_i^{(2)})$. Then (2.18) with ASI can be rewritten as

$$\min_{\alpha_n \in C(\mathbb{R}^2)} \int_{\mathbb{R}^2} \alpha_n^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) \ \text{ s.t. } \int_{\mathbb{R}^2} \alpha_n(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) = y_j, \tag{2.19}$$

where j ranges from 1 to M. Here we minimize over functions α_n in $C(\mathbb{R}^2)$, but since only the values on $(\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)})_{i=1}^n$ are taken into account, we can take any continuous interpolation of $\alpha_n(\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)})$, $i = 1, \dots, n$.

Now we can consider the infinite width limit. Let μ be the probability measure of $(\boldsymbol{\mathcal{W}}, \mathcal{B})$.

By substituting μ for μ_n, we obtain a continuous version of problem (2.19) as follows:

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^2)} \quad & \int_{\mathbb{R}^2} \alpha^2(\mathbf{W}^{(1)}, b)\, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^2} \alpha(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) = y_j, \quad j = 1, \ldots, M. \end{aligned} \tag{2.20}$$

Using that μ_n weakly converges to μ, the following theorem shows that in fact the solution of problem (2.19) converges to the solution of (2.20). The proof is given in Appendix 2.G.

Theorem 12 (Infinite width limit)**.** *Let $(\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)})_{i=1}^n$ be i.i.d. samples from a pair $(\boldsymbol{\mathcal{W}}, \mathcal{B})$ with finite fourth moment. Suppose μ_n is the empirical distribution of $(\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)})_{i=1}^n$ and $\overline{\alpha}_n(\mathbf{W}^{(1)}, b)$ is the solution of (2.19). Let $\overline{\alpha}(\mathbf{W}^{(1)}, b)$ be the solution of (2.20). Then, for any compact set $D \subset \mathbb{R}^d$, we have $\sup_{\mathbf{x} \in D} |g_n(\mathbf{x}, \overline{\alpha}_n) - g(\mathbf{x}, \overline{\alpha})| = O_p(n^{-1/2})$, where $g_n(\mathbf{x}, \alpha_n) = \int_{\mathbb{R}^2} \alpha_n(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b]_+ \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b)$ is the function represented by a network with n hidden neurons after training, and $g(\mathbf{x}, \alpha) = \int_{\mathbb{R}^2} \alpha(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)$ is the function represented by the infinite-width network.*

2.6 Implicit Bias for Univariate Regression

In this section we solve the optimization problem (2.20) in the univariate case, which provides a function space characterization of the implicit bias previously described in parameter space. First we rewrite the problem in terms of breakpoints. Consider the breakpoint $c = -b/W^{(1)}$ of a ReLU with weight $W^{(1)}$ and bias b. We define a corresponding random variable $\mathcal{C} = -\mathcal{B}/\mathcal{W}$ and let ν denote the distribution of $(\mathcal{W}, \mathcal{C})$.[3] Then, writing $\gamma(W^{(1)}, c) = \alpha(W^{(1)}, -cW^{(1)})$, the optimization problem (2.20) is equivalently given as

$$\min_{\gamma \in C(\mathbb{R}^2)} \int_{\mathbb{R}^2} \gamma^2(W^{(1)}, c)\, \mathrm{d}\nu(W^{(1)}, c) \text{ s.t. } \int_{\mathbb{R}^2} \gamma(W^{(1)}, c)[W^{(1)}(x_j - c)]_+ \, \mathrm{d}\nu(W^{(1)}, c) = y_j, \tag{2.21}$$

[3]Here we assume that $\mathbb{P}(\mathcal{W} = 0) = 0$ so that the random variable $\mathcal{C}$ is well defined. This is not an important restriction, since neurons with weight $W^{(1)} = 0$ have a constant output value that can be absorbed in the bias of the output layer.

where j ranges from 1 to M. Let $\nu_{\mathcal{C}}$ denote the distribution of $\mathcal{C} = -\mathcal{B}/\mathcal{W}$, and $\nu_{\mathcal{W}|\mathcal{C}=c}$ the conditional distribution of $\mathcal{W}$ given $\mathcal{C} = c$. Suppose $\nu_{\mathcal{C}}$ has support $\operatorname{supp}(\nu_{\mathcal{C}})$ and a density function $p_{\mathcal{C}}(c)$. Let $g(x, \gamma) = \int_{\mathbb{R}^2} \gamma(W^{(1)}, c)[W^{(1)}(x - c)]_+ \, \mathrm{d}\nu(W^{(1)}, c)$, which again corresponds to the output function of the network. Then, the second derivative g'' with respect to x satisfies $g''(x, \gamma) = p_{\mathcal{C}}(x) \int_{\mathbb{R}} \gamma(W^{(1)}, x)\big|W^{(1)}\big| \, \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)})$ (for details on this see Appendix 2.H.1). This shows that $\gamma(W^{(1)}, c)$ is closely related to $g''(x, \gamma)$. In the following we seek to express (2.21) in terms of $g''(x, \gamma)$. Since $g''(x, \gamma)$ determines $g(x, \gamma)$ only up to linear functions, we consider the following problem:

$$\begin{aligned} \min_{\gamma \in C(\mathbb{R}^2), u \in \mathbb{R}, v \in \mathbb{R}} \quad & \int_{\mathbb{R}^2} \gamma^2(W^{(1)}, c) \, \mathrm{d}\nu(W^{(1)}, c) \\ \text{subject to} \quad & u x_j + v + \int_{\mathbb{R}^2} \gamma(W^{(1)}, c)[W^{(1)}(x_j - c)]_+ \, \mathrm{d}\nu(W^{(1)}, c) = y_j, \quad j = 1, \ldots, M. \end{aligned} \tag{2.22}$$

Here u, v are not included in the cost. They add a linear function to the output of the neural network. If u and v in the solution of (2.22) are small, then the solution is close to the solution of (2.21). [OWS20] also use this trick to simplify the characterization of neural networks in function space. Next we study the solution of (2.22) in function space. This is our main technical result for univariate regression.

Theorem 13 (Implicit bias in function space for univariate regression)**.** *Assume $\mathcal{W}$ and $\mathcal{B}$ are random variables with $\mathbb{P}(\mathcal{W} = 0) = 0$, and let $\mathcal{C} = -\mathcal{B}/\mathcal{W}$. Let ν denote the probability distribution of $(\mathcal{W}, \mathcal{C})$. Suppose $(\overline{\gamma}, \overline{u}, \overline{v})$ is the solution of (2.22), and consider the corresponding output function*

$$g(x, (\overline{\gamma}, \overline{u}, \overline{v})) = \overline{u}x + \overline{v} + \int_{\mathbb{R}^2} \overline{\gamma}(W^{(1)}, c)[W^{(1)}(x - c)]_+ \, \mathrm{d}\nu(W^{(1)}, c). \tag{2.23}$$

Let $\nu_{\mathcal{C}}$ denote the marginal distribution of $\mathcal{C}$ and assume it has a density function $p_{\mathcal{C}}$. Assume that $\mathcal{W}$ has finite second moment. Let $\mathbb{E}(\mathcal{W}^2|\mathcal{C})$ denote the conditional expectation of $\mathcal{W}^2$ given $\mathcal{C}$. Consider the function $\zeta(x) = p_{\mathcal{C}}(x)\mathbb{E}(\mathcal{W}^2|\mathcal{C} = x)$, assume its support contains the input samples, $x_i \in \operatorname{supp}(\zeta)$, $i = 1, \ldots, m$, and let $S = \operatorname{supp}(\zeta) \cap [\min_i x_i, \max_i x_i]$. Then

$g(x, (\overline{\gamma}, \overline{u}, \overline{v}))$ satisfies $g''(x, (\overline{\gamma}, \overline{u}, \overline{v})) = 0$ for $x \notin S$ and for $x \in S$ it is the solution to the following problem:

$$\min_{h \in C^2(S)} \int_S \frac{(h''(x))^2}{\zeta(x)} \, \mathrm{d}x \quad s.t. \quad h(x_j) = y_j, \quad j = 1, \ldots, m. \tag{2.24}$$

The proof is provided in Appendix 2.H.1, where we also present the corresponding statement without ASI.

Finally, we discuss the curvature penalty function. We provide the proof of following propositions in Appendix 2.H.2.

Proposition 14 (Curvature penalty function)**.** *Let $p_{\mathcal{W},\mathcal{B}}$ denote the joint density function of $(\mathcal{W}, \mathcal{B})$ and let $\mathcal{C} = -\mathcal{B}/\mathcal{W}$ so that $p_{\mathcal{C}}$ is the breakpoint density. Then $\zeta(x) = \mathbb{E}(W^2|C = x)p_{\mathcal{C}}(x) = \int_{\mathbb{R}} |W|^3 p_{\mathcal{W},\mathcal{B}}(W, -Wx) \, \mathrm{d}W$.*

We note that if we sample the initial weight and biases from a suitable joint distribution, we can make the curvature penalty $\rho = 1/\zeta$ arbitrary:

Proposition 15 (Constructing any curvature penalty)**.** *Given any function $\varrho \colon \mathbb{R} \to \mathbb{R}_{>0}$, satisfying $Z = \int_{\mathbb{R}} \frac{1}{\varrho} < \infty$, if we set the density of $\mathcal{C}$ as $p_{\mathcal{C}}(x) = \frac{1}{Z} \frac{1}{\varrho(x)}$ and make $\mathcal{W}$ independent of $\mathcal{C}$ with non-vanishing second moment, then $(\mathbb{E}(W^2|C = x)p_{\mathcal{C}}(x))^{-1} = (\mathbb{E}(W^2)p_{\mathcal{C}}(x))^{-1} \propto \varrho(x)$, $x \in \mathbb{R}$.*

2.7 Implicit Bias for Multivariate Regression

In this section we solve the optimization problem (2.20) in the multivariate case. Similar to Section 2.6, we can relax the optimization problem to

$$\begin{aligned} &\min_{\substack{\alpha \in C(\mathbb{R}^d \times \mathbb{R}), \\ \mathbf{u} \in \mathbb{R}^d, v \in \mathbb{R}}} \int_{\mathbb{R}^d \times \mathbb{R}} \alpha^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\ &\text{subject to } \int_{\mathbb{R}^d \times \mathbb{R}} \alpha(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) + \langle \mathbf{u}, \mathbf{x}_j \rangle + v = y_j, \ j = 1, \ldots, M. \end{aligned} \tag{2.25}$$

Let $\mathcal{U} = \|\boldsymbol{\mathcal{W}}\|_2$, $\boldsymbol{\mathcal{V}} = \boldsymbol{\mathcal{W}}/\|\boldsymbol{\mathcal{W}}\|_2$ and $\mathcal{C} = -\mathcal{B}/\|\boldsymbol{\mathcal{W}}\|_2$. Let ν denote the distribution of $(\mathcal{U}, \boldsymbol{\mathcal{V}}, \mathcal{C})$ and $\gamma(u, \boldsymbol{V}, c) = \alpha(u\boldsymbol{V}, -cu)$. Then, after the change of variables, the optimization problem (2.25) is equivalently expressed as

$$\begin{aligned} \min_{\substack{\alpha \in C(\mathbb{R}^+ \times \mathbb{S}^{d-1} \times \mathbb{R}), \\ \mathbf{u} \in \mathbb{R}^d, v \in \mathbb{R}}} \ & \int_{\mathbb{R}^+ \times \mathbb{S}^{d-1} \times \mathbb{R}} \gamma^2(u, \boldsymbol{V}, c) \, \mathrm{d}\nu(u, \boldsymbol{V}, c) \\ \text{subject to } & \int_{\mathbb{R}^+ \times \mathbb{S}^{d-1} \times \mathbb{R}} \gamma(u, \boldsymbol{V}, c) \cdot u \cdot [\langle \boldsymbol{V}, \mathbf{x}_j \rangle - c]_+ \, \mathrm{d}\nu(u, \boldsymbol{V}, c) + \langle \mathbf{u}, \mathbf{x}_j \rangle + v = y_j, \\ & j = 1, \ldots, M. \end{aligned} \tag{2.26}$$

Define the output of the infinite-width network by

$$g(\mathbf{x}, (\gamma, \mathbf{u}, v)) = \int_{\mathbb{R}^+ \times \mathbb{S}^{d-1} \times \mathbb{R}} \gamma(u, \boldsymbol{V}, c) \cdot u \cdot [\langle \boldsymbol{V}, \mathbf{x} \rangle - c]_+ \, \mathrm{d}\nu(u, \boldsymbol{V}, c) + \langle \mathbf{u}, \mathbf{x} \rangle + v.$$

Then the Laplacian $\Delta g(\mathbf{x}, (\gamma, \mathbf{u}, v)) = \sum_{i=1}^d \partial_{x_i}^2 g(\mathbf{x}, (\gamma, \mathbf{u}, v))$ is given by

$$\begin{aligned} \Delta g(\mathbf{x}, (\gamma, \mathbf{u}, v)) &= \int_{\mathbb{R}^+ \times \mathbb{S}^{d-1} \times \mathbb{R}} \gamma(u, \boldsymbol{V}, c) \cdot u \cdot \ \delta(\langle \boldsymbol{V}, \mathbf{x} \rangle - c) \, \mathrm{d}\nu(u, \boldsymbol{V}, c) \\ &= \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \left(\int_{\mathbb{R}^+} \gamma(u, \boldsymbol{V}, c) \cdot u \, \mathrm{d}\nu_{\mathcal{U} | \boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c}(u) \right) \delta(\langle \boldsymbol{V}, \mathbf{x} \rangle - c) \, \mathrm{d}\nu_{\boldsymbol{\mathcal{V}}, \mathcal{C}}(\boldsymbol{V}, c), \end{aligned} \tag{2.27}$$

where $\nu_{\boldsymbol{\mathcal{V}}, \mathcal{C}}$ denotes the joint distribution of $(\boldsymbol{\mathcal{V}}, \mathcal{C})$, and $\nu_{\mathcal{U} | \boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c}$ the conditional distribution of $\mathcal{U}$ given $\boldsymbol{\mathcal{V}} = \boldsymbol{V}$ and $\mathcal{C} = c$. Let $\nu_{C | \boldsymbol{\mathcal{V}} = \boldsymbol{V}}$ denote the conditional distribution of $\mathcal{C}$ given $\boldsymbol{\mathcal{V}} = \boldsymbol{V}$. Suppose $\nu_{C | \boldsymbol{\mathcal{V}} = \boldsymbol{V}}$ has a density function $p_{\mathcal{C} | \boldsymbol{\mathcal{V}} = \boldsymbol{V}}(c)$. Define

$$\kappa(\boldsymbol{V}, c) = \int_{\mathbb{R}^+} \gamma(u, \boldsymbol{V}, c) \cdot u \, \mathrm{d}\nu_{\mathcal{U} | \boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c}(u). \tag{2.28}$$

Then (2.27) becomes

$$\begin{aligned}\Delta g(\mathbf{x},(\alpha,\mathbf{u},v)) &= \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \kappa(\boldsymbol{V},c)\ \delta(\langle \boldsymbol{V},\mathbf{x}\rangle - c)\ \mathrm{d}\nu_{\boldsymbol{\mathcal{V}},\mathcal{C}}(\boldsymbol{V},c) \\ &= \int_{\mathbb{S}^{d-1}} \left(\int_{\mathbb{R}} \kappa(\boldsymbol{V},c)\ \delta(\langle \boldsymbol{V},\mathbf{x}\rangle - c) p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\mathrm{d}c \right) \mathrm{d}\nu_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \\ &= \int_{\mathbb{S}^{d-1}} \kappa(\boldsymbol{V},\langle \boldsymbol{V},\mathbf{x}\rangle)\ p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(\langle \boldsymbol{V},\mathbf{x}\rangle)\ \mathrm{d}\nu_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}),\end{aligned} \tag{2.29}$$

where $\nu_{\boldsymbol{\mathcal{V}}}$ denotes the distribution of $\boldsymbol{\mathcal{V}}$. Assume that $\nu_{\boldsymbol{\mathcal{V}}}$ has a density function $p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})$ with respect to the spherical measure σ^{d-1}. Then (2.29) becomes

$$\Delta g(\mathbf{x},(\alpha,\mathbf{u},v)) = \int_{\mathbb{S}^{d-1}} \kappa(\boldsymbol{V},\langle \boldsymbol{V},\mathbf{x}\rangle)\ p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(\langle \boldsymbol{V},\mathbf{x}\rangle) p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})\ \mathrm{d}\sigma^{d-1}(\boldsymbol{V}). \tag{2.30}$$

Now, defining

$$\beta(\boldsymbol{V},c) = \kappa(\boldsymbol{V},c)\ p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}), \tag{2.31}$$

we observe that

$$\begin{aligned}\Delta g(\mathbf{x},(\alpha,\mathbf{u},v)) &= \int_{\mathbb{S}^{d-1}} \beta(\boldsymbol{V},\langle \boldsymbol{V},\mathbf{x}\rangle)\ \mathrm{d}\boldsymbol{V} \\ &= \mathcal{R}^*\{\beta\}(\mathbf{x}).\end{aligned} \tag{2.32}$$

The right-hand side of (2.32) is precisely the dual Radon transform of β. Let $\beta = \beta^+ + \beta^-$ be the even–odd decomposition of β, where β^+ is even and β^- is odd, i.e., $\beta^+(\boldsymbol{V},c) = \beta^+(-\boldsymbol{V},-c)$ and $\beta^-(\boldsymbol{V},c) = -\beta^-(-\boldsymbol{V},-c)$ for all $(\boldsymbol{V},c) \in \mathbb{S}^{d-1}\times\mathbb{R}$. Since the dual Radon transform annihilates odd functions, we have $\Delta g(\mathbf{x},(\alpha,\mathbf{u},v)) = \int_{\mathbb{S}^{d-1}} \beta^+(\boldsymbol{V},\langle \boldsymbol{V},\mathbf{x}\rangle)\ \mathrm{d}\boldsymbol{V}$. [OWS20] observed that β^+ can be recovered from Δg by using the inversion formula of the dual Radon transform. According to [OWS20, Lemma 3],

$$\beta^+ = -\frac{1}{2(2\pi)^{d-1}} \mathcal{R}\{(-\Delta)^{(d+1)/2} g(\cdot,\alpha)\}, \tag{2.33}$$

where $\mathcal{R}$ is the Radon transform which is defined by

$$\mathcal{R}\{f\}(\boldsymbol{\omega},b) \coloneqq \int_{\langle \boldsymbol{\omega},\mathbf{x}\rangle = b} f(\mathbf{x})\mathrm{d}s(\mathbf{x}), \quad (\boldsymbol{\omega},b) \in \mathbb{S}^{d-1}\times\mathbb{R},$$

where $\mathrm{d}s(\mathbf{x})$ represents integration with respect to the $(d-1)$-dimensional surface measure on the hyperplane $\langle \boldsymbol{\omega}, \mathbf{x} \rangle = b$. The fractional power of the negative Laplacian $(-\Delta)^{(d+1)/2}$ in (2.33) is the operator defined in Fourier domain by

$$\widehat{(-\Delta)^{(d+1)/2} f}(\boldsymbol{\xi}) = \|\boldsymbol{\xi}\|^{d+1} \widehat{f}(\boldsymbol{\xi}).$$

When $d+1$ is a even number, $(-\Delta)^{(d+1)/2}$ is the same as applying the negative Laplacian $(d+1)/2$ times. When $d+1$ is odd, it is a pseudo-differential operator given by convolution with a singular kernel (see [Kwa17]). Then according to (2.33) and the definition of β, we have

$$\begin{aligned} &\mathcal{R}\{(-\Delta)^{(d+1)/2} g(\cdot, \alpha)\}(\boldsymbol{V}, c) - 2(2\pi)^{d-1} \beta^{-} \\ =& -2(2\pi)^{d-1} \kappa(\boldsymbol{V}, c) p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \\ =& -2(2\pi)^{d-1} p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \int_{\mathbb{R}^+} \gamma(u, \boldsymbol{V}, c) \cdot u \, \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}, \mathcal{C}=c}(u). \end{aligned} \tag{2.34}$$

From the above equation, we show how $\gamma(u, \boldsymbol{V}, c)$ is characterized by the network output function, which allows us to study the solution of (2.26) in function space. The following theorem generalizes Theorem 13 to the multivariate case.

Theorem 16 (Implicit bias in function space for multivariate regression)**.** *Assume that (1) $\boldsymbol{\mathcal{W}}$ is a random vector with $\mathbb{P}(\|\boldsymbol{\mathcal{W}}\| = 0) = 0$ and $\mathcal{B}$ is a random variable; (2) the distribution of $(\boldsymbol{\mathcal{W}}, \mathcal{B})$ is symmetric, i.e., $(\boldsymbol{\mathcal{W}}, \mathcal{B})$ and $(-\boldsymbol{\mathcal{W}}, -\mathcal{B})$ have the same distribution; (3) $\|\boldsymbol{\mathcal{W}}\|_2$ and $\mathcal{B}$ both have finite second moments. Let $\mathcal{U} = \|\boldsymbol{\mathcal{W}}\|_2$, $\boldsymbol{\mathcal{V}} = \boldsymbol{\mathcal{W}}/\|\boldsymbol{\mathcal{W}}\|_2$ and $\mathcal{C} = -\mathcal{B}/\|\boldsymbol{\mathcal{W}}\|_2$. Let ν denote the distribution of $(\mathcal{U}, \boldsymbol{\mathcal{V}}, \mathcal{C})$. Suppose $(\overline{\gamma}, \overline{\mathbf{u}}, \overline{v})$ is the solution of (2.26), and assume that (2.26) is feasible, which means*

$$\int_{\mathbb{R}^+ \times \mathbb{S}^{d-1} \times \mathbb{R}} \overline{\gamma}^2(u, \boldsymbol{V}, c) \, \mathrm{d}\nu(u, \boldsymbol{V}, c) < +\infty.$$

Consider the corresponding output function

$$g(\mathbf{x}, (\overline{\gamma}, \overline{\mathbf{u}}, \overline{v})) = \int_{\mathbb{R}^+ \times \mathbb{S}^{d-1} \times \mathbb{R}} \overline{\gamma}(u, \boldsymbol{V}, c) \cdot u \cdot [\langle \boldsymbol{V}, \mathbf{x} \rangle - c]_+ \, \mathrm{d}\nu(u, \boldsymbol{V}, c) + \langle \overline{\mathbf{u}}, \mathbf{x} \rangle + \overline{v}. \tag{2.35}$$

Let $\nu_{\boldsymbol{\mathcal{V}}}$ denote the marginal distribution of $\mathcal{C}$ and assume it has a density function $p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})$. Let $\nu_{C|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}$ denote the conditional distribution of $\mathcal{C}$ given $\boldsymbol{\mathcal{V}}=\boldsymbol{V}$ and assume it has a density function $p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)$. Let $\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c)$ denote the conditional expectation of $\mathcal{U}^2$ given $\mathcal{V}$ and $\mathcal{C}$. Consider the following function $\zeta\colon \mathbb{S}^{d-1}\times\mathbb{R}\to\mathbb{R}$,

$$\zeta(\boldsymbol{V},c)=p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c). \tag{2.36}$$

Then $g(\mathbf{x},(\overline{\gamma},\overline{\mathbf{u}},\overline{v}))$ is the solution of the following problem:

$$\begin{aligned}\min_{h\in \mathrm{Lip}(\mathbb{R}^d)\cap C(\mathbb{R}^d)} \quad & \int_{\mathrm{supp}(\zeta)} \frac{\left(\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)\right)^2}{\zeta(\boldsymbol{V},c)}\ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\ \text{subject to} \quad & h(\mathbf{x}_j)=y_j,\quad j=1,\dots,M, \\ & \mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)=0,\ \forall(\boldsymbol{V},c)\notin\mathrm{supp}(\zeta), \\ & (-\Delta)^{(d+1)/2}h\in L^p(\mathbb{R}^d),\ 1\le p<d/(d-1),\end{aligned} \tag{2.37}$$

where $\mathrm{Lip}(\mathbb{R}^d)$ is the space of Lipschitz continuous function on $\mathbb{R}^d$ and σ^{d-1} is the spherical measure.

The proof of Theorem 16 is provided in Appendix 2.I.1. The optimization problem (2.37) characterizes the implicit bias of the gradient descent in function space for the multivariate setting. [ZXL20] obtained a characterization in terms of the minimization of a kernel norm in function space, which is also valid for multi-dimensional inputs. In Appendix 2.M we prove the equivalence between the kernel norm minimization and our result in the one-dimensional setting. In future work it will be interesting to show that in the multivariate setting, the kernel norm is equivalent to the objective in (2.37) under appropriate conditions.

To conclude this section, we discuss the function ζ in the variational problem (2.37). The proofs of the following statements are presented in Appendix 2.I.4. First we propose an initialization scheme such that ζ is constant over a bounded region.

Proposition 17 (Constant ζ over a bounded region)**.** *If $\boldsymbol{\mathcal{W}}$ is sampled uniformly from the unit sphere and $\mathcal{B}$ from a symmetric interval, i.e., $\boldsymbol{\mathcal{W}}\sim\mathrm{Unif}(\mathbb{S}^{d-1})$ and $\mathcal{B}\sim\mathrm{Unif}(-a,a)$,*

then $\zeta(\boldsymbol{V}, c)$ *is constant over* $\{(\boldsymbol{V}, c) : |c| \leq a\}$ *and* $\zeta(\boldsymbol{V}, c) = 0$ *for* $|c| > a$.

Now we discuss the form of ζ under certain conditions.

Proposition 18 (Penalty function ζ)**.** *Let* $p_{\boldsymbol{\mathcal{W}},\mathcal{B}}$ *denote the joint density function of* $(\boldsymbol{\mathcal{W}}, \mathcal{B})$ *and let* $\mathcal{U} = \|\boldsymbol{\mathcal{W}}\|_2$, $\boldsymbol{\mathcal{V}} = \boldsymbol{\mathcal{W}}/\|\boldsymbol{\mathcal{W}}\|_2$ *and* $\mathcal{C} = -\mathcal{B}/\|\boldsymbol{\mathcal{W}}\|_2$. *Then* $\zeta(\boldsymbol{V}, c) = p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c) = \int_{\mathbb{R}} u^{d+2} p_{\boldsymbol{\mathcal{W}},\mathcal{B}}(u\boldsymbol{V}, -uc)\ \mathrm{d}u$.

Using the above result we compute the explicit form of ζ for Gaussian initialization.

Theorem 19 (Explicit form of ζ for Gaussian initialization)**.** *Assume that* $\boldsymbol{\mathcal{W}}$ *and* $\mathcal{B}$ *are independent,* $\boldsymbol{\mathcal{W}} \sim \mathcal{N}(0, \sigma_w^2 \boldsymbol{I}_d)$ *and* $\mathcal{B} \sim \mathcal{N}(0, \sigma_b^2)$. *Then* ζ *is given by*

$$\zeta(\boldsymbol{V}, c) = \frac{\sigma_w^3 \sigma_b^{d+2}}{\pi^{(d+1)/2}(\sigma_b^2 + c^2\sigma_w^2)^{(d+3)/2}} \Gamma(\frac{d+3}{2}).$$

2.8 Conclusion

We obtained explicit descriptions in function space for the implicit bias of gradient descent in mean squared error regression with wide shallow ReLU networks covering the univariate and multivariate cases. We also presented a generalization to networks with different activation functions and discussed a relaxation related to early stopping and training trajectories in function space.

In the case of univariate regression, our main result shows that the trained network function interpolates the training data while minimizing a weighted 2-norm of the second derivative with respect to the input. Such functions correspond to spatially adaptive interpolating splines. In the case of multivariate regression, our results also characterize the trained network functions. Under specific parameter initialization schemes, these functions correspond to polyharmonic interpolating splines. The spaces of interpolating splines are linear of dimension in the order of the number of data points. Hence, our results imply that, even if the network has many parameters, the complexity of the trained functions will be adjusted to the number of training data points. This can be used to explain why overparametrized networks do not

overfit in practice, as the generalization error can be regarded as the precision of the spline interpolation (see, e.g., [Wen04]).

[ZXL20] described the implicit bias of gradient descent as minimizing a RKHS norm from initialization. Our result can be regarded as making the RKHS norm explicit, providing an interpretable description of the bias in function space. Compared with [ZXL20], our results describe the role of the parameter initialization scheme, which determines the curvature penalty function $1/\zeta$. This gives us a clearer picture of how the initialization affects the implicit bias of gradient descent. This could be used in order to select a good initialization scheme. For instance, one could conduct a pre-assessment of the data to estimate the locations of the input space where the solution should have a high curvature, and choose the parameter initialization accordingly. This is an interesting possibility to experiment with based on our theoretical results.

Our results can also be interpreted in combination with early stopping. The training trajectory is approximated by a smoothing spline, meaning that the network will filter out high frequencies which are usually associated to noise in the training data. This behaviour is sometimes referred to as a spectral bias [RBA19a]. [CFW21] studied spectral bias theoretically and showed that spherical harmonics of low frequency are easier to be learned by over-parameterized neural networks if the input data is uniformly distributed over the unit hypersphere.

Appendix

The appendix is organized as follows.

- In Appendix 2.A we illustrate our theoretical results numerically, and provide details on the numerical implementation.
- In Appendix 2.B we briefly discuss definitions and settings around the parametrization and initialization of neural networks, as well as on the limiting NTK and the linearization

of a neural network.

- In Appendices 2.C, 2.D, 2.E, 2.F, 2.G, we provide proofs and supporting results for the results presented in Sections 2.3, 2.4.1, 2.4.2, and 2.5.
- In Appendices 2.H and 2.I, we provide the proofs of the results in Sections 2.6 and 2.7 for univariate and multivariate regression respectively.
- In Appendix 2.J, we prove a corresponding result for activation functions other than ReLU.
- In Appendix 2.K we discuss the linear adjustment of the training data and why our result still gives a good description of training with the original data for non-linear target functions.
- In Appendix 2.M we show the equivalence between our variational characterization of the implicit bias of gradient descent in function space and the description in terms of a kernel norm minimization problem.
- In Appendix 2.N we discuss the relation between the gradient descent optimization trajectory and a trajectory of spatially adaptive smoothing splines with decreasing smoothness regularization coefficient which converges to the spatially adaptive interpolating spline.
- In Appendix 2.O we give the explicit form of the solution to our variational problem, i.e., the spatially adaptive interpolating spline, which corresponds to the output function after gradient descent training in the infinite width limit.
- In Appendix 2.P we comment on possible extensions and generalizations of the analysis.

2.A Numerical Illustration of the Theoretical Results

Implementation of gradient descent Training is implemented as full-batch gradient descent. In practice we choose the learning rate as follows. We start with a large learning rate and keep decreasing it by half until we observe that the loss function decreases. After

that, we start training with the fixed learning rate we found. We observe that the learning rate we found is inversely proportional to the width n of the neural network. This observation is in accord with Theorem 20 with respect to the upper bound of the learning rate in order to converge.

We note that the implicit bias in parameter space shown in Theorem 20 is independent of the specific step size that is used in the optimization, so long as it is small enough (see Appendix 2.D). The stopping criterion for training of the neural network is that the change in the training loss in consecutive iterations is less than a pre-specified threshold: $|L(\theta_t) - L(\theta_{t-1})| \leq 10^{-8}$.

We use ASI (see Appendix 2.B.2) at initialization. Then the initial output function of the network is $f(\cdot, \theta_0) \equiv 0$. Hence in the figures the network output function is actually equal to the difference from initialization.

For the comparison of the functions $f(\cdot, \theta^*)$ and g^*, the infinity norm $\|f(\cdot, \theta^*) - g^*\|_\infty$ is computed over a discretization of $[-\max_i \|\mathbf{x}_i\|_2, \max_i \|\mathbf{x}_i\|_2]^d$.

Implementation of numerical solutions to the variational problem For univariate regression, the variational problem for cubic splines can be solved explicitly as described in Appendix 2.O. For a general non-constant curvature penalty function $1/\zeta$, we can obtain a numerical solution to problem (2.24) as follows. First we discretize the interval $[-I, I]$ evenly with points $x_j = -I + 2jI/N$, $j = 0, \ldots, N$. For simplicity we suppose that the M input training data points are among these grid points, and we denote them by $x_{j_1}, \ldots, x_{j_M}$. Then we initialize $f(x_j) = 0$ for x_j not in the training data (to be optimized) and $f(x_{j_i}) = y_i$ (fixed values during optimization). We use the central difference to approximate the second derivative, $f''(x_j) = \frac{f(x_{j+1}) - 2f(x_j) + f(x_{j-1})}{h^2}$, where $h = |x_{j+1} - x_j|$. Then the objective function in (2.24) is approximated by $\sum_{j=1}^{N-1} \frac{1}{\zeta(x_j)} \left(\frac{f(x_{j+1}) - 2f(x_j) + f(x_{j-1})}{h^2} \right)^2$. This is a quadratic problem in $f(x_j)$, $j \in \{1, \ldots, N\} \setminus \{j_1, \ldots, j_M\}$. If we equate the gradient to zero, we obtain a linear system. The solution can be written in closed form in terms of the inverse of a design matrix. As with any linear regression problem, in practice we may still prefer to use an iterative

approach to obtain a numerical solution. In our experiment, we discretize the interval $[-2, 2]$ into 200 pieces and use conjugate gradient descent for solving the linear system.

For multivariate regression, it is not straightforward to numerically solve (2.8). Hence we numerically solve (2.25) instead. We discretize the interval $[-I_w, I_w]$ evenly with points $w_j = -I_w + 2jI_w/n_w$, $j = 0, \ldots, n_w$ and the interval $[-I_b, I_b]$ evenly with points $b_j = -I_b + 2jI_b/n_b$, $j = 0, \ldots, n_b$. Let $\alpha_{(i_1,\ldots,i_d,j)} = \alpha((w_{i_1}, \ldots, w_{i_d}), b_j)$, $i_k = 0, \ldots, n_w$, $j = 0, \ldots, n_b$. We use numerical integration to approximate the objective and constraints of (2.25) and then get an optimization problem with search variables $\alpha_{(i_1,\ldots,i_d,j)}$. This is a quadratic programming problem which can be solved using an internal point method.

Gradient descent training and variational problem To illustrate Theorem 1 across different initialization procedures, in Figures 2.3 and 2.4 we show analogous experiments to those in the left panel of Figure 2.1, but using two types of Gaussian initialization instead of the uniform initialization. As we already observed in the right panel of Figure 2.1, here the effect of the curvature penalty function is also visible. In portions of the input space where ζ is peaked, the solution function can have a high curvature, and, conversely, in portions of the input space where ζ takes small values, the solution function has a small second derivative and is more linear.

To verify that the results are stable over different data sets, in Figure 2.5 we show an experiment similar to that of Figure 2.1, but for a larger data set.

Training all layers versus training only the output layer To illustrate Theorem 10, we conduct the following experiment. We use the same training set as in Figure 2.1 and use uniform initialization. Starting from the same initial weights, we train the network in two ways. One way is only training the output layer and another way is training all layers of the network. The result is shown in Figure 2.6. The left panel plots the error between two trained network functions against the number of neurons n. In this experiment the error is of order $n^{-3/2}$, which is even smaller than the upper bound n^{-1} given in Theorem 10. Potentially the

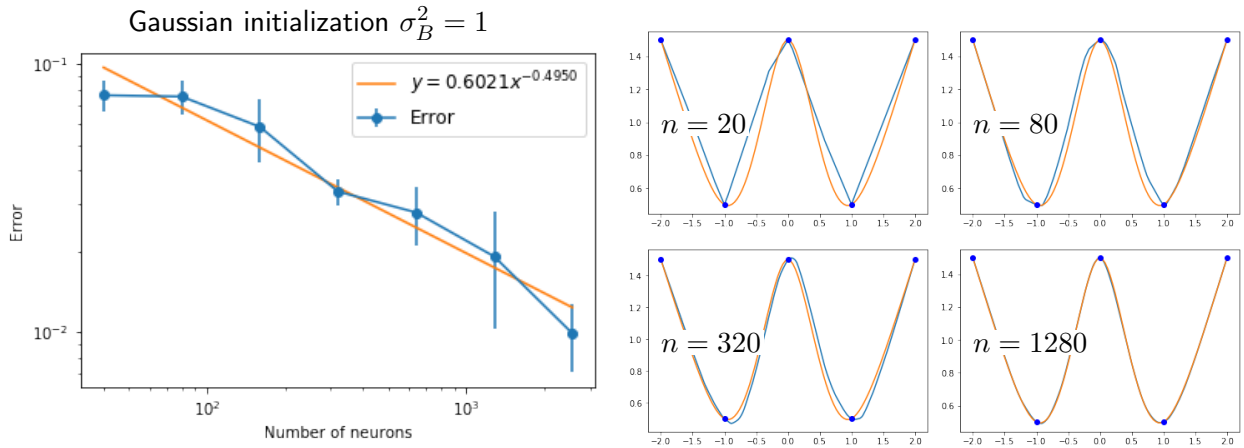

Figure 2.3: Illustration of Theorem 1. Shown is the error between the output function $f(\cdot, \theta^*)$ of the trained neural network and the solution g^* to the variational problem (2.24) against the number of neurons, n. Shown is the average over 5 repetitions, with error bars indicating the standard deviation. Here the training data is fixed, and the parameters were initialized with $W \sim \mathcal{N}(0, 1)$ and $B \sim \mathcal{N}(0, 1)$. The right panel shows the data (dots), trained network functions (blue) with 20, 80, 320, 1280 neurons, and the solution (orange) to the variational problem.

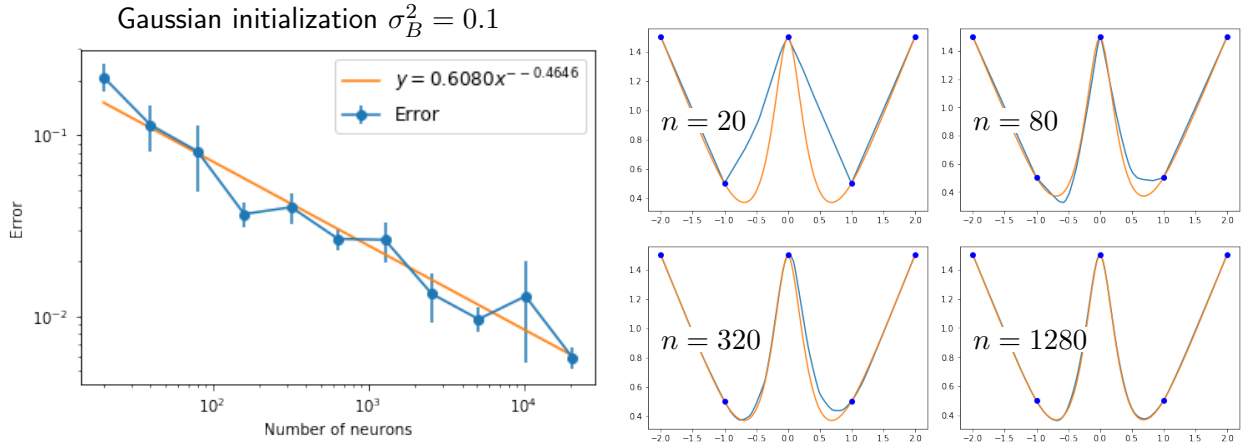

Figure 2.4: Illustration of Theorem 1. Similar to Figure 2.3, but with a different initialization $\mathcal{W} \sim \mathcal{N}(0, 1)$ and $\mathcal{B} \sim \mathcal{N}(0, 0.1)$, which gives rise to a curvature penalty function ζ that is more strongly peaked around $x = 0$ (see Figure 2.1). We observe in particular that the solutions are more curvy around $x = 0$.

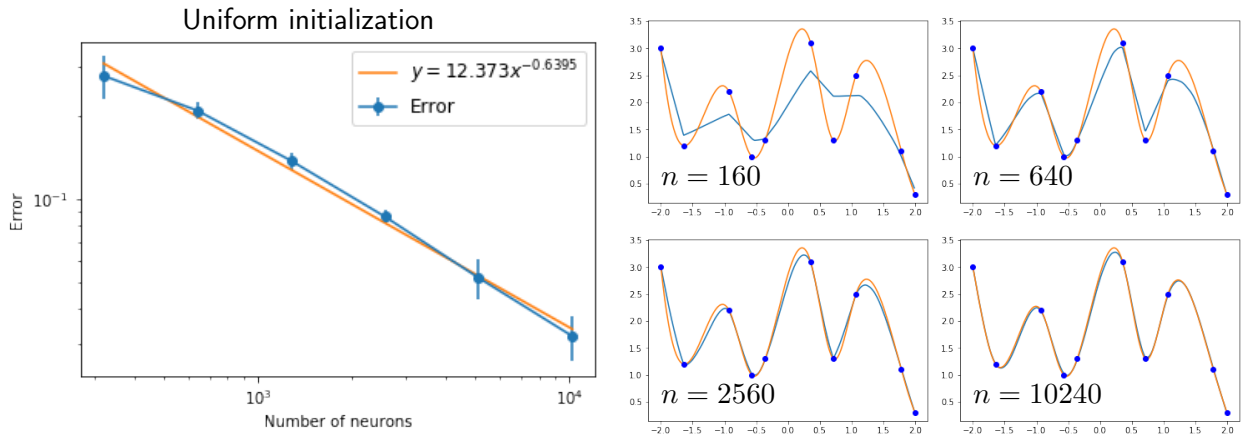

Figure 2.5: Illustration of Theorem 1. Similar to Figure 2.1, with uniform initialization, but with a larger data set and larger networks.

bound can be improved. The right panel plots two trained network functions with 20, 80, 320, 1280 neurons.

Effect of linear function on implicit bias In Theorem 1, since the variational problem defines functions only up to addition of linear functions, we need to adjust training data by subtracting a specific linear function $ux+v$. However, in our previous experiments, we observed that even if we do not adjust the training data, the statement of Theorem 1 still approximately holds. We attribute this to the fact that the linear function can be easily fit by the neural network. We provide details about this in Appendix 2.K. In order evaluate the effect of this linear function on the implicit bias, we conduct the following experiment. Similar to Figure 2.1, we use uniform initialization. We add a linear function $10x + 10$ to the training data in Figure 2.1. So the training data we use are $\{(-2, -8.5), (-1, 0.5), (0, 11.5), (1, 20.5), (2, 31.5)\}$. In Figure 2.7 we show analogous experiments to those in the left panel of Figure 2.1. In order to clearly show the difference between the trained network function and the solution to the variational problem, we subtract $10x + 10$ from these two functions in the right panel of Figure 2.7. From the right panel of Figure 2.7, we see that the difference between plotted two functions is relatively larger than that in Figure 2.1. From the left panel of Figure 2.7, we see that the error between these two functions stops to decrease when number of neurons n

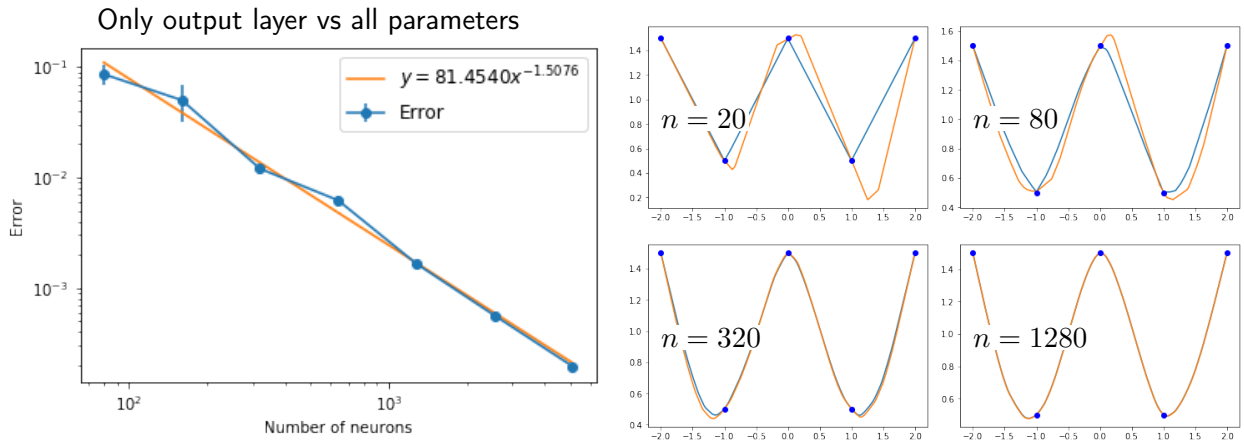

Figure 2.6: Illustration of Theorem 10. Training only output layer vs training all parameters of the network. We use uniform initialization and the same training set as in Figure 2.1. The left panel plots the error between two trained network functions against the number of neurons n. For one network, we only train the output layer while for the another one, we train all layers. The right panel shows the data (dots) and two trained network functions with 20, 80, 320, 1280 neurons.

is larger than 1280. It means that the limit of trained network function as $n \to \infty$ is slightly different from the solution to the variational problem. If we choose bigger u and v, we expect that the difference will become larger.

Experiments for two-dimensional regression problems We illustrate Theorem 6 numerically in Figure 2.2. We conduct experiments similar to Figure 2.1 and Figure 2.3 for the bivariate case. The initialization used in Figure 2.2 is $\boldsymbol{\mathcal{W}} \sim U(\mathbb{S}^1)$ and $\mathcal{B} \sim U(-2, 2)$, thus we can use Theorem 8 to exactly compute the solution to the variational problem (2.8). In close agreement with the theory, the solution to the variational problem captures the solution of gradient descent training uniformly with error of order $n^{-1/2}$.

To verify that the results are stable over different data sets, in Figure 2.8 we show an experiment similar to that of Figure 2.2, but for a larger data set.

To illustrate Theorem 6 across different initialization procedures, in Figures 2.9 and 2.10 we show analogous experiments to Figure 2.2, but using Gaussian initialization instead. The initialization used in Figure 2.9 is $\boldsymbol{\mathcal{W}} \sim \mathcal{N}(0, I_d)$ and $\mathcal{B} \sim \mathcal{N}(0, 1)$, and the initialization used

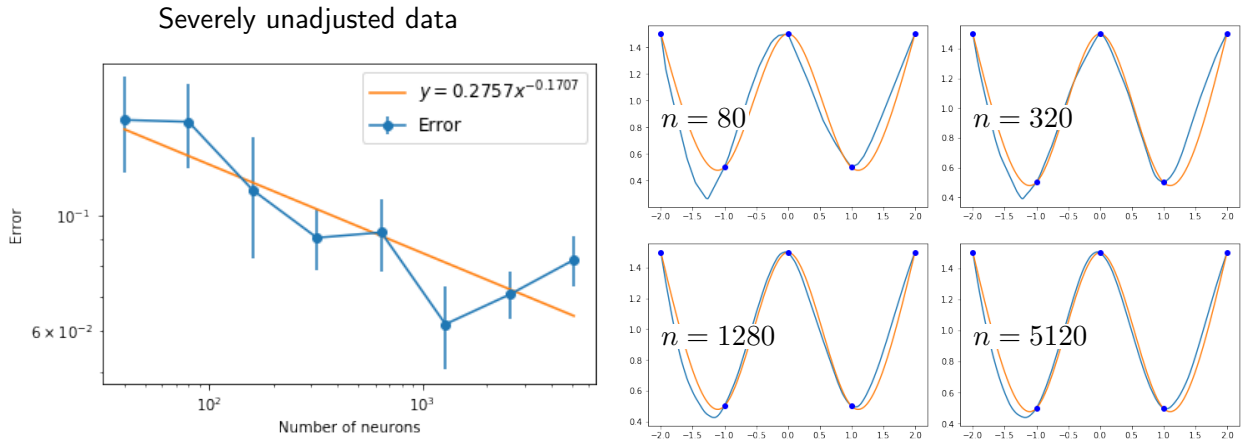

Figure 2.7: Effect of not adjusting the data. We use uniform initialization and add a linear function $10x + 10$ to the training data of Figure 2.1. In order to clearly show the difference between trained network function and the solution to the variational problem, we subtract $10x + 10$ from these two functions in the right panel. In the right panel we see that if we ignore u and v in the variational problem (2.22), the solution is slightly different from (2.24).

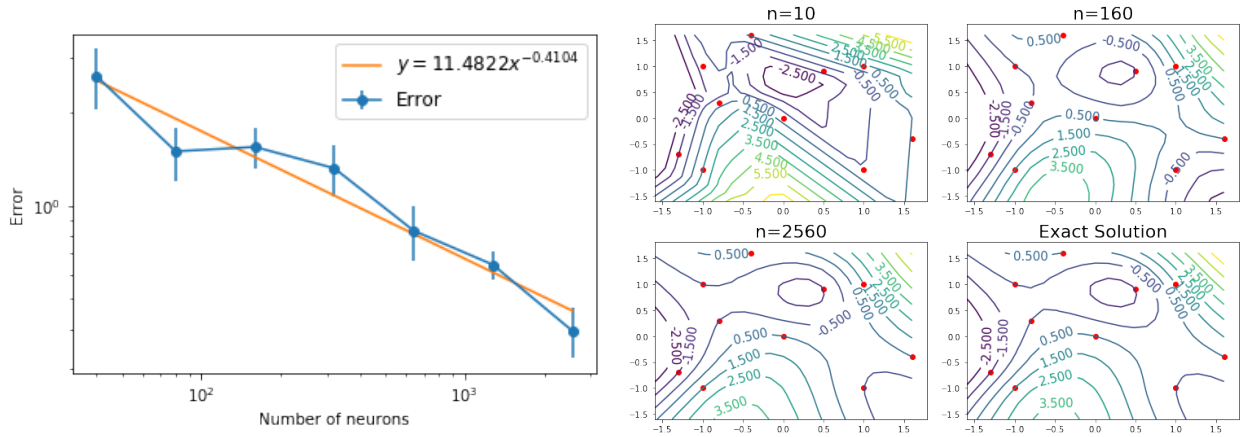

Figure 2.8: Illustration of Theorem 6. Similar to Figure 2.2, with the same initialization, but with a larger data set.

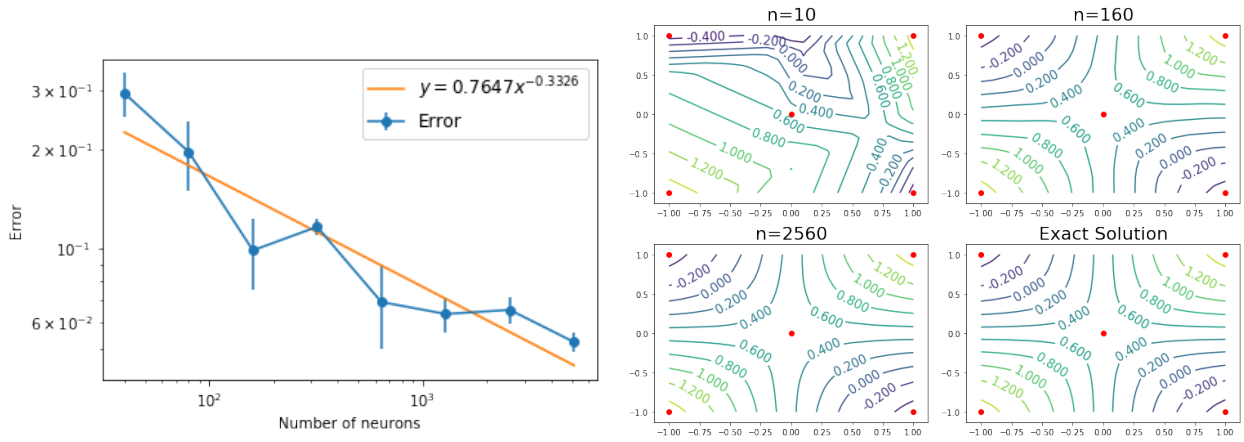

Figure 2.9: Illustration of Theorem 6. Similar to Figure 2.2, but with the Gaussian initialization $\boldsymbol{\mathcal{W}} \sim \mathcal{N}(0, I_d)$ and $\mathcal{B} \sim \mathcal{N}(0, 1)$.

in Figure 2.10 is $\boldsymbol{\mathcal{W}} \sim \mathcal{N}(0, I_d)$ and $\mathcal{B} \sim \mathcal{N}(0, 0.1)$. So we can use Theorem 19 to exactly compute the curvature penalty function and solve the variational problem (2.8) numerically.

2.B Additional Background on the NTK, Initialization, and Parametrization

In this appendix we provide a few additional details on the NTK, ASI initialization, standard vs NTK parametrization, and discuss the difference between our results and weight norm minimization.

2.B.1 NTK Convergence and Positive-definiteness

The convergence of the empirical NTK to a deterministic limiting NTK as the width of the network tends to infinity and the positive-definiteness of this limiting kernel can be ensured whenever the neural network converges to a Gaussian process. The arguments from [JGH18a] to prove convergence and positive definiteness hold in this case. As they mention, the limiting NTK only depends on the choice of the network activation function, the depth of the network, and the variance of the parameters at initialization. They prove positive definiteness when the input data is supported on a sphere. More generally, positive

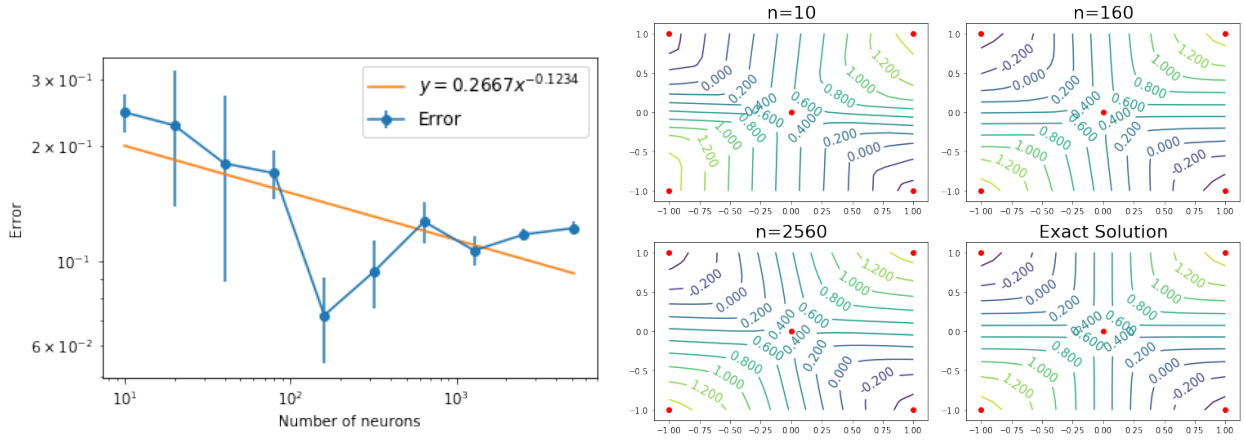

Figure 2.10: Illustration of Theorem 6. Similar to Figure 2.2, but with the Gaussian initialization $\boldsymbol{\mathcal{W}} \sim N(0, I_d)$ and $\mathcal{B} \sim N(0, 0.1)$. Because of the linear adjustment, the exact solution of the variational problem (2.8) is slightly different from the network output with a large number of hidden neurons.

definiteness can be proved based on the structure of the NTK as a covariance matrix. Let $\|f\|_p^2 = \mathbb{E}_{\mathbf{x}\sim p}[f(\mathbf{x})^T f(\mathbf{x})]$, where p denotes the distribution of inputs. The NTK is positive definite when the span of the partial derivatives $\partial_{\theta_i} f(\cdot, \theta)$, $i = 1, \ldots, d$, becomes dense in function space with respect to $\|\cdot\|_p$ as the width of the network tends to infinity [JGH18a]. For a finite data set $\mathbf{x}_1, \ldots, \mathbf{x}_M$, positive definiteness of the corresponding Gram matrix is equivalent to $\partial_{\theta_i} f(\mathbf{x}_j, \cdot)$ being linearly independent [DZP19, Theorem 3.1]. This condition for positive definiteness does not depend on the specific distribution of the parameters, but if anything it only depends on the support of the distribution of parameters and on the input data. The precise value of the least eigenvalue may be affected by changes in the distribution however. The convergence of the network function to a Gaussian process in the limit of infinite width and independent parameter initialization is a classic result [Nea96c]. To verify this Gaussian process assumption it is sufficient that $\sum_i W_i^{(2)} \sigma(\langle \mathbf{W}_i^{(1)}, \mathbf{x} \rangle + b_i)$ is a sum of independent random variables with finite variance.

2.B.2 Anti-Symmetrical Initialization (ASI)

The AntiSymmetrical Initialization (ASI) trick as proposed by [ZXL20] creates duplicate hidden units with opposite output weights, ensuring that $f(\cdot, \theta_0) \equiv 0$. More precisely, ASI defines $f_{\text{ASI}}(\mathbf{x}, \vartheta) = \frac{\sqrt{2}}{2} f(\mathbf{x}, \vartheta') - \frac{\sqrt{2}}{2} f(\mathbf{x}, \vartheta'')$. Here $\vartheta = (\vartheta', \vartheta'')$ is initialized with $\vartheta'_0 = \vartheta''_0$, so that

$$f_{\text{ASI}}(\mathbf{x}, \vartheta_0) = \sum_{i=1}^{n} \frac{\sqrt{2}}{2} \overline{V}_i^{(2)} [\langle \overline{\boldsymbol{V}}_i^{(1)}, \mathbf{x} \rangle + \overline{a}_i^{(1)}]_+ + \sum_{i=1}^{n} -\frac{\sqrt{2}}{2} \overline{V}_i^{(2)} [\langle \overline{\boldsymbol{V}}_i^{(1)}, \mathbf{x} \rangle + \overline{a}_i^{(1)}]_+ \equiv 0.$$

The parameter vector at initialization is thus $\vartheta_0 = \text{vec}(\overline{\boldsymbol{V}}^{(1)}, \overline{\boldsymbol{V}}^{(1)}, \overline{\boldsymbol{a}}^{(1)}, \overline{\boldsymbol{a}}^{(1)}, \frac{\sqrt{2}}{2}\overline{\boldsymbol{V}}^{(2)}, -\frac{\sqrt{2}}{2}\overline{\boldsymbol{V}}^{(2)}, \frac{\sqrt{2}}{2}\overline{a}^{(2)}, -\frac{\sqrt{2}}{2}\overline{a}^{(2)})$.

The basic statistics on the size of the parameters remains like (2.3), even if now there are perfectly correlated pairs of parameters. Hence the analysis and results on limits when the number of hidden units tends to infinity remain valid under ASI. The ASI is not needed for our analysis, which can be used to compare different types of initialization procedures, but it simplifies some of the presentation. One motivation for using ASI in practical applications is that it provides a simple way to implement a simple output function at initialization. Since the output function at initialization directly influences the bias of the gradient descent solution, this is a particular way to control the bias. Manipulating the bias from initialization is also the motivation presented by [ZXL20]. A related discussion also appears in [SDP20].

2.B.3 Standard vs NTK Parametrization

We have focused on the standard parametrization of the neural network. [JGH18a] use a non-standard parametrization which is now known as the NTK parametrization. We briefly discuss the difference. A network with NTK parametrization is described as

$$\begin{cases} \mathbf{h}^{(l+1)} = \sqrt{\frac{1}{n_l}} \mathbf{W}^{(l+1)} \mathbf{x}^l + \mathbf{b}^{(l+1)} \\ \mathbf{x}^{(l+1)} = \phi(\mathbf{h}^{(l+1)}) \end{cases} \quad \text{and} \quad \begin{cases} W_{i,j}^{(l)} \sim \mathcal{N}(0,1) \\ b_j^{(l)} \sim \mathcal{N}(0,1) \end{cases}.$$

In contrast to the standard parametrization, in the NTK parametrization the factor $\sqrt{1/n_l}$ is carried outside of the trainable parameter. In this case, the scaling of the derivatives is $\nabla_{W_{i,j}^{(1)}} f(x, \theta_0) = O(n^{-\frac{1}{2}})$ and $\nabla_{W_i^{(2)}} f(x, \theta_0) = O(n^{-\frac{1}{2}})$. In turn, during training the changes of $W_{i,j}^{(1)}$ and $W_i^{(2)}$ are comparable in magnitude. This implies that we can not ignore the changes of $W_{i,j}^{(1)}$ and approximate the dynamics by that of the linearized model that trains only the output weights as we did in the case of the standard parametrization. In particular, we can not use problem (2.20) to describe the result of gradient descent as $n \to \infty$.

2.B.4 Weight Norm Minimization

[SES19] studied networks of the form $f(x, \theta) = \sum_{i=1}^{n} W_i^{(2)}[W_i^{(1)}x + b_i^{(1)}]_+ + b^{(2)}$ allowing the width to tend to infinity. They showed that the minimum weight norm for approximating a given function g is related to a measure of the smoothness of g by $\lim_{\epsilon \to 0}(\inf_\theta C(\theta) \text{ s.t. } \|f(\cdot, \theta) - g\|_\infty \leq \epsilon) = \max\{\int_{-\infty}^{\infty} |g''(x)| \, \mathrm{d}x, \ |g'(-\infty) + g'(\infty)|\}$, where $C(\theta) = \frac{1}{2}\sum_{i=1}^{n}((W_i^{(2)})^2 + (W_i^{(1)})^2)$. Here the derivatives are understood in the weak sense. This implies that infinite width shallow networks trained with weight norm regularization (sparing biases) represent functions with smallest 1-norm of the second derivative, an example of which are linear splines. (Note that $C(\theta)$ is not strictly convex in the space of all parameters and also the 1-norm of the second derivative is not strictly convex, hence the solution is not unique).

The result of [SES19] is illuminating in that it connects properties of the parameters and properties of the represented functions. However, the result does not necessarily inform us about the functions represented by the network upon gradient descent training without explicit weight norm regularization. Indeed, if we initialize the parameters by (2.3) with sub-Gaussian distribution, the neural network can be approximated by the linearized model. Then by Theorem 20, $\|\omega - \theta_0\|_2$ is minimized rather than $\|\omega\|_2$. But in this case $\|\theta_0\|_2$ is bounded away from zero with high probability and the 2-norm of all parameters (or also of the weights only) is not minimized. On the other hand, if we initialize the parameters with $\|\theta_0\|_2$ close to 0, then the neural network might not be well approximated by the linearized model. This has been observed experimentally by [COB19] and we further illustrate it in

Appendix 2.B.5.

Even if we assume that the linearization of a network at the origin is valid, in order for the network to approximate certain complex functions, the weights necessarily have to be bounded away from zero. This means that reaching zero training error requires to move far from the basis point, where the difference between linearized and non-linearized model could become significant. In turn, the implicit bias description derived from a linearization at the origin may not accurately reflect the implicit bias of gradient descent in the original non-linearized model.

The above paragraphs discuss why the result of [SES19] does not apply to gradient descent training without weight norm regularization. It is also interesting to discuss the difference between our result and the result of [SES19]. In our result, the implicit bias of gradient descent without weight norm regularization is characterized by 2-norm of the second derivative weighted by $1/\zeta$, which is a RKHS-norm. In the result of [SES19], they showed that training with weight norm regularization (sparing biases) leads to functions with smallest 1-norm of the second derivative, which is not a RKHS norm. The reason why training without weight decay gives RKHS norm is because the training trajectory can be approximated by that of a linear model, which corresponds to a certain RKHS. And for training with weight norm regularization, the weight in the first layer is regularized, so it changes the feature space and we can no longer regard that as a linear model. Some works give empirical evidence that minimizing a non-RKHS norm can have better generalization than minimizing an RKHS norm because of the limitation of linear models and the kernel regime. However, as far as we know, there is no theory which shows that a non-RKHS-norm could result in better generalization than a RKHS norm.

The paper by [PN19] follows the approach of [SES19] and generalizes the result of [SES19] to different types of activation functions σ. Then they show that minimizing the weight "norm" of two-layer neural networks with activation function σ is actually minimizing 1-norm of $\mathrm{L}f$ in place of the second derivative, where f is the output function of the neural network. Here L and σ satisfy $L\sigma = \delta$, i.e., σ is a Green's function of L. Such activation functions can

be used in combination with our analysis. We comment further on such generalizations in Appendix 2.J.

2.B.5 Basis Parameter for Linearization of the Model

We discuss how the quality of the approximation of a neural network by a linearized model depends on the basis point. For a feedforward ReLU network and a list $\mathcal{X} = (x_i)_{i=1}^m$ of input data points, the mapping $\theta \mapsto f(\mathcal{X}, \theta) = [f(x_1, \theta), \ldots, f(x_m, \theta)]$ is piecewise multilinear. Each of the pieces is smooth and we can assume that it is approximated reasonably well by its Taylor expansion. However, the quality of the approximation can drop when we cross the boundary between smooth pieces. Consider a single-input network with a layer of n ReLUs and a single output unit. At an input x the prediction is $f(x;\theta) = \sum_{j=1}^n W_j^{(2)}[W_j^{(1)}x + b_j^{(1)}]_+ + b^{(2)}$, where $\theta = \mathrm{vec}(\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{W}^{(2)}, b^{(2)})$. The Jacobian is non-smooth whenever $\theta \in H_{xj} = \{W_j^{(1)}x + b_j^{(1)} = 0\}$ for some $j = 1, \ldots, n$. Hence for m input data points x_i, $i = 1, \ldots, m$, the locus of non-smoothness is given by m central hyperplanes H_{ij}, $i = 1, \ldots, m$ in the parameter space of each hidden unit $j = 1, \ldots, n$. For an individual ReLU, if the parameter θ_0 is drawn from a centrally symmetric probability distribution, the probability p that an ϵ ball around $c\theta_0$ intersects one of the non-linearity hyperplanes H_i, $i = 1, \ldots, m$, behaves roughly as $p = O(mc^{-1})$ as c goes to infinity. Hence we can expect that the prediction function will be better approximated by its linearization $f^{\mathrm{lin}}(x, \theta) = f(x, c\theta_0) + \nabla_\theta f(x, c\theta_0)(\theta - c\theta_0)$ at a point $c\theta_0$ if c is larger. This is well reflected numerically in Figure 2.11. As we see, for larger initialization the model looks more linear. We observed that this qualitative behavior remains same if we try to adjust the size of the window around the initial value.

2.C Proof of Theorem 1 and Theorem 6

The proof of Theorem 1 and Theorem 6 is the compilation of results from Sections 2.4, 2.5, 2.6 and 2.7. Next we give the proof of Theorem 6. Theorem 1 can be similarly proved.

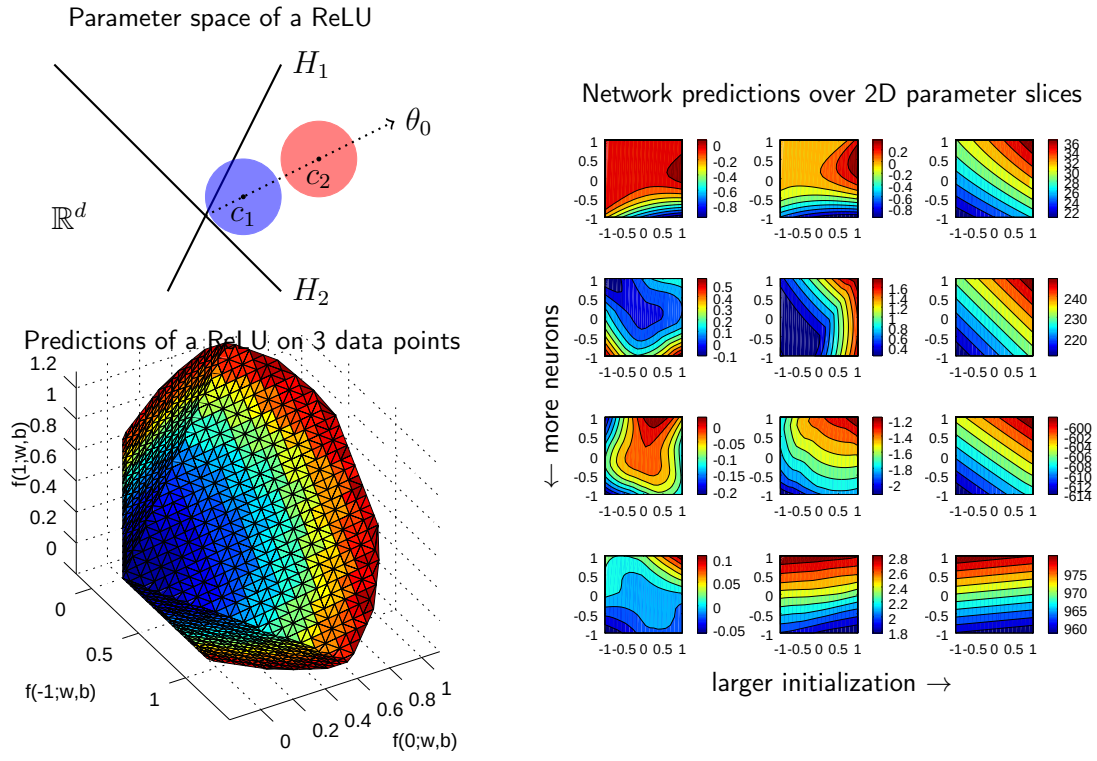

Figure 2.11: Left: For a single ReLU, the map $\theta \mapsto f(\mathcal{X}, \theta)$ from parameters to prediction vectors over a set $\mathcal{X} = \{x_1, \ldots, x_m\}$ of m input data points is piecewise linear, with pieces separated by m central hyperplanes. Right: Shown is the prediction $f(x, \theta)$ of a shallow ReLU network on a fixed input point x, over a 2D slice of parameters $\theta = c\theta_0 + v_1\xi_1 + v_2\xi_2$ spanned by two random orthogonal unit norm vectors v_1, v_2 and parametrized by $(\xi_1, \xi_2) \in [-1, 1]^2$. From top to bottom, the number of hidden units is $n = 1, 5, 25, 125$ and in each row the initial parameter θ_0 is drawn i.i.d. from a standard Gaussian. In each column we use a different scaling constant $c = 0, 0.5, 10$. As we see, for larger scaling c of the initialization the model looks more linear.

Proof of Theorem 6. The convergence to zero training error for ReLU networks is by now a well known result [DZP19, ALS19b]. We proceed with the implicit bias result.

For simplicity, we give out the proof under ASI (see Appendix 2.B.2). In Section 2.7, we relax the optimization problem (2.20) to (2.25). Suppose $(\overline{\alpha}, \overline{\boldsymbol{u}}, \overline{v})$ is the solution of (2.25). The we can adjust the training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^M$ to $\{(\mathbf{x}_i, y_i - \langle \overline{\boldsymbol{u}}, \mathbf{x}_i \rangle - \overline{v})\}_{i=1}^M$. It's easy to see that on the adjusted training samples, $(\overline{\alpha}, \mathbf{0}, 0)$ is the solution of (2.25). Then $\overline{\alpha}$ is the solution of (2.20) on the adjusted data. Furthermore, the solution of (2.20) in function space, $g(\mathbf{x}, \overline{\alpha})$, equals to the solution of (2.25) in function space, $g(\mathbf{x}, (\overline{\alpha}, 0, 0))$, i.e.,

$$g(\mathbf{x}, \overline{\alpha}) = g(\mathbf{x}, (\overline{\alpha}, \mathbf{0}, 0)). \tag{2.38}$$

It we change the variable α to γ as in Section 2.7, we get

$$g(\mathbf{x}, (\overline{\alpha}, \mathbf{0}, 0)) = g(\mathbf{x}, (\overline{\gamma}, \mathbf{0}, 0)), \tag{2.39}$$

On any compact set $D \subset \mathbb{R}^d$, according to Theorem 12,

$$\sup_{\mathbf{x} \in D} |g_n(\mathbf{x}, \overline{\alpha}_n) - g(\mathbf{x}, \overline{\alpha})| = O_p(n^{-1/2}), \tag{2.40}$$

where $g_n(\mathbf{x}, \overline{\alpha}_n)$ is the solution of problem (2.19) in function space. Since problem (2.19) is equivalent to problem (2.18), $g_n(\mathbf{x}, \overline{\alpha}_n)$ is also the solution of (2.18) in function space. According to discussion in Section 2.5, $f^{\text{lin}}(\mathbf{x}, \widetilde{\omega}_\infty)$ is the solution of (2.18). Then we have

$$g_n(\mathbf{x}, \overline{\alpha}_n) = f^{\text{lin}}(\mathbf{x}, \widetilde{\omega}_\infty). \tag{2.41}$$

According to Corollary 11, we get

$$\sup_{\mathbf{x} \in D} |f^{\text{lin}}(\mathbf{x}, \tilde{\omega}_\infty) - f(\mathbf{x}, \theta^*)| = O_p(n^{-\frac{1}{2}}). \tag{2.42}$$

Finally, according to Theorem 16 (to prove Theorem 1, apply Theorem 13 and Proposition

14), $g(\mathbf{x}, (\overline{\gamma}, \mathbf{0}, 0))$ is the solution of (2.8), which is $g^*(\mathbf{x})$. It means that

$$g(\mathbf{x}, (\overline{\gamma}, 0, 0)) = g^*(\mathbf{x}). \tag{2.43}$$

Combining (2.38), (2.39), (2.40), (2.41), (2.42), (2.43), we prove the theorem. □

2.D Implicit Bias in Parameter Space for a Linearized Model

[ZXL20] show that gradient flow converges to the solution with zero empirical loss which is closest to the initial weights. We show a similar result for the case of gradient descent with small enough learning rate.

Theorem 20 (Bias of the linearized model in parameter space)**.** *Consider a convex loss function ℓ with a unique finite minimum and its derivative is K-Lipschitz continuous, i.e., $|\frac{\mathrm{d}}{\mathrm{d}y}\ell(y_1, \hat{y}) - \frac{\mathrm{d}}{\mathrm{d}y}\ell(y_2, \hat{y})| \leq K|y_1 - y_2|$. If $\operatorname{rank}(\nabla_\theta f(\mathcal{X}, \theta_0)) = M$, then the gradient descent iteration (2.14) with learning rate $\eta \leq \frac{M}{Kn\lambda_{\max}(\hat{\Theta}_n)}$ converges to the unique solution of following constrained optimization problem:*

$$\min_{\omega} \|\omega - \theta_0\|_2 \quad s.t.\ f^{\mathrm{lin}}(\mathcal{X}, \omega) = \mathcal{Y}. \tag{2.44}$$

Remark 21 (Remark on Theorem 20, step size)**.** *Note that this statement is valid for the linearization of any set of functions, not only neural networks. The proof remains valid for a changing step size as long as this satisfies the required inequality.*

Remark 22 (Remark on Theorem 20, rank assumption)**.** *The assumption $\nabla_\theta f(\mathcal{X}, \theta_0) = M$ is satisfied in most cases when $n \geq M$ (here n refers to the number of parameters in θ since we use the linearized model). This is because $\nabla_\theta f(\mathcal{X}, \theta_0)$ is a $M \times n$ matrix. The M rows corresponds to M training samples and they are almost always linearly independent.*

Here we give out the proof of Theorem 20. We note that [ZXL20] prove a similar result for gradient flow. Our proof is for finite step size and different from theirs.

Proof of Theorem 20. We use gradient descent to minimize $L^{\text{lin}}(\omega) = \frac{1}{M}\sum_{i=1}^{M} \ell(f^{\text{lin}}(\mathbf{x}_i, \omega), y_i)$. First we prove that $\nabla_\omega L^{\text{lin}}(\omega)$ is Lipschitz continuous as follows:

$$
\begin{aligned}
&\|\nabla_\omega L^{\text{lin}}(\omega_1) - \nabla_\omega L^{\text{lin}}(\omega_2)\|_2 \\
=&\frac{1}{M}\|\nabla_\theta f(\mathcal{X}, \theta_0)^\top \nabla_{f^{\text{lin}}(\mathcal{X},\omega_1)} L - \nabla_\theta f(\mathcal{X}, \theta_0)^\top \nabla_{f^{\text{lin}}(\mathcal{X},\omega_2)} L\|_2 \\
\leq&\frac{1}{M}\|\nabla_\theta f(\mathcal{X}, \theta_0)^\top\|_2 \|\nabla_{f^{\text{lin}}(\mathcal{X},\omega_1)} L - \nabla_{f^{\text{lin}}(\mathcal{X},\omega_2)} L\|_2 \\
=&\frac{1}{M}\|\nabla_\theta f(\mathcal{X}, \theta_0)^\top\|_2 \sqrt{\sum_{i=1}^{M}\left(\frac{\mathrm{d}}{\mathrm{d}y} l(f^{\text{lin}}(\mathbf{x}_i, \omega_1), y_i) - \frac{\mathrm{d}}{\mathrm{d}y} l(f^{\text{lin}}(\mathbf{x}_i, \omega_2), y_i)\right)^2} \\
\leq&\frac{K}{M}\|\nabla_\theta f(\mathcal{X}, \theta_0)^\top\|_2 \|f^{\text{lin}}(\mathcal{X}, \omega_1) - f^{\text{lin}}(\mathcal{X}, \omega_2)\|_2 \quad \text{(K-Lipschitz continuity of } \ell) \\
=&\frac{K}{M}\|\nabla_\theta f(\mathcal{X}, \theta_0)^\top\|_2 \|\nabla_\theta f(\mathcal{X}, \theta_0)(\omega_1 - \omega_2)\|_2 \\
\leq&\frac{K}{M}\|\nabla_\theta f(\mathcal{X}, \theta_0)^\top\|_2 \|\nabla_\theta f(\mathcal{X}, \theta_0)\|_2 \|(\omega_1 - \omega_2)\|_2 \\
\leq&\frac{Kn}{M}\lambda_{\max}(\hat{\Theta}_n)\|\omega_1 - \omega_2\|_2.
\end{aligned}
$$

So $L^{\text{lin}}(\omega)$ is Lipschitz continuous with Lipschitz constant $\frac{Kn}{M}\lambda_{\max}(\hat{\Theta}_n)$. Since L^{lin} is convex over ω, gradient descent with learning rate $\eta = \frac{M}{Kn\lambda_{\max}(\hat{\Theta}_n)}$ converges to a global minimium of $L^{\text{lin}}(\omega)$. By assumption that $\text{rank}(\nabla_\theta f(\mathcal{X}, \theta_0)) = M$, the model can perfectly fit all data. Then the minimium of $L^{\text{lin}}(\omega)$ is zero and gradient descent converges to zero loss.

Let $\omega_\infty = \lim_{t\to\infty} \omega_t$. Then $f^{\text{lin}}(\mathcal{X}, \omega_\infty) = \mathcal{Y}$. According to gradient descent iteration,

$$
\begin{aligned}
\omega_\infty &= \theta_0 - \sum_{t=0}^{\infty} \eta \nabla_\theta f(\mathcal{X}, \theta_0)^T \nabla_{f^{\text{lin}}(\mathcal{X},\omega_t)} L^{\text{lin}} \\
&= \theta_0 - \eta \nabla_\theta f(\mathcal{X}, \theta_0)^T \sum_{t=0}^{\infty} \nabla_{f^{\text{lin}}(\mathcal{X},\omega_t)} L^{\text{lin}}.
\end{aligned}
$$

Since f^{lin} is linear over weights ω and $\|\omega - \theta_0\|_2$ is strongly convex, the constrained optimization problem (2.44) is a strongly convex optimization problem. The first order

optimality condition of the problem is

$$\begin{cases} \omega - \theta_0 + \nabla_\theta f^{\text{lin}}(\mathcal{X}, \theta_0)^T \lambda = 0, \\ f^{\text{lin}}(\mathcal{X}, \omega) = \mathcal{Y}. \end{cases} \tag{2.45}$$

Let $\lambda = \sum_{t=0}^{\infty} \nabla_{f^{\text{lin}}(\mathcal{X},\theta_t)} L$, we can easily check that ω_∞ satisfies condition (2.45). So ω_∞ is the solution of problem (2.44). □

Remark 23 (Remark on Theorem 20)**.** *Making an analogous statement to Theorem 20 to describe the bias in parameter space when training wide networks rather than the linearized model is interesting, but harder, because the gradient direction is no longer constant. [OS19] obtain bounds on the trajectory length in parameter space, putting the final solution within a factor* $4\beta/\alpha$ *of* $\min_\theta \|\theta_0 - \theta\|$*, where* β *and* α *are upper and lower bounds on the singular values of the Jacobian over the relevant region. However, currently it is unclear whether the solution upon gradient optimization is indeed the distance minimizer from initialization.*

Next we discuss the implicit bias of SGD (stochastic gradient descent) in parameter space. Consider the following stochastic gradient descent iteration for the linearized model:

$$\omega_0 = \theta_0, \quad \omega_{t+1} = \omega_t - \eta_t \frac{\mathrm{d}}{\mathrm{d}y} \ell(f^{\text{lin}}(\mathbf{x}_{r(t)}, \omega_t), y_{r(t)}) \nabla_\theta f(\mathbf{x}_{r(t)}, \theta_0), \tag{2.46}$$

where $r(t)$ is evenly chosen from the set $\{1, 2, ..., M\}$ and η_t is the learning rate at the step t. Typically, η_t needs to decay in order for SGD to converge. However, for overparametrized linearized model, we can show that SGD converges for constant learning rate and the implicit bias of SGD is the same as gradient descent under certain conditions. This is shown in the following theorem.

Theorem 24 (Bias of the linearized model in parameter space, SGD)**.** *Consider a convex loss function* ℓ *with a unique finite minimum and its derivative is* K*-Lipschitz continuous, i.e.,* $|\frac{\mathrm{d}}{\mathrm{d}y}\ell(y_1, \hat{y}) - \frac{\mathrm{d}}{\mathrm{d}y}\ell(y_2, \hat{y})| \leq K|y_1 - y_2|$*. If* $\operatorname{rank}(\nabla_\theta f(\mathcal{X}, \theta_0)) = M$*, the stochastic gradient descent iteration* (2.46) *with constant learning rate* $\eta_t = \eta \leq \frac{1}{K \max_j \|\nabla_\theta f(\mathbf{x}_j, \theta_0)\|_2^2}$ *converges to*

the unique solution of following constrained optimization problem with probability 1*:*

$$\min_{\omega} \|\omega - \theta_0\|_2 \quad s.t.\ f^{\text{lin}}(\mathcal{X}, \omega) = \mathcal{Y}. \tag{2.47}$$

Proof of Theorem 24. Let ω^* be the solution to the optimization problem (2.47). Let $\mathbf{z}_j = \nabla_\theta f(\mathbf{x}_j, \theta_0)$. It is easy to see that $\omega_t - \langle \omega_t - \omega^*, \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2}\rangle \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2}$ is the projection of ω_t onto the hyperplane $\{\langle \omega, \mathbf{z}_j\rangle\} = \{\langle \omega^*, \mathbf{z}_j\rangle\}$. So for any $\hat{\eta} \leq 1$, we have

$$\left\|\omega_t - \hat{\eta}\langle \omega_t - \omega^*, \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2}\rangle \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2} - \omega^*\right\|_2^2 = \|\omega_t - \omega^*\|_2^2 - (1 - (1-\hat{\eta})^2)\left|\langle \omega_t - \omega^*, \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2^2}\rangle\right|^2 \tag{2.48}$$

$$\leq \|\omega_t - \omega^*\|_2^2. \tag{2.49}$$

The length of the stochastic gradient in (2.46) can be bounded as follows:

$$\eta_t \frac{\mathrm{d}}{\mathrm{d}y}\ell(f^{\text{lin}}(\mathbf{x}_{r(t)}, \omega_t), y_{r(t)})\|\mathbf{z}_{r(t)}\|_2 \tag{2.50}$$

$$\leq \eta_t K |f^{\text{lin}}(\mathbf{x}_{r(t)}, \omega_t) - y_{r(t)}| \|\mathbf{z}_{r(t)}\|_2 \tag{2.51}$$

$$\leq K \frac{1}{K \max_j \|\nabla_\theta f(\mathbf{x}_j, \theta_0)\|_2^2} |f^{\text{lin}}(\mathbf{x}_{r(t)}, \omega_t) - y_{r(t)}| \|\mathbf{z}_{r(t)}\|_2 \tag{2.52}$$

$$= \frac{1}{\max_j \|\mathbf{z}_j\|_2^2} \langle \omega_t - \omega^*, \mathbf{z}_{r(t)}\rangle \|\mathbf{z}_{r(t)}\|_2 \tag{2.53}$$

$$\leq \frac{1}{\max_j \|\mathbf{z}_j\|_2^2} \|\mathbf{z}_{r(t)}\|_2^2 \langle \omega_t - \omega^*, \frac{\mathbf{z}_{r(t)}}{\|\mathbf{z}_{r(t)}\|_2}\rangle \tag{2.54}$$

$$\leq \langle \omega_t - \omega^*, \frac{\mathbf{z}_{r(t)}}{\|\mathbf{z}_{r(t)}\|_2}\rangle \tag{2.55}$$

Then according to (2.49), we have

$$\left\|\omega_t - \eta_t \frac{\mathrm{d}}{\mathrm{d}y}\ell(f^{\text{lin}}(\mathbf{x}_{r(t)}, \omega_t), y_{r(t)})\|\mathbf{z}_{r(t)}\|_2 \frac{\mathbf{z}_{r(t)}}{\|\mathbf{z}_{r(t)}\|_2} - \omega^*\right\|_2 \leq \|\omega_t - \omega^*\|_2. \tag{2.56}$$

The above equation means that

$$\|\omega_{t+1} - \omega^*\|_2 \leq \|\omega_t - \omega^*\|_2. \tag{2.57}$$

Then $\|\omega_t\|_2$ is bounded and $\lim_{t\to\infty} \|\omega_t - \omega^*\|_2 - \|\omega_{t+1} - \omega^*\|_2 = 0$. Next we show that for any convergent subsequence $\{\omega_{t_k}\}_{k\geq 1}$ of $\{\omega_t\}_{t\geq 1}$, we have $\lim_{k\to\infty} \omega_{t_k} = \omega^*$.

Let $\lim_{k\to\infty} \omega_{t_k} = \bar{\omega}$. Asuume that $\bar{\omega} \neq \omega^*$. According to the first order optimality (2.45), we have that $\omega^* = \theta_0 + \sum_{j=1}^M \lambda_j \mathbf{z}_j$. From the stochastic gradient descent iterations, we have $\omega_t = \theta_0 - \eta \sum_{s=1}^{t-1} \frac{\mathrm{d}}{\mathrm{d}y} \ell(f^{\mathrm{lin}}(\mathbf{x}_{r(s)}, \omega_s), y_{r(s)}) \mathbf{z}_{r(s)}$. Then $\omega_t - \omega^*$ is a linear combination of $\{\omega_j\}_{j=1}^M$. It means that $\bar{\omega} - \omega^*$ is a linear combination of $\{\mathbf{z}_j\}_{j=1}^M$. Since $\bar{\omega} - \omega^*$ is not zero, the set $A = \left\{ j : \left| \langle \bar{\omega} - \omega^*, \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2} \rangle \right| > 0 \right\}$ is not empty. With probability 1, we have that $r(t) \in A$ infinitely many times. So for any given k, we can find $t'_k \geq t_k$ such that $r(t'_k) \in A$ and $r(t) \notin A$ for $t_k \leq t < t'_k$.

When we prove (2.57), we only use the property that $f^{\mathrm{lin}}(\mathbf{x}_{r(t)}, \omega^*) = y_{r(t)}$. When $t_k \leq t < t'_k$, we have $r(t) \notin A$, so $\langle \bar{\omega}, \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2} \rangle = \langle \omega^*, \frac{\mathbf{z}_{r(t)}}{\|\mathbf{z}_{r(t)}\|_2} \rangle$. It means that $f^{\mathrm{lin}}(\mathbf{x}_{r(t)}, \bar{\omega}) = f^{\mathrm{lin}}(\mathbf{x}_{r(t)}, \omega^*) = y_{r(t)}$. Using the same argument as (2.57), we have $\|\omega_{t+1} - \bar{\omega}\|_2 \leq \|\omega_t - \bar{\omega}\|_2$ when $t_k \leq t < t'_k$. Then $\|\omega_{t'_k} - \bar{\omega}\|_2 \leq \|\omega_{t_k} - \bar{\omega}\|_2$. Then $\lim_{k\to\infty} \omega_{t'_k} = \lim_{k\to\infty} \omega_{t_k} = \bar{\omega}$. According to (2.48), we have

$$\|\omega_{t+1} - \omega^*\|_2^2 = \|\omega_t - \omega^*\|_2^2 - (1 - (1 - \tilde{\eta}_t)^2) \left| \langle \omega_t - \omega^*, \frac{\mathbf{z}_{r(t)}}{\|\mathbf{z}_{r(t)}\|_2} \rangle \right|^2 \tag{2.58}$$

$$\text{and } \tilde{\eta}_t = \frac{\eta \frac{\mathrm{d}}{\mathrm{d}y} \ell(f^{\mathrm{lin}}(\mathbf{x}_{r(t)}, \omega_t), y_{r(t)}) \|\mathbf{z}_{r(t)}\|_2}{\left| \langle \omega_t - \omega^*, \frac{\mathbf{z}_{r(t)}}{\|\mathbf{z}_{r(t)}\|_2} \rangle \right|}. \tag{2.59}$$

Since $\lim_{k\to\infty} \omega_{t'_k} = \bar{\omega}$, for sufficiently large k we have

$$\left| \langle \omega_{t'_k} - \omega^*, \frac{\mathbf{z}_{r(t'_k)}}{\|\mathbf{z}_{r(t'_k)}\|_2} \rangle \right|^2 \geq \frac{1}{2} \min_{j\in A} \left| \langle \bar{\omega} - \omega^*, \frac{\mathbf{z}_{r(j)}}{\|\mathbf{z}_{r(j)}\|_2} \rangle \right|^2 \tag{2.60}$$

$$= \Omega(1), \tag{2.61}$$

and

$$\tilde{\eta}_t \geq \frac{1}{2}\frac{\eta \min_{j\in A} \frac{\mathrm{d}}{\mathrm{d}y}\ell(f^{\mathrm{lin}}(\mathbf{x}_j,\bar{\omega}),y_j)\min_{j\in A}\|\mathbf{z}_j\|_2}{\|\bar{\omega}-\omega^*\|_2} \tag{2.62}$$

$$= \Omega(1)\min_{j\in A}\frac{\mathrm{d}}{\mathrm{d}y}\ell(f^{\mathrm{lin}}(\mathbf{x}_j,\bar{\omega}),y_j) \tag{2.63}$$

$$= \Omega(1), \tag{2.64}$$

where (2.64) holds because $f^{\mathrm{lin}}(\mathbf{x}_j,\bar{\omega}) - y_j = \langle \bar{\omega}-\omega^*, \mathbf{z}_j\rangle \neq 0$ for all $j \in A$ and $\frac{\mathrm{d}}{\mathrm{d}y}\ell(y,\hat{y}) = 0$ if and only if $y = \hat{y}$ according to the fact that loss function ℓ has a unique finite minimum. From (2.61) and (2.64) we have $\|\omega_{t'_k} - \omega^*\|_2^2 - \|\omega_{t'_k+1} - \omega^*\|_2^2 = \Omega(1)$. This contradicts the fact that $\lim_{t\to\infty}\|\omega_t-\omega^*\|_2 - \|\omega_{t+1}-\omega^*\|_2 = 0$. Then the assumption $\bar{\omega} \neq \omega^*$ is not true. So for any convergent subsequence $\{\omega_{t_k}\}_{k\geq 1}$ of $\{\omega_t\}_{t\geq 1}$, we have $\lim_{k\to\infty}\omega_{t_k} = \omega^*$. Combining the above statement with the fact that $\|\omega_t\|_2$ is bounded, we have $\lim_{t\to\infty}\omega_t = \omega^*$ □

Remark 25 (Remark on Theorem 24)**.** *Theorem 24 shows that SGD and gradient descent has the same implicit bias in parameter space. Then our main theorem also holds for SGD training.*

2.E Proof of Theorem 10

We note that assumption $\liminf_{n\to\infty}\lambda_{\min}(\hat{\Theta}_n) > 0$ is satisfied if the empirical NTK converges and the limit NTK is positive definite. For details see Appendix 2.B.1.

Proof of Theorem 10. According to (2.14),

$$\omega_{t+1} = \omega_t - \eta\nabla_\theta f(\mathcal{X},\theta_0)^T\nabla_{f^{\mathrm{lin}}(\mathcal{X},\omega_t)}L^{\mathrm{lin}}.$$

Since we use the MSE loss, we have

$$\omega_{t+1} = \omega_t - \eta\nabla_\theta f(\mathcal{X},\theta_0)^T(f^{\mathrm{lin}}(\mathcal{X},\omega_t) - \mathcal{Y}).$$

Using (2.12), we get

$$\begin{aligned} f^{\text{lin}}(\mathcal{X}, \omega_{t+1}) &= f^{\text{lin}}(\mathcal{X}, \omega_t) - \eta \nabla_\theta f(\mathcal{X}, \theta_0) \nabla_\theta f(\mathcal{X}, \theta_0)^T (f^{\text{lin}}(\mathcal{X}, \omega_t) - \mathcal{Y}) \\ &= f^{\text{lin}}(\mathcal{X}, \omega_t) - n\eta \hat{\Theta}_n (f^{\text{lin}}(\mathcal{X}, \omega_t) - \mathcal{Y}). \end{aligned}$$

Then we have

$$f^{\text{lin}}(\mathcal{X}, \omega_{t+1}) - \mathcal{Y} = (I - n\eta \hat{\Theta}_n)(f^{\text{lin}}(\mathcal{X}, \omega_t) - \mathcal{Y}),$$

and

$$\begin{aligned} f^{\text{lin}}(\mathcal{X}, \omega_t) - \mathcal{Y} &= (I - n\eta \hat{\Theta}_n)^t (f^{\text{lin}}(\mathcal{X}, \theta_0) - \mathcal{Y}) \\ &= (I - n\eta \hat{\Theta}_n)^t (f(\mathcal{X}, \theta_0) - \mathcal{Y}). \end{aligned}$$

According to the update rule of ω_t, we know that $\omega_t = \nabla_\theta f(\mathcal{X}, \theta_0)^T \xi + \theta_0$, where ξ is a column vector. Then we have

$$\begin{aligned} f^{\text{lin}}(\mathcal{X}, \omega_t) - \mathcal{Y} &= f^{\text{lin}}(\mathcal{X}, \omega_t) - f(\mathcal{X}, \theta_0) + f(\mathcal{X}, \theta_0) - \mathcal{Y} \\ &= \nabla_\theta f(\mathcal{X}, \theta_0)(\omega_t - \theta_0) + f(\mathcal{X}, \theta_0) - \mathcal{Y} \\ &= \nabla_\theta f(\mathcal{X}, \theta_0) \nabla_\theta f(\mathcal{X}, \theta_0)^T \xi + f(\mathcal{X}, \theta_0) - \mathcal{Y} \\ &= n\hat{\Theta}_n \xi + f(\mathcal{X}, \theta_0) - \mathcal{Y} \\ &= (I - n\eta \hat{\Theta}_n)^t (f(\mathcal{X}, \theta_0) - \mathcal{Y}). \end{aligned}$$

From above equation we can solve for ξ:

$$\xi = -n^{-1} \hat{\Theta}_n^{-1} [I - (I - n\eta \hat{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}).$$

Therefore

$$\omega_t = -n^{-1} \nabla_\theta f(\mathcal{X}, \theta_0)^T \hat{\Theta}_n^{-1} [I - (I - n\eta \hat{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}) + \theta_0. \tag{2.65}$$

For any $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned} f^{\text{lin}}(\mathbf{x}, \omega_t) &= f(\mathbf{x}, \theta_0) + \nabla_\theta f(\mathbf{x}, \theta_0)(\omega_t - \theta_0) \\ &= f(\mathbf{x}, \theta_0) - n^{-1} \nabla_\theta f(\mathbf{x}, \theta_0) \nabla_\theta f(\mathcal{X}, \theta_0)^T \hat{\Theta}_n^{-1} [I - (I - n\eta \hat{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}). \end{aligned} \tag{2.66}$$

For the training process (2.17), we can define the corresponding empirical neural tangent kernel in the following way:

$$\tilde{\Theta}_n = \frac{1}{n} \nabla_{\mathbf{W}^{(2)}} f(\mathcal{X}, \theta_0) \nabla_{\mathbf{W}^{(2)}} f(\mathcal{X}, \theta_0)^T.$$

Using the same argument, we have

$$\widetilde{\mathbf{W}}_t^{(2)} = -n^{-1} \nabla_{\mathbf{W}^{(2)}} f(\mathcal{X}, \theta_0)^T \tilde{\Theta}_n^{-1} [I - (I - n\eta \tilde{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}) + \overline{\mathbf{W}}_0^{(2)} \tag{2.67}$$

and

$$f^{\text{lin}}(\mathbf{x}, \widetilde{\omega}_t) = f(\mathbf{x}, \theta_0) - n^{-1} \nabla_{\mathbf{W}^{(2)}} f(\mathbf{x}, \theta_0) \nabla_{\mathbf{W}^{(2)}} f(\mathcal{X}, \theta_0)^T \tilde{\Theta}_n^{-1} [I - (I - n\eta \tilde{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}). \tag{2.68}$$

According to (2.66) and (2.68), we have

$$\begin{aligned} &|f^{\text{lin}}(\mathbf{x}, \widetilde{\omega}_t) - f^{\text{lin}}(\mathbf{x}, \omega_t)| \\ =&n^{-1} \Big| \nabla_\theta f(\mathbf{x}, \theta_0) \nabla_\theta f(\mathcal{X}, \theta_0)^T \hat{\Theta}_n^{-1} [I - (I - n\eta \hat{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}) \\ &- \nabla_{\mathbf{W}^{(2)}} f(\mathbf{x}, \theta_0) \nabla_{\mathbf{W}^{(2)}} f(\mathcal{X}, \theta_0)^T \tilde{\Theta}_n^{-1} [I - (I - n\eta \tilde{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}) \Big|. \end{aligned} \tag{2.69}$$

The next step is to compute the difference between $\tilde{\Theta}_n$ and $\hat{\Theta}_n$. Let $\Delta\Theta = \hat{\Theta}_n - \tilde{\Theta}_n$, then the ij-th entry of the matrix $\Delta\Theta$ is

$$\begin{aligned} (\Delta\Theta)_{ij} =& \frac{1}{n} \Bigg[\sum_{k=1}^{n} \Big(\Big\langle \nabla_{\mathbf{W}_k^{(1)}} f(\mathbf{x}_i, \theta_0), \nabla_{\mathbf{W}_k^{(1)}} f(\mathbf{x}_j, \theta_0) \Big\rangle + \nabla_{b_k^{(1)}} f(\mathbf{x}_i, \theta_0) \nabla_{b_k^{(1)}} f(\mathbf{x}_j, \theta_0) \Big) \\ &+ \nabla_{b^{(2)}} f(\mathbf{x}_i, \theta_0) \nabla_{b^{(2)}} f(\mathbf{x}_j, \theta_0) \Bigg]. \end{aligned} \tag{2.70}$$

Given $\mathbf{x} \in \mathbb{R}^d$, we have

$$\begin{aligned}
\|\nabla_{\mathbf{W}_k^{(1)}} f(\mathbf{x}, \theta_0)\| =& \|W_k^{(2)} H(\langle \mathbf{W}_k^{(1)}, \mathbf{x} \rangle + b_k^{(1)}) \cdot \mathbf{x}\| \leq |W_k^{(2)}| \|\mathbf{x}\| && (2.71)\\
|\nabla_{b_k^{(1)}} f(\mathbf{x}, \theta_0)| =& |W_k^{(2)} H(\langle \mathbf{W}_k^{(1)}, \mathbf{x} \rangle + b_k^{(1)})| \leq |W_k^{(2)}| && (2.72)\\
|\nabla_{W_k^{(2)}} f(\mathbf{x}, \theta_0)| =& [\langle \mathbf{W}_k^{(1)}, \mathbf{x} \rangle + b_k^{(1)}]_+ \leq \|\mathbf{W}_k^{(1)}\| \|\mathbf{x}\|_2 + b_k^{(1)} && (2.73)\\
|\nabla_{b^{(2)}} f(\mathbf{x}, \theta_0)| =& 1. && (2.74)
\end{aligned}$$

Therefore,

$$\begin{aligned}
|(\Delta\Theta)_{ij}| &\leq \frac{1}{n} \left[\sum_{k=1}^{n} \Big(|W_k^{(2)}|^2 \|\mathbf{x}_i\| \|\mathbf{x}_j\| + |W_k^{(2)}|^2 \Big) + 1 \right] \\
&= \frac{\|\mathbf{x}_i\| \|\mathbf{x}_j\| + 1}{n} \sum_{k=1}^{n} |W_k^{(2)}|^2 + \frac{1}{n}.
\end{aligned} \tag{2.75}$$

According to initialization (2.3), $W_k^{(2)} \overset{d}{=} \sqrt{1/n}\, \mathcal{W}^{(2)}$. Then according to the law of large numbers, $\lim_{n\to\infty} \sum_{k=1}^{n} |W_k^{(2)}|^2 = \mathbb{E}|\mathcal{W}^{(2)}|^2$ almost surely as $n \to \infty$. Then $\sum_{k=1}^{n} |W_k^{(2)}|^2 = O_p(1)$ and $|(\Delta\Theta)_{ij}| = O_p(n^{-1})$.

Since the size of $\Delta\Theta$ is $M \times M$, which does not change as n goes up. So $\|\Delta\Theta\|_2 = O_p(n^{-1})$, which means $\|\hat{\Theta}_n - \tilde{\Theta}_n\|_2 = O_p(n^{-1})$.

Now we measure the difference of each part in (2.69). According to the assumption $\inf_n \lambda_{\min}(\hat{\Theta}_n) > 0$, we have

$$\begin{aligned}
\lambda_{\min}(\hat{\Theta}_n^{-1}) &\geq \frac{1}{\inf_n \lambda_{\min}(\hat{\Theta}_n)} = O_p(1) && (2.76)\\
\lambda_{\min}(\tilde{\Theta}_n^{-1}) &\geq \frac{1}{\inf_n \lambda_{\min}(\hat{\Theta}_n) - O_p(n^{-1})} = O_p(1). && (2.77)
\end{aligned}$$

Therefore

$$\begin{aligned}
\|\hat{\Theta}_n^{-1} - \tilde{\Theta}_n^{-1}\|_2 &= \|\hat{\Theta}_n^{-1}(\tilde{\Theta}_n - \hat{\Theta}_n)\tilde{\Theta}_n^{-1}\|_2 \\
&\leq \|\hat{\Theta}_n^{-1}\|_2 \|\Delta\Theta\|_2 \|\tilde{\Theta}_n^{-1}\|_2 \\
&= O_p(n^{-1}).
\end{aligned} \tag{2.78}$$

The assumption $\eta < \frac{2}{n\lambda_{\max}(\hat{\Theta}_n)}$ implies

$$\|I - n\eta\hat{\Theta}_n\|_2 < 1, \tag{2.79}$$

and

$$\begin{aligned}\|I - n\eta\tilde{\Theta}_n\|_2 &\le \|I - n\eta\hat{\Theta}_n\|_2 + n\eta\|\hat{\Theta}_n - \Theta\|_2 \\ &\le \max\{n\eta\frac{\lambda_{\max}(\Theta)}{2}, 1 - n\eta\lambda_{\min}(\hat{\Theta}_n)\} + O_p(n^{-1}).\end{aligned}$$

For any $\delta > 0$, as n is large enough, we also have $\|I - n\eta\tilde{\Theta}_n\|_2 < 1$ with probability at least $1 - \delta$. Then as n is large enough,

$$\begin{aligned}&\|[I - (I - n\eta\hat{\Theta}_n)^t] - [I - (I - n\eta\tilde{\Theta}_n)^t]\|_2 \\ &= \|(I - n\eta\hat{\Theta}_n)^t - (I - n\eta\tilde{\Theta}_n)^t\|_2 \\ &\le \|[(I - n\eta\hat{\Theta}_n) - (I - n\eta\tilde{\Theta}_n)](I - n\eta\hat{\Theta}_n)^{t-1}\|_2 \\ &\quad + \|(I - n\eta\tilde{\Theta}_n)[(I - n\eta\hat{\Theta}_n) - (I - n\eta\tilde{\Theta}_n)](I - n\eta\hat{\Theta}_n)^{t-2}\|_2 \\ &\quad + \cdots \\ &\quad + \|(I - n\eta\tilde{\Theta}_n)^{t-1}[(I - n\eta\hat{\Theta}_n) - (I - n\eta\tilde{\Theta}_n)]\|_2 \\ &\le \eta\|\hat{\Theta}_n - \tilde{\Theta}_n\|_2\|I - n\eta\hat{\Theta}_n\|_2^{t-1} \\ &\quad + \eta\|I - n\eta\tilde{\Theta}_n\|_2\|\hat{\Theta}_n - \tilde{\Theta}_n\|_2\|I - n\eta\hat{\Theta}_n\|_2^{t-2} \\ &\quad + \cdots \\ &\quad + \eta\|I - n\eta\tilde{\Theta}_n\|_2^{t-1}\|\hat{\Theta}_n - \tilde{\Theta}_n\|_2 \\ &\le \eta\|\hat{\Theta}_n - \tilde{\Theta}_n\|_2 \cdot t \cdot (\max\{\|I - n\eta\hat{\Theta}_n\|_2, \|I - n\eta\tilde{\Theta}_n\|_2\})^{t-1}.\end{aligned}$$

Since $\max\{\|I - n\eta\hat{\Theta}_n\|_2, \|I - n\eta\tilde{\Theta}_n\|_2\} < 1$, $\sup_{t>0} t \cdot (\max\{\|I - n\eta\hat{\Theta}_n\|_2, \|I - n\eta\tilde{\Theta}_n\|_2\})^{t-1}$ is a finite number. Then we have

$$\begin{aligned}\|[I - (I - n\eta\hat{\Theta}_n)^t] - [I - (I - n\eta\tilde{\Theta}_n)^t]\|_2 &\le O(\eta\|\hat{\Theta}_n - \tilde{\Theta}_n\|_2) \\ &= O_p(n^{-1}).\end{aligned} \tag{2.80}$$

Let $\Delta\Theta(\mathbf{x},\mathcal{X}) = n^{-1}(\nabla_\theta f(\mathbf{x},\theta_0)\nabla_\theta f(\mathcal{X},\theta_0)^T - \nabla_{\mathbf{W}^{(2)}} f(\mathbf{x},\theta_0)\nabla_{\mathbf{W}^{(2)}} f(\mathcal{X},\theta_0)^T)$, then the i-th entry of the vector $\Delta\Theta(\mathbf{x},\mathcal{X})$ is

$$\begin{aligned}(\Delta\Theta(\mathbf{x},\mathcal{X}))_i =& \frac{1}{n}\Bigg[\sum_{k=1}^{n}\Big(\nabla_{\mathbf{W}_k^{(1)}} f(\mathbf{x},\theta_0)\nabla_{\mathbf{W}_k^{(1)}} f(\mathbf{x}_i,\theta_0) + \nabla_{b_k^{(1)}} f(\mathbf{x},\theta_0)\nabla_{b_k^{(1)}} f(\mathbf{x}_i,\theta_0)\Big) \\ &+ \nabla_{b^{(2)}} f(\mathbf{x},\theta_0)\nabla_{b^{(2)}} f(\mathbf{x}_i,\theta_0)\Bigg].\end{aligned}$$

Similar to (2.75), we have

$$(\Delta\Theta(\mathbf{x},\mathcal{X}))_i| = O_p(n^{-1}). \tag{2.81}$$

Since the size of $\Delta\Theta(\mathbf{x},\mathcal{X})$ is M, which does not change as n goes up. So

$$\|\Delta\Theta(\mathbf{x},\mathcal{X})\|_2 = O_p(n^{-1}). \tag{2.82}$$

Let $\tilde{\Theta}_n(\mathbf{x},\mathcal{X}) = n^{-1}(\nabla_{\mathbf{W}^{(2)}} f(\mathbf{x},\theta_0)\nabla_{\mathbf{W}^{(2)}} f(\mathcal{X},\theta_0)^T))$, then the i-th entry of the vector $\tilde{\Theta}_n(\mathbf{x},\mathcal{X})$ is

$$\begin{aligned}|(\tilde{\Theta}_n(\mathbf{x},\mathcal{X}))_i| &\le \frac{1}{n}\sum_{k=1}^{n}|\nabla_{W_k^{(2)}} f(\mathbf{x},\theta_0)\nabla_{W_k^{(2)}} f(\mathbf{x}_i,\theta_0)| \\ &\le \frac{1}{n}\sum_{k=1}^{n}|(\|\mathbf{W}_k^{(1)}\|\|\mathbf{x}\|_2 + b_k^{(1)})(\|\mathbf{W}_k^{(1)}\|\|\mathbf{x}_i\|_2 + b_k^{(1)})|.\end{aligned} \tag{2.83}$$

According to initialization (2.3), $(\mathbf{W}_k^{(1)}, b_k^{(1)}) \overset{d}{=} (\boldsymbol{\mathcal{W}}, \mathcal{B})$. Then according to the law of large numbers,

$$|(\tilde{\Theta}_n(\mathbf{x},\mathcal{X}))_i| = O_p(1). \tag{2.84}$$

Since the size of $\tilde{\Theta}_n(\mathbf{x},\mathcal{X})$ is M, which does not change as n goes up. So

$$\|\tilde{\Theta}_n(\mathbf{x},\mathcal{X})\|_2 = O_p(1).$$

[Nea96c], [LSP18] show that as n goes to infinity, the output function at initialization $f(\cdot,\theta_0)$ converges to a Gaussian process in distribution, which means that $f(\mathcal{X},\theta_0) \sim \mathcal{N}(0,\mathcal{K}(\mathcal{X},\mathcal{X}))$. Here $\mathcal{K}(\mathcal{X},\mathcal{X})$ can be computed recursively. Then $f(\mathcal{X},\theta_0)$ is bounded in probability and we

get

$$\|f(\mathcal{X},\theta_0)-\mathcal{Y}\|_2=O_p(1). \tag{2.85}$$

Then following (2.69) and (2.85), we get

$$\begin{aligned}
&|f^{\text{lin}}(\mathbf{x},\widetilde{\omega}_t)-f(\mathbf{x},\theta_t)|\\
=&n^{-1}|\nabla_\theta f(\mathbf{x},\theta_0)\nabla_\theta f(\mathcal{X},\theta_0)^T\hat{\Theta}_n^{-1}[I-(I-n\eta\hat{\Theta}_n)^t](f(\mathcal{X},\theta_0)-\mathcal{Y})\\
&-\nabla_{\mathbf{W}^{(2)}}f(\mathbf{x},\theta_0)\nabla_{\mathbf{W}^{(2)}}f(\mathcal{X},\theta_0)^T\tilde{\Theta}_n^{-1}[I-(I-n\eta\tilde{\Theta}_n)^t](f(\mathcal{X},\theta_0)-\mathcal{Y})|\\
=&n^{-1}\|\nabla_\theta f(\mathbf{x},\theta_0)\nabla_\theta f(\mathcal{X},\theta_0)^T\hat{\Theta}_n^{-1}[I-(I-n\eta\hat{\Theta}_n)^t]\\
&-\nabla_{\mathbf{W}^{(2)}}f(\mathbf{x},\theta_0)\nabla_{\mathbf{W}^{(2)}}f(\mathcal{X},\theta_0)^T\tilde{\Theta}_n^{-1}[I-(I-n\eta\tilde{\Theta}_n)^t]\|_2\|f(\mathcal{X},\theta_0)-\mathcal{Y}\|_2\\
=&n^{-1}\|\nabla_\theta f(\mathbf{x},\theta_0)\nabla_\theta f(\mathcal{X},\theta_0)^T\hat{\Theta}_n^{-1}[I-(I-n\eta\hat{\Theta}_n)^t]\\
&-\nabla_{\mathbf{W}^{(2)}}f(\mathbf{x},\theta_0)\nabla_{\mathbf{W}^{(2)}}f(\mathcal{X},\theta_0)^T\tilde{\Theta}_n^{-1}[I-(I-n\eta\tilde{\Theta}_n)^t]\|_2\cdot O_p(1).
\end{aligned}$$

According to (2.78), (2.79), (2.80), (2.82) and (2.84), we have that

$$\begin{aligned}
&n^{-1}\|\nabla_\theta f(\mathbf{x},\theta_0)\nabla_\theta f(\mathcal{X},\theta_0)^T\hat{\Theta}_n^{-1}[I-(I-n\eta\hat{\Theta}_n)^t]\\
&-\nabla_{\mathbf{W}^{(2)}}f(\mathbf{x},\theta_0)\nabla_{\mathbf{W}^{(2)}}f(\mathcal{X},\theta_0)^T\tilde{\Theta}_n^{-1}[I-(I-n\eta\tilde{\Theta}_n)^t]\|_2\\
\le&n^{-1}\|\nabla_\theta f(\mathbf{x},\theta_0)\nabla_\theta f(\mathcal{X},\theta_0)^T-\nabla_{\mathbf{W}^{(2)}}f(\mathbf{x},\theta_0)\nabla_{\mathbf{W}^{(2)}}f(\mathcal{X},\theta_0)^T\|\|\hat{\Theta}_n^{-1}\|_2\|I-(I-n\eta\hat{\Theta}_n)^t\|_2\\
&+n^{-1}\|\nabla_{\mathbf{W}^{(2)}}f(\mathbf{x},\theta_0)\nabla_{\mathbf{W}^{(2)}}f(\mathcal{X},\theta_0)^T\|_2\|\hat{\Theta}_n^{-1}-\tilde{\Theta}_n^{-1}\|_2\|I-(I-n\eta\hat{\Theta}_n)^t\|_2\\
&+n^{-1}\|\nabla_{\mathbf{W}^{(2)}}f(\mathbf{x},\theta_0)\nabla_{\mathbf{W}^{(2)}}f(\mathcal{X},\theta_0)^T\|_2\|\tilde{\Theta}_n^{-1}\|_2\|[I-(I-n\eta\hat{\Theta}_n)^t]-[I-(I-n\eta\tilde{\Theta}_n)^t]\|_2\\
\le&O_p(n^{-1})O_p(1)O_p(1)+O_p(1)O_p(n^{-1})O_p(1)+O_p(1)O_p(1)O_p(n^{-1})\\
=&O_p(n^{-1}).
\end{aligned}$$

So we have $|f^{\text{lin}}(\mathbf{x},\widetilde{\omega}_t)-f(\mathbf{x},\theta_t)|=O_p(n^{-1})$, and the constants in $O_p(n^{-1})$ do not depend on t. Then we get

$$\sup_t|f^{\text{lin}}(\mathbf{x},\widetilde{\omega}_t)-f^{\text{lin}}(\mathbf{x},\omega_t)|=O_p(n^{-1}),\ \text{as } n\to\infty.$$

For the difference of parameters, we have

$$\widetilde{\omega}_t - \omega_t = \text{vec}(\overline{\mathbf{W}}^{(1)} - \widehat{\mathbf{W}}_t^{(1)}, \overline{\mathbf{b}}^{(1)} - \widehat{\mathbf{b}}_t^{(1)}, \widetilde{\mathbf{W}}_t^{(2)} - \widehat{\mathbf{W}}_t^{(2)}, \overline{b}^{(2)} - \widehat{b}_t^{(2)}).$$

According to (2.65) and (2.67),

$$\begin{aligned}
\|\overline{\mathbf{W}}^{(1)} - \widehat{\mathbf{W}}_t^{(1)}\|_2 &= \|n^{-1}\nabla_{\mathbf{W}^{(1)}} f(\mathcal{X}, \theta_0)^T \hat{\Theta}_n^{-1}[I - (I - n\eta\hat{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y})\|_2 \\
&\leq \|n^{-1}\nabla_{\mathbf{W}^{(1)}} f(\mathcal{X}, \theta_0)^T\|_2 \|\hat{\Theta}_n^{-1}\|_2 \|I - (I - n\eta\hat{\Theta}_n)^t\|_2 \|f(\mathcal{X}, \theta_0) - \mathcal{Y}\|_2 \\
&\leq n^{-1}\|\nabla_{\mathbf{W}^{(1)}} f(\mathcal{X}, \theta_0)^T\|_2 \cdot O_p(1).
\end{aligned}$$

Here $\nabla_{\mathbf{W}^{(1)}} f(\mathcal{X}, \theta_0)^T$ is a $n \times M$ matrix, the ij-th entry of the matrix is $\nabla_{\mathbf{W}_i^{(1)}} f(\mathbf{x}_j, \theta_0)$. According to (2.71), we have $\nabla_{\mathbf{W}_i^{(1)}} f(\mathbf{x}_j, \theta_0) = O_p(n^{-1/2})$. Then we get $\|\nabla_{\mathbf{W}^{(1)}} f(\mathcal{X}, \theta_0)^T\|_2 = O_p(1)$ by the law of large numbers. So we have $\|\overline{\mathbf{W}}^{(1)} - \widehat{\mathbf{W}}_t^{(1)}\|_2 = O_p(n^{-1})$, and $O_p(n^{-1})$ does not contain any constant factor which is related to t. Then we get

$$\sup_t \|\overline{\mathbf{W}}^1 - \widehat{\mathbf{W}}_t^1\|_2 = O_p(n^{-1}), \text{ as } n \to \infty.$$

Similarly we can prove

$$\sup_t \|\overline{\mathbf{b}}^1 - \widehat{\mathbf{b}}_t^1\|_2 = O_p(n^{-1}), \text{ as } n \to \infty, \tag{2.86}$$

$$\sup_t \|\overline{b}^2 - \widehat{b}_t^2\| = O_p(n^{-1}), \text{ as } n \to \infty. \tag{2.87}$$

For $\widetilde{\mathbf{W}}_t^{(2)} - \widehat{\mathbf{W}}_t^{(2)}$, we have

$$\begin{aligned}
\|\overline{\mathbf{W}}^{(2)} - \widehat{\mathbf{W}}_t^{(2)}\|_2 &= \|n^{-1}\nabla_{\mathbf{W}^{(2)}} f(\mathcal{X}, \theta_0)^T \Big(\hat{\Theta}_n^{-1}[I - (I - n\eta\hat{\Theta}_n)^t] - \\
&\qquad \tilde{\Theta}_n^{-1}[I - (I - n\eta\tilde{\Theta}_n)^t]\Big)(f(\mathcal{X}, \theta_0) - \mathcal{Y})\|_2 \\
&\leq \|n^{-1}\nabla_{\mathbf{W}^{(2)}} f(\mathcal{X}, \theta_0)^T\|_2 \Big(\|\hat{\Theta}_n^{-1} - \tilde{\Theta}_n^{-1}\|_2 \|I - (I - n\eta\hat{\Theta}_n)^t\|_2 + \\
&\qquad \|\tilde{\Theta}_n^{-1}\|_2 \|[I - (I - n\eta\hat{\Theta})^t] - [I - (I - n\eta\tilde{\Theta})^t]\|_2\Big)\|f(\mathcal{X}, \theta_0) - \mathcal{Y}\|_2 \\
&\leq n^{-1}\|\nabla_{\mathbf{W}^{(2)}} f(\mathcal{X}, \theta_0)^T\|_2 (O_p(n^{-1})O_p(1) + O_p(1)O_p(n^{-1})) \cdot O_p(1) \\
&= O_p(n^{-2})\|\nabla_{\mathbf{W}^{(2)}} f(\mathcal{X}, \theta_0)^T\|_2.
\end{aligned}$$

Here $\nabla_{\mathbf{W}^{(2)}} f(\mathcal{X}, \theta_0)^T$ is a $n \times M$ matrix, the ij-th entry of the matrix is $\nabla_{W_i^{(2)}} f(\mathbf{x}_j, \theta_0)$. According to (2.73), we have $\nabla_{W_i^{(2)}} f(\mathbf{x}_j, \theta_0) = O_p(1)$. Then $\|\nabla_{\mathbf{W}^{(2)}} f(\mathcal{X}, \theta_0)^T\|_2 = O_p(n^{1/2})$ by the law of large numbers. So we have $\|\widetilde{\mathbf{W}}_t^{(2)} - \widehat{\mathbf{W}}_t^{(2)}\|_2 = O_p(n^{-3/2})$, and $O_p(n^{-3/2})$ does not contain any constant factor which is related to t. Then we get

$$\sup_t \|\widetilde{\mathbf{W}}_t^2 - \widehat{\mathbf{W}}_t^2\|_2 = O_p(n^{-3/2}), \text{ as } n \to \infty.$$

□

2.F Training Only the Output Layer Approximates Training a Wide Shallow Network

Corollary 11 is obtained by combining Theorem 10 and the fact that training a linearized model approximates training a wide network [LXS19b, Theorem H.1]. Although [LXS19b, Theorem H.1] consider Gaussian initialization, the arguments extend to sub-Gaussian initialization.

Proof of Corollary 11. Using Theorem 10, we have that

$$\sup_t |f^{\text{lin}}(\mathbf{x}, \widetilde{\omega}_t) - f^{\text{lin}}(\mathbf{x}, \omega_t)| = O_p(n^{-1}), \text{ as } n \to \infty. \tag{2.88}$$

According to [LXS19b, Theorem H.1], in the case of Gaussian initialization, we have

$$\sup_t |f^{\text{lin}}(\mathbf{x}, \omega_t) - f(\mathbf{x}, \theta)| = O_p(n^{-\frac{1}{2}}), \text{ as } n \to \infty.$$

Under our neural network setting, which is a one-input network with a single hidden layer of n ReLUs and a linear output, we can generalize the above result to sub-Gaussian initialization. In the remark of Theorem 10, we illustrate that the empirical NTK converges to analytic NTK for initialization with finite variance distribution. Then for sub-Gaussian initialization the empirical NTK still converges to analytic NTK. Then the only part we need to adapt in the proof of [LXS19b, Theorem H.1] is the following theorem [LXS19b, Theorem G.3]:

Theorem 26. *Let A be an $N \times n$ random matrix whose entries are independent standard normal random variables. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-t^2/2)$ one has*

$$\|A\|_{\text{op}} \leq \sqrt{N} + \sqrt{n} + t.$$

Then [LXS19b] applies the above theorem to weight matrices in the neural network. In our case, we use sub-Gaussian initialization and next we derive the similar bound for $\|\mathbf{W}^{(1)}\|_{\text{op}}$ and $\|\mathbf{W}^{(2)}\|_{\text{op}}$. Since $W_{ij}^{(1)}$ is sub-Gaussian, $\mathbb{P}(|W_{ij}^{(1)}| \geq t) \leq 2\exp(-t^2/2\sigma^2)$ for some positive σ. Then $(W_{ij}^{(1)})^2$ is sub-exponential. Using the property of sub-Gaussian exponential, we have $\mathbb{E}\exp(|W_{ij}^{(1)}|^2/\lambda) \leq 2$ for some positive λ. Using [Ver18, Theorem 1.4.1], we have

$$\mathbb{P}\left(|\sum_{i=1}^{n}\sum_{j=1}^{d}(W_{ij}^{(1)})^2 - \mathbb{E}\sum_{i=1}^{n}\sum_{j=1}^{d}(W_{ij}^{(1)})^2| \geq t\right) \leq 2\exp\left[-c\min\left(\frac{t^2}{nd\lambda^2}, \frac{t}{\lambda}\right)\right],$$

where c is a constant. Let $t = nd\lambda$, then we have

$$\mathbb{P}\left(\sum_{i=1}^{n}\sum_{j=1}^{d}(W_{ij}^{(1)})^2 \geq \mathbb{E}\sum_{i=1}^{n}\sum_{j=1}^{d}(W_{ij}^{(1)})^2 + nd\lambda\right) \leq 2\exp(-cnd).$$

The above equation means that with probability at least $1 - 2\exp(-cnd)$,

$$\begin{aligned}\|\mathbf{W}^{(1)}\|_F^2 &= \sum_{i=1}^{n}\sum_{j=1}^{d}(W_{ij}^{(1)})^2 \\ &\leq \mathbb{E}\sum_{i=1}^{n}\sum_{j=1}^{d}(W_{ij}^{(1)})^2 + nd\lambda \\ &= nd\mathbb{E}(W_{ij}^{(1)})^2 + nd\lambda \\ &= O(n).\end{aligned}$$

So $\|\mathbf{W}^{(1)}\|_{\text{op}} \leq \|\mathbf{W}^{(1)}\|_F = O_p(\sqrt{n})$. For a similar reason, $\|\mathbf{W}^{(2)}\|_{\text{op}} = O_p(1)$. Then following similar arguments as [LXS19b] we can show that

$$\sup_t |f^{\text{lin}}(\mathbf{x}, \omega_t) - f(\mathbf{x}, \theta)| = O_p(n^{-\frac{1}{2}}), \text{ as } n \to \infty.$$

Combining the above equation with (2.88) concludes the proof. □

2.G Proof of Theorem 12

Proof of Theorem 12. The Lagrangian of problem (2.19) is

$$L(\alpha_n, \lambda^{(n)}) = \int_{\mathbb{R}^2} \alpha_n^2(\mathbf{W}^{(1)}, b)\ \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) + \sum_{j=1}^{M} \lambda_j^{(n)}(g_n(\mathbf{x}_j, \alpha_n) - y_j).$$

The optimal condition is $\nabla_{\alpha_n} L = 0$, which means

$$\nabla_{\alpha_n} L = 2\alpha_n(\mathbf{W}^{(1)}, b) + \sum_{j=1}^{M} \lambda_j^{(n)}[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+ = 0 \text{ when } (\mathbf{W}^{(1)}, b) = (\mathbf{W}_i^{(1)}, b_i),\ i = 1, \ldots, k.$$

Then we get

$$\overline{\alpha}_n(\mathbf{W}^{(1)}, b) = -\frac{1}{2}\sum_{j=1}^{M} \lambda_j^{(n)}[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+ \text{ when } (\mathbf{W}^{(1)}, b) = (\mathbf{W}_i^{(1)}, b_i),\ i = 1, \ldots, k.$$

Since only function values on $(\mathbf{W}_i^{(1)}, b_i)_{i=1}^M$ are taken into account in problem (2.19), we can let

$$\overline{\alpha}_n(\mathbf{W}^{(1)}, b) = -\frac{1}{2}\sum_{j=1}^{M} \lambda_j^{(n)}[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+ \quad \forall (\mathbf{W}^{(1)}, b) \in \mathbb{R}^{d+1} \tag{2.89}$$

without changing $\int_{\mathbb{R}^2} \overline{\alpha}_n^2(\mathbf{W}^{(1)}, b)\ \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b)$ and $g_n(\mathbf{x}, \overline{\alpha}_n)$.

Here $\lambda_j^{(n)}$, $j = 1, \ldots, M$ are chosen to make $g_n(\mathbf{x}_i, \overline{\alpha}_n) = y_i$, $i = 1, \ldots, M$. This means that

$$-\frac{1}{2}\sum_{j=1}^{M} \lambda_j^{(n)} \int_{\mathbb{R}^2} [\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+ [\langle \mathbf{W}^{(1)}, \mathbf{x}_i\rangle + b]_+\ \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) = y_i,\ i = 1, \ldots, M. \tag{2.90}$$

Similarly, the Lagrangian of problem (2.20) is

$$\widetilde{L}(\alpha, \lambda) = \int_{\mathbb{R}^2} \alpha^2(\mathbf{W}^{(1)}, b)\ \mathrm{d}\mu(\mathbf{W}^{(1)}, b) + \sum_{j=1}^{M} \lambda_j (g(\mathbf{x}_j, \alpha) - y_j).$$

The optimality condition is $\nabla_\alpha \widetilde{L} = 0$, which means

$$\nabla_\alpha \widetilde{L} = 2\alpha(\mathbf{W}^{(1)}, b) + \sum_{j=1}^{M} \lambda_j [\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+ = 0 \quad \forall (\mathbf{W}^{(1)}, b) \in \mathbb{R}^{d+1}.$$

Then we get

$$\overline{\alpha}(\mathbf{W}^{(1)}, b) = -\frac{1}{2}\sum_{j=1}^{M} \lambda_j [\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+ \quad \forall (\mathbf{W}^{(1)}, b) \in \mathbb{R}^{d+1}. \tag{2.91}$$

Here λ_j, $j = 1, \ldots, M$ are chosen to make $g(\mathbf{x}, \alpha) = y_i$, $i = 1, \ldots, M$. This means that

$$-\frac{1}{2}\sum_{j=1}^{M} \lambda_j \int_{\mathbb{R}^2} [\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+ [\langle \mathbf{W}^{(1)}, \mathbf{x}_i\rangle + b]_+\ \mathrm{d}\mu(\mathbf{W}^{(1)}, b) = y_i, \quad i = 1, \ldots, M. \tag{2.92}$$

Compare (2.90) and (2.92). Since the number of samples is finite, $\mathbf{x}_i$ is also bounded. Then by the assumption that $\mathcal{W}$ and $\mathcal{B}$ have finite fourth moments, we have that $[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle +$

$b]_+[\langle \mathbf{W}^{(1)}, \mathbf{x}_i\rangle + b]_+$ has finite variance. According to central limit theorem, as $n \to \infty$, $\int_{\mathbb{R}^2}[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+[\langle \mathbf{W}^{(1)}, \mathbf{x}_i\rangle + b]_+ \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b)$ tends to a Gaussian distribution with variance $O(n^{-1})$. This implies that $\forall i = 1, \ldots, M, \ \forall j = 1, \ldots, M$,

$$\begin{aligned}
&\Big| \int_{\mathbb{R}^2}[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+[\langle \mathbf{W}^{(1)}, \mathbf{x}_i\rangle + b]_+ \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) \\
&- \int_{\mathbb{R}^2}[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+[\langle \mathbf{W}^{(1)}, \mathbf{x}_i\rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)\Big| \\
&= O_p(n^{-1/2})
\end{aligned}$$

Since (2.90) and (2.92) are systems of linear equations and coefficients of (2.90) converge to coefficients of (2.92) at the rate of $O_p(n^{-1/2})$, then we get

$$|\lambda_j^n - \lambda_j| = O_p(n^{-1/2}), \quad j = 1, \ldots, M. \tag{2.93}$$

Compare (2.89) and (2.91). Given $(\mathbf{W}^{(1)}, b)$, we have

$$|\overline{\alpha}_n(\mathbf{W}^{(1)}, b) - \overline{\alpha}(\mathbf{W}^{(1)}, b)| = O_p(n^{-1/2}). \tag{2.94}$$

Next we want to prove that $\sup_{\mathbf{x}\in D} |g_n(\mathbf{x}, \overline{\alpha}_n) - g(\mathbf{x}, \overline{\alpha})| = O_p(n^{-1/2})$. Firstly, we prove that $\sup_{\mathbf{x}\in D} |g_n(\mathbf{x}, \overline{\alpha}) - g(\mathbf{x}, \overline{\alpha})| = O_p(n^{-1/2})$. Note that

$$\begin{aligned}
g_n(\mathbf{x}, \overline{\alpha}) &= \int_{\mathbb{R}^2} \overline{\alpha}(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}\rangle + b]_+ \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) \\
g(\mathbf{x}, \overline{\alpha}) &= \int_{\mathbb{R}^2} \overline{\alpha}(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}\rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}(g_n(\mathbf{x}, \overline{\alpha})) &= g(\mathbf{x}, \overline{\alpha}) \\
\mathrm{Var}(g_n(\mathbf{x}, \overline{\alpha})) &= \frac{1}{n} \int_{\mathbb{R}^2} [\overline{\alpha}(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}\rangle + b]_+ - g(\mathbf{x}, \overline{\alpha})]^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b).
\end{aligned} \tag{2.95}$$

Here the expectation and the variance are with respect to $(\mathbf{W}_i^{(1)}, b_i)_{i=1}^n$. According to (2.91)

and the assumption that $\mathcal{W}$ and $\mathcal{B}$ have finite fourth moments, the integral in (2.95) is bounded on D. So $\sup_{\mathbf{x}\in D} \operatorname{Var} g_n(\mathbf{x}, \overline{\alpha}) = O(n^{-1})$. According to central limit theorem, as $n \to \infty$, $g_n(\mathbf{x}, \overline{\alpha})$ tends to Gaussian distribution of variance $O(n^{-1})$ for any $\mathbf{x} \in D$. Then $|g_n(\mathbf{x}, \overline{\alpha}) - g(\mathbf{x}, \overline{\alpha})| = O_p(n^{-1/2})$ pointwise on D.Then we only need to prove that the sequence of functions $\{g_n(\mathbf{x}, \overline{\alpha})\}_{n=1}^{\infty}$ is uniformly equicontinuous. Actually, $\forall \mathbf{x}_1, \mathbf{x}_2 \in D$

$$
\begin{aligned}
&|g_n(\mathbf{x}_1, \overline{\alpha}) - g_n(\mathbf{x}_2, \overline{\alpha})| \\
\leq &\int_{\mathbb{R}^2} \left|\overline{\alpha}(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}_1\rangle + b]_+ - \overline{\alpha}(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}_2\rangle + b]_+\right| \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) \\
\leq &\int_{\mathbb{R}^2} \left|\overline{\alpha}(\mathbf{W}^{(1)}, b)\right| \left|\mathbf{W}_i^{(1)}\right| |\mathbf{x}_1 - \mathbf{x}_2| \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) \\
\leq &\int_{\mathbb{R}^2} \left|\overline{\alpha}(\mathbf{W}^{(1)}, b)\right| \left|\mathbf{W}_i^{(1)}\right| \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) |\mathbf{x}_1 - \mathbf{x}_2|.
\end{aligned}
$$

Notice that $\int_{\mathbb{R}^2} \left|\overline{\alpha}(\mathbf{W}^{(1)}, b)\right| \left|\mathbf{W}_i^{(1)}\right| \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) \to \int_{\mathbb{R}^2} \left|\overline{\alpha}(\mathbf{W}^{(1)}, b)\right| \left|\mathbf{W}_i^{(1)}\right| \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)$ with probability 1 according to the law of large numbers. Hence $\int_{\mathbb{R}^2} \left|\overline{\alpha}(\mathbf{W}^{(1)}, b)\right| \left|\mathbf{W}_i^{(1)}\right| \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b)$ is bounded and the bound is independent of n. So $\{g_n(\mathbf{x}, \overline{\alpha})\}_{n=1}^{\infty}$ is uniformly equicontinuous. Then by similar arguments to the Arzela-Ascoli theorem,

$$
\sup_{\mathbf{x}\in D} |g_n(\mathbf{x}, \overline{\alpha}) - g(\mathbf{x}, \overline{\alpha})| = O_p(n^{-1/2}). \tag{2.96}
$$

Finally, we prove that $\sup_{\mathbf{x}\in D}|g_n(\mathbf{x},\overline{\alpha}_n)-g_n(\mathbf{x},\overline{\alpha})|=O_p(n^{-1/2})$. Since $\forall\mathbf{x}\in D$

$$
\begin{aligned}
&|g_n(\mathbf{x},\overline{\alpha}_n)-g_n(\mathbf{x},\overline{\alpha})|\\
\leq&\int_{\mathbb{R}^2}\left|\overline{\alpha}_n(\mathbf{W}^{(1)},b)[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+-\overline{\alpha}(\mathbf{W}^{(1)},b)[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\right|\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)\\
\leq&\int_{\mathbb{R}^2}\left|\overline{\alpha}_n(\mathbf{W}^{(1)},b)-\overline{\alpha}(\mathbf{W}^{(1)},b)\right|[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)\\
\leq&\int_{\mathbb{R}^2}\left|-\frac{1}{2}\sum_{j=1}^{M}(\lambda_j^n-\lambda_j)[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+\right|[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)\\
\leq&\frac{1}{2}\sum_{j=1}^{M}|\lambda_j^n-\lambda_j|\int_{\mathbb{R}^2}[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)\\
\leq&\frac{1}{2}\left(\max_{\mathbf{x}\in D}\int_{\mathbb{R}^2}[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)\right)\sum_{j=1}^{M}|\lambda_j^n-\lambda_j|.
\end{aligned}
$$

Because D is compact and $\int_{\mathbb{R}^2}[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)$ converges according to the law of large numbers, we have that $\max_{\mathbf{x}\in D}\int_{\mathbb{R}^2}[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)$ is bounded by a finite number independent of n. Then according to (2.93),

$$\sup_{\mathbf{x}\in D}|g_n(\mathbf{x},\overline{\alpha}_n)-g_n(\mathbf{x},\overline{\alpha})|=O_p(n^{-1/2}).$$

Combined with (2.96), we have

$$\sup_{\mathbf{x}\in D}|g_n(\mathbf{x},\overline{\alpha}_n)-g(\mathbf{x},\overline{\alpha})|=O_p(n^{-1/2}).$$

This concludes the proof. □

2.H Proofs of Results for Univariate Regression

2.H.1 Proof of Theorem 13

The second derivative g'' is given by

$$g''(x,\gamma) = p_{\mathcal{C}}(x)\int_{\mathbb{R}} \gamma(W^{(1)},x)\big|W^{(1)}\big|\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}). \tag{2.97}$$

The detailed calculation of (2.97) is as follows:

$$\begin{aligned} g''(x,\gamma) &= \int_{\mathbb{R}^2} \gamma(W^{(1)},c)\big|W^{(1)}\big|\delta(x-c)\ \mathrm{d}\nu(W^{(1)},c) \\ &= \int_{\mathrm{supp}(\nu_{\mathcal{C}})}\left(\int_{\mathbb{R}} \gamma(W^{(1)},c)\big|W^{(1)}\big|\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right)\delta(x-c)\ \mathrm{d}\nu_{\mathcal{C}}(c) \\ &= \int_{\mathrm{supp}(\nu_{\mathcal{C}})}\left(\int_{\mathbb{R}} \gamma(W^{(1)},c)\big|W^{(1)}\big|\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right)\delta(x-c)p_{\mathcal{C}}(c)\mathrm{d}c \\ &= p_{\mathcal{C}}(x)\int_{\mathbb{R}} \gamma(W^{(1)},x)\big|W^{(1)}\big|\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}). \end{aligned} \tag{2.98}$$

Proof of Theorem 13. First, if $x \notin \mathrm{supp}(\zeta)$, similar to (2.97), we have

$$\begin{aligned} g(x,(\overline{\gamma},\overline{u},\overline{v})) &= p_{\mathcal{C}}(x)\int_{\mathbb{R}} \gamma(W^{(1)},x)\big|W^{(1)}\big|\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}) \\ &= 0. \end{aligned}$$

Next, we prove that $g(x,(\overline{\gamma},\overline{u},\overline{v}))$ restricted on $\mathrm{supp}(\zeta)$ is the solution of the following problem:

$$\begin{aligned} \min_{h\in C^2(\mathrm{supp}(\zeta))} \quad & \int_{\mathrm{supp}(\zeta)} \frac{(h''(x))^2}{\zeta(x)}\ \mathrm{d}x \\ \text{subject to} \quad & h(x_j) = y_j, \quad j = 1,\ldots,m. \end{aligned} \tag{2.99}$$

Let $L(f) = \int_{\mathrm{supp}(\zeta)} \frac{(f''(x))^2}{p(x)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=x)}\mathrm{d}x$. Then the functional $L(f)$ is strictly convex on space $\{f \in C^2(\mathbb{R}^2) | f(x_i) = y_i,\ i = 1,\ldots,m\}$ when $m \geq 2$. This means that the minimizer of problem (2.99) is unique.

Suppose $h(x)$ is the minimizer of problem (2.99) and $h(x)$ is different from $g(x,(\overline{\gamma},\overline{u},\overline{v}))$

restricted on supp(ζ). Then by uniqueness of the solution,

$$L(h) < L(g(\cdot, (\overline{\gamma}, \overline{u}, \overline{v}))). \tag{2.100}$$

Now our goal is to find a different (γ, u, v) with smaller cost in problem (2.22). Then $(\overline{\gamma}, \overline{u}, \overline{v})$ is not the solution of (2.22), which is a contradiction. We set

$$\gamma(W^{(1)}, c) = \frac{h''(c)|W^{(1)}|}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}, \quad c \in \text{supp}(\zeta).$$

Then according to (2.97),

$$\begin{aligned}
g''(x,\gamma) &= p(x)\int_{\mathbb{R}} \gamma(W^{(1)}, x)\big|W^{(1)}\big| \, \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}) \\
&= p(x)\int_{\mathbb{R}} \frac{h''(x)|W^{(1)}|}{p(x)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=x)}\big|W^{(1)}\big| \, \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}) \\
&= \frac{h''(x)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=x)}\int_{\mathbb{R}} \big|W^{(1)}\big|^2 \, \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}) \\
&= \frac{h''(x)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=x)}\mathbb{E}(\mathcal{W}^2|\mathcal{C}=x) \\
&= h''(x), \quad x \in \text{supp}(\zeta).
\end{aligned}$$

This means that we can find $u, v \in \mathbb{R}$ such that $ux + v + g(x, \gamma) \equiv h(x)$. Then we find (γ, u, v) such that $g(x, (\gamma, u, v)) = ux + v + g(x, \gamma) = h(x)$ on supp(ζ). So $g(x_j, (\gamma, u, v)) = h(x_j) = y_j$. It means that (γ, u, v) satisfies the condition in problem (2.22). Next we compute the cost of

(γ, u, v):

$$
\begin{aligned}
&\int_{\mathbb{R}^2} \gamma^2(W^{(1)}, c)\ \mathrm{d}\nu(W^{(1)}, c) \\
=&\int_{\mathbb{R}^2} \left(\frac{h''(c)|W^{(1)}|}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}\right)^2 \mathrm{d}\nu(W^{(1)}, c) \\
=&\int_{\mathrm{supp}(\zeta)} \left(\int_{\mathbb{R}} \left(\frac{h''(c)|W^{(1)}|}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}\right)^2 \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right) \mathrm{d}\nu_{\mathcal{C}}(c) \\
=&\int_{\mathrm{supp}(\zeta)} \left(\frac{h''(c)}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}\right)^2 \left(\int_{\mathbb{R}} |W^{(1)}|^2\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right) p_{\mathcal{C}}(c)\mathrm{d}c \\
=&\int_{\mathrm{supp}(\zeta)} \left(\frac{h''(c)}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}\right)^2 \left(\int_{\mathbb{R}} |W^{(1)}|^2\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right) p_{\mathcal{C}}(c)\mathrm{d}c \\
=&\int_{\mathrm{supp}(\zeta)} \frac{(h''(c))^2}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}\mathrm{d}x \\
=&L(h).
\end{aligned}
\tag{2.101}
$$

On the other hand, the cost of $(\overline{\gamma}, \overline{u}, \overline{v})$ is

$$
\begin{aligned}
&\int_{\mathbb{R}^2} \overline{\gamma}^2(W^{(1)}, c)\ \mathrm{d}\nu(W^{(1)}, c) \\
=&\int_{\mathrm{supp}(\zeta)} \left(\int_{\mathbb{R}} \overline{\gamma}^2(W^{(1)}, c)\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right) p_{\mathcal{C}}(c)\mathrm{d}c \\
\geq&\int_{\mathrm{supp}(\zeta)} \frac{\left(\int_{\mathbb{R}} \overline{\gamma}(W^{(1)}, c)|W^{(1)}|\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}\right)^2}{\int_{\mathbb{R}} |W^{(1)}|^2\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}} p_{\mathcal{C}}(c)\mathrm{d}c \quad \text{(Cauchy-Schwarz inequality)} \\
=&\int_{\mathrm{supp}(\zeta)} \frac{\left(g''(c, \overline{\gamma})/p_{\mathcal{C}}(c))\right)^2}{\int_{\mathbb{R}} |W^{(1)}|^2\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}} p_{\mathcal{C}}(c)\mathrm{d}c \quad \text{(according to (2.97))} \\
=&\int_{\mathrm{supp}(\zeta)} \frac{\left(g''(c, \overline{\gamma})\right)^2}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}\mathrm{d}c \\
=&L(g(\cdot, \overline{\gamma})) \\
=&L(g(\cdot, (\overline{\gamma}, \overline{u}, \overline{v}))) \quad (g(\cdot, (\overline{\gamma}, \overline{u}, \overline{v})) \text{ has the same second derivative as } g(\cdot, \overline{\gamma})).
\end{aligned}
\tag{2.102}
$$

From this we have

$$\begin{aligned}\int_{\mathbb{R}^2} \gamma^2(W^{(1)}, c)\, \mathrm{d}\nu(W^{(1)}, c) &= L(h) \quad \text{(according to (2.101))} \\ &< L(g(\cdot, (\overline{\gamma}, \overline{u}, \overline{v}))) \quad \text{(according to (2.100))} \\ &\leq \int_{\mathbb{R}^2} \overline{\gamma}^2(W^{(1)}, c)\, \mathrm{d}\nu(W^{(1)}, c) \quad \text{(according to (2.102))}.\end{aligned}$$

It means that the cost of (γ, u, v) is smaller than the cost of $(\overline{\gamma}, \overline{u}, \overline{v})$. So $(\overline{\gamma}, \overline{u}, \overline{v})$ is not the solution of (2.99), which is a contradiction. So our assumption is wrong. So $h(x) \equiv g(x, (\overline{\gamma}, \overline{u}, \overline{v}))$ on $\mathrm{supp}(\zeta)$, and $g(x, (\overline{\gamma}, \overline{u}, \overline{v}))$ is the solution of problem (2.99). In the last step we prove that $g''(x, (\overline{\gamma}, \overline{u}, \overline{v})) = 0$ when $x \notin [\min_i x_i, \max_i x_i]$ and $g(x, (\overline{\gamma}, \overline{u}, \overline{v}))$ restricted on $\mathrm{supp}(\zeta) \cap [\min_i x_i, \max_i x_i]$ is the solution of (2.99). We only need to prove these statements for $h(x)$, which is the solution of (2.99).

Since $|x_i| \in [\min_i x_i, \max_i x_i]$, the function values on $(-\infty, \min_i x_i)$ and $(\max_i x_i, \infty)$ are not related to constraints of problem (2.99), so $h(x)$ can be replaced by following $\tilde{h}(x)$ which also satisfies the constraints of problem (2.99):

$$\tilde{h}(x) = \begin{cases} h(x) & x \in [\min_i x_i, \max_i x_i] \\ h'(\min_i x_i)(x - \min_i x_i) + h(\min_i x_i) & x \in (-\infty, \min_i x_i) \\ h'(\max_i x_i)(x - \max_i x_i) + h(\max_i x_i) & x \in (\max_i x_i, \infty). \end{cases}$$

Then we get

$$\tilde{h}''(x) = \begin{cases} h''(x) & x \in [\min_i x_i, \max_i x_i] \\ 0 & x \in (-\infty, \min_i x_i) \\ 0 & x \in (\max_i x_i, \infty). \end{cases}$$

So the cost of $\tilde{h}(x)$ is less than that of $h(x)$. Then the fact $h(x)$ is the minimizer of (2.99) tell us that $h(x) \equiv \tilde{h}(x)$. So $h(x)$ should be linear on $(-\infty, \min_i x_i)$ and $(\max_i x_i, \infty)$. Then $h''(x) = 0$ when $x \notin [\min_i x_i, \max_i x_i]$. Let $h(x)|_S$ denote the function $h(x)$ restricted on $S = \mathrm{supp}(\zeta) \cap [\min_i x_i, \max_i x_i]$. Since $h(x)$ is the solution to problem (2.99), we get $h(x)|_S$

is the solution to problem (2.99). This concludes the proof. □

In the case of not using ASI, problem (2.22) becomes

$$\begin{aligned} \min_{\gamma\in C(\mathbb{R}^2),u\in\mathbb{R},v\in\mathbb{R}} \quad & \int_{\mathbb{R}^2} \gamma^2(W^{(1)},c)\ \mathrm{d}\nu(W^{(1)},c) \\ \text{subject to} \quad & ux_j + v + \int_{\mathbb{R}^2} \gamma(W^{(1)},c)[W^{(1)}(x_j-c)]_+\ \mathrm{d}\nu(W^{(1)},c) = y_j - f(x_j,\theta_0), \\ & j=1,\dots,M. \end{aligned} \tag{2.103}$$

Then Theorem 13 without ASI is stated as follows.

Theorem 27 (Theorem 13 without ASI)**.** *Suppose* $(\overline{\gamma},\overline{u},\overline{v})$ *is the solution of* (2.103)*, and consider the corresponding output function*

$$g(x,(\overline{\gamma},\overline{u},\overline{v})) = \overline{u}x + \overline{v} + \int_{\mathbb{R}^2} \overline{\gamma}(W^{(1)},c)[W^{(1)}(x-c)]_+\ \mathrm{d}\nu(W^{(1)},c) + f(x,\theta_0). \tag{2.104}$$

Then $g(x,(\overline{\gamma},\overline{u},\overline{v}))$ *satisfies* $g''(x,(\overline{\gamma},\overline{u},\overline{v})) = f''(x,\theta_0)$ *for* $x\notin S$ *and for* $x\in S$ *it is the solution of the following problem:*

$$\begin{aligned} \min_{h\in C^2(S)} \quad & \int_S \frac{(h''(x)-f''(x,\theta_0))^2}{\zeta(x)}\ \mathrm{d}x \\ \text{subject to} \quad & h(x_j)=y_j, \quad j=1,\dots,M. \end{aligned} \tag{2.105}$$

2.H.2 Proof of Proposition 14 and Remarks to Proposition 15

Proof of Proposition 14. Let $p_{\mathcal{W},\mathcal{C}}$ and $p_{\mathcal{W},\mathcal{B}}$ denote the joint density functions of $(\mathcal{W},\mathcal{C})$ and $(\mathcal{W},\mathcal{B})$, respectively. We have

$$p_{\mathcal{W},\mathcal{C}}(W,c) = \left|\frac{\partial(W,-Wc)}{\partial(W,c)}\right| p_{\mathcal{W},\mathcal{B}}(W,-Wc) = |W| p_{\mathcal{W},\mathcal{B}}(W,-Wc),$$

and

$$\begin{aligned}\mathbb{E}(W^2|C=x)p_{\mathcal{C}}(x) &= \int_{\mathbb{R}} W^2 p_{\mathcal{W}|\mathcal{C}=x}(W)\ \mathrm{d}W \cdot p_{\mathcal{C}}(x) \\ &= \int_{\mathbb{R}} W^2 p_{\mathcal{W},\mathcal{C}}(W,x)\ \mathrm{d}W \\ &= \int_{\mathbb{R}} |W|^3 p_{\mathcal{W},\mathcal{B}}(W,-Wx)\ \mathrm{d}W.\end{aligned} \tag{2.106}$$

□

Proof of Proposition 15. The construction is given in the statement of the proposition. □

Remark 28 (Remark to Proposition 15, sampling the initial parameters)**.** *The variables* $(\mathcal{W},\mathcal{B})$ *can be sampled by first sampling* C *from* $p_{\mathcal{C}}(x)=\frac{1}{Z}\frac{1}{\varrho(x)}$*, then independently sampling* W *from a standard Gaussian distribution and setting* $B=-WC$*. In this construction, in general* $\mathcal{W}$ *and* $\mathcal{B}$ *are not independent.*

Intuitively, if we want the output function to be smooth at a certain point x_0, we can let the conditional distribution of $\mathcal{W}$ given $\mathcal{C}$ be concentrated around zero for $\mathcal{C}=x_0$, or we can let the probability density function of $\mathcal{C}$ to be small at $\mathcal{C}=x_0$. Note that $p_{\mathcal{C}}$ is the breakpoint density at initialization. The form of this has been studied for uniform initialization by [SDP20]. We provide the explicit form of the smoothness penalty function for several types of initialization in Appendix 2.H.3.

Remark 29 (Remark to Proposition 15, independent initialization)**.** *Note that constructing an arbitrary curvature penalty function will necessitate in general a non-independent joint distribution of* $\mathcal{W}$ *and* $\mathcal{B}$*. If* $\mathcal{W}$ *and* $\mathcal{B}$ *are required to be independent random variables,* (2.106) *gives*

$$\zeta(x)=\mathbb{E}(W^2|C=x)p_{\mathcal{C}}(x)=\int_{\mathbb{R}}|W|^3 p_{\mathcal{W}}(W)p_{\mathcal{B}}(-Wx)\ \mathrm{d}W.$$

Given a desired function for the left hand side, we can still try to solve for the parameter densities. This type of integral equation problem has been studied [Nas73] and one can write a formal solution, although it is not always clear whether it will be a density.

2.H.3 Proof of Theorem 2

We prove the statement for the three considered types of initialization distributions in turn.

Proof of Theorem 2 for Gaussian initialization. Using (2.106), we have

$$\begin{aligned}\mathbb{E}(W^2|C=x)p_{\mathcal{C}}(x) &= \int_{\mathbb{R}} |W|^3 p_{\mathcal{W}}(W) p_{\mathcal{B}}(-Wx) \mathrm{d}W \\ &= \int_{\mathbb{R}} |W|^3 \frac{1}{\sqrt{2\pi}\sigma_w} e^{-\frac{W^2}{2\sigma_w^2}} \frac{1}{\sqrt{2\pi}\sigma_b} e^{-\frac{W^2x^2}{2\sigma_b^2}} \mathrm{d}W \\ &= \frac{1}{2\pi\sigma_w\sigma_b} \int_{\mathbb{R}} |W|^3 e^{-(\frac{1}{2\sigma_w^2}+\frac{x^2}{2\sigma_b^2})W^2} \mathrm{d}W.\end{aligned}$$

Let $\sigma^2 = 1/\left(\frac{1}{\sigma_w^2} + \frac{x^2}{\sigma_b^2}\right)$, then we get

$$\begin{aligned}\mathbb{E}(W^2|C=x)p_{\mathcal{C}}(x) &= \frac{\sigma}{\sqrt{2\pi}\sigma_w\sigma_b} \int_{\mathbb{R}} |W|^3 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{W^2}{2\sigma^2}} \mathrm{d}W \\ &= \frac{\sigma}{\sqrt{2\pi}\sigma_w\sigma_b} \sigma^3 \cdot 2 \cdot \sqrt{\frac{2}{\pi}} \\ &= \frac{2\sigma^4}{\pi\sigma_w\sigma_b} \\ &= \frac{2\sigma_w^3\sigma_b^3}{\pi(\sigma_b^2 + x^2\sigma_w^2)^2}.\end{aligned}$$

Then we have

$$\begin{aligned}\zeta(x) &= \mathbb{E}(W^2|C=x)p_{\mathcal{C}}(x) \\ &= \frac{2\sigma_w^3\sigma_b^3}{\pi(\sigma_b^2 + x^2\sigma_w^2)^2}.\end{aligned}$$

□

Proof of Theorem 2 for binary-uniform initialization. Since $\mathcal{W}$ is either -1 or 1, $\mathbb{E}(\mathcal{W}^2|\mathcal{C} = x) = 1$ for any $x \in \mathrm{supp}(\nu_{\mathcal{C}})$. Since $\mathcal{B} \sim \mathrm{Unif}(-a_b, a_b)$, it is easy to check $-\mathcal{B}/\mathcal{W} \sim \mathrm{Unif}(-a_b, a_b)$. So $\zeta(x) = 1/2a_b,\ x \in [-a_b, a_b]$. □

Proof of Theorem 2 for uniform initialization. According to Theorem 1 in [SDP20], the den-

sity function $p_{\mathcal{C}}(c)$ of $\nu_{\mathcal{C}}$ is

$$p_{\mathcal{C}}(c) = \frac{1}{4a_w a_b}\left(\min\left\{\frac{a_b}{|c|}, a_w\right\}\right)^2, \quad c \in \text{supp}(\nu_{\mathcal{C}}).$$

When $|c| \leq \frac{a_b}{a_w}$, then $p_{\mathcal{C}}(c) = \frac{1}{4a_w a_b}(a_w)^2$. It means that $p_{\mathcal{C}}(c)$ is constant when $|c| \leq \frac{a_b}{a_w}$.

Let $p_{\mathcal{W},\mathcal{B}}(W^{(1)}, b)$ denote the density function of μ, $p_{\mathcal{W},\mathcal{C}}(W^{(1)}, c)$ denote the density function of ν, so

$$\begin{aligned} p_{\mathcal{W},\mathcal{C}}(W^{(1)}, c) &= p_{\mathcal{W},\mathcal{B}}(W^{(1)}, -cW^{(1)})\frac{\partial b}{\partial c} \\ &= \frac{1}{4a_w a_b}\mathbb{1}_{W^{(1)}\in[-a_w,a_w]} \cdot \mathbb{1}_{-cW^{(1)}\in[-a_b,a_b]} \cdot (-W^{(1)}). \end{aligned}$$

Here $\mathbb{1}_a$ is the indicator function which equals to 1 when condition a is true, and 0 otherwise. Then density function $p_{\mathcal{W}|\mathcal{C}}(W^{(1)}|c)$ of the conditional distribution $\nu_{\mathcal{W}|\mathcal{C}=c}$ is

$$\begin{aligned} p_{\mathcal{W}|\mathcal{C}}(W^{(1)}|c) &= \frac{p_{\mathcal{W},\mathcal{C}}(W^{(1)}, c)}{p_{\mathcal{C}}(c)} \\ &= \frac{\frac{1}{4a_w a_b}\mathbb{1}_{W^{(1)}\in[-a_w,a_w]} \cdot \mathbb{1}_{-cW^{(1)}\in[-a_b,a_b]} \cdot (-W^{(1)})}{p_{\mathcal{C}}(c)}. \end{aligned}$$

When $|c| \leq \frac{a_b}{a_w}$, $|-cW^{(1)}| \leq \frac{a_b}{a_w}a_w = a_b$. So $-cW^{(1)} \in [-a_b, a_b]$ is true and $\mathbb{1}_{-cW^{(1)}\in[-a_b,a_b]} = 1$. Combined with the fact that $p_{\mathcal{C}}(c)$ is constant when $|c| \leq \frac{a_b}{a_w}$, we have $p_{\mathcal{W}|\mathcal{C}}(W^{(1)}|c)$ is independent of c when $|c| \leq \frac{a_b}{a_w}$. So $\mathbb{E}(\mathcal{W}^2|\mathcal{C} = c)$ is constant when $|c| \leq \frac{a_b}{a_w}$. Since $\frac{a_b}{a_w} \geq I$, $\mathbb{E}(\mathcal{W}^2|\mathcal{C} = c)$ and $p_{\mathcal{C}}(c)$ are constant when $c \in [-I, I]$. Then $\zeta(x) = \mathbb{E}(W^2|C = x)p_{\mathcal{C}}(x)$ is constant when $c \in [-I, I]$. □

2.I Proofs of Results for Multivariate Regression

2.I.1 Proof of Theorem 16

In this section, we prove Theorem 16. We will need the following lemmas:

Lemma 30. *Let $f \in \text{Lip}(\mathbb{R}^d)$ be considered as a tempered distribution and $(-\Delta)^s f \equiv 0$,*

$s > 0$. *Then f is linear, i.e., $f = \langle \mathbf{u}, \mathbf{x} \rangle + v$.*

Proof of Lemma 30. In the following proof we regard f as a tempered distribution, thus the fractional Laplacian and Fourier transform of f can be defined. We first give a brief introduction of tempered distribution.

The space of tempered distributions $S'(\mathbb{R}^d)$ is the space of continuous linear functionals on the space of Schwartz test functions $S(\mathbb{R}^d)$. The space of Schwartz test functions on $\mathbb{R}^d$ is the rapidly decreasing function space

$$S(\mathbb{R}^d) := \left\{ \psi \in C^\infty(\mathbb{R}^d) \mid \forall \alpha, \beta \in \mathbb{N}^d, \sup_{\mathbf{x} \in \mathbb{R}^d} |\mathbf{x}^\beta D^\alpha \psi(\mathbf{x})| < \infty \right\}. \tag{2.107}$$

The details of defining norms and the topology on $S(\mathbb{R}^d)$ is shown in [MU08, Chapter 1].

For any $f \in \mathrm{Lip}(\mathbb{R}^d)$, we can define a corresponding tempered distribution T_f by

$$T_f : S(\mathbb{R}^d) \mapsto \mathbb{R},\ T_f(\psi) = \int_{\mathbb{R}^d} f \psi \mathrm{d}\mathbf{x}. \tag{2.108}$$

So any $f \in \mathrm{Lip}(\mathbb{R}^d)$ can be naturally regarded as a tempered distribution T_f.

Let $\mathcal{F}$ be the Fourier transform. Since $\mathcal{F}$ and its adjoint maps a Schwartz function to a Schwartz function, we can define the Fourier transform of a tempered distribution by

$$\mathcal{F} : S'(\mathbb{R}^d) \mapsto S'(\mathbb{R}^d),\ (\mathcal{F} T_f)(\psi) = \int_{\mathbb{R}^d} f \cdot \mathcal{G} \psi \mathrm{d}\mathbf{x}, \tag{2.109}$$

where $\mathcal{G}$ is the adjoint of $\mathcal{F}$. Details of Fourier transform on tempered distributions can be found in [MU08, Chapter 1.7].

Similarly the fractional Laplacian of a tempered distribution is defined by

$$(-\Delta)^s : S'(\mathbb{R}^d) \mapsto S'(\mathbb{R}^d),\ ((-\Delta)^s T_f)(\psi) = \int_{\mathbb{R}^d} f \cdot (-\Delta)^s \psi \mathrm{d}\mathbf{x}, \tag{2.110}$$

Since $(-\Delta)^s f \equiv 0$, in Fourier domain we have $\|\boldsymbol{\xi}\|^{2s} \mathcal{F} f \equiv 0$. It means that the support

of $\mathcal{F}f$ is $\{0\}$. From [Fol99b, Chapter 9], we know that $\widehat{f}$ is the linear combinations of δ (Dirac's Delta) and its derivatives. Then f is a polynomial. Since f is Lipschitz continuous, we conclude that f is linear. □

Lemma 31. *Let $\alpha \in L^2(\mathbb{S}^{d-1} \times \mathbb{R})$. Suppose that $\alpha = \alpha^+ + \alpha^-$ where α^+ is even and α^- is odd. Then $\|\alpha\|_2 \geq \|\alpha^+\|_2$ and $\|\alpha\|_2 \geq \|\alpha^-\|_2$.*

Proof of Lemma 31. Since

$$\begin{aligned}
\|\alpha\|_2^2 &= \|\alpha^+ + \alpha^-\|_2^2 \\
&= \|\alpha^+\|_2^2 + \|\alpha^-\|_2^2 + 2\langle \alpha^+, \alpha^+ \rangle \\
&= \|\alpha^+\|_2^2 + \|\alpha^-\|_2^2 + 2\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \alpha^+ \cdot \alpha^- \, \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
&= \|\alpha^+\|_2^2 + \|\alpha^-\|_2^2,
\end{aligned}$$

where the last equality holds true since $\alpha^+ \cdot \alpha^-$ is odd. Then we have $\|\alpha\|_2 \geq \|\alpha^+\|_2$ and $\|\alpha\|_2 \geq \|\alpha^-\|_2$. □

The next lemma shows that the output of the infinite-width network is Lipschitz continuous. This is also observed in [OWS20, Proposition 8].

Lemma 32. *Assume that (1) the norm of the random vector $\|\boldsymbol{\mathcal{W}}\|$ has the finite second moment; (2) $\int_{\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}} \gamma^2(u, \boldsymbol{V}, c) \, \mathrm{d}\nu(u, \boldsymbol{V}, c) < +\infty$; (3) $\mathbf{u} \in \mathbb{R}^d$ and $v \in \mathbb{R}$. Then $g(\mathbf{x}, (\gamma, \mathbf{u}, v))$ is Lipschitz continuous.*

Proof of Lemma 32. Let $\alpha(u\boldsymbol{V}, -cu) = \gamma(u, \boldsymbol{V}, c)$. For all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$, we have

$$
\begin{aligned}
&|g(\mathbf{x}_1, (\gamma, \mathbf{u}, v)) - g(\mathbf{x}_2, (\gamma, \mathbf{u}, v))| \\
\leq &\left| \int_{\mathbb{R}^d \times \mathbb{R}} |\alpha(\mathbf{W}^{(1)}, b)| \big| [\langle \mathbf{W}^{(1)}, \mathbf{x}_1 \rangle + b]_+ - [\langle \mathbf{W}^{(1)}, \mathbf{x}_2 \rangle + b]_+ \big| \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \right| + |\langle \mathbf{u}, \mathbf{x}_1 - \mathbf{x}_2 \rangle| \\
\leq &\left| \int_{\mathbb{R}^d \times \mathbb{R}} |\alpha(\mathbf{W}^{(1)}, b)| \big| \langle \mathbf{W}^{(1)}, \mathbf{x}_1 - \mathbf{x}_2 \rangle \big| \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \right| + |\langle \mathbf{u}, \mathbf{x}_1 - \mathbf{x}_2 \rangle| \\
\leq &\left(\int_{\mathbb{R}^d \times \mathbb{R}} |\alpha(\mathbf{W}^{(1)}, b)| \|\mathbf{W}^{(1)}\| \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) + \|\mathbf{u}\| \right) \|\mathbf{x}_1 - \mathbf{x}_2\| \\
\leq &\left(\int_{\mathbb{R}^d \times \mathbb{R}} \alpha^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \cdot \int_{\mathbb{R}^d \times \mathbb{R}} \|\mathbf{W}^{(1)}\|^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) + \|\mathbf{u}\| \right) \|\mathbf{x}_1 - \mathbf{x}_2\| \\
\leq &\left(\int_{\mathbb{R}^d \times \mathbb{R}} \alpha^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \cdot \mathbb{E}(\|\boldsymbol{\mathcal{W}}\|^2) + \|\mathbf{u}\| \right) \|\mathbf{x}_1 - \mathbf{x}_2\|.
\end{aligned}
$$

According to the assumptions, $\int_{\mathbb{R}^d \times \mathbb{R}} \alpha^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)$, $\mathbb{E}(\|\boldsymbol{\mathcal{W}}\|^2)$ and $\|\mathbf{u}\|$ are all finite. Then $g(\mathbf{x}, (\gamma, \mathbf{u}, v))$ is Lipschitz continuous. □

Lemma 33. *Given a function $h \in \mathrm{Lip}(\mathbb{R}^d) \cap C(\mathbb{R}^d)$. Define $\psi : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ by $\psi \coloneqq -\frac{1}{2(2\pi)^{d-1}} \mathcal{R}\{(-\Delta)^{(d+1)/2} h\}$. Assume that (1) $\int_{\mathrm{supp}(\zeta)} (\psi(\boldsymbol{V}, c))^2 / \zeta(\boldsymbol{V}, c) \, \mathrm{d}\sigma^{d-1}(\boldsymbol{V}) \mathrm{d}c < +\infty$, where $\zeta(\boldsymbol{V}, c)$ is define in (2.36), and $\psi(\boldsymbol{V}, c) = 0, \ \forall (\boldsymbol{V}, c) \notin \mathrm{supp}(\zeta)$; (2) $\|\boldsymbol{\mathcal{W}}\|_2$ and $\mathcal{B}$ both have finite second moments; (3) $(-\Delta)^{(d+1)/2} h \in L^p(\mathbb{R}^d), \ 1 \leq p < d/(d-1)$. Then there exist $\mathbf{u} \in \mathbb{R}^d$ and $v \in \mathbb{R}$ such that $h(\mathbf{x}) = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \psi(\boldsymbol{V}, c) [\langle \boldsymbol{V}, \mathbf{x} \rangle - c]_+ \, \mathrm{d}\sigma^{d-1}(\boldsymbol{V}) \mathrm{d}c + \langle \mathbf{u}, \mathbf{x} \rangle + v$.*

Proof of Lemma 33. Since $\|\boldsymbol{\mathcal{W}}\|_2$ and $\mathcal{B}$ both have finite second moments, we have

$$
\begin{aligned}
&\int_{\mathrm{supp}(\zeta)} \zeta(\boldsymbol{V}, c) \, \mathrm{d}\sigma^{d-1}(\boldsymbol{V}) \mathrm{d}c \\
&= \int_{\mathrm{supp}(\zeta)} p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) \, p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \mathbb{E}(\mathcal{U}^2 | \boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c) \, \mathrm{d}\sigma^{d-1}(\boldsymbol{V}) \mathrm{d}c \\
&= \mathbb{E}\big(\mathbb{E}(\mathcal{U}^2 | \boldsymbol{\mathcal{V}}, \mathcal{C})\big) \\
&= \mathbb{E}(\mathcal{U}^2) \\
&= \mathbb{E}(\|\boldsymbol{\mathcal{W}}\|_2^2) \\
&< +\infty,
\end{aligned}
$$

and

$$
\begin{aligned}
&\int_{\mathrm{supp}(\zeta)} \zeta(\boldsymbol{V}, c) \cdot c^2 \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
&= \int_{\mathrm{supp}(\zeta)} p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c) \cdot c^2 \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
&= \mathbb{E}\big(\mathbb{E}(\mathcal{U}^2\mathcal{C}^2|\boldsymbol{\mathcal{V}}, \mathcal{C})\big) \\
&= \mathbb{E}(\mathcal{U}^2\mathcal{C}^2) \\
&= \mathbb{E}(\mathcal{B}^2) \\
&< +\infty.
\end{aligned}
$$

Let $\tilde{h}(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \psi(\boldsymbol{V}, c)[\langle \boldsymbol{V}, \mathbf{x}\rangle - c]_+ \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c$. For any $\mathbf{x} \in \mathbb{R}^d$, we have

$$
\begin{aligned}
&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} |\psi(\boldsymbol{V}, c)|[\langle \boldsymbol{V}, \mathbf{x}\rangle - c]_+ \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
\leq &\int_{\mathrm{supp}(\zeta)} |\psi(\boldsymbol{V}, c)|(\|\boldsymbol{V}\|_2\|\mathbf{x}\|_2 + |c|) \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
\leq &\int_{\mathrm{supp}(\zeta)} |\psi(\boldsymbol{V}, c)|(\|\mathbf{x}\|_2 + |c|) \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
\leq &\|\mathbf{x}\|_2 \sqrt{\int_{\mathrm{supp}(\zeta)} \frac{\big(\psi(\boldsymbol{V}, c)\big)^2}{\zeta(\boldsymbol{V}, c)} \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \cdot \int_{\mathrm{supp}(\zeta)} \zeta(\boldsymbol{V}, c) \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c} \\
&+ \sqrt{\int_{\mathrm{supp}(\zeta)} \frac{\big(\psi(\boldsymbol{V}, c)\big)^2}{\zeta(\boldsymbol{V}, c)} \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \cdot \int_{\mathrm{supp}(\zeta)} \zeta(\boldsymbol{V}, c) \cdot c^2 \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c} \\
&< +\infty.
\end{aligned}
$$

So $\tilde{h}(\mathbf{x})$ is well-defined. The above inequality also implies that the Lipschitz constant of $\tilde{h}(\mathbf{x})$ is bounded by $\int_{\mathrm{supp}(\zeta)} |\psi(\boldsymbol{V}, c)| \|\boldsymbol{V}\|_2 \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c$, which is finite. So $\tilde{h}(\mathbf{x})$ is Lipschitz continuous. Then we have

$$
\begin{aligned}
(-\Delta)^{(d+1)/2}\tilde{h} &= -(-\Delta)^{(d-1)/2} \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \psi(\boldsymbol{V}, c)\delta(\langle \boldsymbol{V}, \mathbf{x}\rangle - c) \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
&= -(-\Delta)^{(d-1)/2} \int_{\mathbb{S}^{d-1}} \psi(\boldsymbol{V}, \langle \boldsymbol{V}, \mathbf{x}\rangle) \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V}) \\
&= -(-\Delta)^{(d-1)/2} \mathcal{R}^*\{\psi\}.
\end{aligned} \tag{2.111}
$$

Since $(-\Delta)^{(d+1)/2}h \in L^p(\mathbb{R}^d),\ 1 \leq p < d/(d-1)$, we can apply the inversion formula of the Radon transform [Sol87]:

$$\begin{aligned}(-\Delta)^{(d+1)/2}h &= \frac{1}{2(2\pi)^{d-1}}(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}\} \\ &= -(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\psi\} \\ &= (-\Delta)^{(d+1)/2}\tilde{h}.\end{aligned}$$

According to Lemma 30, we have that $h - \tilde{h}$ is linear, which gives the claim. □

Lemma 33 immediately gives the following corollary:

Corollary 34. *If $\mathcal{R}\{(-\Delta)^{(d+1)/2}g\} \equiv \mathcal{R}\{(-\Delta)^{(d+1)/2}h\}$, and $(-\Delta)^{(d+1)/2}g, (-\Delta)^{(d+1)/2}h \in L^p(\mathbb{R}^d),\ 1 \leq p < d/(d-1)$, then $g - h$ is linear.*

Next lemma shows that the minimizer $h(\mathbf{x})$ of problem (2.37) satisfies that $\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}$ is compactly supported.

Lemma 35. *Consider the training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^M$. Let R be the maximum 2-norm of training inputs, i.e., $R = \max_i \|\mathbf{x}_i\|_2$. Suppose $h(\mathbf{x})$ is the solution of the optimization problem (2.37). Then $\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V}, c) = 0,\ \forall(\boldsymbol{V}, c) \notin \mathbb{S}^{d-1} \times [-R, R]$.*

Proof of Lemma 35. Define $\psi : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ by $\psi \coloneqq -\frac{1}{2(2\pi)^{d-1}}\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}$. Then we construct the function $\overline{\psi} : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ as follows:

$$\overline{\psi}(\boldsymbol{V}, c) = \begin{cases} \psi(\boldsymbol{V}, c), & \text{for } |c| \leq R \\ 0. & \text{for } |c| > R. \end{cases}$$

Since the Radon transform is even, we have that ψ and $\overline{\psi}$ are both even. Since h is the solution of (2.37), ψ satisfies all assumptions of Lemma 33. Then according to Lemma 33, $h(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \psi(\boldsymbol{V}, c)[\langle \boldsymbol{V}, \mathbf{x}\rangle - c]_+ \,\mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c + \langle \mathbf{u}, \mathbf{x}\rangle + v$. Let $\overline{h}(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \overline{\psi}(\boldsymbol{V}, c)[\langle \boldsymbol{V}, \mathbf{x}\rangle - c]_+ \,\mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c$. Then $\overline{h}(\mathbf{x}) - h(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} (\psi - \overline{\psi})(\boldsymbol{V}, c)[\langle \boldsymbol{V}, \mathbf{x}\rangle - c]_+ \,\mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c + \langle \mathbf{u}, \mathbf{x}\rangle + v$.

When $|c| \leq R$, $\psi - \overline{\psi} = 0$. When $|c| > R$, $[\langle \boldsymbol{V}, \mathbf{x}\rangle - c]_+$ is linear with respect to x on $\{\mathbf{x} : \|\mathbf{x}\|_2 \leq R\}$. It means that $\overline{h}(\mathbf{x}) - h(\mathbf{x})$ is linear on $\{\mathbf{x} : \|\mathbf{x}\|_2 \leq R\}$. Then we can find out $\overline{u}$ and $\overline{v}$ such that $h(\mathbf{x}) = \overline{h}(\mathbf{x}) + \langle \overline{\mathbf{u}}, \mathbf{x}\rangle + \overline{v}$ on $\{\mathbf{x} : \|\mathbf{x}\|_2 \leq R\}$. Let $\widetilde{h}(\mathbf{x}) = \overline{h}(\mathbf{x}) + \langle \overline{\mathbf{u}}, \mathbf{x}\rangle + \overline{v}$. Since all training inputs satisfy $\|\mathbf{x}_i\| \leq R$, we have that $\widetilde{h}(\mathbf{x})$ fits all training data. Similar to (2.111), we have that $\Delta\widetilde{h} = \mathcal{R}^*\{\overline{\psi}\}$. Since $\overline{\psi}$ has compact support, the inversion formula of the Radon transform [Sol87] gives that $\overline{\psi} = -\frac{1}{2(2\pi)^{d-1}}\mathcal{R}\{(-\Delta)^{(d+1)/2}\widetilde{h}\}$. Since the support of $\overline{\psi}$ is contained in the support of ψ, we have $\mathcal{R}\{(-\Delta)^{(d+1)/2}\widetilde{h}\}(\boldsymbol{V}, c) = 0,\ \forall(\boldsymbol{V}, c) \notin \mathrm{supp}(\zeta)$. Since $(-\Delta)^{(d+1)/2}\widetilde{h} = -(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\overline{\psi}\}$ and $\overline{\psi}$ is compactly supported, we have $(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\overline{\psi}\} \in L^p(\mathbb{R}^d),\ 1 \leq p < d/(d-1)$ according to [Sol87, Lemma 4.1]. The above argument shows that $\widetilde{h}$ satisfies all constrains of the problem (2.37). Since h is the solution of (2.37), we have $\int_{\mathrm{supp}(\zeta)} \frac{(\psi(\boldsymbol{V},c))^2}{\zeta(\boldsymbol{V},c)}\,\mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \leq \int_{\mathrm{supp}(\zeta)} \frac{\left(\overline{\psi}(\boldsymbol{V},c)\right)^2}{\zeta(\boldsymbol{V},c)}\,\mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c$. It means that $\psi(\boldsymbol{V}, c) = 0$ when $|c| > R$, which gives the claim. □

The proof of Lemma 35 also applies to the optimization problem without the constraint $\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V}, c) = 0,\ \forall(\boldsymbol{V}, c) \notin \mathrm{supp}(\zeta)$. Then we have the following corollary.

Corollary 36. *Consider the training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^M$. Let R be the maximum 2-norm of training inputs, i.e., $R = \max_i \|\mathbf{x}_i\|_2$. Suppose $h(\mathbf{x})$ is the solution of the following optimization problem:*

$$\begin{aligned} \min_{h\in \mathrm{Lip}(\mathbb{R}^d)\cap C(\mathbb{R}^d)} \quad & \int_{\mathrm{supp}(\zeta)} \frac{\left(\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V}, c)\right)^2}{\zeta(\boldsymbol{V}, c)}\,\mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\ \text{subject to} \quad & h(\mathbf{x}_j) = y_j, \quad j = 1, \ldots, M, \\ & (-\Delta)^{(d+1)/2}h \in L^p(\mathbb{R}^d),\ 1 \leq p < d/(d-1). \end{aligned} \tag{2.112}$$

Then $\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V}, c) = 0,\ \forall(\boldsymbol{V}, c) \notin \mathbb{S}^{d-1} \times [0, R]$. It means that if $\mathbb{S}^{d-1} \times [0, R] \subset \mathrm{supp}(\zeta)$, $h(\mathbf{x})$ is also the solution of (2.37).

Now we are ready to prove Theorem 16. We use the proof technique of Theorem 13 and (2.34).

Proof of Theorem 16. First, according to (2.28) and (2.34), if $(\boldsymbol{V}, c) \notin \text{supp}(\zeta)$, we have

$$\begin{aligned}
&|\mathcal{R}\{(-\Delta)^{(d+1)/2} g(\cdot, (\overline{\gamma}, \mathbf{u}, v))\}(\boldsymbol{V}, c)| \\
=&|2(2\pi)^{d-1} \int_{\mathbb{R}} \gamma(u, \boldsymbol{V}, c) \cdot u \ \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}, \mathcal{C}=c}(u) \cdot p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) \ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})| \\
\leq&|2(2\pi)^{d-1} \int_{\mathbb{R}} \gamma^2(u, \boldsymbol{V}, c) \ \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}, \mathcal{C}=c}(u) \cdot \mathbb{E}(\mathcal{U}^2 | \boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c) p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) \ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})| \\
=&0.
\end{aligned} \tag{2.113}$$

By Lemma 32, we have that $g(\mathbf{x}, (\overline{\gamma}, \overline{\mathbf{u}}, \overline{v}))$ is Lipschitz continuous, thus $g(\mathbf{x}, (\overline{\gamma}, \overline{\mathbf{u}}, \overline{v}))$ satisfies all constraints of (2.37). Next, we prove that $g(\mathbf{x}, (\overline{\gamma}, \overline{\mathbf{u}}, \overline{v}))$ is the solution of (2.37).

Let $L(f) = \int_{\text{supp}(\zeta)} \frac{\left(\mathcal{R}\{(-\Delta)^{(d+1)/2} g\}(\boldsymbol{V}, c)\right)^2}{\zeta(\boldsymbol{V}, c)} \ \mathrm{d}\sigma^{d-1}(\boldsymbol{V}) \mathrm{d}c$. We first show that when $m \geq d+1$, the functional $L(f)$ is strictly convex on the feasible set, which means that the minimizer of problem (2.37) is unique.

Suppose f_1, f_2 are two different functions in the feasible set of (2.37). So $\mathcal{R}\{(-\Delta)^{(d+1)/2} f_1\}$ and $\mathcal{R}\{(-\Delta)^{(d+1)/2} f_2\}$ should be different. Otherwise, according to Corollary 34, $f_1 - f_2$ is a linear function. We know that $(f_1 - f_2)(\mathbf{x}_i) = 0$, $i = 1, \ldots, m$. So $f_1 = f_2$ on at least $d+1$ points. Then $f_1 - f_2 \equiv 0$ and this is a contradiction. Since $\mathcal{R}\{(-\Delta)^{(d+1)/2}(f_1)\}$ and $\mathcal{R}\{(-\Delta)^{(d+1)/2}(f_2)\}$ are different, by strict convexity of the square function, we have that $L(f)$ is strictly convex on the feasible set.

Suppose $h(\mathbf{x})$ is the minimizer of problem (2.37) and $h(\mathbf{x})$ is different from $g(\mathbf{x}, (\overline{\gamma}, \overline{\mathbf{u}}, \overline{v}))$. Then by uniqueness of the solution,

$$L(h) < L(g(\mathbf{x}, (\overline{\gamma}, \overline{\mathbf{u}}, \overline{v}))). \tag{2.114}$$

Now our goal is to find a different $(\gamma, \mathbf{u}, v)$ with smaller cost in problem (2.26). Then

$(\overline{\gamma}, \overline{\mathbf{u}}, \overline{v})$ is not the solution of (2.26), which is a contradiction. We set

$$\gamma(u, \boldsymbol{V}, c) = \begin{cases} \dfrac{\mathcal{R}\{(-\Delta)^{(d+1)/2} h\}(\boldsymbol{V}, c) \cdot u}{-2(2\pi)^{d-1} \zeta(\boldsymbol{V}, c)}, & (\boldsymbol{V}, c) \in \operatorname{supp}(\zeta), \\ 0, & (\boldsymbol{V}, c) \notin \operatorname{supp}(\zeta). \end{cases}$$

According to (2.32), we have $\Delta g(\cdot, (\gamma, \mathbf{0}, 0)) = \mathcal{R}^*\{\beta\}$ where β is defined in (2.28) and (2.31). Using Lemma 35, we know that $\mathcal{R}\{(-\Delta)^{(d+1)/2} h\}$ is compactly supported. Then we can easily verify that β is also compactly supported. According to [Sol87, Lemma 4.1], $(-\Delta)^{(d-1)/2} \mathcal{R}^*\{\beta\} \in L^p(\mathbb{R}^d),\ 1 \leq p < d/(d-1)$, which means that $g(\cdot, (\gamma, \mathbf{0}, 0))$ satisfies the third constraint of the optimization problem (2.37).

Since the Radon transform is an even function, we have $\gamma(u, \boldsymbol{V}, c) = \gamma(u, -\boldsymbol{V}, -c)$. Since the distribution of $(\boldsymbol{\mathcal{W}}, \mathcal{B})$ is symmetric, we have that $\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}, \mathcal{C}=c}$ is the same probability measure as $\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=-\boldsymbol{V}, \mathcal{C}=-c}$ and $p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) p_{-\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) = p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(-c) p_{\boldsymbol{\mathcal{V}}}(-\boldsymbol{V})$. From the definition of κ (2.28) and β (2.31), we have that κ and β are even. Then the odd part β^- of β is 0. According to (2.34),

$$\begin{aligned} &\mathcal{R}\{(-\Delta)^{(d+1)/2} g(\cdot, (\gamma, \mathbf{0}, 0)))\}(\boldsymbol{V}, c) \\ =& -2(2\pi)^{d-1} p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \int_{\mathbb{R}^+} \gamma(u, \boldsymbol{V}, c) \cdot u \, \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}, \mathcal{C}=c}(u) \\ =& -2(2\pi)^{d-1} p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \int_{\mathbb{R}^+} \frac{\mathcal{R}\{(-\Delta)^{(d+1)/2} h\}(\boldsymbol{V}, c) \cdot u^2}{-2(2\pi)^{d-1} \zeta(\boldsymbol{V}, c)} \, \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}, \mathcal{C}=c}(u) \qquad (2.115) \\ =& -2(2\pi)^{d-1} p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \frac{\mathcal{R}\{(-\Delta)^{(d+1)/2} h\}(\boldsymbol{V}, c) \cdot \mathbb{E}(\mathcal{U}^2 | \boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c)}{-2(2\pi)^{d-1} \zeta(\boldsymbol{V}, c)} \\ =& \mathcal{R}\{(-\Delta)^{(d+1)/2} h\}(\boldsymbol{V}, c), \quad (\boldsymbol{V}, c) \in \operatorname{supp}(\zeta). \end{aligned}$$

It is not difficult to show that if $(\boldsymbol{V}, c) \notin \operatorname{supp}(\zeta)$, then $\mathcal{R}\{(-\Delta)^{(d+1)/2} g(\cdot, (\gamma, \mathbf{0}, 0)))\}(\boldsymbol{V}, c) = 0$ as in (2.113). Then, according to (2.115), $\mathcal{R}\{(-\Delta)^{(d+1)/2} g(\cdot, (\gamma, \mathbf{0}, 0)))\} \equiv \mathcal{R}\{(-\Delta)^{(d+1)/2} h\}$. According to Corollary 34, we have that $g(\cdot, (\gamma, \mathbf{0}, 0))) - h$ is a linear function. This means that we can find $\mathbf{u} \in \mathbb{R}^d, v \in \mathbb{R}$ such that $\langle \mathbf{u}, \mathbf{x} \rangle + v + g(\mathbf{x}, (\gamma, \mathbf{0}, 0))) \equiv h(\mathbf{x})$. Then we find $(\gamma, \mathbf{u}, v)$ such that $g(\mathbf{x}, (\gamma, \mathbf{u}, v)) = \langle \mathbf{u}, \mathbf{x} \rangle + v + g(\mathbf{x}, (\gamma, \mathbf{0}, 0))) = h(\mathbf{x})$ on $\operatorname{supp}(\zeta)$. So

$g(\mathbf{x}_j, (\gamma, \mathbf{u}, v)) = h(\mathbf{x}_j) = y_j$. This means that $(\gamma, \mathbf{u}, v)$ satisfies the condition in problem (2.26). Next we compute the cost of $(\gamma, \mathbf{u}, v)$:

$$
\begin{aligned}
&\int_{\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}} \gamma^2(u, \boldsymbol{V}, c)\ \mathrm{d}\nu(u, \boldsymbol{V}, c) \\
=&\int_{\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}} \left(\frac{\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V}, c)\cdot u}{-2(2\pi)^{d-1}\zeta(\boldsymbol{V}, c)}\right)^2 \mathrm{d}\nu(u, \boldsymbol{V}, c) \\
=&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \left(\frac{\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V}, c)}{\zeta(\boldsymbol{V}, c)}\right)^2 \left(\frac{\int_{\mathbb{R}^+} u^2\ \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c}(u)}{4(2\pi)^{2(d-1)}}\right) \mathrm{d}\nu_{\boldsymbol{\mathcal{V}},\mathcal{C}}(\boldsymbol{V}, c) \\
=&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \left(\frac{\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V}, c)}{\zeta(\boldsymbol{V}, c)}\right)^2 \frac{\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}}=\boldsymbol{V}, \mathcal{C}=c) p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})}{4(2\pi)^{2(d-1)}}\ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
=&\frac{1}{4(2\pi)^{2(d-1)}}\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \frac{\left(\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V}, c)\right)^2}{\zeta(\boldsymbol{V}, c)}\ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
=&\frac{1}{4(2\pi)^{2(d-1)}}L(h).
\end{aligned}
\tag{2.116}
$$

According to (2.34), the cost of $(\overline{\gamma}, \overline{\mathbf{u}}, \overline{v})$ is

$$
\begin{aligned}
&\int_{\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}} \overline{\gamma}^2(u, \boldsymbol{V}, c)\ \mathrm{d}\nu(u, \boldsymbol{V}, c) \\
=&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \left(\int_{\mathbb{R}^+} \overline{\gamma}^2(u, \boldsymbol{V}, c)\ \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c}(u)\right) \mathrm{d}\nu_{\boldsymbol{\mathcal{V}},\mathcal{C}}(\boldsymbol{V}, c) \\
\geq&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \frac{\left(\int_{\mathbb{R}^+} \overline{\gamma}(u, \boldsymbol{V}, c)\cdot u\ \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c}(u)\right)^2}{\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}}=\boldsymbol{V}, \mathcal{C}=c)}\ \mathrm{d}\nu_{\boldsymbol{\mathcal{V}},\mathcal{C}}(\boldsymbol{V}, c) \\
=&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \left(\frac{\mathcal{R}\{(-\Delta)^{(d+1)/2}g(\cdot, (\overline{\gamma}, \mathbf{0}, 0))\}(\boldsymbol{V}, c) - 2(2\pi)^{d-1}\beta^-}{2(2\pi)^{d-1}p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})}\right)^2 \frac{\mathrm{d}\nu_{\boldsymbol{\mathcal{V}},\mathcal{C}}(\boldsymbol{V}, c)}{\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}}=\boldsymbol{V}, \mathcal{C}=c)} \\
\geq&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \left(\frac{\mathcal{R}\{(-\Delta)^{(d+1)/2}g(\cdot, (\overline{\gamma}, \mathbf{0}, 0))\}(\boldsymbol{V}, c)}{2(2\pi)^{d-1}p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})}\right)^2 \frac{p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})}{\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}}=\boldsymbol{V}, \mathcal{C}=c)}\ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
=&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \frac{\left(\mathcal{R}\{(-\Delta)^{(d+1)/2}g(\cdot, (\overline{\gamma}, \mathbf{0}, 0))\}(\boldsymbol{V}, c)\right)^2}{4(2\pi)^{2(d-1)}p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}}=\boldsymbol{V}, \mathcal{C}=c)}\ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
=&\frac{1}{4(2\pi)^{2(d-1)}}L(g(\cdot, (\overline{\gamma}, \mathbf{0}, 0))) \\
=&\frac{1}{4(2\pi)^{2(d-1)}}L(g(\cdot, (\overline{\gamma}, \mathbf{u}, v))) \quad \text{(since } (-\Delta)^{(d+1)/2} \text{ is invariant up to a linear function)},
\end{aligned}
\tag{2.117}
$$

where the first inequality is by the Cauchy-Schwarz inequality and the second inequality

is by the Lemma 31 and the fact that $\frac{\mathcal{R}\{(-\Delta)^{(d+1)/2}g(\cdot,(\overline{\gamma},\mathbf{0},0))\}(\boldsymbol{V},c)}{2(2\pi)^{d-1}p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})}$ is an even function and $\frac{-2(2\pi)^{d-1}\beta^{-}}{2(2\pi)^{d-1}p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})}$ is an odd function. Then we have

$$
\begin{aligned}
&\int_{\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}} \gamma^2(u,\boldsymbol{V},c)\ \mathrm{d}\nu(u,\boldsymbol{V},c)\\
=&\frac{1}{4(2\pi)^{2(d-1)}}L(h) \quad \text{(according to (2.116))}\\
<&\frac{1}{4(2\pi)^{2(d-1)}}L(g(\cdot,(\overline{\gamma},\mathbf{u},v))) \quad \text{(according to (2.114))}\\
\leq&\int_{\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}} \frac{1}{4(2\pi)^{2(d-1)}}\overline{\gamma}^2(u,\boldsymbol{V},c)\ \mathrm{d}\nu(u,\boldsymbol{V},c) \quad \text{(according to (2.117))}.
\end{aligned}
$$

This means that the cost of $(\gamma,\mathbf{u},v)$ is smaller than the cost of $(\overline{\gamma},\overline{\mathbf{u}},\overline{v})$. This implies that $(\overline{\gamma},\overline{\mathbf{u}},\overline{v})$ is not the solution of (2.26), which is a contradiction and hence the assumption cannot be true. In turn, $h(\mathbf{x})\equiv g(\mathbf{x},(\overline{\gamma},\overline{\mathbf{u}},\overline{v}))$, and $g(x,(\overline{\gamma},\overline{\mathbf{u}},\overline{v})$ is the solution of problem (2.26). This concludes the proof. □

2.I.2 Proof of Theorem 7

Proof of Theorem 7. To simplify the analysis, we let $f(\mathbf{x},\theta_0)\equiv 0$. The analysis still holds without this simplification. It is easy to verify that $\mathrm{supp}(\zeta)=\mathbb{S}^{d-1}\times[-a_b,a_b]$ and $\zeta(\boldsymbol{V},c)$ is constant over $\mathrm{supp}(\zeta)$ according to Proposition 17. According to Corollary 36, we have that the variational problem (2.8) is equivalent to the following variational problem:

$$
\begin{aligned}
\min_{h\in\mathrm{Lip}(\mathbb{R}^d)\cap C(\mathbb{R}^d)} \quad & \int_{\mathrm{supp}(\zeta)} \left(\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)\right)^2\ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c\\
\text{subject to} \quad & h(\mathbf{x}_j)=y_j, \quad j=1,\ldots,M,\\
& (-\Delta)^{(d+1)/2}h\in L^p(\mathbb{R}^d),\ 1\leq p<d/(d-1).
\end{aligned}
\tag{2.118}
$$

The solution $h(\mathbf{x})$ of (2.118) satisfies that $\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)=0,\ \forall(\boldsymbol{V},c)\notin\mathbb{S}^{d-1}\times[0,\max_i\|\mathbf{x}_i\|_2]$. The assumption $a_b\geq\max_i\|\mathbf{x}_i\|_2$ means that $\mathbb{S}^{d-1}\times[0,\max_i\|\mathbf{x}_i\|_2]\subset\mathrm{supp}(\zeta)$. So $\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)=0,\ \forall(\boldsymbol{V},c)\notin\mathrm{supp}(\zeta)$, which means that $h(\mathbf{x})$ is also the solution

of the following variational problem:

$$\begin{aligned} \min_{h\in \mathrm{Lip}(\mathbb{R}^d)\cap C(\mathbb{R}^d)} \quad & \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \big(\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)\big)^2 \,\mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\ \text{subject to} \quad & h(\mathbf{x}_j)=y_j, \quad j=1,\dots,M. \\ & (-\Delta)^{(d+1)/2}h\in L^p(\mathbb{R}^d),\ 1\le p<d/(d-1). \end{aligned} \tag{2.119}$$

So it is sufficient to prove that if $h\in \mathrm{Lip}(\mathbb{R}^d)$ and $(-\Delta)^{(d+1)/2}h\in L^p(\mathbb{R}^d),\ 1\le p<d/(d-1)$, we have

$$\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \big(\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)\big)^2 \,\mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c = \int_{\mathbb{R}^d} \big((-\Delta)^{(d+3)/4}h(\mathbf{x})\big)^2 \,\mathrm{d}\mathbf{x}.$$

Given $f:\mathbb{S}^{d-1}\times\mathbb{R}\to\mathbb{R}$, let $\widetilde{f}$ be the Fourier transform over affine parameter:

$$\widetilde{f}(\boldsymbol{V},\tau)=\int_{\infty}^{\infty} f(\boldsymbol{V},c)e^{-ic\tau}\mathrm{d}c.$$

According to [Sol87, Lemma 4.5], we have

$$\begin{aligned} \widetilde{\mathcal{R}}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},\tau) &= \widehat{(-\Delta)^{(d+1)/2}h}(\tau\boldsymbol{V}) \\ &= \|\tau\|^{d+1}\widehat{h}(\tau\boldsymbol{V}) \quad \text{a.e.,} \end{aligned}$$

where $\widehat{h}$ is the Fourier transform of h. Then we have

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \left(\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)\right)^2 \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
=&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \left(\widetilde{\mathcal{R}}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},\tau)\right)^2 \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}\tau \\
=&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \left(\|\tau\|^{d+1}\widehat{h}(\tau\boldsymbol{V})\right)^2 \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}\tau \\
=&\int_{\mathbb{R}^d} \left(\|\tau\|^{(d+3)/2}\widehat{h}(\mathbf{x})\right)^2 \mathrm{d}\mathbf{x} \\
=&\int_{\mathbb{R}^d} \left(\widehat{(-\Delta)^{(d+3)/4}h(\mathbf{x})}\right)^2 \mathrm{d}\mathbf{x} \\
=&\int_{\mathbb{R}^d} \left((-\Delta)^{(d+3)/4}h(\mathbf{x})\right)^2 \mathrm{d}\mathbf{x}.
\end{aligned}$$

□

2.I.3 Proof of Theorem 8

In order to prove Theorem 8, we need following lemmas:

Lemma 37. *For any $d \geq 2$ and $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$, we have $(-\Delta)^{(d+1)/2}(\|\mathbf{x}-\mathbf{x}_1\|^3 - \|\mathbf{x}-\mathbf{x}_2\|^3) = C_d(\Gamma(\mathbf{x}-\mathbf{x}_1) - \Gamma(\mathbf{x}-\mathbf{x}_2))$, where C_d is a constant.*

Proof of Lemma 37. In order to prove the lemma, we need the following simple fact that

$$(-\Delta)\|\mathbf{x}\|^p = \tilde{C}_p\|\mathbf{x}\|^{p-2}, \tag{2.120}$$

where $\tilde{C}_p$ is a constant depends on p.

For $d \geq 3$, we can actually prove that $(-\Delta)^{(d+1)/2}\|\mathbf{x}\|^3 = C_d(\Gamma(\mathbf{x}))$. We discuss the cases of odd d and even d separately. If d is odd, we apply (2.120) for $(d+1)/2$ times and get

$$\begin{aligned}
(-\Delta)^{(d+1)/2}\|\mathbf{x}\|^3 &= \bar{C}\|\mathbf{x}\|^{3-(d+1)} \\
&= C_d(\Gamma(\mathbf{x})),
\end{aligned}$$

where C_d and $\bar{C}$ are some constant.

If d is even, we apply (2.120) for $d/2$ times and get

$$(-\Delta)^{(d+1)/2}\|\mathbf{x}\|^3 = \bar{C}(-\Delta)^{1/2}\|\mathbf{x}\|^{3-d}.$$

Then we only need to prove that $(-\Delta)^{1/2}\|\mathbf{x}\|^{3-d} = C\|\mathbf{x}\|^{2-d}$ for some constant C. Let $g(\mathbf{x}) = (-\Delta)^{1/2}\|\mathbf{x}\|^{3-d}$. Since the fraction Laplacian can be written a singular integral, we have

$$g(\mathbf{x}) = C_1 \int_{\mathbb{R}^d} \frac{\|\mathbf{x}\|^{3-d} - \|\mathbf{y}\|^{3-d}}{\|\mathbf{x}-\mathbf{y}\|^{d+1}} \mathrm{d}\mathbf{y},$$

where C_1 is some constant. It is easy to see that $g(\mathbf{x})$ is radially symmetric. For any positive number $k > 0$, we have

$$\begin{aligned} g(k\mathbf{x}) &= C_1 \int_{\mathbb{R}^d} \frac{\|k\mathbf{x}\|^{3-d} - \|\mathbf{y}\|^{3-d}}{\|k\mathbf{x}-\mathbf{y}\|^{d+1}} \mathrm{d}\mathbf{y} \\ &= C_1 \int_{\mathbb{R}^d} k^d \cdot \frac{\|k\mathbf{x}\|^{3-d} - k\|\mathbf{y}\|^{3-d}}{\|k\mathbf{x}-k\mathbf{y}\|^{d+1}} \mathrm{d}\mathbf{y} \\ &= C_1 \int_{\mathbb{R}^d} k^{2-d} \cdot \frac{\|\mathbf{x}\|^{3-d} - \|\mathbf{y}\|^{3-d}}{\|\mathbf{x}-\mathbf{y}\|^{d+1}} \mathrm{d}\mathbf{y} \\ &= k^{2-d} g(\mathbf{x}). \end{aligned}$$

Combining the above equation with the fact that $g(\mathbf{x})$ is radially symmetric, we show that $g(\mathbf{x}) = C\|x\|^{2-d}$ for some constant C.

Now we have proved the lemma for $d \geq 3$. Next we consider the case when $d = 2$. Since $(-\Delta)^{3/2}(\|\mathbf{x}-\mathbf{x}_1\|^3 - \|\mathbf{x}-\mathbf{x}_2\|^3) = (-\Delta)^{1/2}(\|\mathbf{x}-\mathbf{x}_1\| - \|\mathbf{x}-\mathbf{x}_2\|)$, we only need to prove that $(-\Delta)^{1/2}(\|\mathbf{x}-\mathbf{x}_1\| - \|\mathbf{x}-\mathbf{x}_2\|) = C(\log\|\mathbf{x}-\mathbf{x}_1\| - \log\|\mathbf{x}-\mathbf{x}_2\|)$, where C is a constant.

Using the singular integral definition of fractional Laplacian, we get

$$
\begin{aligned}
&(-\Delta)^{1/2}(\|\mathbf{x}-\mathbf{x}_1\|-\|\mathbf{x}-\mathbf{x}_2\|)\\
=&C_1\int_{\mathbb{R}^d}\frac{\|\mathbf{x}-\mathbf{x}_1\|-\|\mathbf{x}-\mathbf{x}_2\|-\|\mathbf{y}-\mathbf{x}_1\|+\|\mathbf{y}-\mathbf{x}_2\|}{\|\mathbf{x}-\mathbf{y}\|^3}\mathrm{d}\mathbf{y}\\
=&C_1\lim_{R\to\infty}\int_{B(\mathbf{x}_1,R)\cup B(\mathbf{x}_2,R)}\frac{\|\mathbf{x}-\mathbf{x}_1\|-\|\mathbf{x}-\mathbf{x}_2\|-\|\mathbf{y}-\mathbf{x}_1\|+\|\mathbf{y}-\mathbf{x}_2\|}{\|\mathbf{x}-\mathbf{y}\|^3}\mathrm{d}\mathbf{y}\\
=&C_1\lim_{R\to\infty}\int_{B(\mathbf{x}_1,R)}\frac{\|\mathbf{x}-\mathbf{x}_1\|-\|\mathbf{y}-\mathbf{x}_1\|}{\|\mathbf{x}-\mathbf{y}\|^3}\mathrm{d}\mathbf{y}-C_1\lim_{R\to\infty}\int_{B(\mathbf{x}_2,R)}\frac{\|\mathbf{x}-\mathbf{x}_2\|-\|\mathbf{y}-\mathbf{x}_2\|}{\|\mathbf{x}-\mathbf{y}\|^3}\mathrm{d}\mathbf{y}\\
+&C_1\lim_{R\to\infty}\int_{B(\mathbf{x}_2,R)\backslash B(\mathbf{x}_1,R)}\frac{\|\mathbf{x}-\mathbf{x}_1\|-\|\mathbf{y}-\mathbf{x}_1\|}{\|\mathbf{x}-\mathbf{y}\|^3}\mathrm{d}\mathbf{y}\\
-&C_1\lim_{R\to\infty}\int_{B(\mathbf{x}_1,R)\backslash B(\mathbf{x}_2,R)}\frac{\|\mathbf{x}-\mathbf{x}_2\|-\|\mathbf{y}-\mathbf{x}_2\|}{\|\mathbf{x}-\mathbf{y}\|^3}\mathrm{d}\mathbf{y}.
\end{aligned}
$$

Since for $\mathbf{y}\in B(\mathbf{x}_2,R)\backslash B(\mathbf{x}_1,R)$, we have $\|\mathbf{y}\|\geq R-\|\mathbf{x}_1\|$. And the area of $B(\mathbf{x}_2,R)\backslash B(\mathbf{x}_1,R)$ is at most $2R\|\mathbf{x}_1-\mathbf{x}_2\|$. So

$$
\begin{aligned}
&\lim_{R\to\infty}\int_{B(\mathbf{x}_2,R)\backslash B(\mathbf{x}_1,R)}\left|\frac{\|\mathbf{x}-\mathbf{x}_1\|-\|\mathbf{y}-\mathbf{x}_1\|}{\|\mathbf{x}-\mathbf{y}\|^3}\right|\mathrm{d}\mathbf{y}\\
\leq&\lim_{R\to\infty}2R\|\mathbf{x}_1-\mathbf{x}_2\|\cdot\frac{\|\mathbf{x}-\mathbf{x}_1\|+R+\|\mathbf{x}_1\|+\|\mathbf{x}_2\|}{(R-\|\mathbf{x}\|-\|\mathbf{x}_1\|)^3}\\
=&0.
\end{aligned}
$$

Similarly we have $\lim_{R\to\infty}\int_{B(\mathbf{x}_1,R)\backslash B(\mathbf{x}_2,R)}\frac{\|\mathbf{x}-\mathbf{x}_2\|-\|\mathbf{y}-\mathbf{x}_2\|}{\|\mathbf{x}-\mathbf{y}\|^3}\mathrm{d}\mathbf{y}=0$. Then we get

$$
\begin{aligned}
&(-\Delta)^{1/2}(\|\mathbf{x}-\mathbf{x}_1\|-\|\mathbf{x}-\mathbf{x}_2\|)\\
=&C_1\lim_{R\to\infty}\int_{B(\mathbf{x}_1,R)}\frac{\|\mathbf{x}-\mathbf{x}_1\|-\|\mathbf{y}-\mathbf{x}_1\|}{\|\mathbf{x}-\mathbf{y}\|^3}\mathrm{d}\mathbf{y}-C_1\lim_{R\to\infty}\int_{B(\mathbf{x}_2,R)}\frac{\|\mathbf{x}-\mathbf{x}_2\|-\|\mathbf{y}-\mathbf{x}_2\|}{\|\mathbf{x}-\mathbf{y}\|^3}\mathrm{d}\mathbf{y}\\
=&C_1\lim_{R\to\infty}\int_{B(0,R)}\frac{\|\mathbf{x}-\mathbf{x}_1\|-\|\mathbf{y}\|}{\|\mathbf{x}-\mathbf{x}_1-\mathbf{y}\|^3}\mathrm{d}\mathbf{y}-C_1\lim_{R\to\infty}\int_{B(0,R)}\frac{\|\mathbf{x}-\mathbf{x}_2\|-\|\mathbf{y}\|}{\|\mathbf{x}-\mathbf{x}_2-\mathbf{y}\|^3}\mathrm{d}\mathbf{y}.
\end{aligned}
$$

Let $f(\mathbf{x},R)=\int_{B(0,R)}\frac{\|\mathbf{x}\|-\|\mathbf{y}\|}{\|\mathbf{x}-\mathbf{y}\|^3}\mathrm{d}\mathbf{y}$. Then $(-\Delta)^{1/2}(\|\mathbf{x}-\mathbf{x}_1\|-\|\mathbf{x}-\mathbf{x}_2\|)=\lim_{R\to\infty}f(\mathbf{x}-\mathbf{x}_1,R)-$

$f(\mathbf{x}-\mathbf{x}_2, R)$. Next we show that $f(\lambda\mathbf{x}, \lambda R) = f(\mathbf{x}, R)$ for any $\lambda > 0$. Actually

$$\begin{aligned} f(\lambda\mathbf{x}, \lambda R) &= \int_{B(0,\lambda R)} \frac{\|\lambda\mathbf{x}\| - \|\mathbf{y}\|}{\|\lambda\mathbf{x} - \mathbf{y}\|^3} \mathrm{d}\mathbf{y} \\ &= \int_{B(0,R)} \frac{\lambda\|\mathbf{x}\| - \lambda\|\mathbf{y}\|}{\|\lambda\mathbf{x} - \lambda\mathbf{y}\|^3} \lambda^d \mathrm{d}\mathbf{y} \\ &= \int_{B(0,R)} \frac{\|\mathbf{x}\| - \|\mathbf{y}\|}{\|\mathbf{x} - \mathbf{y}\|^3} \mathrm{d}\mathbf{y} = f(\mathbf{x}, R). \end{aligned}$$

Also it is easy to see that $f(\mathbf{x}, R)$ is radially symmetric over $\mathbf{x}$. So $f(\mathbf{x}, R) = f(\|\mathbf{x}\|\mathbf{u}, R)$ for any unit vector $\mathbf{u} \in \mathbb{R}^2$. Then we get

$$\begin{aligned} \lim_{R\to\infty} f(\mathbf{x}-\mathbf{x}_1, R) - f(\mathbf{x}-\mathbf{x}_2, R) &= \lim_{R\to\infty} f(\mathbf{u}, \frac{R}{\|\mathbf{x}-\mathbf{x}_1\|}) - f(\mathbf{u}, \frac{R}{\|\mathbf{x}-\mathbf{x}_2\|}) \\ &= \lim_{R\to\infty} \int_{B(0,\frac{R}{\|\mathbf{x}-\mathbf{x}_1\|})\setminus B(0,\frac{R}{\|\mathbf{x}-\mathbf{x}_2\|})} \frac{\|\mathbf{u}\| - \|\mathbf{y}\|}{\|\mathbf{u} - \mathbf{y}\|^3} \mathrm{d}\mathbf{y} \\ &= \lim_{R\to\infty} \int_{B(0,\frac{R}{\|\mathbf{x}-\mathbf{x}_1\|})\setminus B(0,\frac{R}{\|\mathbf{x}-\mathbf{x}_2\|})} \frac{-\|\mathbf{y}\|}{\|-\mathbf{y}\|^3} \mathrm{d}\mathbf{y} \\ &= -\lim_{R\to\infty} \int_{[\frac{R}{\|\mathbf{x}-\mathbf{x}_2\|}, \frac{R}{\|\mathbf{x}-\mathbf{x}_1\|}]} \frac{2\pi}{r} \mathrm{d}r \\ &= -2\pi \lim_{R\to\infty} \log \frac{R}{\|\mathbf{x}-\mathbf{x}_1\|} - \log \frac{R}{\|\mathbf{x}-\mathbf{x}_2\|} \\ &= 2\pi(\log\|\mathbf{x}-\mathbf{x}_1\| - \log\|\mathbf{x}-\mathbf{x}_2\|). \end{aligned}$$

So we proved the lemma for case $d = 2$. □

The problem (2.37) is over the Lipschitz continuous function space, which is hard to analyse. The following Lemma shows that we can consider the optimization problem over $-\Delta h$.

Lemma 38. *Suppose $h(\mathbf{x})$ is the solution of the variational problem* (2.37)*. Then there exist*

$\mathbf{u} \in \mathbb{R}^d, v \in \mathbb{R}$ *such that* $(-\Delta h(\mathbf{x}), \mathbf{u}, v)$ *is the solution of the following variational problem:*

$$\begin{aligned}
\min_{\substack{f\in C(\mathbb{R}^d),\\ \mathbf{u}\in\mathbb{R}^d, v\in\mathbb{R}}} \quad & \int_{\mathrm{supp}(\zeta)} \frac{\left(\mathcal{R}\{(-\Delta)^{(d-1)/2} f\}(\boldsymbol{V}, c)\right)^2}{\zeta(\boldsymbol{V}, c)} \, \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
\text{subject to} \quad & \int_{\mathbb{R}^d} f(\mathbf{s})[\Gamma(\mathbf{x}_j - \mathbf{s}) - \Gamma(-\mathbf{s}) - \langle \mathbf{x}_j, \nabla\Gamma(-\mathbf{s})\rangle]\mathrm{d}\mathbf{s} + \langle \mathbf{u}, \mathbf{x}_j\rangle + v = y_j, \ j = 1, \ldots, M, \\
& \mathcal{R}\{(-\Delta)^{(d-1)/2} f\}(\boldsymbol{V}, c) = 0, \ \forall (\boldsymbol{V}, c) \notin \mathrm{supp}(\zeta), \\
& (-\Delta)^{(d-1)/2} f \in L^p(\mathbb{R}^d), \ 1 \le p < d/(d-1), \\
& \sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{x}\| \cdot |f(\mathbf{x})| < \infty,
\end{aligned} \tag{2.121}$$

where $\Gamma(\mathbf{x})$ *is the fundamental solution of the Laplace equation* $-\Delta\Gamma(\mathbf{x}) = \delta(\mathbf{x})$*. The closed form of* $\Gamma(\mathbf{x})$ *is*

$$\Gamma(\mathrm{x}) = \begin{cases} -\frac{1}{2\pi}\log\|\mathbf{x}\|, & d = 2, \\ \frac{1}{d(d-2)V_d\|\mathbf{x}\|^{d-2}}, & d \ge 3, \end{cases}$$

where V_d *is the volume of the unit ball in* $\mathbb{R}^d$*.*

Proof of Lemma 38. First we prove that $\sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{x}\| \cdot |-\Delta h| < \infty$. According to Lemma 35 and Lemma 33, we have $-\Delta h = \mathcal{R}^*\{\psi\}$, where ψ is tightly supported. Then [Sol87, Corollary 3.6] shows that $\mathcal{R}^*\{\psi\} = O(\|\mathbf{x}\|^{-1})$, which gives that $\sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{x}\| \cdot |-\Delta h| < \infty$.

Now it is sufficient to prove that for any $\overline{h} \in \mathrm{Lip}(\mathbb{R}^d)$ satisfying that $-\Delta\overline{h} \in C(\mathbb{R}^d)$ and $\sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{x}\| \cdot |-\Delta\overline{h}(\mathbf{x})| < \infty$, there exist $\mathbf{u} \in \mathbb{R}^d, v \in \mathbb{R}$ such that

$$\int_{\mathbb{R}^d} [-\Delta\overline{h}(\mathbf{s})][\Gamma(\mathbf{x} - \mathbf{s}) - \Gamma(-\mathbf{s}) - \langle \mathbf{x}, \nabla\Gamma(-\mathbf{s})\rangle]\mathrm{d}\mathbf{s} + \langle \mathbf{u}, \mathbf{x}\rangle + v = \overline{h}(\mathbf{x}).$$

Let $\overline{g}(\mathbf{x}) = \int_{\mathbb{R}^d} [-\Delta\overline{h}(\mathbf{s})][\Gamma(\mathbf{x} - \mathbf{s}) - \Gamma(-\mathbf{s}) - \langle \mathbf{x}, \nabla\Gamma(-\mathbf{s})\rangle]\mathrm{d}\mathbf{s}$. First we show that $\overline{g}(\mathbf{x})$ is well-defined. Since $\int_{\|\mathbf{s}\|\ge 1} \|\mathbf{s}\|^{-(d+1)}\mathrm{d}\mathbf{s} < \infty$, we only need to prove that $\Gamma(\mathbf{x} - \mathbf{s}) - \Gamma(-\mathbf{s}) - \langle \mathbf{x}, \nabla\Gamma(-\mathbf{s})\rangle = O(\|\mathbf{s}\|^{-d})$ as $\|\mathbf{s}\| \to \infty$ for any given $\mathbf{x}$. Using Taylor's expansion, we have

$$\Gamma(\mathbf{x} - \mathbf{s}) - \Gamma(-\mathbf{s}) - \langle \mathbf{x}, \nabla\Gamma(-\mathbf{s})\rangle = \mathbf{x}^T H_\Gamma(c\mathbf{x} - \mathbf{s})\mathbf{x} \text{ for some } c \in [0, 1], \tag{2.122}$$

where H_Γ is the Hessian matrix of Γ. Since

$$\frac{\partial \Gamma}{\partial s_i \partial s_j}(\mathbf{s}) = -\frac{\delta_{ij}\|\mathbf{s}\|^2 - d s_i s_j}{dV_d \|\mathbf{s}\|^{d+2}} = O(\|\mathbf{s}\|^{-d}), \tag{2.123}$$

where $\delta_{ij} = 1$ when $i = j$, and $\delta_{ij} = 0$ otherwise. According to (2.122) we have $\Gamma(\mathbf{x}-\mathbf{s}) - \Gamma(-\mathbf{s}) - \langle \mathbf{x}, \nabla\Gamma(-\mathbf{s})\rangle = O(\|\mathbf{s}\|^{-d})$ as $\|\mathbf{s}\| \to \infty$. Then we proved that $\overline{g}(\mathbf{x})$ is well-defined.

Next we prove that $\|\nabla \overline{g}(\mathbf{x})\| = O(\log \|\mathbf{x}\|)$. We only need to consider the large enough $\mathbf{x}$. Suppose $\|\mathbf{x}\| \geq 2$. The partial derivative of $\overline{g}$ is given by

$$\frac{\partial \overline{g}}{\partial x_i}(\mathbf{x}) = \int_{\mathbb{R}^d} -\frac{1}{dV_d}[-\Delta \overline{h}(\mathbf{s})]\left[\frac{x_i - s_i}{\|\mathbf{x}-\mathbf{s}\|^d} + \frac{s_i}{\|\mathbf{s}\|^d}\right] \mathrm{d}\mathbf{s}. \tag{2.124}$$

Since $\sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{x}\| \cdot |-\Delta\overline{h}(\mathbf{x})| < \infty$, we have $\|-\Delta\overline{h}(\mathbf{x})\| \leq C \cdot \min\{1, \frac{1}{\mathbf{x}})\}$ for some constant C. It is easy to see that the integrand of (2.124) is $O(\|\mathbf{s}\|^{d+1})$. So $|\frac{\partial \overline{g}}{\partial x_i}(\mathbf{x})| < \infty$. Next we estimate the integral (2.124) on $\mathbb{R}^d \backslash B(0, \|\mathbf{x}\|/2)$:

$$\begin{aligned}
&\left|\int_{\|\mathbf{s}\|>\|\mathbf{x}\|/2} [-\Delta\overline{h}(\mathbf{s})]\left[\frac{x_i - s_i}{\|\mathbf{x}-\mathbf{s}\|^d} + \frac{s_i}{\|\mathbf{s}\|^d}\right] \mathrm{d}\mathbf{s}\right| \\
\leq &\left|\int_{\|\mathbf{s}\|>\|\mathbf{x}\|/2} \frac{1}{\|\mathbf{s}\|}\left[\frac{x_i - s_i}{\|\mathbf{x}-\mathbf{s}\|^d} + \frac{s_i}{\|\mathbf{s}\|^d}\right] \mathrm{d}\mathbf{s}\right| \\
= &\left|\int_{\|\mathbf{s}\|>1/2} \frac{1}{\|\mathbf{s}\|}\left[\frac{x_i/\|\mathbf{x}\| - s_i}{\|\mathbf{x}/\|\mathbf{x}\| - \mathbf{s}\|^d} + \frac{s_i}{\|\mathbf{s}\|^d}\right] \mathrm{d}\mathbf{s}\right| \\
\leq &\max_{\|\widehat{\mathbf{x}}\|=1}\left|\int_{\|\mathbf{s}\|>1/2} \frac{1}{\|\mathbf{s}\|}\left[\frac{\widehat{x}_i - s_i}{\|\widehat{\mathbf{x}}-\mathbf{s}\|^d} + \frac{s_i}{\|\mathbf{s}\|^d}\right] \mathrm{d}\mathbf{s}\right|.
\end{aligned} \tag{2.125}$$

Since $\left|\int_{\|\mathbf{s}\|>1/2} \frac{1}{\|\mathbf{s}\|}\left[\frac{\widehat{x}_i - s_i}{\|\widehat{\mathbf{x}}-\mathbf{s}\|^d} + \frac{s_i}{\|\mathbf{s}\|^d}\right] \mathrm{d}\mathbf{s}\right|$ is well-defined and continuous function over $\widehat{\mathbf{x}}$. Then $\max_{\|\widehat{\mathbf{x}}\|=1}\left|\int_{\|\mathbf{s}\|>1/2} \frac{1}{\|\mathbf{s}\|}\left[\frac{\widehat{x}_i - s_i}{\|\widehat{\mathbf{x}}-\mathbf{s}\|^d} + \frac{s_i}{\|\mathbf{s}\|^d}\right] \mathrm{d}\mathbf{s}\right|$ is a finite number.

Next we estimate the integral (2.124) on $B(0, \|\mathbf{x}\|/2)$:

$$
\begin{aligned}
&\left|\int_{\|\mathbf{s}\|\le\|\mathbf{x}\|/2}[-\Delta\overline{h}(\mathbf{s})]\left[\frac{x_i-s_i}{\|\mathbf{x}-\mathbf{s}\|^{d-1}}+\frac{s_i}{\|\mathbf{s}\|^d}\right]\mathrm{d}\mathbf{s}\right| \\
\le&\left|\int_{\|\mathbf{s}\|\le 1} C\left[\frac{1}{\|\mathbf{x}-\mathbf{s}\|^d}+\frac{1}{\|\mathbf{s}\|^{d-1}}\right]\mathrm{d}\mathbf{s}\right|+\left|\int_{1<\|\mathbf{s}\|\le\|\mathbf{x}\|/2}\frac{C}{\|\mathbf{s}\|}\left[\frac{1}{\|\mathbf{x}-\mathbf{s}\|^{d-1}}+\frac{1}{\|\mathbf{s}\|^{d-1}}\right]\mathrm{d}\mathbf{s}\right| \\
\le&\left|\int_{\|\mathbf{s}\|\le 1} C\left[\frac{2^d}{\|\mathbf{x}\|^d}+\frac{1}{\|\mathbf{s}\|^{d-1}}\right]\mathrm{d}\mathbf{s}\right|+\left|\int_{1<\|\mathbf{s}\|\le\|\mathbf{x}\|/2}\frac{C}{\|\mathbf{s}\|}\left[\frac{2^d}{\|\mathbf{x}\|^{d-1}}+\frac{1}{\|\mathbf{s}\|^{d-1}}\right]\mathrm{d}\mathbf{s}\right| \\
\le&C_1+\frac{2^dC}{\|\mathbf{x}\|^{d-1}}\left|\int_{1<\|\mathbf{s}\|\le\|\mathbf{x}\|/2}\frac{1}{\|\mathbf{s}\|}\mathrm{d}\mathbf{s}\right|+C\left|\int_{1<\|\mathbf{s}\|\le\|\mathbf{x}\|/2}\frac{1}{\|\mathbf{s}\|^d}\mathrm{d}\mathbf{s}\right| \\
\le&C_1+\frac{2^dC_2}{\|\mathbf{x}\|^{d-1}}\|\mathbf{x}\|^{d-1}+C_3\log\|\mathbf{x}\| \\
\le&C_4+C_3\log\|\mathbf{x}\|,
\end{aligned}
\tag{2.126}
$$

where C_1, C_2, C_3 and C_4 are some constants. Combining (2.125) and (2.126) we proved that $\|\nabla\overline{g}(\mathbf{x})\| = O(\log\|\mathbf{x}\|)$.

In our last step, we prove that $\overline{g}-\overline{h}$ is linear. Because of the property of the fundamental solution, we have $-\Delta(\overline{g}-\overline{h})\equiv 0$. Since $\overline{h}$ is Lipschitz continuous and $\|\nabla\overline{g}(\mathbf{x})\| = O(\log\|\mathbf{x}\|)$, we have $\nabla(\overline{g}-\overline{h}) = O(\log\|\mathbf{x}\|)$. So we can regard $\overline{g}-\overline{h}$ as a tempered distribution. Using the proof technique of Lemma 30, we have that $\overline{g}-\overline{h}$ is a polynomial. Since $\nabla(\overline{g}-\overline{h}) = O(\log\|\mathbf{x}\|)$, $\overline{g}-\overline{h}$ must be a linear function, which gives the claim. □

Proof of Theorem 8. To simplify the proof, we let $f(\mathbf{x},\theta_0)\equiv 0$. The analysis still holds without this simplification. Let $h(\mathbf{x})$ be the solution of (2.9). Then Lemma 38 tell us that there exist $\mathbf{u}\in\mathbb{R}^d, v\in\mathbb{R}$ such that $(-\Delta h(\mathbf{x}),\mathbf{u},v)$ is the solution of the following variational

problem:

$$\begin{aligned}
\min_{\substack{f\in C(\mathbb{R}^d),\\ \mathbf{u}\in\mathbb{R}^d, v\in\mathbb{R}}} \quad & \int_{\mathbb{R}^d} \left((-\Delta)^{(d-1)/4} f(\mathbf{x})\right)^2 \mathrm{d}\mathbf{x} \\
\text{subject to} \quad & \int_{\mathbb{R}^d} f(\mathbf{s})[\Gamma(\mathbf{x}_j-\mathbf{s})-\Gamma(-\mathbf{s})-\langle \mathbf{x}_j, \nabla\Gamma(-\mathbf{s})\rangle]\mathrm{d}\mathbf{s}+\langle \mathbf{u}, \mathbf{x}_j\rangle + v = y_j,\ j=1,\dots,M \\
& (-\Delta)^{(d-1)/2} f \in L^p(\mathbb{R}^d),\ 1\le p < d/(d-1) \\
& \sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{x}\|\cdot|f(\mathbf{x})| < \infty,
\end{aligned} \tag{2.127}$$

Suppose that $f(\mathbf{x})$ is the solution of (2.127). Let $J(f,\mathbf{u},v)=\int_{\mathbb{R}^d}\left((-\Delta)^{(d-1)/4}f(\mathbf{x})\right)^2\mathrm{d}\mathbf{x}$ and $G_j(f,\mathbf{u},v)=\int_{\mathbb{R}^d} f(\mathbf{s})[\Gamma(\mathbf{x}_j-\mathbf{s})-\Gamma(-\mathbf{s})-\langle\mathbf{x}_j,\nabla\Gamma(-\mathbf{s})\rangle]\mathrm{d}\mathbf{s}+\langle\mathbf{u},\mathbf{x}_j\rangle+v$. For any function φ in Schwartz space $\mathcal{S}(\mathbb{R}^d)$,[4] $\tilde{\mathbf{u}}\in\mathbb{R}^d$ and $\tilde{v}\in\mathbb{R}$, we consider the perturbation $(\epsilon\varphi,\epsilon\tilde{\mathbf{u}},\epsilon\tilde{v})$ to the solution $(-\Delta h,\mathbf{u},v)$. It is easy to verify that $-\Delta h+\epsilon\varphi$ satisfies that $(-\Delta)^{(d-1)/2}(-\Delta h+\epsilon\varphi)\in L^p(\mathbb{R}^d),\ 1\le p<d/(d-1)$ and $\sup_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{x}\|\cdot|(-\Delta h+\epsilon\varphi)(\mathbf{x})|<\infty$. Next we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\epsilon}J(-\Delta h+\epsilon\varphi,\mathbf{u}+\epsilon\tilde{\mathbf{u}},v+\epsilon\tilde{v}) &= 2\int_{\mathbb{R}^d}\left((-\Delta)^{(d-1)/4}(-\Delta h)\right)\left((-\Delta)^{(d-1)/4}\varphi\right)\mathrm{d}\mathbf{x} \\
&= 2\int_{\mathbb{R}^d}\varphi\cdot\left((-\Delta)^{(d-1)/2}(-\Delta h)\right)\mathrm{d}\mathbf{x},
\end{aligned}$$

The last equality holds because $\varphi\in\mathcal{S}(\mathbb{R}^d)$. Also we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\epsilon}G_j(-\Delta h+\epsilon\varphi,\mathbf{u}+\epsilon\tilde{\mathbf{u}},v+\epsilon\tilde{v}) \\
&=\int_{\mathbb{R}^d}\varphi(\mathbf{s})[\Gamma(\mathbf{x}_j-\mathbf{s})-\Gamma(-\mathbf{s})-\langle\mathbf{x}_j,\nabla\Gamma(-\mathbf{s})\rangle]\mathrm{d}\mathbf{s}+\langle\tilde{\mathbf{u}},\mathbf{x}_j\rangle+\tilde{v}.
\end{aligned}$$

Then according to the first-order optimality condition, there are scalars $\bar{\lambda}_1,\dots,\bar{\lambda}_M$ such that

$$\begin{cases}
(-\Delta)^{(d-1)/2}(-\Delta h(\mathbf{x})) = \sum_{j=1}^M \bar{\lambda}_j[\Gamma(\mathbf{x}_j-\mathbf{x})-\Gamma(-\mathbf{x})-\langle\mathbf{x}_j,\nabla\Gamma(-\mathbf{x})\rangle] \\
\sum_{j=1}^M \bar{\lambda}_j = 0 \\
\sum_{j=1}^M \bar{\lambda}_j\mathbf{x}_j = \mathbf{0}
\end{cases},$$

[4]The Schwartz functions on $\mathbb{R}^d$ is the function space $\mathcal{S}(\mathbb{R}^d)=\{f\in\mathcal{C}^\infty(\mathbb{R}^d):\forall\alpha,\beta\in\mathbb{N}^d,\sup_{\mathbf{x}\in\mathbb{R}^d}\mathbf{x}^\alpha(D^\beta f)(\mathbf{x})<\infty\}$, where α and β are multi-indices.

which can be simplified to

$$\begin{cases} (-\Delta)^{(d+1)/2} h(\mathbf{x}) = \sum_{j=1}^{M} \bar{\lambda}_j [\Gamma(\mathbf{x} - \mathbf{x}_j) - \Gamma(\mathbf{x})] \\ \sum_{j=1}^{M} \bar{\lambda}_j = 0 \\ \sum_{j=1}^{M} \bar{\lambda}_j \mathbf{x}_j = \mathbf{0} \end{cases} . \tag{2.128}$$

According to Lemma 37 and Lemma 30, we can find out $\mathbf{u} \in \mathbb{R}^d, v \in \mathbb{R}$ such that

$$\begin{aligned} h(\mathbf{x}) &= \frac{1}{C_d} \sum_{j=1}^{M} \bar{\lambda}_j \left[\|\mathbf{x} - \mathbf{x}_j\|^3 - \|\mathbf{x}\|^3 \right] + \langle \mathbf{u}, \mathbf{x} \rangle + v \\ &= \frac{1}{C_d} \sum_{j=1}^{M} \bar{\lambda}_j \|\mathbf{x} - \mathbf{x}_j\|^3 + \langle \mathbf{u}, \mathbf{x} \rangle + v, \end{aligned}$$

which gives (2.10) after substituting $\frac{\bar{\lambda}_j}{C_d}$ by λ_j. Since $h(\mathbf{x})$ should fit all training data and λ_j should satisfy (2.128), the coefficients λ_j, $\mathbf{u}$ and v satisfy (2.11). Now $h(\mathbf{x})$ satisfies the first-order optimality condition and fits all training data. Since the variational problem (2.119) is convex, we only need to check that $h \in \mathrm{Lip}(\mathbb{R}^d)$ and $(-\Delta)^{(d+1)/2} h \in L^p(\mathbb{R}^d)$, $1 \leq p < d/(d-1)$ then we can conclude that $h(\mathbf{x})$ is the solution of (2.119). Using (2.122), we have

$$\begin{aligned} (-\Delta)^{(d+1)/2} h(\mathbf{x}) &= \sum_{j=1}^{M} \bar{\lambda}_j [\Gamma(\mathbf{x}_j - \mathbf{x}) - \Gamma(-\mathbf{x}) - \langle \mathbf{x}_j, \nabla \Gamma(-\mathbf{x}) \rangle] \\ &= \sum_{j=1}^{M} \bar{\lambda}_j \mathbf{x}_j^T H_\Gamma(c\mathbf{x}_j - \mathbf{x}) \mathbf{x}_j \text{ for some } c \in [0, 1]. \end{aligned}$$

According to (2.123), we get that $(-\Delta)^{(d+1)/2} h(\mathbf{x}) = O(\|\mathbf{x}\|^{-d})$. We set $p = (d+1)/d$ which satisfies $1 \leq p < d/(d-1)$. It is easy to verify that $\int_{B(\mathbf{x}_j, \epsilon)} \Gamma^p(\mathbf{x}_j - \mathbf{x}) \mathrm{d}\mathbf{x}$ is integrable for small enough ϵ and $\int_{\mathbb{R}^d \setminus B(0,1)} \|\mathbf{x}\|^{-pd} \mathrm{d}\mathbf{x}$ is integrable. Then $(-\Delta)^{(d+1)/2} h \in L^p(\mathbb{R}^d)$.

Similarly we have

$$\begin{aligned} h(\mathbf{x}) &= \sum_{j=1}^{M} \bar{\lambda}_j \left[\|\mathbf{x}_j - \mathbf{x}\|^3 - \| - \mathbf{x}\|^3 - \langle \mathbf{x}_j, \nabla(\|\cdot\|^3)(-\mathbf{x}) \rangle \right] \\ &= \sum_{j=1}^{M} \bar{\lambda}_j \mathbf{x}_j^T H_{\|\cdot\|^3}(c\mathbf{x}_j - \mathbf{x})\mathbf{x}_j \text{ for some } c \in [0,1], \end{aligned}$$

where $H_{\|\cdot\|^3}$ is the Hessian matrix of $\|\mathbf{x}\|^3$. As $\|\mathbf{x}\| \to \infty$, we have

$$\frac{\partial \|\cdot\|^3}{\partial x_i \partial x_j}(\mathbf{x}) = 3\delta_{ij}\|\mathbf{x}\| - 3\frac{x_i x_j}{\|\mathbf{x}\|} = O(\|\mathbf{x}\|), \tag{2.129}$$

where $\delta_{ij} = 1$ when $i = j$, and $\delta_{ij} = 0$ otherwise. Then we have $h \in \text{Lip}(\mathbb{R}^d)$. □

2.I.4 Explicit Form of the Curvature Penalty Function

Proof of Proposition 17. Since $\boldsymbol{\mathcal{W}} \sim U(\mathbb{S}^{d-1})$, we have that $p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})$ is constant over $\mathbb{S}^{d-1}$ and $\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c) = 1$ because $U = \|\boldsymbol{\mathcal{W}}\| = 1$. Since $\mathcal{B} \sim U(-a, a)$ and $\boldsymbol{\mathcal{W}}$ and $\mathcal{B}$ are independent, we have $p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) = \frac{1}{2a}\mathbb{1}_{[-a,a]}(c)$. Then we get

$$\begin{aligned} \zeta(\boldsymbol{V}, c) &= p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c) \\ &= C_1 \mathbb{1}_{[-a,a]}(c), \end{aligned}$$

where C_1 is a constant. □

Proof of Proposition 18. Let $p_{\boldsymbol{\mathcal{W}},\mathcal{B}}$ and $p_{\mathcal{U},\boldsymbol{\mathcal{V}},\mathcal{C}}$ denote the joint density functions of $(\boldsymbol{\mathcal{W}}, \mathcal{B})$ and $(\mathcal{U}, \boldsymbol{\mathcal{V}}, \mathcal{C})$, respectively. We have

$$p_{\mathcal{U},\boldsymbol{\mathcal{V}},\mathcal{C}}(u, \boldsymbol{V}, c) = \left| \frac{\partial(u\boldsymbol{V}, -uc)}{\partial(u, \boldsymbol{V}, c)} \right| p_{\boldsymbol{\mathcal{W}},\mathcal{B}}(u\boldsymbol{V}, -uc) = u^d p_{\boldsymbol{\mathcal{W}},\mathcal{B}}(u\boldsymbol{V}, -uc),$$

and

$$
\begin{aligned}
&p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c)\\
=&p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})\cdot\int_{\mathbb{R}^+}u^2 p_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c}(u)\ \mathrm{d}u\\
=&\int_{\mathbb{R}^+}u^2 p_{\mathcal{U},\boldsymbol{\mathcal{V}},\mathcal{C}}(u,\boldsymbol{V},c)\ \mathrm{d}u\\
=&\int_{\mathbb{R}^+}u^{d+2} p_{\boldsymbol{\mathcal{W}},\mathcal{B}}(u\boldsymbol{V},-uc)\ \mathrm{d}u.
\end{aligned}
\tag{2.130}
$$

□

Proof of Theorem 19. Using (2.130), we have

$$
\begin{aligned}
\zeta(\boldsymbol{V},c)&=\int_{\mathbb{R}^+}u^{d+2}p_{\boldsymbol{\mathcal{W}},\mathcal{B}}(u\boldsymbol{V},-uc)\ \mathrm{d}u\\
&=\int_{\mathbb{R}^+}u^{d+2}\frac{1}{\sqrt{(2\pi)^d}\sigma_w^d}e^{-\frac{\|u\boldsymbol{V}\|_2^2}{2\sigma_w^2}}\frac{1}{\sqrt{2\pi}\sigma_b}e^{-\frac{u^2c^2}{2\sigma_b^2}}\ \mathrm{d}u\\
&=\frac{1}{(2\pi)^{(d+1)/2}\sigma_w^d\sigma_b}\int_{\mathbb{R}^+}u^{d+2}e^{-(\frac{1}{2\sigma_w^2}+\frac{c^2}{2\sigma_b^2})u^2}\mathrm{d}u.
\end{aligned}
$$

Let $\sigma^2=1/\left(\frac{1}{\sigma_w^2}+\frac{c^2}{\sigma_b^2}\right)$, then we have

$$
\begin{aligned}
\zeta(\boldsymbol{V},c)&=\frac{\sigma}{(2\pi)^{d/2}\sigma_w^d\sigma_b}\int_{\mathbb{R}^+}u^{d+2}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{u^2}{2\sigma^2}}\mathrm{d}u\\
&=\frac{\sigma}{(2\pi)^{d/2}\sigma_w^d\sigma_b}\sigma^{d+2}\cdot 2^{d/2}\cdot\frac{\Gamma(\frac{d+3}{2})}{\sqrt{\pi}}\\
&=\frac{\sigma^{d+3}}{\pi^{(d+1)/2}\sigma_w^d\sigma_b}\Gamma(\frac{d+3}{2})\\
&=\frac{1}{\pi^{(d+1)/2}\sigma_w^d\sigma_b\left(\frac{1}{\sigma_w^2}+\frac{c^2}{\sigma_b^2}\right)^{(d+3)/2}}\Gamma(\frac{d+3}{2})\\
&=\frac{\sigma_w^3\sigma_b^{d+2}}{\pi^{(d+1)/2}(\sigma_b^2+c^2\sigma_w^2)^{(d+3)/2}}\Gamma(\frac{d+3}{2}).
\end{aligned}
$$

□

2.J Other Activation Functions for Univariate Regression

We have focused on networks with ReLUs. The ReLU is special in that the second derivative of ReLU is a delta function. For other activation functions the variational problem on function space will look different.

The paper by [PN19] considers different types of activation functions σ. These are then related to different types of linear operators L in the definition of the smoothness regularizer. Here L and σ satisfy $\mathrm{L}\sigma = \delta$, i.e., σ is a Green's function of L. Suppose σ is homogeneous. Then [PN19] show that minimizing the weight "norm"[5] of two-layer neural networks with activation function σ is actually minimizing 1-norm of $\mathrm{L}f$ where f is the output function of the neural network.

The approach in [PN19] can be combined with our analysis. So if for example we replace the ReLU by another homogeneous activation, we can replace the operator accordingly and get an analogous result.

Proof of Corollary 4. Use the same notation as in Section 2.5, and let σ be the activation function, where we assume that σ is a Green's function of a linear operator L. Then optimization problem (2.19) becomes:

$$\begin{aligned} \min_{\alpha_n \in C(\mathbb{R}^2)} \quad & \int_{\mathbb{R}^2} \alpha_n^2(W^{(1)}, b) \, \mathrm{d}\mu_n(W^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^2} \alpha_n(W^{(1)}, b)\sigma(W^{(1)}x_j + b) \, \mathrm{d}\mu_n(W^{(1)}, b) = y_j, \quad j = 1, \ldots, M. \end{aligned} \tag{2.131}$$

The limit of the problem (2.131) as width $n \to \infty$ is

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^2)} \quad & \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b) \, \mathrm{d}\mu(W^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^2} \alpha(W^{(1)}, b)\sigma(W^{(1)}x_j + b) \, \mathrm{d}\mu(W^{(1)}, b) = y_j, \quad j = 1, \ldots, M. \end{aligned} \tag{2.132}$$

[5]Here the form of "norm" depends on the degree of homogeneity of the activation σ. We use quotation marks because it is a generalized notion of norm which may not satisfy the property of a norm.

As in Section 2.6, we can change the variables and relax the optimization problem (2.132) to

$$\begin{aligned} \min_{\substack{\gamma\in C(\mathbb{R}^2),\\ p\in C(\mathbb{R})}} \quad & \int_{\mathbb{R}^2} \gamma^2(W^{(1)},c)\ \mathrm{d}\nu(W^{(1)},c) \\ \text{subject to} \quad & p(x_j)+\int_{\mathbb{R}^2}\gamma(W^{(1)},c)\sigma\big(W^{(1)}(x_j-c)\big)\ \mathrm{d}\nu(W^{(1)},c)=y_j, \quad j=1,\dots,M \\ & \mathrm{L}\ p\equiv 0. \end{aligned} \tag{2.133}$$

If the activation function σ is ReLU, p is a linear function. Then (2.133) becomes the optimization problem (2.22). Define the output function g of the neural network by

$$g(x,(\gamma,p))=p(x)+\int_{\mathbb{R}^2}\gamma(W^{(1)},c)[W^{(1)}(x-c)]_+\ \mathrm{d}\nu(W^{(1)},c).$$

Assume that the activation function σ is homogeneous of degree k, i.e., $\sigma(ax)=a^k\sigma(x)$ for all $a>0$. Similar to (2.98), we have

$$\begin{aligned} (\mathrm{L}g)(x,(\gamma,p)) &= \mathrm{L}\bigg(\int_{\mathbb{R}^2}\gamma(W^{(1)},c)\big|W^{(1)}\big|^k\sigma\big(\mathrm{sign}(W^{(1)})\cdot(x-c)\big)\ \mathrm{d}\nu(W^{(1)},c)\bigg) \\ &= \int_{\mathbb{R}^2}\gamma(W^{(1)},c)\big|W^{(1)}\big|^k\delta(x-c)\ \mathrm{d}\nu(W^{(1)},c) \\ &= \int_{\mathrm{supp}(\nu_{\mathcal{C}})}\bigg(\int_{\mathbb{R}}\gamma(W^{(1)},c)\big|W^{(1)}\big|^k\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\bigg)\delta(x-c)\ \mathrm{d}\nu_{\mathcal{C}}(c) \\ &= \int_{\mathrm{supp}(\nu_{\mathcal{C}})}\bigg(\int_{\mathbb{R}}\gamma(W^{(1)},c)\big|W^{(1)}\big|^k\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\bigg)\delta(x-c)p_{\mathcal{C}}(c)\mathrm{d}c \\ &= p_{\mathcal{C}}(x)\int_{\mathbb{R}}\gamma(W^{(1)},x)\big|W^{(1)}\big|^k\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}). \end{aligned} \tag{2.134}$$

Then similar to Theorem 13, we show that the solution of (2.133) in function space actually solves the following optimization problem:

$$\min_{h\in C^2(S)}\int_S\frac{\big((\mathrm{L}h)(x)\big)^2}{\zeta(x)}\ \mathrm{d}x \quad \text{s.t.} \quad h(x_j)=y_j, \quad j=1,\dots,m, \tag{2.135}$$

where $\zeta(x)=p_{\mathcal{C}}(x)\mathbb{E}(\mathcal{W}^{2k}|\mathcal{C}=x)$ and $S=\mathrm{supp}(\zeta)\cap[\min_i x_i,\max_i x_i]$. Then Corollary 4 can be shown by using (2.135) and the technique used in proof of Theorem 1. □

2.K Effect of Linear Adjustment of the Training Data

In this section, we show that the solution of the variational problem with linearly adjusted training data (2.25) is close to the solution of training with the original training data (2.20). This means that our characterization of the implicit bias in Theorem 1 gives a close description of the solution of gradient descent training with the original data set. The high level intuition is that fitting a linear function only requires a very small adjustment of the parameters of the network in comparison with the parameter adjustment needed to fit a non-linear function.

For the reader's convenience, we restate the continuous version of the problem (2.20):

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^d \times \mathbb{R})} \quad & \int_{\mathbb{R}^d \times \mathbb{R}} \alpha^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^d \times \mathbb{R}} \alpha(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) = y_j, \quad j = 1, \dots, M, \end{aligned} \tag{2.136}$$

and the linearly adjusted variational problem:

$$\begin{aligned} \min_{\substack{\alpha \in C(\mathbb{R}^d \times \mathbb{R}), \\ \mathbf{u} \in \mathbb{R}^d, v \in \mathbb{R}}} \quad & \int_{\mathbb{R}^d \times \mathbb{R}} \alpha^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^d \times \mathbb{R}} \alpha(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) + \langle \mathbf{u}, \mathbf{x}_j \rangle + v = y_j, \ j = 1, \dots, M. \end{aligned} \tag{2.137}$$

In this chapter, our main focus is on the variational problem (2.137), thus we derive our main result Theorem 1 and Theorem 6 which are statements on linearly adjusted training data. In this section, we try to analyze the difference between solutions of variational problems (2.136) and (2.137), and thus show that to what extent the variational problem (2.5) and (2.8) in Theorem 1 and Theorem 6 describes the implicit bias of gradient descent on original training data.

Suppose the solution of problem (2.136) is $\bar{\alpha}_1$, and the corresponding output function is

$$g(\mathbf{x}, \bar{\alpha}_1) = \int_{\mathbb{R}^2} \bar{\alpha}_1(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b).$$

The solution of problem (2.137) is $(\bar{\alpha}_2, \bar{\mathbf{u}}, \bar{v})$ and the corresponding output function is:

$$g(\mathbf{x}, (\bar{\alpha}_2, \bar{\mathbf{u}}, \bar{v})) = \langle \bar{\mathbf{u}}, \mathbf{x} \rangle + \bar{v} + \int_{\mathbb{R}^2} \bar{\alpha}_2(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b).$$

Our goal is to show that $g(\mathbf{x}, \bar{\alpha}_1)$ and $g(\mathbf{x}, (\bar{\alpha}_2, \bar{\mathbf{u}}, \bar{v}))$ are close to each other.

Suppose that the linear function $\langle \bar{\mathbf{u}}, \mathbf{x} \rangle + \bar{v}$ can be fitted by an infinite width network with parameters α_s, i.e.,

$$\int_{\mathbb{R}^2} \alpha_s(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) = \langle \bar{\mathbf{u}}, \mathbf{x} \rangle + \bar{v}. \tag{2.138}$$

Then $\bar{\alpha}_2 + \alpha_s$ is a feasible solution of the problem (2.136). It is easy to show that $g(\mathbf{x}, \bar{\alpha}_2 + \alpha_s) = g(\mathbf{x}, (\bar{\alpha}_2, \bar{\mathbf{u}}, \bar{v}))$. So we only need to measure the difference between $g(\mathbf{x}, \bar{\alpha}_1)$ and $g(\mathbf{x}, \bar{\alpha}_2 + \alpha_s)$. The next theorem characterizes the relative difference between $\bar{\alpha}_1$ and $\bar{\alpha}_2 + \alpha_s$.

Theorem 39. *Suppose that the solution of the optimization problem* (2.136) *is* $\bar{\alpha}_1$ *and the solution of the optimization problem* (2.137) *is* $(\bar{\alpha}_2, \bar{\mathbf{u}}, \bar{v})$. *Suppose that* α_s *satisfies* (2.138). *Then we have*

$$\frac{\int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s)^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)} \leq 2\sqrt{\frac{\int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}} + \frac{\int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}.$$

Proof of Theorem 39. Since $(\bar{\alpha}_2, \bar{u}, \bar{v})$ is the minimizer of (2.137), we have that $(\bar{\alpha}_1, 0, 0)$ is a feasible solution of (2.136) but not optimal, which means

$$\int_{\mathbb{R}^2} \bar{\alpha}_1^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \geq \int_{\mathbb{R}^2} \bar{\alpha}_2^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b). \tag{2.139}$$

From the optimality of α_1, we have

$$\int_{\mathbb{R}^2} \bar{\alpha}_1^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \leq \int_{\mathbb{R}^2} (\bar{\alpha}_2 + \alpha_s)^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b).$$

Using the first order optimality condition on the problem (2.136), we have that there exist

$\lambda_j \in \mathbb{R}$ such that

$$\alpha_1(\mathbf{W}^{(1)}, b) = \sum_{j=1}^{M} \lambda_j [\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b]_+. \tag{2.140}$$

Since both $\bar{\alpha}_1$ and $\bar{\alpha}_2 + \alpha_s$ are the feasible solutions of the problem (2.132),

$$\int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s) \cdot [\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) = 0, \quad j = 1, \ldots, M. \tag{2.141}$$

Using (2.140) and (2.141), we have

$$\begin{aligned}
&\int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s) \bar{\alpha}_1 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
=&\int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s) \sum_{j=1}^{M} \lambda_j [\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
=&\sum_{j=1}^{M} \lambda_j \int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s) \cdot [\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
=&0.
\end{aligned} \tag{2.142}$$

Then we measure the difference between $\bar{\alpha}_1$ and $\bar{\alpha}_2 + \alpha_s$:

$$
\begin{aligned}
&\int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s)^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
=&\int_{\mathbb{R}^2} (\bar{\alpha}_2 + \alpha_s)^2 - (2\bar{\alpha}_2 + 2\alpha_s - \bar{\alpha}_1)\bar{\alpha}_1 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
=&\int_{\mathbb{R}^2} (\bar{\alpha}_2 + \alpha_s)^2 - \bar{\alpha}_1^2 + (2\bar{\alpha}_2 + 2\alpha_s - 2\bar{\alpha}_1)\bar{\alpha}_1 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
=&\int_{\mathbb{R}^2} (\bar{\alpha}_2 + \alpha_s)^2 - \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \quad \text{(use (2.142))} \\
=&\int_{\mathbb{R}^2} (\bar{\alpha}_2^2 + 2\bar{\alpha}_2\alpha_s + \alpha_s^2) - \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
\leq&\int_{\mathbb{R}^2} (\bar{\alpha}_1^2 + 2\bar{\alpha}_2\alpha_s + \alpha_s^2) - \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \quad \text{(use (2.139))} \\
\leq&\int_{\mathbb{R}^2} 2\bar{\alpha}_2\alpha_s + \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
\leq&2\sqrt{\int_{\mathbb{R}^2} \bar{\alpha}_2^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \cdot \int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)} + \int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
\leq&2\sqrt{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \cdot \int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)} + \int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \quad \text{(use (2.139))}.
\end{aligned}
$$

Then we bound the relative difference between $\bar{\alpha}_1$ and $\bar{\alpha}_2 + \alpha_s$:

$$
\begin{aligned}
&\frac{\int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s)^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)} \\
\leq&\frac{2\sqrt{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \cdot \int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)} + \int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)} \\
=&2\sqrt{\frac{\int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}} + \frac{\int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}.
\end{aligned}
$$

□

The above theorem means that if $\int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)$ is much smaller than $\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)$, the relative difference between $\bar{\alpha}_1$ and $\bar{\alpha}_2 + \alpha_s$ is quite small. Here α_s fits a linear function and $\bar{\alpha}_1$ fits the original training data. Since it is much easier for a neural network to fit a linear function than a non-linear function, in practice we observe that $\int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)$ is

	dimension of inputs	training input set $\mathcal{X}$	training output $\mathcal{Y}$	distribution of $(\boldsymbol{\mathcal{W}}, \mathcal{B})$
Setting 1	1	-2, -1.6, 0.3, 0.6, 2	1.5, 0.5, 1.5, 0.5, 1.5	$W \sim U(-1,1)$ $B \sim U(-2,2)$
Setting 2	2	$(-1,-1)$, $(1,1)$, $(0,0)$, $(-1,1)$, $(1,-1)$	1.5, 1.5, 0.5, -0.5, -0.5	$\boldsymbol{\mathcal{W}} \sim U(\mathbb{S}^1)$ $B \sim U(-2,2)$
Setting 3	2	$(-1,1)$, $(1,1)$, $(0.5,0.9)$, $(-1,-1)$, $(1,-1)$, $(0,0)$, $(-1.3,-0.7)$, $(-0.8,0.3)$, $(-0.4,1.6)$, $(1.6,-0.4)$	1.5, 1.5, 0.5, -0.5, -0.5, -1.5, -1.5, 0.5, 0.5, 0.5	$\boldsymbol{\mathcal{W}} \sim U(\mathbb{S}^1)$ $B \sim U(-2,2)$

Table 2.1: Experimental settings.

indeed much smaller than $\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)$ when the training data is not highly linearly correlated. This is shown in the right panel of Figure 2.12.

Generally speaking, the relative difference between $g(\mathbf{x}, \bar{\alpha}_1)$ and $g(\mathbf{x}, (\bar{\alpha}_2, \bar{\mathbf{u}}, \bar{v}))$ can be related to the relative difference between $\bar{\alpha}_1$ and $\bar{\alpha}_2 + \alpha_s$, which can be bounded by using $D_1 \coloneqq \frac{\int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}{\int_{\mathbb{R}^2} \alpha_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}$. In experiments, the relative difference between $g(\mathbf{x}, \bar{\alpha}_1)$ and $g(\mathbf{x}, (\bar{\alpha}_2, \bar{\mathbf{u}}, \bar{v}))$ is measured by $D \coloneqq \frac{\int_{[-R,R]^d} (g(\mathbf{x}, \bar{\alpha}_1) - g(\mathbf{x}, (\bar{\alpha}_2, \bar{\mathbf{u}}, \bar{v})))^2 \, \mathrm{d}\mathbf{x}}{\int_{[-R,R]^d} (g(\mathbf{x}, \bar{\alpha}_1))^2 \, \mathrm{d}\mathbf{x}}$, where R is the minimal positive number such that $[-R, R]^d$ includes all training samples. In order to compute $\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)$ we only need to solve the optimization problem (2.136) and get α_1. To compute $\int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)$, we first need to solve the optimization problem (2.137) and get $(\bar{\alpha}_2, \bar{\mathbf{u}}, \bar{v})$. Then we need to find out α_s which satisfies (2.138). We can give out an easy form of α_s if we assume that the distribution of $(\boldsymbol{\mathcal{W}}, \mathcal{B})$ is symmetric over each component, i.e., $(\mathcal{W}_1, \ldots, \mathcal{W}_i, \ldots, \mathcal{W}_d, \mathcal{B})$ and $(\mathcal{W}_1, \ldots, -\mathcal{W}_i, \ldots, \mathcal{W}_d, \mathcal{B})$ have the same distribution for $i = 1, \ldots, d$. In this case we can choose $\alpha_s(\mathbf{W}^{(1)}, b) = C_1 \langle \mathbf{W}^{(1)}, \bar{u} \rangle + C_2 \bar{v}$ where C_1, C_2 are constants which is determined by (2.138).

Next, we conduct some experiments to verify the above argument. We try three different settings and they are summarized in Table 2.1. For each setting, we add different linear functions to training data and compute corresponding D_1 and D. In order to verify the idea that D_1 is small if training data is not highly correlated, we compute the coefficient of determination R^2 of the training data and then compare it with D_1. In Figure 2.12 we plot

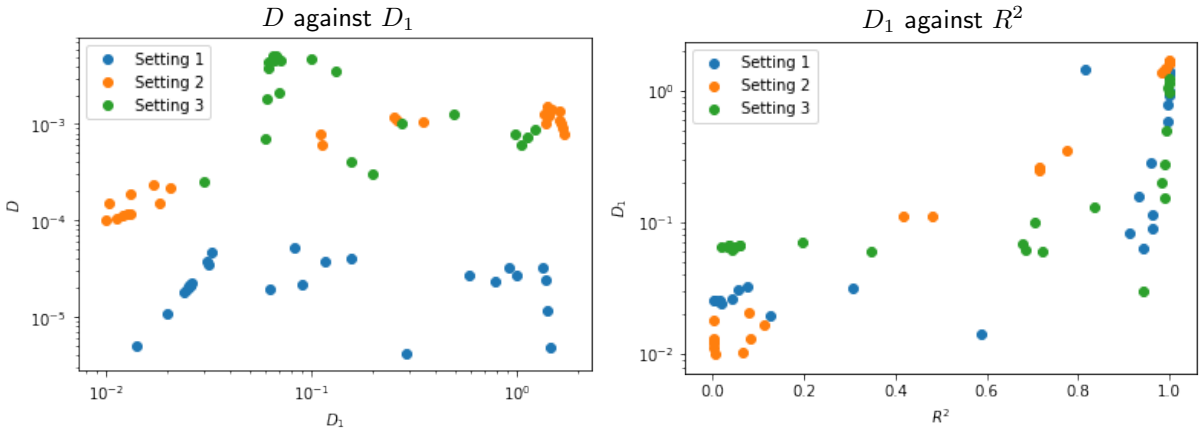

Figure 2.12: Scatter plots of D_1, D and R^2. The left panel is the scatter plot of D against D_1, which shows that D_1 is a very loose upper bound of D. Even when D_1 is around 1, D is still around 10^{-3}. The right panel is the scatter plot of D_1 against R^2, which shows that D_1 is small when training data are not highly linearly correlated and D_1 is large when training data are highly linearly correlated.

D against D_1 and D_1 against R^2. We observe that D_1 is small when R^2 is small and D_1 is a loose upper bound of D. Actually, D is very small even if D_1 is relatively large, which implies that the relative difference between solutions of (2.136) and (2.137) is small in practice.

2.L Neural Networks with Skip Connections

For any given input $\mathbf{x} \in \mathbb{R}^d$, the output of the network with skip connections from the inputs to the outputs is

$$f(\mathbf{x}, \theta) = \sum_{i=1}^{n} W_i^{(2)} \phi(\langle \mathbf{W}_i^{(1)}, \mathbf{x} \rangle + b_i^{(1)}) + \langle \mathbf{u}, \mathbf{x} \rangle + v. \tag{2.143}$$

The initializations of $\mathbf{W}_i^{(1)}, b_i^{(1)}, W_i^{(2)}$ are the same as (2.3). The parameters of skip connections are initialized by zero. We also train this network by gradient descent. The learning rate of parameters $\mathbf{W}_i^{(1)}, b_i^{(1)}, W_i^{(2)}$ is η_r and the learning rate of parameters of skip connections $\mathbf{u}, v$ is η_s. Let $\theta_0 = \mathrm{vec}(\overline{\mathbf{W}}^{(1)}, \overline{\mathbf{b}}^{(1)}, \overline{\mathbf{W}}^{(2)}, \mathbf{0}, 0)$ be the parameters at initialization and $\theta_t = \mathrm{vec}(\mathbf{W}_t^{(1)}, \mathbf{b}_t^{(1)}, \mathbf{W}_t^{(2)}, \mathbf{u}_t, v_t)$ be the parameters after t steps of gradient descent. Then the

gradient descent iterations are

$$\begin{aligned}
&\mathbf{W}_0^{(1)} = \overline{\mathbf{W}}^{(1)}, \quad \mathbf{W}_{t+1}^{(1)} = \mathbf{W}_t^{(1)} - \eta_r \nabla_{\mathbf{W}^{(1)}} L^{\text{lin}}(\theta_t) \\
&\mathbf{b}_0^{(1)} = \overline{\mathbf{b}}^{(1)}, \qquad \mathbf{b}_{t+1}^{(1)} = \mathbf{b}_t^{(1)} - \eta_r \nabla_{\mathbf{b}^{(1)}} L^{\text{lin}}(\theta_t) \\
&\mathbf{W}_0^{(2)} = \overline{\mathbf{W}}^{(2)}, \quad \mathbf{W}_{t+1}^{(2)} = \mathbf{W}_t^{(2)} - \eta_r \nabla_{\mathbf{W}^{(2)}} L^{\text{lin}}(\theta_t) \\
&\mathbf{u}_0 = \mathbf{0}, \qquad\qquad \mathbf{u}_{t+1} = \mathbf{u}_t - \eta_s \nabla_{\mathbf{u}} L^{\text{lin}}(\theta_t) \\
&v_0 = 0, \qquad\qquad v_{t+1} = v_t - \eta_s \nabla_v L^{\text{lin}}(\theta_t)
\end{aligned} \tag{2.144}$$

Let $\widetilde{\omega}_t = \text{vec}(\overline{\mathbf{W}}^{(1)}, \overline{\mathbf{b}}^{(1)}, \widetilde{\mathbf{W}}_t^{(2)}, \widetilde{\mathbf{u}}, \widetilde{v})$ be the parameters at time t under the update rule where $\overline{\mathbf{W}}^{(1)}, \overline{\mathbf{b}}^{(1)}$ are kept fixed at their initial values, and

$$\begin{aligned}
&\widetilde{\mathbf{W}}_0^{(2)} = \overline{\mathbf{W}}^{(2)}, \quad \widetilde{\mathbf{W}}_{t+1}^{(2)} = \widetilde{\mathbf{W}}_t^{(2)} - \eta_r \nabla_{\mathbf{W}^{(2)}} L^{\text{lin}}(\widetilde{\omega}_t) \\
&\widetilde{\mathbf{u}}_0 = \mathbf{0}, \qquad\qquad \widetilde{\mathbf{u}}_{t+1} = \widetilde{\mathbf{u}}_t - \eta_s \nabla_{\mathbf{u}} L^{\text{lin}}(\widetilde{\omega}_t) \\
&\widetilde{v}_0 = 0, \qquad\qquad \widetilde{v}_{t+1} = \widetilde{v}_t - \eta_s \nabla_v L^{\text{lin}}(\widetilde{\omega}_t)
\end{aligned} \tag{2.145}$$

Let $\Psi = \sum_{j=1}^M (\mathbf{x}_j, 1)^T (\mathbf{x}_j, 1)$. Using the similar argument in Section 2.4, we can show that training all parameters can be approximated by training only output weights and skip connections parameters, which is actually a linearized model. Then we can apply Theorem 2.44 with some modifications and show that gradient descent training of the output weights (2.145) on mean squared loss with $\eta_r \leq \frac{M}{4n\lambda_{\max}(\hat{\Theta}_n)}, \eta_s \leq \frac{M}{4\lambda_{\max}(\Psi)}$, achieves zero loss and solves the following optimization problem:

$$\begin{aligned}
\min_{\mathbf{W}^{(2)}} \quad & \frac{1}{\eta_r} \|\mathbf{W}^{(2)} - \overline{\mathbf{W}}^{(2)}\|_2^2 + \frac{1}{\eta_s} \big(\|\mathbf{u}\|_2^2 + v^2 \big) \\
\text{s.t.} \quad & \sum_{i=1}^n (W_i^{(2)} - \overline{W}_i^{(2)}) [\langle \overline{\mathbf{W}}_i^{(1)}, \mathbf{x}_j \rangle + \overline{b}_i^{(1)}]_+ + \langle \mathbf{u}, \mathbf{x}_j \rangle + v = y_j - f(\mathbf{x}_j, \theta_0), \ j = 1, \ldots, M.
\end{aligned} \tag{2.146}$$

Similar to Section 2.5, we let $f^{\text{lin}}(\mathbf{x}, \theta_0) \equiv 0$ by using the Anti-Symmetrical Initialization (ASI) trick. Let μ_n denote the empirical distribution of the samples $(\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)})_{i=1}^n$, i.e., $\mu_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A \Big((\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)}) \Big)$, where $\mathbb{1}_A$ denotes the indicator function for measurable

subsets A in $\mathbb{R}^2$. We further consider a function $\alpha_n \colon \mathbb{R}^2 \to \mathbb{R}$, $\alpha_n(\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)}) = n(W_i^{(2)} - \overline{W}_i^{(2)})$. Then (2.146) with ASI can be rewritten as

$$\begin{aligned} \min_{\alpha_n \in C(\mathbb{R}^2)} \ & \int_{\mathbb{R}^2} \alpha_n^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) + \frac{n\eta_r}{\eta_s}\left(\|\mathbf{u}\|_2^2 + v^2\right) \\ \text{s.t.} \ & \int_{\mathbb{R}^2} \alpha_n(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+ \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) + \langle \mathbf{u}, \mathbf{x}_j\rangle + v = y_j, \ j = 1, \ldots, M. \end{aligned} \tag{2.147}$$

Now we can consider the infinite width limit. Let μ be the probability measure of $(\boldsymbol{\mathcal{W}}, \mathcal{B})$. Assume that $\eta_r \leq n^{-1.5}\eta_s$. Then $\frac{n\eta_r}{\eta_s} = o(1)$ as $n \to \infty$, thus it can be ignored in the infinite width limit. By substituting μ for μ_n, we obtain a continuous version of problem (2.147) as follows:

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^2)} \ & \int_{\mathbb{R}^2} \alpha^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\ \text{s.t.} \ & \int_{\mathbb{R}^2} \alpha(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) + \langle \mathbf{u}, \mathbf{x}_j\rangle + v = y_j, \ j = 1, \ldots, M. \end{aligned} \tag{2.148}$$

Using that μ_n weakly converges to μ, we show that in fact the solution of problem (2.147) converges to the solution of (2.148) in Theorem 40.

Theorem 40 (Infinite width limit for network with skip connections)**.** *Let $(\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)})_{i=1}^n$ be i.i.d. samples from a pair $(\boldsymbol{\mathcal{W}}, \mathcal{B})$ with finite fourth moment. Suppose μ_n is the empirical distribution of $(\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)})_{i=1}^n$ and $(\overline{\alpha}_n, \overline{\mathbf{u}}_n, \overline{v}_n)$ is the solution of (2.147). Let $(\overline{\alpha}, \overline{\mathbf{u}}, \overline{v})$ be the solution of (2.148). Assume that $\eta_r \leq n^{-1.5}\eta_s$. Then, for any compact set $D \subset \mathbb{R}^d$, we have $\sup_{\mathbf{x}\in D} |g_n(\mathbf{x}, (\overline{\alpha}_n, \overline{\mathbf{u}}_n, \overline{v}_n)) - g(\mathbf{x}, (\overline{\alpha}, \overline{\mathbf{u}}, \overline{v}))| = O_p(n^{-1/2})$, where $g_n(\mathbf{x}, (\overline{\alpha}_n, \overline{\mathbf{u}}_n, \overline{v}_n)) = \int_{\mathbb{R}^2} \alpha_n(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}\rangle + b]_+ \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) + \langle \mathbf{u}_n, \mathbf{x}\rangle + v_n$ is the function represented by a network with n hidden neurons and skip connections after training, and $g(\mathbf{x}, (\overline{\alpha}, \overline{\mathbf{u}}, \overline{v})) = \int_{\mathbb{R}^2} \alpha(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}\rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) + \langle \mathbf{u}, \mathbf{x}\rangle + v$ is the function represented by the infinite-width network with skip connections.*

The proof of Theorem 40 is provided at the end of the section. In Section 2.6 and Section 2.7, we show that the optimization problem (2.148) is equivalent to (2.24) in univariate

case and equivalent to (2.37) in multivariate case. Then we can prove our main theorems for networks with skip connections without adjusting the training data.

Theorem 41 (Implicit bias of networks with skip connections, univariate)**.** *Consider a two-layer feedforward network with skip connections* (2.143)*. Assume parameter initialization* (2.3)*, which means for each hidden unit the input weight and bias are initialized from a sub-Gaussian* $(\mathcal{W}, \mathcal{B})$ *with joint density* $p_{\mathcal{W},\mathcal{B}}$*. Then, for any finite data set* $\{(x_j, y_j)\}_{j=1}^M$ *and sufficiently large* n*, the optimization of the mean squared error on the training data* $\{(x_j, y_j)\}_{j=1}^M$ *by gradient descent iterations* (2.144) *with learning rate* $\eta_s \leq \frac{M}{4\lambda_{\max}(\Psi)}, \eta_r \leq n^{-1.5}\eta_s$ *converges to a parameter* θ^* *for which the output function* $f(x, \theta^*)$ *attains zero training error. Furthermore, letting* $\zeta(x) = \int_{\mathbb{R}} |W|^3 p_{\mathcal{W},\mathcal{B}}(W, -Wx)\ \mathrm{d}W$ *and* $S = \mathrm{supp}(\zeta) \cap [\min_j x_j, \max_j x_j]$*, we have* $\sup_{x\in S} \|f(x, \theta^*) - g^*(x)\|_2 = O_p(n^{-\frac{1}{2}})$*over the random initialization* θ_0*, where* g^* *solves following variational problem:*

$$\begin{aligned} \min_{g\in C^2(S)} \quad & \int_S \frac{1}{\zeta(x)} (g''(x) - f''(x, \theta_0))^2\ \mathrm{d}x \\ \text{subject to} \quad & g(x_j) = y_j - ux_j - v, \quad j = 1, \ldots, M. \end{aligned} \tag{2.149}$$

Theorem 42 (Implicit bias of networks with skip connections, multivariate)**.** *Consider the same network settings as in Theorem 41 except with* d *input units instead of a single input unit. Assume that* $\boldsymbol{\mathcal{W}}$ *is a random vector with* $\mathbb{P}(\|\boldsymbol{\mathcal{W}}\| = 0) = 0$ *and* $\mathcal{B}$ *is a random variable; the distribution of* $(\boldsymbol{\mathcal{W}}, \mathcal{B})$ *is symmetric, i.e.,* $(\boldsymbol{\mathcal{W}}, \mathcal{B})$ *and* $(-\boldsymbol{\mathcal{W}}, -\mathcal{B})$ *have the same distribution; and* $\|\boldsymbol{\mathcal{W}}\|_2$ *and* $\mathcal{B}$ *are both sub-Gaussian. Then, for any finite data set* $\{(\mathbf{x}_j, y_j)\}_{i=1}^M$ *and sufficiently large* n*, the optimization of the mean squared error on the training data* $\{(\mathbf{x}_j, y_j)\}_{j=1}^M$ *by gradient descent iterations* (2.144) *with learning rate* $\eta_s \leq \frac{M}{4\lambda_{\max}(\Psi)}, \eta_r \leq n^{-1.5}\eta_s$ *converges to a parameter* θ^* *for which* $f(\mathbf{x}, \theta^*)$ *attains zero training error. Furthermore, let* $\mathcal{U} = \|\boldsymbol{\mathcal{W}}\|_2$*,* $\boldsymbol{\mathcal{V}} = \boldsymbol{\mathcal{W}}/\|\boldsymbol{\mathcal{W}}\|_2$*,* $\mathcal{C} = -\mathcal{B}/\|\boldsymbol{\mathcal{W}}\|_2$ *and* $\zeta(\boldsymbol{V}, c) = p_{\boldsymbol{\mathcal{V}},\mathcal{C}}(\boldsymbol{V}, c)\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c)$*, where* $p_{\boldsymbol{\mathcal{V}},\mathcal{C}}$ *is the joint density of* $(\boldsymbol{\mathcal{V}}, \mathcal{C})$*. Then, for any compact set* $D \subset \mathbb{R}^d$*, we have* $\sup_{\mathbf{x}\in D} \|f(\mathbf{x}, \theta^*) - g^*(\mathbf{x})\|_2 = O_p(n^{-\frac{1}{2}})$ *over the random initialization* θ_0*, where* g^* *solves*

following variational problem:

$$\begin{aligned}
\min_{g\in \mathrm{Lip}(\mathbb{R}^d)} \quad & \int_{\mathrm{supp}(\zeta)} \frac{\big(\mathcal{R}\{(-\Delta)^{(d+1)/2}(g-f(\cdot,\theta_0))\}(\boldsymbol{V},c)\big)^2}{\zeta(\boldsymbol{V},c)}\,\mathrm{d}\boldsymbol{V}\mathrm{d}c \\
\text{subject to} \quad & g(\mathbf{x}_j)=y_j, \quad j=1,\dots,M \\
& \mathcal{R}\{(-\Delta)^{(d+1)/2}(g-f(\cdot,\theta_0))\}(\boldsymbol{V},c)=0, \quad (\boldsymbol{V},c)\notin \mathrm{supp}(\zeta) \\
& (-\Delta)^{(d+1)/2}(g-f(\cdot,\theta_0))\in L^p(\mathbb{R}^d),\ 1\le p<d/(d-1).
\end{aligned} \tag{2.150}$$

Proof of Theorem 40. The Lagrangian of problem (2.147) is

$$\begin{aligned}
&L((\alpha_n,\mathbf{u}_n,v_n),\lambda^{(n)}) \\
&=\int_{\mathbb{R}^2}\alpha_n^2(\mathbf{W}^{(1)},b)\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)+\frac{n\eta_r}{\eta_s}\big(\|\mathbf{u}_n\|_2^2+v_n^2\big)+\sum_{j=1}^M\lambda_j^{(n)}(g_n(\mathbf{x}_j,\alpha_n)-y_j).
\end{aligned}$$

The optimal condition is $\nabla_{\alpha_n}L=0$, which means

$$\begin{aligned}
2\alpha_n(\mathbf{W}^{(1)},b)+\sum_{j=1}^M\lambda_j^{(n)}[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+=0 \text{ when } (\mathbf{W}^{(1)},b)=(\mathbf{W}_i^{(1)},b_i),\ i=1,\dots,k \\
\frac{2n\eta_r}{\eta_s}\mathbf{u}_n+\sum_{j=1}^M\lambda_j^{(n)}\mathbf{x}_j=0 \\
\frac{2n\eta_r}{\eta_s}v_n+\sum_{j=1}^M\lambda_j^{(n)}=0.
\end{aligned}$$

Since only function values on $(\mathbf{W}_i^{(1)},b_i)_{i=1}^M$ are taken into account in problem (2.147), we can let

$$\overline{\alpha}_n(\mathbf{W}^{(1)},b)=-\frac{1}{2}\sum_{j=1}^M\lambda_j^{(n)}[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+ \quad \forall(\mathbf{W}^{(1)},b)\in\mathbb{R}^{d+1} \tag{2.151}$$

without changing $\int_{\mathbb{R}^2}\overline{\alpha}_n^2(\mathbf{W}^{(1)},b)\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)$ and $g_n(\mathbf{x},\overline{\alpha}_n)$.

Here $\lambda_j^{(n)}$, $j=1,\dots,M$ are chosen to make $g_n(\mathbf{x}_i,\overline{\alpha}_n)=y_i$, $i=1,\dots,M$. So we get a

system of linear equations in variables $\{\lambda_j^{(n)}\}_{j=1}^M, \mathbf{u}_n$ and v_n:

$$\begin{aligned} -\frac{1}{2}\sum_{j=1}^{M}\lambda_j^{(n)}\int_{\mathbb{R}^2}[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+[\langle \mathbf{W}^{(1)}, \mathbf{x}_i\rangle + b]_+ \,\mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) + \langle \mathbf{u}_n, \mathbf{x}_i\rangle + v_n = y_i, \\ \sum_{j=1}^{M}\lambda_j^{(n)}\mathbf{x}_j + \frac{2n\eta_r}{\eta_s}\mathbf{u}_n = 0, \\ \sum_{j=1}^{M}\lambda_j^{(n)} + \frac{2n\eta_r}{\eta_s}v_n = 0. \end{aligned} \tag{2.152}$$

for any $i = 1, \ldots, M$. Similarly, the Lagrangian of problem (2.148) is

$$\widetilde{L}(\alpha, \lambda) = \int_{\mathbb{R}^2}\alpha^2(\mathbf{W}^{(1)}, b)\,\mathrm{d}\mu(\mathbf{W}^{(1)}, b) + \sum_{j=1}^{M}\lambda_j(g(\mathbf{x}_j, \alpha) - y_j).$$

The optimality condition is $\nabla_\alpha \widetilde{L} = 0$, which means

$$\begin{gathered} 2\alpha(\mathbf{W}^{(1)}, b) + \sum_{j=1}^{M}\lambda_j^{(n)}[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+ = 0 \quad \forall(\mathbf{W}^{(1)}, b) \in \mathbb{R}^{d+1} \\ 0 \cdot \mathbf{u} + \sum_{j=1}^{M}\lambda_j^{(n)}\mathbf{x}_j = 0 \\ 0 \cdot v + \sum_{j=1}^{M}\lambda_j^{(n)} = 0. \end{gathered}$$

Then we get

$$\overline{\alpha}(\mathbf{W}^{(1)}, b) = -\frac{1}{2}\sum_{j=1}^{M}\lambda_j[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+ \quad \forall(\mathbf{W}^{(1)}, b) \in \mathbb{R}^2. \tag{2.153}$$

Here λ_j, $j = 1, \ldots, M$ are chosen to make $g(\mathbf{x}, \alpha) = y_i$, $i = 1, \ldots, M$. This means that

$$\begin{aligned} -\frac{1}{2}\sum_{j=1}^{M}\lambda_j \int_{\mathbb{R}^2}[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+[\langle \mathbf{W}^{(1)}, \mathbf{x}_i\rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) + \langle \mathbf{u}, \mathbf{x}_i\rangle + v = y_i, \ i = 1, \ldots, M \\ \sum_{j=1}^{M}\lambda_j \mathbf{x}_j + 0 \cdot \mathbf{u} = 0 \\ \sum_{j=1}^{M}\lambda_j + 0 \cdot v = 0 \\ . \end{aligned} \tag{2.154}$$

Compare (2.152) and (2.154). Since the number of samples is finite, $\mathbf{x}_i$ is also bounded. Then by the assumption that $\mathcal{W}$ and $\mathcal{B}$ have finite fourth moments, we have that $[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+[\langle \mathbf{W}^{(1)}, \mathbf{x}_i\rangle + b]_+$ has finite variance. According to central limit theorem, as $n \to \infty$, $\int_{\mathbb{R}^2}[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+[\langle \mathbf{W}^{(1)}, \mathbf{x}_i\rangle + b]_+ \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b)$ tends to a Gaussian distribution with variance $O(n^{-1})$. This implies that $\forall i = 1, \ldots, M, \ \forall j = 1, \ldots, M$,

$$\begin{aligned} &\Big| \int_{\mathbb{R}^2}[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+[\langle \mathbf{W}^{(1)}, \mathbf{x}_i\rangle + b]_+ \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) \\ &- \int_{\mathbb{R}^2}[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+[\langle \mathbf{W}^{(1)}, \mathbf{x}_i\rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)\Big| \\ &= O_p(n^{-1/2}) \end{aligned}$$

Also according to the assumption $\eta_r \leq n^{-1.5}\eta_s$, we have $\frac{2n\eta_r}{\eta_s} = O(n^{-1/2})$. So coefficients of (2.152) converge to coefficients of (2.154) at the rate of $O_p(n^{-1/2})$, then we get

$$|\lambda_j^n - \lambda_j| = O_p(n^{-1/2}), \quad j = 1, \ldots, M. \tag{2.155}$$

Compare (2.151) and (2.153). Given $(\mathbf{W}^{(1)}, b)$, we have

$$|\overline{\alpha}_n(\mathbf{W}^{(1)}, b) - \overline{\alpha}(\mathbf{W}^{(1)}, b)| = O_p(n^{-1/2}). \tag{2.156}$$

Next we want to prove that $\sup_{\mathbf{x}\in D}|g_n(\mathbf{x}, (\overline{\alpha}_n, \overline{\mathbf{u}}_n, \overline{v}_n)) - g(\mathbf{x}, (\overline{\alpha}, \overline{\mathbf{u}}, \overline{v}))| = O_p(n^{-1/2})$. Firstly,

we prove that $\sup_{\mathbf{x}\in D}|g_n(\mathbf{x},(\overline{\alpha},\overline{\mathbf{u}},\overline{v}))-g(\mathbf{x},(\overline{\alpha},\overline{\mathbf{u}},\overline{v}))|=O_p(n^{-1/2})$. Note that $|g_n(\mathbf{x},(\overline{\alpha},\overline{\mathbf{u}},\overline{v}))-g(\mathbf{x},(\overline{\alpha},\overline{\mathbf{u}},\overline{v}))|=|g_n(\mathbf{x},(\overline{\alpha},\mathbf{0},0))-g(\mathbf{x},(\overline{\alpha},\mathbf{0},0))|$. According to (2.96) in the proof of Theorem 12 in Appendix 2.G, we have $\sup_{\mathbf{x}\in D}|g_n(\mathbf{x},(\overline{\alpha},\mathbf{0},0))-g(\mathbf{x},(\overline{\alpha},\mathbf{0},0))|=O_p(n^{-1/2})$. Then we have

$$\sup_{\mathbf{x}\in D}|g_n(\mathbf{x},(\overline{\alpha},\overline{\mathbf{u}},\overline{v}))-g(\mathbf{x},(\overline{\alpha},\overline{\mathbf{u}},\overline{v}))|=O_p(n^{-1/2}). \tag{2.157}$$

Finally, we prove that $\sup_{\mathbf{x}\in D}|g_n(\mathbf{x},(\overline{\alpha}_n,\overline{\mathbf{u}}_n,\overline{v}_n))-g_n(\mathbf{x},(\overline{\alpha},\overline{\mathbf{u}},\overline{v}))|=O_p(n^{-1/2})$. Since $\forall\mathbf{x}\in D$

$$\begin{aligned}
&|g_n(\mathbf{x},(\overline{\alpha}_n,\overline{\mathbf{u}}_n,\overline{v}_n))-g_n(\mathbf{x},(\overline{\alpha},\overline{\mathbf{u}},\overline{v}))|\\
\leq&\int_{\mathbb{R}^2}\left|\overline{\alpha}_n(\mathbf{W}^{(1)},b)[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+-\overline{\alpha}(\mathbf{W}^{(1)},b)[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\right|\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)\\
&+\|\mathbf{x}\|_2\|\overline{\mathbf{u}}_n-\overline{\mathbf{u}}\|_2+|\overline{v}_n-\overline{v}|\\
\leq&\int_{\mathbb{R}^2}\left|\overline{\alpha}_n(\mathbf{W}^{(1)},b)-\overline{\alpha}(\mathbf{W}^{(1)},b)\right|[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)+\|\mathbf{x}\|_2\|\overline{\mathbf{u}}_n-\overline{\mathbf{u}}\|_2+|\overline{v}_n-\overline{v}|\\
\leq&\int_{\mathbb{R}^2}\left|-\frac{1}{2}\sum_{j=1}^{M}(\lambda_j^n-\lambda_j)[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+\right|[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)\\
&+\|\mathbf{x}\|_2\|\overline{\mathbf{u}}_n-\overline{\mathbf{u}}\|_2+|\overline{v}_n-\overline{v}|\\
\leq&\frac{1}{2}\sum_{j=1}^{M}|\lambda_j^n-\lambda_j|\int_{\mathbb{R}^2}[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)\\
&+\|\mathbf{x}\|_2\|\overline{\mathbf{u}}_n-\overline{\mathbf{u}}\|_2+|\overline{v}_n-\overline{v}|\\
\leq&\frac{1}{2}\left(\max_{\mathbf{x}\in D}\int_{\mathbb{R}^2}[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)\right)\sum_{j=1}^{M}|\lambda_j^n-\lambda_j|\\
&+\max_{\mathbf{x}\in D}\|\mathbf{x}\|_2\|\overline{\mathbf{u}}_n-\overline{\mathbf{u}}\|_2+|\overline{v}_n-\overline{v}|.
\end{aligned}$$

Because D is compact and $\int_{\mathbb{R}^2}[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)$ converges according to the law of large numbers, we have that $\max_{\mathbf{x}\in D}\int_{\mathbb{R}^2}[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+[\langle\mathbf{W}^{(1)},\mathbf{x}\rangle+b]_+\,\mathrm{d}\mu_n(\mathbf{W}^{(1)},b)$ and $\max_{\mathbf{x}\in D}\|\mathbf{x}\|_2$ is bounded by a finite number independent of n. Then

according to (2.155),

$$\sup_{\mathbf{x}\in D} |g_n(\mathbf{x}, (\overline{\alpha}_n, \overline{\mathbf{u}}_n, \overline{v}_n)) - g_n(\mathbf{x}, (\overline{\alpha}, \overline{\mathbf{u}}, \overline{v}))| = O_p(n^{-1/2}).$$

Combined with (2.157), we have

$$\sup_{\mathbf{x}\in D} |g_n(\mathbf{x}, (\overline{\alpha}_n, \overline{\mathbf{u}}_n, \overline{v}_n)) - g(\mathbf{x}, (\overline{\alpha}, \overline{\mathbf{u}}, \overline{v}))| = O_p(n^{-1/2}).$$

This concludes the proof. □

2.M Equivalence of Our Characterization and NTK Norm Minimization for Univariate Regression

In this section we demonstrate that NTK norm minimization [ZXL20], which characterizes the implicit bias of training a linearized model by gradient descent, is equivalent to our characterization in Section 2.5 and Section 2.6. For simplicity, we only consider univariate regression in this section. Following [JGH18a], [ZXL20] show that gradient descent can be regarded as a kernel gradient descent in function space, whereby the kernel is given by the NTK. Then for a linearized model, gradient descent finds the global minimum that is closest to the initial output function in the corresponding reproducing kernel Hilbert space (RKHS). Let $\tilde{\Theta}_n$ be the empirical neural tangent kernel of training only the output layer, i.e.,

$$\begin{aligned}
\tilde{\Theta}_n(x_1, x_2) &= \frac{1}{n} \nabla_{W^{(2)}} f(\mathbf{x}_1, \theta_0) \nabla_{W^{(2)}} f(x_2, \theta_0)^T \\
&= \frac{1}{n} \sum_{i=1}^{n} \nabla_{W_i^{(2)}} f(x_1, \theta_0) \nabla_{W_i^{(2)}} f(x_2, \theta_0) \\
&= \frac{1}{n} \sum_{i=1}^{n} [W_i^{(1)} x_1 + b_i^{(1)}]_+ [W_i^{(1)} x_2 + b_i^{(1)}]_+.
\end{aligned}$$

As $n \to \infty$, $\tilde{\Theta}_n \to \tilde{\Theta}$, where

$$\tilde{\Theta}(x_1, x_2) = \int_{\mathbb{R}^2} [W^{(1)} x_1 + b^{(1)}]_+ [W^{(1)} x_2 + b^{(1)}]_+ \, \mathrm{d}\mu(W^{(1)}, b). \tag{2.158}$$

Equivalently, using the notation in Section 2.6, we have

$$\tilde{\Theta}(x_1, x_2) = \int_{\mathbb{R}^2} [W^{(1)}(x_1 - c)]_+ [W^{(1)}(x_2 - c)]_+ \, \mathrm{d}\nu(W^{(1)}, c). \tag{2.159}$$

Next, [ZXL20] construct a RKHS $\mathcal{H}_{\tilde{\Theta}}(S)$ by kernel $\tilde{\Theta}$, and the inner product of the RKHS is denoted by $\langle \cdot, \cdot \rangle_{\tilde{\Theta}}$. Then $\mathcal{H}_{\tilde{\Theta}}(S)$ satisfies:

$$\text{(i)} \quad \forall x \in S, \tilde{\Theta}(\cdot, x) \in \mathcal{H}_{\tilde{\Theta}}(S); \tag{2.160}$$

$$\text{(ii)} \quad \forall x \in S, \forall f \in \mathcal{H}_{\tilde{\Theta}}, \langle f(\cdot), \tilde{\Theta}(\cdot, x) \rangle_{\tilde{\Theta}} = f(x); \tag{2.161}$$

$$\text{(iii)} \quad \forall x, y \in S, \langle \tilde{\Theta}(\cdot, x), \tilde{\Theta}(\cdot, y) \rangle_{\tilde{\Theta}} = \tilde{\Theta}(x, y). \tag{2.162}$$

Here the domain is $S = \mathrm{supp}(\zeta) \cap [\min_i x_i, \max_i x_i]$, which is the same as in Theorem 1 and Theorem 13. Using the reproducing kernel Hilbert space, [ZXL20] prove that $f^{\mathrm{lin}}(x, \widetilde{\omega}_\infty)$ (defined in Section 2.4.2) is the solution of the following optimization problem:

$$\min_{g \in \mathcal{H}_{\tilde{\Theta}}(S)} \|g\|_{\tilde{\Theta}_n} \quad \text{s.t. } g(x_j) = y_j, \ j = 1, \ldots, M.$$

As the width n tends to infinity, the above optimization problem becomes

$$\min_{g \in \mathcal{H}_{\tilde{\Theta}}(S)} \|g\|_{\tilde{\Theta}} \quad \text{s.t. } g(x_j) = y_j, \ j = 1, \ldots, M. \tag{2.163}$$

In Section 2.5, we show that $f^{\mathrm{lin}}(x, \widetilde{\omega}_\infty)$ is the solution of the optimization problem (2.19) in function space. As width n tends to infinity, the optimization problem (2.19) becomes (2.20),

which we repeat below:

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^2)} \quad & \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b)\ \mathrm{d}\mu(W^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^2} \alpha(W^{(1)}, b)[W^{(1)} x_j + b]_+\ \mathrm{d}\mu(W^{(1)}, b) = y_j, \quad j = 1, \ldots, M. \end{aligned} \tag{2.164}$$

Since optimization problems (2.163) and (2.164) both characterize the implicit bias of training a linearized model by gradient descent, they must have the same solution in function space. We express this formally in the following theorem:

Theorem 43 (Equivalence of our variational problem and NTK norm minimization)**.** *Assume that optimization problems* (2.163) *and* (2.164) *are both feasible. Suppose* $\overline{\alpha}$ *is the solution of* (2.164)*, and consider the corresponding output function:*

$$\overline{g}(x) = \int_{\mathbb{R}^2} \overline{\alpha}(W^{(1)}, b)[W^{(1)} x + b]_+\ \mathrm{d}\mu(W^{(1)}, b). \tag{2.165}$$

Then $\overline{g}(x)$ *restricted on* S *is the solution of the optimization problem* (2.163)*.*

Next, we give a standalone proof of this theorem using the property of kernel norm. The proof gives us an idea of what the kernel norm actually looks like.

Proof of Theorem 43. Since $\overline{\alpha}(W^{(1)}, b)$ is the solution of (2.164), according to (2.91) in the proof of Theorem 12,

$$\overline{\alpha}(W^{(1)}, b) = -\frac{1}{2} \sum_{j=1}^{M} \lambda_j [W^{(1)} x_j + b]_+ \quad \forall (W^{(1)}, b) \in \mathbb{R}^2$$

for some constants $\lambda_j, j = 1, \ldots, M$. Then we write $\overline{\alpha}(W^{(1)}, b)$ in the following form:

$$\overline{\alpha}(W^{(1)}, b) = \int_S h(x)[W^{(1)} x + b]_+ \mathrm{d}x, \tag{2.166}$$

where $h(x)$ can be a combination of Dirac delta functions. Then substitute (2.166) into the

expression of $\overline{g}(x)$ (2.165) to obtain

$$\begin{aligned}\overline{g}(x) &= \int_{\mathbb{R}^2 \times S} h(\tilde{x})[W^{(1)}\tilde{x} + b]_+[W^{(1)}x + b]_+ \, \mathrm{d}\mu(W^{(1)}, b)\mathrm{d}\tilde{x} \\ &= \int_S h(\tilde{x})\tilde{\Theta}(x, \tilde{x})\mathrm{d}\tilde{x},\end{aligned} \tag{2.167}$$

where we use the expression of the NTK in equation (2.158). Then we get

$$\begin{aligned}\langle g(x), g(x)\rangle_{\tilde{\Theta}} &= \langle g(x), \int_S h(\tilde{x})\tilde{\Theta}(x, \tilde{x})\mathrm{d}\tilde{x}\rangle_{\tilde{\Theta}}\mathrm{d}\tilde{x} \\ &= \int_S h(\tilde{x})\langle g(x), \tilde{\Theta}(x, \tilde{x})\rangle_{\tilde{\Theta}}\mathrm{d}\tilde{x} \\ &= \int_S h(\tilde{x})g(\tilde{x})\mathrm{d}\tilde{x} \quad \text{(here we use the property of RKHS norm (2.161))} \\ &= \int_{S\times S} h(\tilde{x})h(\bar{x})\tilde{\Theta}(\tilde{x}, \bar{x})\mathrm{d}\tilde{x}\mathrm{d}\bar{x} \quad \text{(use (2.167))}.\end{aligned} \tag{2.168}$$

On the other hand, using (2.166), the objective of (2.164) becomes

$$\begin{aligned}&\int_{S^2} \overline{\alpha}^2(W^{(1)}, b) \, \mathrm{d}\mu(W^{(1)}, b) \\ =&\int_{S\times S\times\mathbb{R}^2} h(\tilde{x})[W^{(1)}\tilde{x} + b]_+ h(\bar{x})[W^{(1)}\bar{x} + b]_+ \, \mathrm{d}\tilde{x}\mathrm{d}\bar{x}\mathrm{d}\mu(W^{(1)}, b) \\ =&\int_{S\times S} h(\tilde{x})h(\bar{x}) \int_{\mathbb{R}^2} [W^{(1)}\tilde{x} + b]_+[W^{(1)}\bar{x} + b]_+\mathrm{d}\mu(W^{(1)}, b) \, \mathrm{d}\tilde{x}\mathrm{d}\bar{x} \\ =&\int_{S\times S} h(\tilde{x})h(\bar{x})\tilde{\Theta}(\bar{x}, \tilde{x}) \, \mathrm{d}\tilde{x}\mathrm{d}\bar{x} \quad \text{(use (2.158))}.\end{aligned} \tag{2.169}$$

Comparing (2.168) and (2.169), we have that optimization problems (2.163) and (2.164) are equivalent if $\alpha(W^{(1)}, b)$ has the form (2.166) and $g(x)$ has the form (2.167). Moreover, if every function $g \in \mathcal{H}_{\tilde{\Theta}}(S)$ can be approximated by the shallow network, we can find $\alpha(W^{(1)}, b)$ in form of (2.166) such that $g(x)$ is expressed in the form of (2.167). In this sense we show that optimization problems (2.163) and (2.164) are equivalent. □

In Section 2.6, we relax the optimization problem (2.21) to (2.22) in order to characterize the implicit bias in function space. This relaxation can also be done in the NTK norm

minimization setting. It means that we can equivalently relax the problem (2.163) to the following problem:

$$\min_{g\in\mathcal{H}_{\tilde{\Theta}}(S),u\in\mathbb{R},v\in\mathbb{R}} \|g-ux-v\|_{\tilde{\Theta}} \quad \text{s.t. } g(x_j)=y_j,\ j=1,\dots,M. \tag{2.170}$$

Then the optimization problems (2.22) and (2.170) are equivalent. Theorem 13 shows that (2.22) and (2.24) have the same solution on the set $S=\mathrm{supp}(\zeta)\cap[\min_i x_i,\max_i x_i]$. Then we have that optimization problems (2.170) and (2.24) are equivalent, which means that

$$\min_{u\in\mathbb{R},v\in\mathbb{R}} \|g-ux-v\|_{\tilde{\Theta}} = \int_S \frac{(g''(x))^2}{\zeta(x)}\,\mathrm{d}x, \quad \forall g\in\mathcal{H}_{\tilde{\Theta}}(S). \tag{2.171}$$

Next, we directly prove the above equation (2.171). Given function $g\in\mathcal{H}_{\tilde{\Theta}}(S)$, let $h=\mathrm{argmin}_{h\in\mathcal{H}_{\tilde{\Theta}}(S)}\|h\|_{\tilde{\Theta}}$, s.t. $h=g-ux-v$ for some $u\in\mathbb{R},v\in\mathbb{R}$. Then according to optimality of h, we have $\langle h,x\rangle_{\tilde{\Theta}}=0$ and $\langle h,1\rangle_{\tilde{\Theta}}=0$. Consider the space $G=\{h\in\mathcal{H}_{\tilde{\Theta}}(S):\langle h,x\rangle_{\tilde{\Theta}}=0,\langle h,1\rangle_{\tilde{\Theta}}=0\}$, which is the orthogonal complement of $\mathrm{span}\{1,x\}$. Then h is the projection of g on G. Since $h=g-ux-v$, $h''=g''$. So we can reformulate the equation (2.171) which we want to prove in the following theorem:

Theorem 44 (Explicit form of the kernel norm)**.** *The kernel norm on the space* $G=\{h\in\mathcal{H}_{\tilde{\Theta}}(S):\langle h,x\rangle_{\tilde{\Theta}}=0,\langle h,1\rangle_{\tilde{\Theta}}=0\}$ *is given as follows:*

$$\|h\|_{\tilde{\Theta}}^2 = \int_S \frac{(h''(x))^2}{\zeta(x)}\,\mathrm{d}x, \quad \forall h\in G. \tag{2.172}$$

This theorem gives the explicit form of the kernel norm in a subspace of $\mathcal{H}_{\tilde{\Theta}}(S)$. Next we prove the above theorem using the property of kernel norm.

Proof of Theorem 44. Let $\tilde{\Theta}_x(\cdot)=\tilde{\Theta}(\cdot,x)$. We can find the orthogonal projection of $\tilde{\Theta}_x$ on space G, which is denoted by $\tilde{\Theta}_{x,G}$. Then we only need to prove that $\langle h,\tilde{\Theta}_{x,G}\rangle_{\tilde{\Theta}}=\int_S\frac{h''(y)\tilde{\Theta}''_{x,G}(y)}{\zeta(y)}\,\mathrm{d}y$ for any $h\in G$ and $x\in S$.

First, $\tilde{\Theta}_{x,G}=\tilde{\Theta}_x-ux-v$ for some constant $u,v\in\mathbb{R}$. Since $h\in G$, $\langle h,1\rangle_{\tilde{\Theta}}=0$ and

$\langle h, x\rangle_{\tilde{\Theta}} = 0$. Then we have

$$\begin{aligned}\langle h, \tilde{\Theta}_{x,G}\rangle_{\tilde{\Theta}} &= \langle h, \tilde{\Theta}_x - ux - v\rangle_{\tilde{\Theta}} \\ &= \langle h, \tilde{\Theta}_x\rangle_{\tilde{\Theta}} - u\langle h, x\rangle_{\tilde{\Theta}} - v\langle h, 1\rangle_{\tilde{\Theta}} \\ &= \langle h, \tilde{\Theta}_x\rangle_{\tilde{\Theta}} \\ &= h(x) \quad \text{(use the reproducing property of the kernel (2.161))}.\end{aligned} \tag{2.173}$$

Next, using the notation from Section 2.6 we have

$$\begin{aligned}\tilde{\Theta}''_{x,G}(y) &= (\tilde{\Theta}_x(y) - uy - v)'' = \tilde{\Theta}_x(y)'' = \frac{\partial^2}{\partial y^2}\tilde{\Theta}(x, y) \\ &= \frac{\partial^2}{\partial y^2}\int_{\mathbb{R}^2}[W^{(1)}(x-c)]_+[W^{(1)}(y-c)]_+ \,\mathrm{d}\nu(W^{(1)}, c) \quad \text{(use (2.159))} \\ &= \frac{\partial^2}{\partial y^2}\int_{\mathbb{R}^2}(W^{(1)})^2[\mathrm{sign}(W^{(1)})(x-c)]_+[\mathrm{sign}(W^{(1)})(y-c)]_+ \,\mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\mathrm{d}\nu_{\mathcal{C}}(c) \\ &= \frac{\partial^2}{\partial y^2}\int_{\mathbb{R}}\big(\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}\geq 0)|\mathcal{C}=c)[x-c]_+[y-c]_+ \\ &\qquad +\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}<0)|\mathcal{C}=c)[c-x]_+[c-y]_+\big)p_{\mathcal{C}}(c)\,\mathrm{d}c \\ &= \int_{\mathbb{R}}\bigg(\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}\geq 0)|\mathcal{C}=c)[x-c]_+\frac{\partial^2}{\partial y^2}[y-c]_+ \\ &\quad +\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}<0)|\mathcal{C}=c)[c-x]_+\frac{\partial^2}{\partial y^2}[c-y]_+\bigg)p_{\mathcal{C}}(c)\,\mathrm{d}c \\ &= \int_{\mathbb{R}}\big(\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}\geq 0)|\mathcal{C}=c)[x-c]_+\delta(y-c) \\ &\quad +\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}<0)|\mathcal{C}=c)[c-x]_+\delta(y-c)\big)p_{\mathcal{C}}(c)\,\mathrm{d}c \\ &= \big(\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}\geq 0)|\mathcal{C}=y)[x-y]_+ + \mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}<0)|\mathcal{C}=y)[y-x]_+\big)p_{\mathcal{C}}(y).\end{aligned}$$

Then we have

$$
\begin{aligned}
&\int_S \frac{h''(y)\tilde{\Theta}''_{x,G}(y)}{\zeta(y)}\,\mathrm{d}y \\
=&\int_S \frac{h''(y)(\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}\geq 0)|\mathcal{C}=y)[x-y]_+ + \mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}<0)|\mathcal{C}=y)[y-x]_+)p_{\mathcal{C}}(y)}{\zeta(y)}\,\mathrm{d}y \\
=&\int_S \frac{h''(y)(\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}\geq 0)|\mathcal{C}=y)[x-y]_+ + \mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}<0)|\mathcal{C}=y)[y-x]_+)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=y)}\,\mathrm{d}y \\
=&\int_S \frac{\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}\geq 0)|\mathcal{C}=y)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=y)}h''(y)[x-y]_+ + \frac{\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}<0)|\mathcal{C}=y)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=y)}h''(y)[y-x]_+\,\mathrm{d}y.
\end{aligned}
$$

Now, if we regard $\int_S \frac{h''(y)\tilde{\Theta}''_{x,G}(y)}{\zeta(y)}\,\mathrm{d}y$ as a function of x, then we get

$$
\begin{aligned}
&\frac{\partial^2}{\partial x^2}\int_S \frac{h''(y)\tilde{\Theta}''_{x,G}(y)}{\zeta(y)}\,\mathrm{d}y \\
=&\frac{\partial^2}{\partial x^2}\int_S \frac{\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}\geq 0)|\mathcal{C}=y)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=y)}h''(y)[x-y]_+ + \frac{\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}<0)|\mathcal{C}=y)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=y)}h''(y)[y-x]_+\,\mathrm{d}y \\
=&\int_S \frac{\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}\geq 0)|\mathcal{C}=y)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=y)}h''(y)\delta(x-y) + \frac{\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}<0)|\mathcal{C}=y)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=y)}h''(y)\delta(y-x)\,\mathrm{d}y \\
=&\frac{\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}\geq 0)|\mathcal{C}=x)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=x)}h''(x) + \frac{\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}<0)|\mathcal{C}=x)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=x)}h''(x) \\
=&h''(x).
\end{aligned}
$$

From the definition of the space G, we see that the second derivative uniquely determines the element in G. Since $h\in G$, in order to show that $\int_S \frac{h''(y)\tilde{\Theta}''_{x,G}(y)}{\zeta(y)}\,\mathrm{d}y = h(x)$, we only need to show $\int_S \frac{h''(y)\tilde{\Theta}''_{x,G}(y)}{\zeta(y)}\,\mathrm{d}y \in G$, i.e., $\langle\int_S \frac{h''(y)\tilde{\Theta}''_{x,G}(y)}{\zeta(y)}\,\mathrm{d}y, 1\rangle_{\tilde{\Theta}} = 0$ and $\langle\int_S \frac{h''(y)\tilde{\Theta}''_{x,G}(y)}{\zeta(y)}\,\mathrm{d}y, x\rangle_{\tilde{\Theta}} = 0$.

Then we get

$$
\begin{aligned}
\langle \int_S \frac{h''(y)\tilde{\Theta}''_{x,G}(y)}{\zeta(y)}\ \mathrm{d}y, 1\rangle_{\tilde{\Theta}} =& \langle \int_S \frac{h''(y)\frac{\partial^2}{\partial y^2}\tilde{\Theta}(x,y)}{\zeta(y)}\ \mathrm{d}y, 1\rangle_{\tilde{\Theta}} \\
=& \langle \int_S \frac{h''(y)\lim_{h\to 0}\frac{\tilde{\Theta}(x,y+h)-2\tilde{\Theta}(x,y)+\tilde{\Theta}(x,y-h)}{h^2}}{\zeta(y)}\ \mathrm{d}y, 1\rangle_{\tilde{\Theta}} \\
=& \lim_{h\to 0}\langle \int_S \frac{h''(y)\frac{\tilde{\Theta}(x,y+h)-2\tilde{\Theta}(x,y)+\tilde{\Theta}(x,y-h)}{h^2}}{\zeta(y)}\ \mathrm{d}y, 1\rangle_{\tilde{\Theta}} \\
=& \lim_{h\to 0}\int_S \frac{h''(y)\frac{\langle\tilde{\Theta}(x,y+h),1\rangle_{\tilde{\Theta}}-2\langle\tilde{\Theta}(x,y),1\rangle_{\tilde{\Theta}}+\langle\tilde{\Theta}(x,y-h),1\rangle_{\tilde{\Theta}}}{h^2}}{\zeta(y)}\ \mathrm{d}y \\
=& \lim_{h\to 0}\int_S \frac{h''(y)\frac{y+h-2y+y-h}{h^2}}{\zeta(y)}\ \mathrm{d}y \\
=& 0.
\end{aligned}
$$

Similarly we can show that $\langle \int_S \frac{h''(y)\tilde{\Theta}''_{x,G}(y)}{\zeta(y)}\ \mathrm{d}y, x\rangle_{\tilde{\Theta}} = 0$. This concludes the proof. □

2.N Gradient Descent Trajectory and Trajectory of Smoothing Splines for Univariate Regression

In the following we discuss the relation between the trajectory of functions obtained by gradient descent training of a neural network and a trajectory of solutions to the variational problem with the data fitting constraints replaced by a MSE for decreasing smoothness regularization strength. This Lagrange version of the variational problem is solved by so-called smoothing splines. Smoothing splines have been studied intensively in the literature and in particular they can be written explicitly. We give the explicit form of the solution for the trajectory in the context of our discussion.

2.N.1 Regularized Regression and Early Stopping

[Bis95] shows that for linear regression with quadratic loss, early stopping and L_2 regularization lead to similar solutions. Let us recall some details of his analysis, before proceeding

with our particular setting. He considers the loss function $E(\mathbf{w}) = \|X\mathbf{w} - \mathbf{y}\|_2^2$, where $X = [\mathbf{x}_1, \ldots, \mathbf{x}_M]^T$ is the matrix of training inputs, $\mathbf{y} = [y_1, \ldots, y_M]^T$ is the vector of training outputs, and $\mathbf{w}$ is the weight vector of the linear model. Next the loss function can be written in the form of a quadratic function:

$$\begin{aligned} E(W) &= \|X\mathbf{w} - \mathbf{y}\|_2^2 \\ &= \mathbf{w}^T X^T X\mathbf{w} - 2\mathbf{y}^T X\mathbf{w} + \mathbf{y}^T\mathbf{y} \\ &= \mathbf{w}^T X^T X\mathbf{w} - 2\mathbf{y}^T X\mathbf{w} + \mathbf{y}^T\mathbf{y} \\ &= \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T H(\mathbf{w} - \mathbf{w}^*) + E_0, \end{aligned}$$

where $H = 2X^T X$, E_0 is the minimum of the loss function, and $\mathbf{w}^*$ is the minimizer. The eigenvalues and eigenvectors of H are as follows:

$$H\mathbf{u}_j = \lambda_j \mathbf{u}_j.$$

Then expand $\mathbf{w}$ and $\mathbf{w}^*$ in terms of the eigenvectors of H:

$$\mathbf{w} = \sum_j w_j \mathbf{u}_j, \qquad \mathbf{w}^* = \sum_j w_j^* \mathbf{u}_j.$$

For the L_2 regularized regression problem, consider the regularized loss function $\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + c\|\mathbf{w}\|_2^2$. Denote the minimizer by $\mathbf{w} = \tilde{\mathbf{w}}$ and consider its expansion as $\tilde{\mathbf{w}} = \sum_j \tilde{w}_j \mathbf{u}_j$. [Bis95] shows that

$$\tilde{w}_j = \frac{\lambda_j}{\lambda_j + c} w_j^*. \tag{2.174}$$

For early stopping, consider the gradient descent on $E(\mathbf{w})$ with zero initial weight vector:

$$\begin{aligned} \mathbf{w}^{(\tau)} &= \mathbf{w}^{(\tau-1)} - \eta \nabla E \\ &= \mathbf{w}^{(\tau-1)} - \eta H(\mathbf{w}^{(\tau-1)} - \mathbf{w}^*), \\ \mathbf{w}^{(0)} &= \mathbf{0}. \end{aligned}$$

Writing $\mathbf{w}^{(\tau)} = \sum_j w_j^{(\tau)} \mathbf{u}_j$, we have

$$w_j^{(\tau)} = (1 - (1 - \eta\lambda_j)^\tau) w_j^*.$$

Note that $1 - (1 - \eta\lambda_j)^\tau \to 1 - e^{-\eta\tau\lambda_j}$ as $\eta \to 0$. Hence choosing a sufficiently small learning rate, approximately we have

$$w_j^{(\tau)} = (1 - e^{-\eta\tau\lambda_j}) w_j^*. \tag{2.175}$$

From (2.174) and (2.175), [Bis95] observes that if c is much larger than λ_j, then the regularized solution has coordinate $\tilde{w}_j$ close to 0, and similarly if $1/(\eta\tau)$ is much larger than λ_j, then the early-stopping solution has coordinate $w_j^{(\tau)}$ close to the initial value 0. We note that analogous observations apply when the regularization term has a reference point different from zero, $c\|\mathbf{w} - \overline{\mathbf{w}}\|_2^2$, and the gradient descent iteration is initialized at a point different from zero, $\mathbf{w}^{(0)} = \overline{\mathbf{w}}$.

Now we want to take a closer look at the trajectories. Consider the following two functions:

$$h_1(x) = \frac{\lambda_j}{\lambda_j + x}, \qquad h_2(x) = 1 - e^{-\lambda_j/x}.$$

Actually we can verify that $h_1(0) = h_2(0) = 1$ and $\lim_{x\to\infty} \frac{h_1(x)}{h_2(x)} = 1$. It implies that these two functions are close to each other on $[0, \infty)$. Figure 2.13 shows the plot of functions $h_1(x)$ and $h_2(x)$.

Now we choose the coefficient of regularization $c = \frac{1}{\eta\tau}$. Comparing (2.174) and (2.175), and using the fact that $h_1(x)$ and $h_2(x)$ are close to each other on $[0, \infty)$, we show that early stopping and L_2 regularization lead to similar solutions across different values of $c = \frac{1}{\eta\tau}$.

Back to our problem, we repeat the gradient descent procedures (2.17) here:

$$\widetilde{W}_0^{(2)} = \overline{W}^{(2)}, \quad \widetilde{W}_{t+1}^{(2)} = \widetilde{W}_t^{(2)} - \eta \nabla_{W^{(2)}} L^{\text{lin}}(\widetilde{\omega}_t).$$

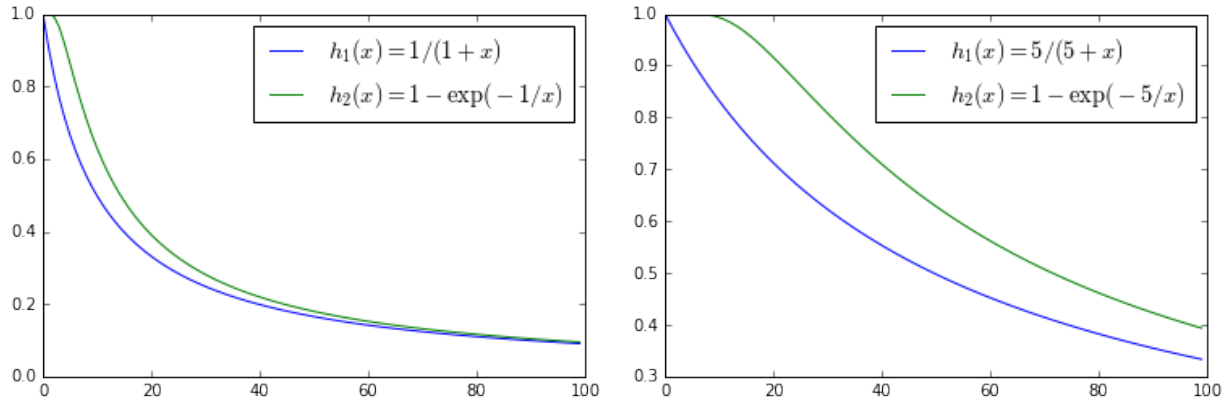

Figure 2.13: Plot of functions $h_1(x)$ and $h_2(x)$. The left panel plots the two function when $\lambda_j = 1$. The right panel plots the two function when $\lambda_j = 5$.

It is actually minimizing the following loss function of $W^{(2)} - \overline{W}$:

$$E(W^{(2)} - \overline{W}) = \sum_{j=1}^{M} \left(\sum_{i=1}^{n} (W_i^{(2)} - \overline{W}_i^{(2)})[W_i^{(1)} x_j + b_i]_+ - (y_j - f(x_j, \theta_0)) \right)^2.$$

Here we change the variable from $W^{(2)}$ to $W^{(2)} - \overline{W}$. Then $W_t^{(2)} - \overline{W} = 0$ when $t = 0$, so that gradient descent starts from the zero initial weight vector. Since the above model is linear with respect to $W^{(2)} - \overline{W}$, we can apply the above argument about early stopping and L_2 regularization. Suppose that we use learning rate μ_n for the neural network of width n. We show that the solution $\widetilde{W}_t^{(2)}$ at iteration t is close to the minimizer of the following regularized optimization problem:

$$\min_{W^{(2)}} \sum_{j=1}^{M} \left(\sum_{i=1}^{n} (W_i^{(2)} - \overline{W}_i^{(2)})[W_i^{(1)} x_j + b_i]_+ - (y_j - f(x_j, \theta_0)) \right)^2 + c\|W^{(2)} - \overline{W}\|_2^2, \quad (2.176)$$

where $c = \frac{1}{\eta_n t}$. Using the same approach and notation as in Section 2.5, the optimization problem (2.176) is equivalent to

$$\begin{aligned} \min_{\alpha_n \in C(\mathbb{R}^2)} \quad & \sum_{j=1}^{M} \left(\int_{\mathbb{R}^2} \alpha_n(W^{(1)}, b)[W^{(1)} x_j + b]_+ \, \mathrm{d}\mu_n(W^{(1)}, b) - y_j \right)^2 \\ & + \frac{1}{n\eta_n t} \int_{\mathbb{R}^2} \alpha_n^2(W^{(1)}, b) \, \mathrm{d}\mu_n(W^{(1)}, b), \end{aligned} \quad (2.177)$$

where we use the ASI trick (see Appendix 2.B.2). Here (2.177) has an extra factor $\frac{1}{n}$ compared to (2.176). This is because we define $\alpha_n(W_i^{(1)}, b_i) = n(W_i^{(2)} - \overline{W}_i^{(2)})$. According to Theorem 20, $\eta_n \leq \frac{M}{Kn\lambda_{\max}(\hat{\Theta}_n)}$ is sufficient in order to ensure convergence. Then we suppose that $\eta_n = \bar{\eta}/n$, where $\bar{\eta}$ is a constant so that the requirement on the learning rate in Theorem 20 is satisfied. The limit of the optimization problem (2.177) as the width n tends to infinity is:

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^2)} \quad & \sum_{j=1}^{M} \left(\int_{\mathbb{R}^2} \alpha(W^{(1)}, b)[W^{(1)} x_j + b]_+ \, \mathrm{d}\mu(W^{(1)}, b) - y_j \right)^2 \\ & + \frac{1}{\bar{\eta} t} \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b) \, \mathrm{d}\mu(W^{(1)}, b). \end{aligned} \tag{2.178}$$

Following the same reasoning of Section 2.6, we relax the optimization problem (2.178) to the following one:

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^2), u \in \mathbb{R}, v \in \mathbb{R}} \quad & \sum_{j=1}^{M} \left(u x_j + v + \int_{\mathbb{R}^2} \alpha(W^{(1)}, b)[W^{(1)} x_j + b]_+ \, \mathrm{d}\mu(W^{(1)}, b) - y_j \right)^2 \\ & + \frac{1}{\bar{\eta} t} \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b) \, \mathrm{d}\mu(W^{(1)}, b). \end{aligned} \tag{2.179}$$

Using the same technique and notation as in Theorem 13, we can prove that the solution of (2.179) actually solves the following optimization problem:

$$\min_{h \in C^2(S)} \sum_{j=1}^{M} [h(x_j) - y_j]^2 + \frac{1}{\bar{\eta} t} \int_S \frac{(h''(x))^2}{\zeta(x)} \, \mathrm{d}x. \tag{2.180}$$

Then in order to study the trajectory of gradient descent, we can study the optimization problem (2.180) with varying t. Figure 2.14 illustrates smoothing spline and gradient descent trajectories. The solution of (2.180) is called spatially adaptive smoothing spline. Here the curvature penalty function is $\frac{1}{\bar{\eta} t} \frac{1}{\zeta(x)}$, with time dependent smoothness regularization coefficient $\frac{1}{\bar{\eta} t}$. Next, we give out the solution of (2.180) in the following two cases: (1) uniform case (ζ is constant over domain S); (2) spatially adaptive case (ζ is not constant over domain S).

Remark 45 (Spectral bias)**.** *We have thus that the gradient descent optimization trajectory*

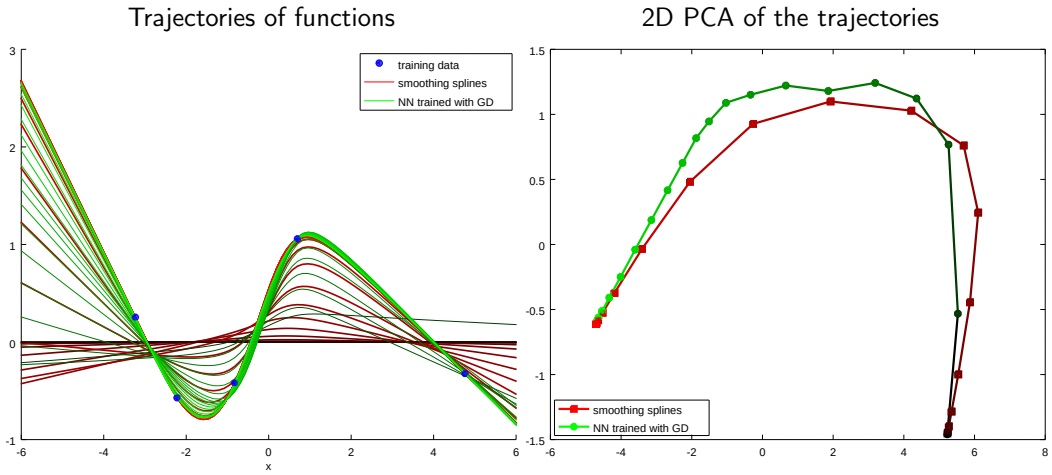

Figure 2.14: Trajectories of functions obtained by gradient descent training a neural network and by smoothing splines of the training data with decreasing regularization strength (from dark to bright). The left panel plots 20 functions along each trajectory. The right panel shows the same functions in a two dimensional PCA representation. With asymmetric initialization of the network parameters and adjusting the training data by ordinary linear regression, both trajectories start at the zero function. The trajectories are not equivalent, but are close, and both converge to the same (spatially adaptive) cubic spline interpolation of the training data (in the limit of infinite wide networks). Here we used a large network with $n = 2000$ hidden units and Gaussian initialization $\mathcal{W} \sim \mathcal{N}(0, 1)$, $\mathcal{B} \sim \mathcal{N}(0, 1)$. The results are similar for smaller networks and different initializations.

can be described approximately by a trajectory of smoothing splines which gradually relaxes the smoothness regularization (relative to initialization) until perfectly fitting the training data. If the function at initialization is at the zero function, e.g., by ASI, then the regularization is on the function itself. Hence the result provides a theoretical explanation for the spectral bias phenomenon that has been observed by [RBA19b]. The spectral bias is that lower frequencies are learned first.

2.N.2 Trajectory of Smoothing Splines with Uniform Curvature Penalty

Suppose the reciprocal curvature penalty is constant $\zeta(x) \equiv z$ on the domain S. Let $\lambda = \frac{1}{\bar{\eta} t z}$. Then (2.180) becomes the following optimization problem:

$$\min_{h \in C^2(S)} \sum_{j=1}^{M} [h(x_j) - y_j]^2 + \lambda \int_S (h''(x))^2 \, \mathrm{d}x. \tag{2.181}$$

[Ger01] gives the explicit form of the minimizer $\hat{h}$ of (2.181), which is called a smoothing spline. The minimizer $\hat{h}$ is a natural cubic spline with knots at the sample points $x_1, \ldots, x_M$. The smoothing spline does not fit the training data exactly, but rather it balances fitting and smoothness. The smoothing parameter $\lambda \geq 0$ controls the trade off between fitting and roughness. The values of the smoothing spline at the knots can be obtained as

$$(\hat{h}(x_1), \ldots, \hat{h}(x_M))^\top = (I + \lambda A)^{-1} Y. \tag{2.182}$$

The matrix A has entries $A_{ij} = \int_S h_i''(x) h_j''(x) \, \mathrm{d}x$, where h_i are spline basis functions which satisfy $h_i(x_j) = 0$ for $j \neq i$ and $h_i(x_j) = 1$ for $j = i$. [Ger01] gives out a rather explicit form of matrix A, which is an $M \times M$ matrix given by $A = \Delta^T W^{-1} \Delta$. Here Δ is an $(M-2) \times M$ matrix of second differences with elements:

$$\Delta_{ii} = \frac{1}{h_i}, \quad \Delta_{i,i+1} = -\frac{1}{h_i} - \frac{1}{h_{i+1}}, \quad \Delta_{i,i+2} = \frac{1}{h_{i+1}}.$$

And W is an $(M-2)\times(M-2)$ symmetric tri-diagonal matrix with elements:

$$W_{i-1,i}=W_{i,i-1}=\frac{h_i}{6},\quad W_{i,i}=\frac{h_i+h_{i+1}}{3},\ \text{here } h_i=x_{i+1}-x_i.$$

As $\lambda\to 0$, the smoothing spline converges to the interpolating spline, and as $\lambda\to\infty$, it converges to the linear least squares estimate.

2.N.3 Trajectory of Spatially Adaptive Smoothing Splines

Let the curvature penalty $\rho(x)=\frac{1}{\bar{\eta}t}\frac{1}{\zeta(x)}\frac{1}{M}$. Then (2.180) can be written as

$$\min_{h\in W_2(S)}\frac{1}{M}\sum_{i=1}^{M}[h(x_j)-y_j]^2+\int_S\rho(x)(h''(x))^2\,\mathrm{d}x, \tag{2.183}$$

where $W_2(S)=\{f\colon f,f' \text{ absolutely continuous and } f''\in L^2(S)\}$, with $L^2(S)$ the square integrable functions over the domain S. [AS96a, PSH06] give out the solution of (2.183) explicitly, which is called a spatially adaptive smoothing spline.

According to [PSH06], the solution can be derived in terms of an appropriate RKHS representation of W_2^0 with inner product $\langle f,g\rangle_\rho=\int f''(x)g''(x)\rho(x)\,\mathrm{d}x$. Here $W_0^2(S)=W_2(S)\cap B_2(S)$, where $W_2(S)$ is defined above, and $B_2(S)=\{f: f(0)=f'(0)=0\}$. Notice that when defining $B_2(S)$ we need $0\in S$. Actually we can choose any point in S. [PSH06] define $B_2(S)$ in this way just for simplicity. Then the kernel of the space $W_0^2(S)$ is given by

$$K_\rho(x_1,x_2)=\int_S\rho(u)^{-1}[x_1-u]_+[x_2-u]_+\mathrm{d}u. \tag{2.184}$$

Then the minimizer $\hat{h}$ of (2.183) is given by

$$\hat{h}(x)=\sum_{j=1}^{M}c_jK_\rho(x_j,x)+a+bx. \tag{2.185}$$

Now define the $M \times M$ matrix

$$\Sigma_\rho = \{K_\rho(x_i, x_j)\}_{i,j=1,\dots,M}, \tag{2.186}$$

and the $M \times 2$ matrix

$$T = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_M \end{bmatrix}. \tag{2.187}$$

Denote the vector of coefficients $\mathbf{c} = (c_1, \dots, c_M)^T$ and the vector of output values $\mathbf{y} = (y_1, \dots, y_M)^T$. Then the coefficients in (2.185) satisfy the following conditions:

$$\Sigma_\rho \left[(\Sigma_\rho + MI)\mathbf{c} + T \begin{pmatrix} a \\ b \end{pmatrix} \right] = \Sigma_\rho \mathbf{y} \quad \text{and} \quad T^\top \left[\Sigma_\rho \mathbf{c} + T \begin{pmatrix} a \\ b \end{pmatrix} \right] = T^\top \mathbf{y}. \tag{2.188}$$

After solving for (2.188), we get the values of $\mathbf{c}$, a and b. Plug them into (2.185), then we get the exact form of the minimizer of (2.183).

2.O Solution to the Variational Problems for Univariate Regression after Training

2.O.1 Interpolating Splines with Uniform Curvature Penalty

Theorem 2 (b) and (c) show that for certain distributions of $(\mathcal{W}, \mathcal{B})$, ζ is constant. In this case problem (2.5) with ASI is solved by the cubic spline interpolation of the data with natural boundary conditions [ANW67].

Theorem 46 ([ANW67]). *For training samples $\{(x_i, y_i)\}_{i=1}^M$, suppose $x_j \in S,\ j = 1, \dots, M$. Then cubic spline interpolation of data $\{(x_i, y_i)\}_{i=1}^M$ with natural boundary condition is the*

solution of

$$\min_{h\in C^2(S)} \quad \int_S (h''(x))^2 \mathrm{d}x$$
$$\text{subject to} \quad h(x_j) = y_j, \quad j = 1, \ldots, m.$$

As already mentioned in Appendix 2.N, cubic spline interpolation is a finite dimensional linear problem and can be solved exactly. A cubic spline is a piecewise polynomial of order 3 with $(M-1)$ pieces. The j-th piece has the form $S_j(x) = a_j + b_j x + c_j x^2 + d_j x^3$, $j = 1, \ldots, M-1$. These $(M-1)$ pieces satisfy equations $S_i(x_i) = y_i$, $S_i(x_{i+1}) = y_{i+1}$, $i = 1, \ldots, M-1$ and $S_i'(x_{i+1}) = S_{i+1}'(x_{i+1})$, $S_i''(x_{i+1}) = S_{i+1}''(x_{i+1})$, $i = 1, \ldots, M-2$, and $S_1''(x_1) = S_{M-1}''(x_M) = 0$. Hence computing the spline amounts to solving a linear system in $4(M-1)$ indeterminates.

2.O.2 Spatially Adaptive Interpolating Splines

In the case that ζ is not constant, we can still give out the form of the solution to the variational problem (2.5) with ASI by using the result in Appendix 2.N. We multiply by a coefficient λ the regularization term in the optimization problem (2.183) and choose $\rho(x) = \frac{1}{\zeta(x)}$. Then we get

$$\min_{h\in W_2(S)} \frac{1}{M}\sum_{i=1}^{M}[h(x_j) - y_j]^2 + \lambda \int_S \frac{1}{\zeta(x)}(h''(x))^2 \, \mathrm{d}x. \tag{2.189}$$

As $\lambda \to 0$, the minimizer of (2.189) converges to the solution of the following optimization problem:

$$\min_{h\in W^2(S)} \int_S \frac{(h''(x))^2}{\zeta(x)} \, \mathrm{d}x \quad \text{s.t.} \quad h(x_j) = y_j, \quad j = 1, \ldots, m,$$

which is the variational problem (2.5) with ASI. According to Appendix 2.N, the solution of (2.189) is given by:

$$\hat{h}^{(\lambda)}(x) = \sum_{j=1}^{M} c_j^{(\lambda)} K_{\frac{\lambda}{\zeta}}(x_j, x) + a^{(\lambda)} + b^{(\lambda)} x. \tag{2.190}$$

And the vector $\mathbf{c}^{(\lambda)} = (c_1^{(\lambda)}, \ldots, c_M^{(\lambda)})^T$, $a^{(\lambda)}$ and $b^{(\lambda)}$ satisfy the following conditions:

$$\Sigma_{\frac{\lambda}{\zeta}} \left[(\Sigma_{\frac{\lambda}{\zeta}} + MI)\mathbf{c}^{(\lambda)} + T \begin{pmatrix} a^{(\lambda)} \\ b^{(\lambda)} \end{pmatrix} \right] = \Sigma_{\frac{\lambda}{\zeta}} \mathbf{y} \quad \text{and} \quad T^\top \left[\Sigma_{\frac{\lambda}{\zeta}} \mathbf{c}^{(\lambda)} + T \begin{pmatrix} a^{(\lambda)} \\ b^{(\lambda)} \end{pmatrix} \right] = T^\top \mathbf{y}, \quad (2.191)$$

where $K_{\frac{\lambda}{\zeta}}$, $\Sigma_{\frac{\lambda}{\zeta}}$ and T are defined in (2.184), (2.186) and (2.187). Next we show that $K_{\frac{\lambda}{\zeta}}$ is inversely proportional to λ:

$$\begin{aligned} K_{\frac{\lambda}{\zeta}}(x_1, x_2) &= \int_S \left(\frac{\lambda}{\zeta} \right)^{-1} [x_1 - u]_+ [x_2 - u]_+ \mathrm{d}u \\ &= \lambda^{-1} \int_S \left(\frac{1}{\zeta} \right)^{-1} [x_1 - u]_+ [x_2 - u]_+ \mathrm{d}u \\ &= \lambda^{-1} K_{\frac{1}{\zeta}}(x_1, x_2). \end{aligned} \quad (2.192)$$

Also $\Sigma_{\frac{\lambda}{\zeta}} = \lambda^{-1} \Sigma_{\frac{1}{\zeta}}$. Then we let $\bar{c}_j^{(\lambda)} = \lambda^{-1} c_j^{(\lambda)}$ and $\bar{\mathbf{c}}^{(\lambda)} = \lambda^{-1} \mathbf{c}^{(\lambda)}$. So we can rewrite (2.190) and (2.191) as

$$\hat{h}^{(\lambda)}(x) = \sum_{j=1}^{M} \bar{c}_j^{(\lambda)} K_{\frac{1}{\zeta}}(x_j, x) + a^{(\lambda)} + b^{(\lambda)} x, \quad (2.193)$$

where $\bar{\mathbf{c}}^{(\lambda)}$, $a^{(\lambda)}$ and $b^{(\lambda)}$ satisfy the following conditions:

$$\Sigma_{\frac{1}{\zeta}} \left[(\Sigma_{\frac{1}{\zeta}} + \lambda MI)\bar{\mathbf{c}}^{(\lambda)} + T \begin{pmatrix} a^{(\lambda)} \\ b^{(\lambda)} \end{pmatrix} \right] = \Sigma_{\frac{1}{\zeta}} \mathbf{y} \quad \text{and} \quad T^\top \left[\Sigma_{\frac{1}{\zeta}} \bar{\mathbf{c}}^{(\lambda)} + T \begin{pmatrix} a^{(\lambda)} \\ b^{(\lambda)} \end{pmatrix} \right] = T^\top \mathbf{y}, \quad (2.194)$$

Now, as $\lambda \to 0$, (2.193) and (2.194) become:

$$\hat{h}^{(0^+)}(x) = \sum_{j=1}^{M} \bar{c}_j^{(0^+)} K_{\frac{1}{\zeta}}(x_j, x) + a^{(0^+)} + b^{(0^+)} x, \quad (2.195)$$

where $\bar{\mathbf{c}}^{(0^+)}$, $a^{(0^+)}$, and $b^{(0^+)}$ satisfy the following conditions:

$$\Sigma_{\frac{1}{\zeta}} \left[\Sigma_{\frac{1}{\zeta}} \bar{\mathbf{c}}^{(0^+)} + T \begin{pmatrix} a^{(0^+)} \\ b^{(0^+)} \end{pmatrix} \right] = \Sigma_{\frac{1}{\zeta}} \mathbf{y} \quad \text{and} \quad T^\top \left[\Sigma_{\frac{1}{\zeta}} \bar{\mathbf{c}}^{(\lambda)} + T \begin{pmatrix} a^{(0^+)} \\ b^{(0^+)} \end{pmatrix} \right] = T^\top \mathbf{y}. \quad (2.196)$$

The expressions (2.195) and (2.196) give the solution of (2.189) as $\lambda \to 0$, which is also the solution to the variational problem (2.24).

2.P Possible Generalizations

2.P.1 Deep Networks and Other Architectures

For deep networks with L layers, if we only train the output layer, then we actually train a linear model. We can actually write down the exact form of the NTK. However it is unclear whether we can write the explicit form of implicit bias in this case.

In the case of shallow networks, we show that training only the output layer is similar to training all parameters. Our analysis of shallow networks is based on this. However, in the case of a deep network, training only the output layer is no longer similar to training all parameters. If we train all model parameters, the results from [LXS19b] show that the model still is approximated by a linearized model. The result on kernel norm minimization [ZXL20] holds in this case. It will be interesting to study the explicit form of the kernel norm, and extensions of our analysis to the case of training all parameters of deep networks.

2.P.2 Other Loss Functions

We have focused on the implicit bias of gradient descent for regression. For this type of problems, one often considers a loss function (per example) which has a single finite minimum. Roughly speaking, our description of the bias is in terms of smoothness properties of the solution functions. There are various works on the implicit bias of gradient descent for classification problems, e.g., [SHN18]. In this case, the implicit bias is often formulated in terms of maximum margins.

In our analysis, some theorems require that the loss function is mean square error (MSE). In Theorem 10, the gradient flow is a linear differential equation if we use MSE. If we use a different loss, this will be more complicated. However, we think that the results can be

generalized. We are also using the result from [LSP18], which is based on MSE. According to them it is not clear whether their result will still apply for other loss functions. Theorems 12 and 13 are about a variational problem that is derived from Theorem 20, in relation to the minimization of $\|\theta - \theta_2\|_2$. Theorem 20 remains valid for other loss functions beside MSE. To sum up, if we can generalize the Theorem 10 and the result of [LSP18] to other loss functions, then we can generalize our main result in Theorem 1 to other loss functions as well.

2.P.3 Other Optimization Procedures

It would be interesting to extend the analysis to modifications of the basic gradient descent optimization procedure. The implicit bias of different optimization methods has been studied by [GLS18a] covering some instances of mirror descent, natural gradient descent, Adam, and steepest descent with respect to different potentials and norms. In particular, they show that the implicit bias of coordinate descent corresponds to the minimization of the 1-norm of the weights. It will be interesting to work out the explicit form of these descriptions in function space.

CHAPTER 3

Learning Curves for Gaussian Process Regression with Power-law Priors and Targets *

3.1 Introduction

Gaussian processes (GPs) provide a flexible and interpretable framework for learning and adaptive inference, and are widely used for constructing prior distributions in non-parametric Bayesian learning. From an application perspective, one crucial question is how fast do GPs learn, i.e., how much training data is needed to achieve a certain level of generalization performance. Theoretically, this is addressed by analyzing so-called "learning curves", which describe the generalization error as a function of the training set size n. The rate at which the curve approaches zero determines the difficulty of learning tasks and conveys important information about the asymptotic performance of GP learning algorithms. In this chapter, we study the learning curves for Gaussian process regression. Our main result characterizes the asymptotics of the generalization error in cases where the eigenvalues of the GP kernel and the coefficients of the eigenexpansion of the target function have a power-law decay. In the remainder of this introductory section, we review related work and outline our main contributions.

Gaussian processes A GP model is a probabilistic model on an infinite-dimensional parameter space [WR06, OT10]. In GP regression (GPR), for example, this space can be the set of all continuous functions. Assumptions about the learning problem are encoded by way

*This chapter is adapted from [JBM22], with the permission from coauthors.

of a prior distribution over functions, which gets transformed into a posterior distribution given some observed data. The mean of the posterior is then used for prediction. The model uses only a finite subset of the available parameters to explain the data and this subset can grow arbitrarily large as more data are observed. In this sense, GPs are "non-parametric" and contrast with parametric models, where there is a fixed number of parameters. For regression with Gaussian noise, a major appeal of the GP formalism is that the posterior is analytically tractable. GPs are also one important part in learning with kernel machines [KHS18] and modeling using GPs has recently gained considerable traction in the neural network community.

Neural networks and kernel learning From a GP viewpoint, there exists a well known correspondence between kernel methods and infinite neural networks (NNs) first studied by [Nea96a]. Neal showed that the outputs of a randomly initialized one-hidden layer neural network (with appropriate scaling of the variance of the initialization distribution) converges to a GP over functions in the limit of an infinite number of hidden units. Follow-up work extended this correspondence with analytical expressions for the kernel covariance for shallow NNs by [Wil97], and more recently for deep fully-connected NNs [LSP18, GHR18], convolutional NNs with many channels [NXB19, GRA19], and more general architectures [Yan19]. The correspondence enables *exact* Bayesian inference in the associated GP model for infinite-width NNs on regression tasks and has led to some recent breakthroughs in our understanding of overparameterized NNs [JGH18b, LXS19a, ADH19b, BMM18, DFS16a, YS19a, BM19]. The most prominent kernels associated with infinite-width NNs are the Neural Network Gaussian Process (NNGP) kernel [LSP18, GHR18], and the Neural Tangent Kernel (NTK) [JGH18b]. Empirical studies have shown that inference with such infinite network kernels is competitive with standard gradient descent-based optimization for fully-connected architectures [LSP20].

Learning curves A large-scale empirical characterization of the generalization performance of state-of-the-art deep NNs showed that the associated learning curves often follow a power law of the form $n^{-\beta}$ with the exponent β ranging between 0.07 and 0.35 depending on the data

and the algorithm [HNA17, SGW20]. Power-law asymptotics of learning curves have been theoretically studied in early works for the Gibbs learning algorithm [AFS92, AM93, HKS96] that showed a generalization error scaling with exponent $\beta = 0.5$, 1 or 2 under certain assumptions. More recent results from statistical learning theory characterize the shape of learning curves depending on the properties of the hypothesis class [BHM21]. In the context of GPs, approximations and bounds on learning curves have been investigated in several works [Sol99, SH02, Sol01, OV99, OM02, WV00, MO01b, MO01a, SKF08, VV11, LG15], with recent extensions to kernel regression from a spectral bias perspective [BCP20, CBP21]. For a review on learning curves in relation to its shape and monotonicity, see [LVM19, VML19, VL21]. A related but complementary line of work studies the convergence rates and posterior consistency properties of Bayesian non-parametric models [Bar98, SKF08, VV11].

Power-law decay of the GP kernel eigenspectrum The rate of decay of the eigenvalues of the GP kernel conveys important information about its smoothness. Intuitively, if a process is "rough" with more power at high frequencies, then the eigenspectrum decays more slowly. On the other hand, kernels that define smooth processes have a fast-decaying eigenspectrum [Ste12, WR06]. The precise eigenvalues $(\lambda_p)_{p\geq 1}$ of the operators associated to many kernels and input distributions are not known explicitly, except for a few special cases [WR06]. Often, however, the asymptotic properties are known. The asymptotic rate of decay of the eigenvalues of stationary kernels for input distributions with bounded support is well understood [Wid63, RWW95]. [RJK19] showed that for inputs distributed uniformly on a hypersphere, the eigenfunctions of the arc-cosine kernel are spherical harmonics and the eigenvalues follow a power-law decay. The spectral properties of the NTK are integral to the analysis of training convergence and generalization of NNs, and several recent works empirically justify and rely on a power law assumption for the NTK spectrum [BDK21, CBP21, LSP20, NS21]. [VY21b] showed that the asymptotics of the NTK of infinitely wide shallow ReLU networks follows a power-law that is determined primarily by the singularities of the kernel and has the form $\lambda_p \propto p^{-\alpha}$ with $\alpha = 1 + \frac{1}{d}$, where d is the input dimension.

Asymptotics of the generalization error of kernel ridge regression (KRR) There is a well known equivalence between GPR and KRR with the additive noise in GPR playing the role of regularization in KRR [KHS18]. Analysis of the decay rates of the excess generalization error of KRR has appeared in several works, e.g, in the noiseless case with constant regularization [BCP20, SGW20, JCO19], and the noisy optimally regularized case [CD07, SHS09, FS20] under the assumption that the kernel eigenspectrum, and the eigenexpansion coefficients of the target function follow a power law. These assumptions, which are often called resp. the *capacity* and *source* conditions are related to the effective dimension of the problem and the difficulty of learning the target function [CD07, BM18]. [CLK21] present a unifying picture of the excess error decay rates under the capacity and source conditions in terms of the interplay between noise and regularization illustrating their results with real datasets.

Contributions In this chapter, we characterize the asymptotics of the generalization error of GPR and KRR under the capacity and source conditions. Our main contributions are as follows:

- When the eigenspectrum of the prior decays with rate α and the eigenexpansion coefficients of the target function decay with rate β, we show that with high probability over the draw of n input samples, the negative log-marginal likelihood behaves as $\Theta(n^{\max\{\frac{1}{\alpha}, \frac{1-2\beta}{\alpha}+1\}})$ (Theorem 53) and the generalization error behaves as $\Theta(n^{\max\{\frac{1}{\alpha}-1, \frac{1-2\beta}{\alpha}\}})$ (Theorem 55). In the special case that the model is correctly specified, i.e., the GP prior is the true one from which the target functions are actually generated, our result implies that the generalization error behaves as $O(n^{\frac{1}{\alpha}-1})$ recovering as a special case a result due to [SH02] (vide Remark 56).

- Under similar assumptions as in the previous item, we leverage the equivalence between GPR and KRR to show that the excess generalization error of KRR behaves as $\Theta(n^{\max\{\frac{1}{\alpha}-1, \frac{1-2\beta}{\alpha}\}})$ (Theorem 58). In the noiseless case with constant regularization, our result implies that the generalization error behaves as $\Theta(n^{\frac{1-2\beta}{\alpha}})$ recovering as a special case a result due to

[BCP20]. Specializing to the case of KRR with Gaussian design, we recover as a special case a result due to [CLK21] (vide Remark 60).

For the unrealizable case, i.e., when the target function is outside the span of the eigenfunctions with positive eigenvalues, we show that the generalization error converges to a constant.

- We present a few toy experiments demonstrating the theory for GPR with arc-cosine kernel without biases (resp. with biases) which is the conjugate kernel of an infinitely wide shallow network with two inputs and one hidden layer without biases (resp. with biases) [CS09, RJK19].

3.2 Bayesian Learning and Generalization Error for GPs

In GP regression, our goal is to learn a target function $f : \Omega \mapsto \mathbb{R}$ between an input $x \in \Omega$ and output $y \in \mathbb{R}$ based on training samples $D_n = \{(x_i, y_i)\}_{i=1}^n$. We consider an additive noise model $y_i = f(x_i) + \epsilon_i$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_{\text{true}})$. If ρ denotes the marginal density of the inputs x_i, then the pairs (x_i, y_i) are generated according to the density $q(x, y) = \rho(x)q(y|x)$, where $q(y|x) = \mathcal{N}(y|f(x), \sigma^2_{\text{true}})$. We assume that there is a prior distribution Π_0 on f which is defined as a zero-mean GP with continuous and bounded covariance function $k : \Omega \times \Omega \to \mathbb{R}$, i.e., $f \sim \mathcal{GP}(0, k)$. This means that for any finite set $\mathbf{x} = (x_1, \ldots, x_n)^T$, the random vector $f(\mathbf{x}) = (f(x_1), \ldots, f(x_n))^T$ follows the multivariate normal distribution $\mathcal{N}(0, K_n)$ with covariance matrix $K_n = (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$. By Bayes' rule, the posterior distribution over f given the training data is given by

$$d\Pi_n(f|D_n) = \frac{1}{Z(D_n)} \prod_{i=1}^n \mathcal{N}(y_i|f(x_i), \sigma^2_{\text{model}}) d\Pi_0(f),$$

where Π_0 is the prior distribution, $Z(D_n) = \int \prod_{i=1}^n \mathcal{N}(y_i|f(x_i), \sigma^2_{\text{model}}) d\Pi_0(f)$ is the *marginal likelihood* or *model evidence* and σ_{model} is the sample variance used in GPR. In practice, we do not know the exact value of σ_{true} and so our choice of σ_{model} can be different from σ_{true}.

The GP prior and the Gaussian noise assumption allows for exact Bayesian inference and the posterior distribution over functions is again a GP with mean and covariance function given by

$$\bar{m}(x) = K_{\mathbf{x}x}^T (K_n + \sigma_{\text{model}}^2 I_n)^{-1} \mathbf{y}, \ x \in \Omega \tag{3.1}$$

$$\bar{k}(x, x') = k(x, x') - K_{\mathbf{x}x}^T (K_n + \sigma_{\text{model}}^2 I_n)^{-1} K_{\mathbf{x}x'}, \ x, x' \in \Omega, \tag{3.2}$$

where $K_{\mathbf{x}x} = (k(x_1, x), \ldots, k(x_n, x))^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ [WR06, Eqs. 2.23-24].

The performance of GPR depends on how well the posterior approximates f as the number of training samples n tends to infinity. The distance of the posterior to the ground truth can be measured in various ways. We consider two such measures, namely the Bayesian generalization error [SKF08, HO97, OV99] and the excess mean squared error [SH02, LG15, BCP20, CLK21].

Definition 47 (Bayesian generalization error)**.** *The Bayesian generalization error is defined as the Kullback-Leibler divergence between the true density $q(y|x)$ and the Bayesian predictive density $p_n(y|x, D_n) = \int \mathcal{N}(y|f(x), \sigma_{\text{model}}^2) d\Pi_n(f|D_n)$,*

$$G(D_n) = \int q(x, y) \log \frac{q(y|x)}{p_n(y|x, D_n)} dxdy. \tag{3.3}$$

A related quantity of interest is the *stochastic complexity* (SC), also known as the *free energy*, which is just the negative log-marginal likelihood. We shall primarily be concerned with a normalized version of the stochastic complexity which is defined as follows:

$$F^0(D_n) = -\log \frac{Z(D_n)}{\prod_{i=1}^n q(y_i|x_i)} = -\log \frac{\int \prod_{i=1}^n \mathcal{N}(y_i|f(x_i), \sigma_{\text{model}}^2) d\Pi_0(f)}{\prod_{i=1}^n q(y_i|x_i)}. \tag{3.4}$$

The generalization error (3.3) can be expressed in terms of the normalized SC as follows [Wat09, Theorem 1.2]:

$$G(D_n) = \mathbb{E}_{(x_{n+1}, y_{n+1})} F^0(D_{n+1}) - F^0(D_n), \tag{3.5}$$

where $D_{n+1} = D_n \cup \{(x_{n+1}, y_{n+1})\}$ is obtained by augmenting D_n with a test point (x_{n+1}, y_{n+1}).

If we only wish to measure the performance of the mean of the Bayesian posterior, then we can use the excess mean squared error:

Definition 48 (Excess mean squared error)**.** *The excess mean squared error is defined as*

$$M(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(\bar{m}(x_{n+1}) - y_{n+1})^2 - \sigma^2_{\text{true}} = \mathbb{E}_{x_{n+1}}(\bar{m}(x_{n+1}) - f(x_{n+1}))^2. \tag{3.6}$$

Proposition 49 (Normalized stochastic complexity for GPR)**.** *Assume that* $\sigma^2_{\text{model}} = \sigma^2_{\text{true}} = \sigma^2$*. The normalized SC* $F^0(D_n)$ (3.4) *for GPR with prior* $\mathcal{GP}(0, k)$ *is given as*

$$F^0(D_n) = \tfrac{1}{2}\log\det(I_n + \tfrac{K_n}{\sigma^2}) + \tfrac{1}{2\sigma^2}\mathbf{y}^T(I_n + \tfrac{K_n}{\sigma^2})^{-1}\mathbf{y} - \tfrac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T(\mathbf{y} - f(\mathbf{x})), \tag{3.7}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$*. The expectation of the normalized SC w.r.t. the noise* $\boldsymbol{\epsilon}$ *is given as*

$$\mathbb{E}_{\boldsymbol{\epsilon}}F^0(D_n) = \tfrac{1}{2}\log\det\big(I_n + \tfrac{K_n}{\sigma^2}\big) - \tfrac{1}{2}\mathrm{Tr}\Big(I_n - \big(I_n + \tfrac{K_n}{\sigma^2}\big)^{-1}\Big) + \tfrac{1}{2\sigma^2}f(\mathbf{x})^T\big(I_n + \tfrac{K_n}{\sigma^2}\big)^{-1}f(\mathbf{x}). \tag{3.8}$$

This is a basic result and has applications in relation to model selection in GPR [WR06]. For completeness, we give a proof of Proposition 49 in Appendix 3.B. [SKF08, Theorem 1] gave an upper bound on the normalized stochastic complexity for the case when f lies in the reproducing kernel Hilbert space (RKHS) of the GP prior. It is well known, however, that sample paths of GP almost surely fall outside the corresponding RKHS [VV11] limiting the applicability of the result.

We next derive the asymptotics of $\mathbb{E}_{\boldsymbol{\epsilon}}F^0(D_n)$, the expected generalization error $\mathbb{E}_{\boldsymbol{\epsilon}}G(D_n) = \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{(x_{n+1},y_{n+1})}F^0(D_n + 1) - \mathbb{E}_{\boldsymbol{\epsilon}}F^0(D_n)$, and the excess mean squared error $\mathbb{E}_{\boldsymbol{\epsilon}}M(D_n)$.

3.3 Asymptotic Analysis of GP Regression with Power-law Priors

3.3.1 Notations and Assumptions

We assume that $f \in L^2(\Omega, \rho)$. By the generalization of Mercer's theorem [SS12, Corollary 3.2], the covariance function of the GP prior can be decomposed as $k(x_1, x_2) = \sum_{p=1}^{\infty} \lambda_p \phi_p(x_1)\phi_p(x_2)$ ρ-almost surely, where $(\phi_p(x))_{p\geq 1}$ are the eigenfunctions of the operator $L_k \colon L^2(\Omega, \rho) \mapsto L^2(\Omega, \rho)$; $(L_k f)(x) = \int_\Omega k(x, s) f(s) \mathrm{d}\rho(s)$, and $(\lambda_p)_{p\geq 1}$ are the corresponding positive eigenvalues. We index the sequence of eigenvalues in decreasing order, that is $\lambda_1 \geq \lambda_2 \geq \cdots > 0$. The target function $f(x)$ is decomposed into the orthonormal set $(\phi_p(x))_{p\geq 1}$ and its orthogonal complement $\{\phi_p(x) : p \geq 1\}^\perp$ as

$$f(x) = \sum_{p=1}^{\infty} \mu_p \phi_p(x) + \mu_0 \phi_0(x) \in L^2(\Omega, \rho), \tag{3.9}$$

where $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_p, \ldots)^T$ are the coefficients of the decomposition, and $\phi_0(x)$ satisfies $\|\phi_0(x)\|_2 = 1$ and $\phi_0(x) \in \{\phi_p(x) : p \geq 1\}^\perp$. For given sample inputs $\mathbf{x}$, let $\phi_p(\mathbf{x}) = (\phi_p(x_1), \ldots, \phi_p(x_n))^T$, $\Phi = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_p(\mathbf{x}), \ldots)$ and $\Lambda = \mathrm{diag}\{0, \lambda_1, \ldots, \lambda_p, \ldots\}$. Then the covariance matrix K_n can be written as $K_n = \Phi\Lambda\Phi^T$, and the function values on the sample inputs can be written as $f(\mathbf{x}) = \Phi\boldsymbol{\mu}$.

We shall make the following assumptions in order to derive the power-law asymptotics of the normalized stochastic complexity and the generalization error of GPR:

Assumption 50 (Power law decay of eigenvalues)**.** *The eigenvalues $(\lambda_p)_{p\geq 1}$ follow the power law*

$$\underline{C_\lambda} p^{-\alpha} \leq \lambda_p \leq \overline{C_\lambda} p^{-\alpha}, \ \forall p \geq 1 \tag{3.10}$$

where $\underline{C_\lambda}$, $\overline{C_\lambda}$ and α are three positive constants which satisfy $0 < \underline{C_\lambda} \leq \overline{C_\lambda}$ and $\alpha > 1$.

As mentioned in the introduction, this assumption, called the capacity condition, is fairly standard in kernel learning and is adopted in many recent works [BCP20, CBP21, JCO19, BVB21, CLK21]. [VY21b] derived the exact value of the exponent α when the kernel function

has a homogeneous singularity on its diagonal, which is the case for instance for the arc-cosine kernel.

Assumption 51 (Power law decay of coefficients of decomposition)**.** *Let $C_\mu, \underline{C_\mu} > 0$ and $\beta > 1/2$ be positive constants and let $\{p_i\}_{i\geq 1}$ be an increasing integer sequence such that $\sup_{i\geq 1}(p_{i+1} - p_i) < \infty$. The coefficients $(\mu_p)_{p\geq 1}$ of the decomposition* (3.9) *of the target function follow the power law*

$$|\mu_p| \leq C_\mu p^{-\beta}, \ \forall p \geq 1 \quad \textit{and} \quad |\mu_{p_i}| \geq \underline{C_\mu} {p_i}^{-\beta}, \ \forall i \geq 1. \tag{3.11}$$

Since $f \in L^2(\Omega, \rho)$, we have $\sum_{p=0}^{\infty} \mu_p^2 < \infty$. The condition $\beta > 1/2$ in Assumption 51 ensures that the sum $\sum_{p=0}^{\infty} \mu_p^2$ does not diverge. When the orthonormal basis $(\phi_p(x))_p$ is the Fourier basis or the spherical harmonics basis, the coefficients $(\mu_p)_p$ decay at least as fast as a power law so long as the target function $f(x)$ satisfies certain smoothness conditions [BM19]. [VY21b] gave examples of some natural classes of functions for which Assumption 51 is satisfied, such as functions that have a bounded support with smooth boundary and are smooth on the interior of this support, and derived the corresponding exponents β.

Assumption 52 (Boundedness of eigenfunctions)**.** *The eigenfunctions $(\phi_p(x))_{p\geq 0}$ satisfy*

$$\|\phi_0\|_\infty \leq C_\phi \quad \textit{and} \quad \|\phi_p\|_\infty \leq C_\phi p^\tau, \ p \geq 1, \tag{3.12}$$

where C_ϕ and τ are two positive constants which satisfy $\tau < \frac{\alpha-1}{2}$.

The second condition in (3.12) appears, for example, in [Val18, Hypothesis H_1] and is less restrictive than the assumption of uniformly bounded eigenfunctions that has appeared in several other works in the GP literature, see, e.g., [Bra06, CPB19, VKP21].

Define

$$T_1(D_n) = \tfrac{1}{2}\log\det\Big(I_n + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2}\Big) - \tfrac{1}{2}\mathrm{Tr}\Big(I_n - \Big(I_n + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2}\Big)^{-1}\Big), \tag{3.13}$$

$$T_2(D_n) = \tfrac{1}{2\sigma^2} f(\mathbf{x})^T \Big(I_n + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2}\Big)^{-1} f(\mathbf{x}), \tag{3.14}$$

$$G_1(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(T_1(D_{n+1}) - T_1(D_n)), \tag{3.15}$$

$$G_2(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(T_2(D_{n+1}) - T_2(D_n)). \tag{3.16}$$

Using (3.8) and (3.5), we have $\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n) = T_1(D_n) + T_2(D_n)$ and $\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n) = G_1(D_n) + G_2(D_n)$. Intuitively, G_1 corresponds to the effect of the noise on the generalization error irrespective of the target function f, whereas G_2 corresponds to the ability of the model to fit the target function. As we will see next in Theorems 55 and 57, if α is large, then the error associated with the noise is smaller. When f is contained in the span of the eigenfunctions $\{\phi_p\}_{p\geq 1}$, G_2 decreases with increasing n, but if f contains an orthogonal component, then the error remains constant and GP regression is not able to learn the target function.

3.3.2 Asymptotics of the Normalized Stochastic Complexity

We derive the asymptotics of the normalized SC (3.8) for the following two cases: $\mu_0 = 0$ and $\mu_0 > 0$. When $\mu_0 = 0$, the target function $f(x)$ lies in the span of all eigenfunctions with positive eigenvalues.

Theorem 53 (Asymptotics of the normalized SC, $\mu_0 = 0$)**.** *Assume that $\mu_0 = 0$ and $\sigma^2_{\text{model}} = \sigma^2_{\text{true}} = \sigma^2 = \Theta(1)$. Under Assumptions 50, 51 and 52, with probability of at least $1 - n^{-q}$ over sample inputs $(x_i)_{i=1}^n$, where $0 \leq q < \min\{\frac{(2\beta-1)(\alpha-1-2\tau)}{4\alpha^2}, \frac{\alpha-1-2\tau}{2\alpha}\}$, the expected normalized SC* (3.8) *has the asymptotic behavior:*

$$\begin{aligned}\mathbb{E}_\epsilon F^0(D_n) &= \big[\tfrac{1}{2}\log\det(I + \tfrac{n}{\sigma^2}\Lambda) - \tfrac{1}{2}\,\mathrm{Tr}\big(I - (I + \tfrac{n}{\sigma^2}\Lambda)^{-1}\big) + \tfrac{n}{2\sigma^2}\boldsymbol{\mu}^T(I + \tfrac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu}\big](1 + o(1)) \\ &= \Theta(n^{\max\{\frac{1}{\alpha}, \frac{1-2\beta}{\alpha}+1\}}).\end{aligned} \tag{3.17}$$

The complete proof of Theorem 53 is given in Appendix 3.D.1. We give a sketch of the proof below. In the sequel, we use the notations O and Θ to denote the standard mathematical orders and the notation $\tilde{O}$ to suppress logarithmic factors.

Proof sketch of Theorem 53. By (3.8), (3.13) and (3.14) we have $\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n) = T_1(D_n) + T_2(D_n)$. In order to analyze the terms $T_1(D_n)$ and $T_2(D_n)$, we will consider truncated versions of these quantities and bound the corresponding residual errors. Given a truncation parameter $R \in \mathbb{N}$, let $\Phi_R = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_R(\mathbf{x})) \in \mathbb{R}^{n \times R}$ be the truncated matrix of eigenfunctions evaluated at the data points, $\Lambda_R = \mathrm{diag}(0, \lambda_1, \ldots, \lambda_R) \in \mathbb{R}^{(R+1)\times(R+1)}$ and $\boldsymbol{\mu}_R = (\mu_0, \mu_1, \ldots, \mu_R) \in \mathbb{R}^{R+1}$. We define the truncated version of $T_1(D_n)$ as follows:

$$T_{1,R}(D_n) = \tfrac{1}{2} \log\det\Big(I_n + \tfrac{\Phi_R \Lambda_R \Phi_R^T}{\sigma^2}\Big) - \tfrac{1}{2}\mathrm{Tr}\Big(I_n - (I_n + \tfrac{\Phi_R \Lambda_R \Phi_R^T}{\sigma^2})^{-1}\Big). \tag{3.18}$$

Similarly, define $\Phi_{>R} = (\phi_{R+1}(\mathbf{x}), \phi_{R+2}(\mathbf{x}), \ldots, \phi_p(\mathbf{x}), \ldots)$, $\Lambda_{>R} = \mathrm{diag}(\lambda_{R+1}, \ldots, \lambda_p, \ldots)$, $f_R(x) = \sum_{p=1}^{R} \mu_p \phi_p(x)$, $f_R(\mathbf{x}) = (f_R(x_1), \ldots, f_R(x_n))^T$, $f_{>R}(x) = f(x) - f_R(x)$, and $f_{>R}(\mathbf{x}) = (f_{>R}(x_1), \ldots, f_{>R}(x_n))^T$. The truncated version of $T_2(D_n)$ is then defined as

$$T_{2,R}(D_n) = \tfrac{1}{2\sigma^2} f_R(\mathbf{x})^T (I_n + \tfrac{\Phi_R \Lambda_R \Phi_R^T}{\sigma^2})^{-1} f_R(\mathbf{x})^T. \tag{3.19}$$

The proof consists of three steps:

- **Approximation step:** In this step, we show that the asymptotics of $T_{1,R}$ resp. $T_{2,R}$ dominates that of the residuals, $|T_{1,R}(D_n) - T_1(D_n)|$ resp. $|T_{2,R}(D_n) - T_2(D_n)|$ (see Lemma 78). This builds upon first showing that $\|\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\|_2 = \tilde{O}(\max\{nR^{-\alpha}, n^{\frac{1}{2}} R^{\frac{1-2\alpha}{2}}, R^{1-\alpha}\})$ (see Lemma 71) and then choosing $R = n^{\frac{1}{\alpha}+\kappa}$ where $0 < \kappa < \frac{\alpha-1-2\tau}{2\alpha^2}$ when we have $\|\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\|_2 = o(1)$. Intuitively, the choice of the truncation parameter R is governed by the fact that $\lambda_R = \Theta(R^{-\alpha}) = n^{-1+\kappa\alpha} = o(n^{-1})$.

- **Decomposition step:** In this step, we decompose $T_{1,R}$ into a term independent of Φ_R and a series involving $\Phi_R^T\Phi_R - nI_R$, and likewise for $T_{2,R}$ (see Lemma 80). This builds

upon first showing using the Woodbury matrix identity [WR06, §A.3] that

$$T_{1,R}(D_n) = \tfrac{1}{2}\log\det(I_R + \tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R) - \tfrac{1}{2}\mathrm{Tr}\Phi_R(\sigma^2 I_R + \Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T, \tag{3.20}$$

$$T_{2,R}(D_n) = \tfrac{1}{2\sigma^2}\boldsymbol{\mu}_R^T\Phi_R^T\Phi_R(\sigma^2 I_R + \Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R, \tag{3.21}$$

and then Taylor expanding the matrix inverse $(\sigma^2 I_R + \Lambda_R\Phi_R^T\Phi_R)^{-1}$ in (3.20) and (3.21) to show that the Φ_R-independent terms in the decomposition of $T_{1,R}$ and $T_{2,R}$ are, respectively, $\frac{1}{2}\log\det(I_R + \frac{n}{\sigma^2}\Lambda_R) - \frac{1}{2}\mathrm{Tr}\big(I_R - (I_R + \frac{n}{\sigma^2}\Lambda_R)^{-1}\big)$, and $\frac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I_R + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R$.

- **Concentration step:** Finally, we use concentration inequalities to show that these Φ_R-independent terms dominate the series involving $\Phi_R^T\Phi_R - nI_R$ (see Lemma 81) when we have

$$T_{1,R}(D_n) = \big(\tfrac{1}{2}\log\det(I_R + \tfrac{n}{\sigma^2}\Lambda_R) - \tfrac{1}{2}\,\mathrm{Tr}\big(I_R - (I_R + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\big)\big)(1 + o(1)) = \Theta(n^{\frac{1}{\alpha}}),$$

$$T_{2,R}(D_n) = \big(\tfrac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I_R + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\big)(1 + o(1)) = \begin{cases} \Theta(n^{\max\{0, \frac{1-2\beta}{\alpha}+1\}}), & \alpha \neq 2\beta - 1, \\ \Theta(\log n), & \alpha = 2\beta - 1. \end{cases}$$

The key idea is to consider the matrix $\Lambda_R^{1/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Phi_R^T\Phi_R(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}$ and show that it concentrates around $n\Lambda_R(I + \frac{n}{\sigma^2})^{-1}$ (see Corollary 68). Note that an ordinary application of the matrix Bernstein inequality to $\Phi_R^T\Phi_R - nI_R$ yields $\|\Phi_R^T\Phi_R - nI\|_2 = O(R\sqrt{n})$, which is not sufficient for our purposes, since this would give $O(R\sqrt{n}) = o(n)$ only when $\alpha > 2$. In contrast, our results are valid for $\alpha > 1$ and cover cases of practical interest, e.g., the NTK of infinitely wide shallow ReLU network [VY21b] and the arc-cosine kernels over high-dimensional hyperspheres [RJK19] that have $\alpha = 1 + O(\frac{1}{d})$, where d is the input dimension. □

For $\mu_0 > 0$, we note the following result:

Theorem 54 (Asymptotics of the normalized SC, $\mu_0 > 0$)**.** *Assume $\mu_0 > 0$ and $\sigma^2_{\text{model}} = \sigma^2_{\text{true}} = \sigma^2 = \Theta(1)$. Under Assumptions 50, 51 and 52, with probability of at least $1 - n^{-q}$*

over sample inputs $(x_i)_{i=1}^n$*, where* $0 \leq q < \min\{\frac{2\beta-1}{2}, \alpha\} \cdot \min\{\frac{\alpha-1-2\tau}{2\alpha^2}, \frac{2\beta-1}{\alpha^2}\}$*. the expected normalized SC* (3.8) *has the asymptotic behavior:* $\mathbb{E}_\epsilon F^0(D_n) = \frac{1}{2\sigma^2}\mu_0^2 n + o(n)$*.*

The proof of Theorem 54 is given in Appendix 3.D.1 and follows from showing that when $\mu_0 > 0$, $T_{2,R}(D_n) = \big(\frac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I_R + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\big)(1 + o(1)) = \frac{1}{2\sigma^2}\mu_0^2 n + o(n)$ (see Lemma 84), which dominates $T_1(D_n)$ and the residual $|T_{2,R}(D_n) - T_2(D_n)|$.

3.3.3 Asymptotics of the Bayesian Generalization Error

In this section, we derive the asymptotics of the expected generalization error $\mathbb{E}_{\boldsymbol{\epsilon}}G(D_n)$ by analyzing the asymptotics of the components $G_1(D_n)$ and $G_2(D_n)$ in resp. (3.15) and (3.16) for the following two cases: $\mu_0 = 0$ and $\mu_0 > 0$. First, we consider the case $\mu_0 = 0$.

Theorem 55 (Asymptotics of the Bayesian generalization error, $\mu_0 = 0$)**.** *Let Assumptions 50, 51, and 52 hold. Assume that* $\mu_0 = 0$ *and* $\sigma^2_{\text{model}} = \sigma^2_{\text{true}} = \sigma^2 = \Theta(n^t)$ *where* $1 - \frac{\alpha}{1+2\tau} < t < 1$*. Then with probability of at least* $1 - n^{-q}$ *over sample inputs* $(x_i)_{i=1}^n$ *where* $0 \leq q < \frac{[\alpha-(1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$*, the expectation of the Bayesian generalization error* (3.3) *w.r.t. the noise* $\boldsymbol{\epsilon}$ *has the asymptotic behavior:*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}}G(D_n) &= \tfrac{1+o(1)}{2\sigma^2}\big(\mathrm{Tr}(I + \tfrac{n}{\sigma^2}\Lambda)^{-1}\Lambda - \|\Lambda^{1/2}(I + \tfrac{n}{\sigma^2}\Lambda)^{-1}\|_F^2 + \|(I + \tfrac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu}\|_2^2\big) \\ &= \tfrac{1}{\sigma^2}\Theta(n^{\max\{\frac{(1-\alpha)(1-t)}{\alpha}, \frac{(1-2\beta)(1-t)}{\alpha}\}}). \end{aligned} \tag{3.22}$$

The proof of Theorem 55 is given in Appendix 3.D.2. Intuitively, for a given t, the exponent $\frac{(1-\alpha)(1-t)}{\alpha}$ in (3.22) captures the rate at which the model suppresses the noise, while the exponent $\frac{(1-2\beta)(1-t)}{\alpha}$ captures the rate at which the model learns the target function. A larger β implies that the exponent $\frac{(1-2\beta)(1-t)}{\alpha}$ is smaller and it is easier to learn the target. A larger α implies that the exponent $\frac{(1-\alpha)(1-t)}{\alpha}$ is smaller and the error associated with the noise is smaller as well. A larger α, however, also implies that the exponent $\frac{(1-2\beta)(1-t)}{\alpha}$ is larger (recall that $\alpha > 1$ and $\beta > 1/2$ by Assumptions 50 and 51, resp.), which means that it is harder to learn the target.

Remark 56. *If $f \sim \mathcal{GP}(0, k)$, then using the Karhunen-Loève expansion we have $f(x) = \sum_{p=1}^{\infty} \sqrt{\lambda_p}\omega_p\phi_p(x)$, where $(\omega_p)_{p=1}^{\infty}$ are i.i.d. standard Gaussian variables. We can bound ω_p almost surely as $|\omega_p| \leq C \log p$, where $C = \sup_{p\geq 1} \frac{|\omega_p|}{\log p}$ is a finite constant. Comparing with the expansion of $f(x)$ in (3.9), we find that $\mu_p = \sqrt{\lambda_p}\omega_p = O(p^{-\alpha/2} \log p) = O(p^{-\alpha/2+\varepsilon})$ where $\varepsilon > 0$ is arbitrarily small. Choosing $\beta = \alpha/2 - \varepsilon$ in (3.22), we have $\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n) = O(n^{\frac{1}{\alpha}-1+\frac{2\varepsilon}{\alpha}})$. This rate matches that of an earlier result due to [SH02], where it is shown that the asymptotic learning curve (as measured by the expectation of the excess mean squared error, $\mathbb{E}_f M(D_n)$) scales as $n^{\frac{1}{\alpha}-1}$ when the model is correctly specified, i.e., f is a sample from the same Gaussian process $\mathcal{GP}(0, k)$, and the eigenvalues decay as a power law for large i, $\lambda_i \sim i^{\alpha}$.*

For $\mu_0 > 0$, we note the following result:

Theorem 57 (Asymptotics of the Bayesian generalization error, $\mu_0 > 0$)**.** *Let Assumptions 50, 51, and 52 hold. Assume that $\mu_0 > 0$ and $\sigma^2_{\text{model}} = \sigma^2_{\text{true}} = \sigma^2 = \Theta(n^t)$ where $1 - \frac{\alpha}{1+2\tau} < t < 1$. Then with probability of at least $1 - n^{-q}$ over sample inputs $(x_i)_{i=1}^n$, where $0 \leq q < \frac{[\alpha-(1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$, the expectation of the Bayesian generalization error (3.3) w.r.t. the noise $\boldsymbol{\epsilon}$ has the asymptotic behavior: $\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n) = \frac{1}{2\sigma^2}\mu_0^2 + o(1)$.*

In general, if $\mu_0 > 0$, then the generalization error remains constant when $n \to \infty$. This means that if the target function contains a component in the kernel of the operator L_k, then GP regression is not able to learn the target function. The proof of Theorem 57 is given in Appendix 3.D.2.

3.3.4 Asymptotics of the Excess Mean Squared Error

In this section we derive the asymptotics of the excess mean squared error in Definition 48.

Theorem 58 (Asymptotics of excess mean squared error)**.** *Let Assumptions 50, 51, and 52 hold. Assume $\sigma^2_{\text{model}} = \Theta(n^t)$ where $1 - \frac{\alpha}{1+2\tau} < t < 1$. Then with probability of at least $1 - n^{-q}$ over sample inputs $(x_i)_{i=1}^n$, where $0 \leq q < \frac{[\alpha-(1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$, the excess mean squared error*

(3.6) *has the asymptotic:*

$$\mathbb{E}_{\boldsymbol{\epsilon}} M(D_n) = (1+o(1))\Bigg[\tfrac{\sigma^2_{\text{true}}}{\sigma^2_{\text{model}}}\Big(\operatorname{Tr}(I+\tfrac{n}{\sigma^2_{\text{model}}}\Lambda)^{-1}\Lambda - \|\Lambda^{1/2}(I+\tfrac{n}{\sigma^2_{\text{model}}}\Lambda)^{-1}\|_F^2\Big) + \|(I+\tfrac{n}{\sigma^2_{\text{model}}}\Lambda)^{-1}\boldsymbol{\mu}\|_2^2\Bigg] = \Theta\Bigg(\max\{\sigma^2_{\text{true}} n^{\frac{1-\alpha-t}{\alpha}}, n^{\frac{(1-2\beta)(1-t)}{\alpha}}\}\Bigg)$$

when $\mu_0 = 0$*, and* $\mathbb{E}_{\boldsymbol{\epsilon}} M(D_n) = \mu_0^2 + o(1)$*, when* $\mu_0 > 0$*.*

The proof of Theorem 58 uses similar techniques as Theorem 55 and is given in Appendix 3.D.3.

Remark 59 (Correspondence with kernel ridge regression)**.** *The kernel ridge regression (KRR) estimator arises as a solution to the optimization problem*

$$\hat{f} = \arg\min_{f\in\mathcal{H}_k} \frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2 + \lambda\|f\|^2_{\mathcal{H}_k}, \tag{3.23}$$

where the hypothesis space $\mathcal{H}_k$ *is chosen to be an RKHS, and* $\lambda > 0$ *is a regularization parameter. The solution to* (3.23) *is unique as a function, and is given by* $\hat{f}(x) = K_{\mathbf{x}x}^T(K_n + n\lambda I_n)^{-1}\mathbf{y}$*, which coincides with the posterior mean function* $\bar{m}(x)$ *of the GPR* (3.1) *if* $\sigma^2_{\text{model}} = n\lambda$ *[KHS18, Proposition 3.6]. Thus, the additive Gaussian noise in GPR plays the role of regularization in KRR. Leveraging this well known equivalence between GPR and KRR we observe that Theorem 58 also describes the generalization error of KRR as measured by the excess mean squared error.*

Remark 60. *[CLK21] derived the asymptotics of the expected excess mean-squared error for different regularization strengths and different scales of noise. In particular, for KRR with Gaussian design where* $\Lambda_R^{1/2}(\phi_1(x),\dots,\phi_R(x)))$ *is assumed to follow a Gaussian distribution* $\mathcal{N}(0,\Lambda_R)$*, and regularization* $\lambda = n^{t-1}$ *where* $1-\alpha \le t$*, [CLK21, Eq. 10] showed that*

$$\mathbb{E}_{\{x_i\}_{i=1}^n}\mathbb{E}_{\boldsymbol{\epsilon}} M(D_n) = O\Big(\max\{\sigma^2_{\text{true}} n^{\frac{1-\alpha-t}{\alpha}}, n^{\frac{(1-2\beta)(1-t)}{\alpha}}\}\Big). \tag{3.24}$$

Let $\delta = n^{-q}$*, where* $0 \le q < \frac{[\alpha-(1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$*. By Markov's inequality, this im-*

plies that with probability of at least $1-\delta$, $\mathbb{E}_{\boldsymbol{\epsilon}} M(D_n) = O(\frac{1}{\delta}\max\{\sigma_{\text{true}}^2 n^{\frac{1-\alpha-t}{\alpha}}, n^{\frac{(1-2\beta)(1-t)}{\alpha}}\}) = O(n^q \max\{\sigma_{\text{true}}^2 n^{\frac{1-\alpha-t}{\alpha}}, n^{\frac{(1-2\beta)(1-t)}{\alpha}}\})$. *Theorem 58 improves upon this by showing that with probability of at least* $1-\delta$, *we have an optimal bound* $\mathbb{E}_{\boldsymbol{\epsilon}} M(D_n) = \Theta(\max\{\sigma_{\text{true}}^2 n^{\frac{1-\alpha-t}{\alpha}}, n^{\frac{(1-2\beta)(1-t)}{\alpha}}\})$. *Furthermore, in contrast to the approach by [CLK21], we have no requirement on the distribution of* $\phi_p(x)$, *and hence our result is more generally applicable. For example, Theorem 58 can be applied to KRR with the arc-cosine kernel when the Gaussian design assumption is not valid. In the noiseless setting (*$\sigma_{\text{true}} = 0$*) with constant regularization (*$t = 0$*), Theorem 58 implies that the mean squared error behaves as* $\Theta(n^{\frac{1-2\beta}{\alpha}})$. *This recovers a result in [BCP20, §2.2].*

Our upper bound in Theorem 58 matches with the ones derived in [SHS09, FS20]. [SHS09] and [FS20] also derived algorithm independent minmax lower bounds. In contrast to their results, our Theorem 58 gives lower bounds for different regularization strengths λ.

3.4 Experiments

We illustrate our theory on a few toy experiments. We let the input x be uniformly distributed on a unit circle, i.e., $\Omega = S^1$ and $\rho = \mathcal{U}(S^1)$. The points on S^1 can be represented by $x = (\cos\theta, \sin\theta)$ where $\theta \in [-\pi, \pi)$. We use the first order arc-cosine kernel function without bias, $k^{(1)}_{\text{w/o bias}}(x_1, x_2) = \frac{1}{\pi}(\sin\psi + (\pi - \psi)\cos\psi)$, where $\psi = \langle x_1, x_2\rangle$ is the angle between x_1 and x_2. Hence Assumption 50 is satisfied with $\alpha = 4$. We consider the target functions in Table 3.1, which satisfy Assumption 51 with the indicated β, and μ_0 indicates whether the function lies in the span of eigenfunctions of the kernel. For each target we conduct GPR 20 times and report the mean and standard deviation of the normalized SC and the Bayesian generalization error in Figure 3.1, which agree with the asymptotics predicted in Theorems 53 and 55. The details of the experiments appear in Appendix 3.A, where we also show more experiments confirming our theory for zero- and second- order arc-cosine kernels, with and without biases.

	function value	β	μ_0	$\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n)$	$\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n)$
f_1	$\cos 2\theta$	$+\infty$	0	$\Theta(n^{1/4})$	$\Theta(n^{-3/4})$
f_2	θ^2	2	> 0	$\Theta(n)$	$\Theta(1)$
f_3	$(\lvert\theta\rvert - \pi/2)^2$	2	0	$\Theta(n^{1/4})$	$\Theta(n^{-3/4})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	0	$\Theta(n^{3/4})$	$\Theta(n^{-1/4})$

Table 3.1: Target functions used in the experiments for the first order arc-cosine kernel without bias $k^{(1)}_{\text{w/o bias}}$, their values of β and μ_0, and theoretical rates for the normalized SC and the Bayesian generalization error from our theorems.

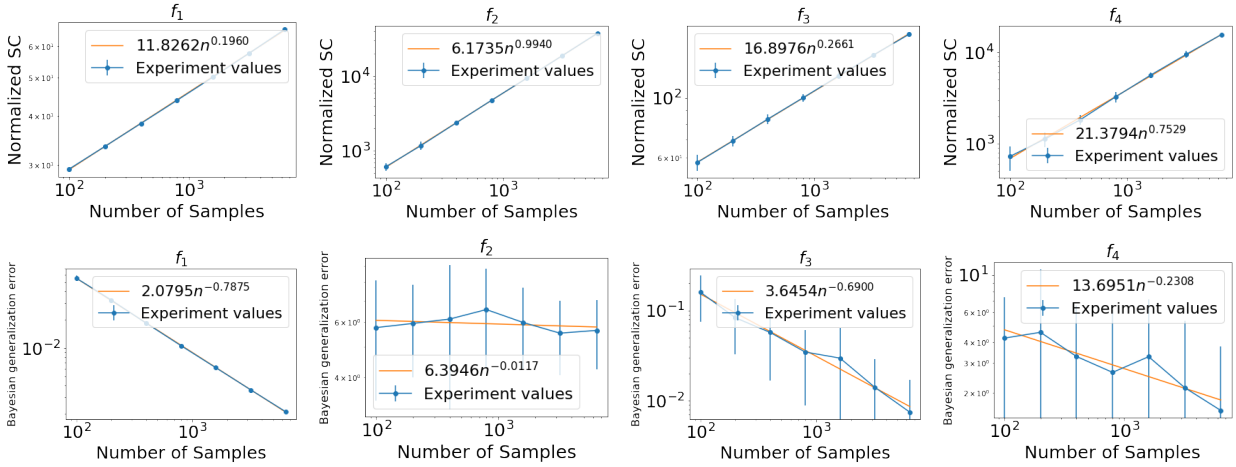

Figure 3.1: Normalized SC (top) and Bayesian generalization error (bottom) for GPR with the kernel $k^{(1)}_{\text{w/o bias}}$ and the target functions in Table 3.1. The orange curves show the linear regression fit for the experimental values (in blue) of the log Bayesian generalization error as a function of log n.

3.5 Conclusion

We described the learning curves for GPR for the case that the kernel and target function follow a power law. This setting is frequently encountered in kernel learning and relates to recent advances on neural networks. Our approach is based on a tight analysis of the concentration of the inner product of empirical eigenfunctions $\Phi^T\Phi$ around nI. This allowed us to obtain more general results with more realistic assumptions than previous works. In particular, we recovered some results on learning curves for GPR and KRR previously obtained under more restricted settings (vide Remarks 56 and 60).

We showed that when $\beta \geq \alpha/2$, meaning that the target function has a compact representation in terms of the eigenfunctions of the kernel, the learning rate is as good as in the correctly specified case. In addition, our result allows us to interpret β from a spectral bias perspective. When $\frac{1}{2} < \beta \leq \frac{\alpha}{2}$, the larger the value of β, the faster the decay of the generalization error. This implies that low-frequency functions are learned faster in terms of the number of training data points.

By leveraging the equivalence between GPR and KRR, we obtained a result on the generalization error of KRR. In the infinite-width limit, training fully-connected deep NNs with gradient descent and infinitesimally small learning rate under least-squared loss is equivalent to solving KRR with respect to the NTK [JGH18b, LXS19a, Dom20], which in several cases is known to have a power-law spectrum [VY21b]. Hence our methods can be applied to study the generalization error of infinitely wide neural networks. In future work, it would be interesting to estimate the values of α and β for the NTK and the NNGP kernel of deep fully-connected or convolutional NNs and real data distributions and test our theory in these cases. Similarly, it would be interesting to consider extensions to finite width kernels.

Appendix

3.A Experiments for Arc-Cosine Kernels of Different Orders

In our experiment, the input space and input distribution are $\Omega = S^1$ and $\rho = \mathcal{U}(S^1)$, and we use the first order arc-cosine kernel function. [CS09] showed that this kernel is the conjugate kernel of an infinitely wide shallow ReLU network with two inputs and no biases in the hidden layer. GP regression with prior $\mathcal{GP}(0, k)$ corresponds to Bayesian training of this network [LSP18]. Under this setting, the eigenvalues and eigenfunctions are $\lambda_1 = \frac{4}{\pi^2}$, $\lambda_2 = \lambda_3 = \frac{1}{4}$, $\lambda_{2p} = \lambda_{2p+1} = \frac{4}{\pi^2((2p-2)^2-1)^2}$, $p \geq 2$ and $\phi_1(\theta) = 1$, $\phi_2(\theta) = \frac{\sqrt{2}}{2}\cos\theta$, $\phi_3(\theta) = \frac{\sqrt{2}}{2}\sin\theta$, $\phi_{2p}(\theta) = \frac{\sqrt{2}}{2}\cos(2p-2)\theta$,$\phi_{2p+1}(\theta) = \frac{\sqrt{2}}{2}\sin(2p-2)\theta$, $p \geq 2$. Hence Assumption 50 is satisfied with $\alpha = 4$, and the second part of Assumption 52 is satisfied with $\|\phi_p\| \leq \frac{\sqrt{2}}{2}$, $p \geq 1$.

The training and test data are generated as follows: We independently sample training inputs $x_1, \ldots, x_n$ and test input x_{n+1} from $\mathcal{U}(S^1)$ and training outputs y_i, $i = 1, \ldots, n$ from $\mathcal{N}(f(x_i), \sigma^2)$, where we choose $\sigma = 0.1$. The Bayesian predictive density conditioned on the test point x_{n+1} $\mathcal{N}(\bar{m}(x_{n+1}), \bar{k}(x_{n+1}, x_{n+1}))$ is obtained by (3.1) and (3.2). We compute the normalized SC by (3.7) and the Bayesian generalization error by the Kullback-Leibler divergence between $\mathcal{N}(f(x_{n+1}), \sigma^2)$ and $\mathcal{N}(\bar{m}(x_{n+1}), \bar{k}(x_{n+1}, x_{n+1}))$.

Next we present experiment results for arc-cosine kernels of different orders and arc-cosine kernels with biases. Consider the first order arc-cosine kernel function with biases,

$$k^{(1)}_{\text{w/ bias}}(x_1, x_2) = \frac{1}{\pi}(\sin\bar{\psi} + (\pi - \bar{\psi})\cos\bar{\psi}), \text{ where } \bar{\psi} = \arccos\big(\tfrac{1}{2}(\langle x_1, x_2\rangle + 1)\big). \tag{3.25}$$

[RJK19] showed that this kernel is the conjugate kernel of an infinitely wide shallow ReLU network with two inputs and one hidden layer with biases, whose eigenvalues satisfy Assumption 50 with $\alpha = 4$. The eigenfunctions of this kernel are the same as that of the first-order arc-cosine kernel without biases, $k^{(1)}_{\text{w/o bias}}$ in Section 3.4. We consider the target functions in Table 3.3, which satisfy Assumption 5 with the indicated β, and μ_0 indicates whether the function lies in the span of eigenfunctions of the kernel. For each target we conduct GPR 20

times and report the mean and standard deviation of the normalized SC and the Bayesian generalization error in Figure 3.2, which agree with the asymptotics predicted in Theorems 53 and 55.

Table 3.2 summarizes all the different kernel functions that we consider in our experiments with pointers to the corresponding tables and figures.

	kernel function	α	activation	bias	pointer
$k^{(1)}_{\text{w/o bias}}$	$\frac{1}{\pi}(\sin\psi + (\pi-\psi)\cos\psi)$	4	$\max\{0,x\}$	no	Table 3.1/Figure 3.1
$k^{(1)}_{\text{w/ bias}}$	$\frac{1}{\pi}(\sin\bar{\psi} + (\pi-\bar{\psi})\cos\bar{\psi})$	4	$\max\{0,x\}$	yes	Table 3.3/Figure 3.2
$k^{(2)}_{\text{w/o bias}}$	$\frac{1}{\pi}(3\sin\psi\cos\psi + (\pi-\psi)(1+2\cos^2\psi))$	6	$(\max\{0,x\})^2$	no	Table 3.4/Figure 3.3
$k^{(2)}_{\text{w/ bias}}$	$\frac{1}{\pi}(3\sin\bar{\psi}\cos\bar{\psi} + (\pi-\bar{\psi})(1+2\cos^2\bar{\psi}))$	6	$(\max\{0,x\})^2$	yes	Table 3.5/Figure 3.4
$k^{(0)}_{\text{w/o bias}}$	$\frac{1}{\pi}(\sin\psi + (\pi-\psi)\cos\psi)$	2	$\frac{1}{2}(1+\text{sign}(x))$	no	Table 3.6/Figure 3.5
$k^{(0)}_{\text{w/ bias}}$	$\frac{1}{\pi}(\sin\bar{\psi} + (\pi-\psi)\cos\bar{\psi})$	2	$\frac{1}{2}(1+\text{sign}(x))$	yes	Table 3.7/Figure 3.6

Table 3.2: The different kernel functions used in our experiments, their values of α, the corresponding neural network activation function along with a pointer to the tables showing the target functions used for the kernels and the corresponding figures.

Summarizing the observations from these experiments, we see that the smoothness of the activation function (which is controlled by the order of the arc-cosine kernel) influences the decay rate α of the eigenvalues. In general, when the activation function is smoother, the decay rate α is larger. Theorem 55 then implies that smooth activation functions are more capable in suppressing noise but slower in learning the target. We also observe that networks with biases are more capable at learning functions compared to networks without bias. For example, the function $\cos(2\theta)$ cannot be learned by the zero order arc-cosine kernel without biases (see Table 3.6 and Figure 3.5), but it can be learned by the zero order arc-cosine kernel with biases (see Table 3.7 and Figure 3.6).

	function value	β	μ_0	$\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n)$	$\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n)$
f_1	$\cos 2\theta$	$+\infty$	0	$\Theta(n^{1/4})$	$\Theta(n^{-3/4})$
f_2	θ^2	2	0	$\Theta(n^{1/4})$	$\Theta(n^{-3/4})$
f_3	$(\lvert\theta\rvert - \pi/2)^2$	2	0	$\Theta(n^{1/4})$	$\Theta(n^{-3/4})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	0	$\Theta(n^{3/4})$	$\Theta(n^{-1/4})$

Table 3.3: Target functions used in the experiments for the first order arc-cosine kernel with bias $k^{(1)}_{\text{w/ bias}}$, their values of β and μ_0, and theoretical rates for the normalized SC and the Bayesian generalization error from our theorems.

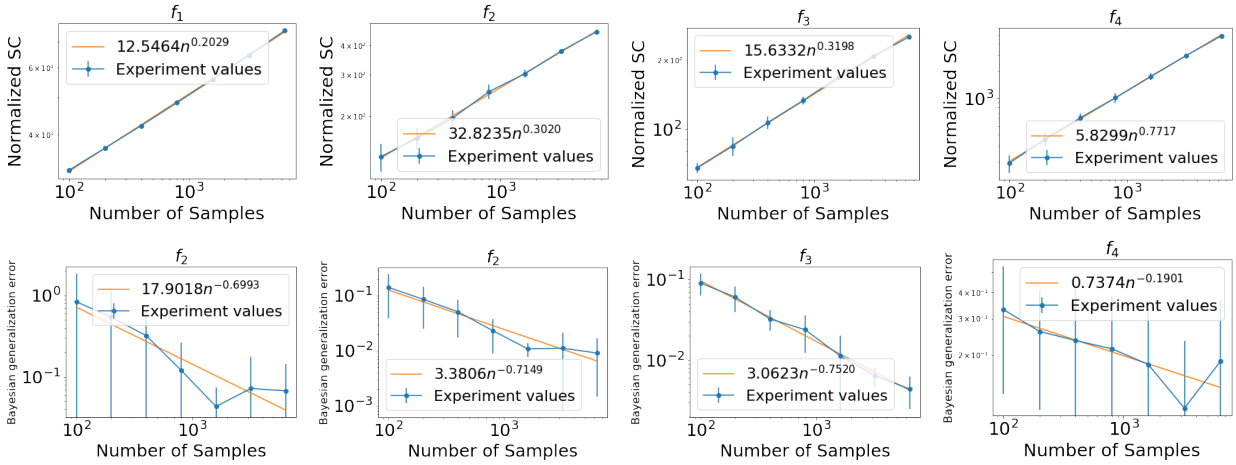

Figure 3.2: Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(1)}_{\text{w/ bias}}$ and the target functions in Table 3.3. The orange curves show the linear regression fit for the experimental values (in blue) of the log Bayesian generalization error as a function of log n.

	function value	β	μ_0	$\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n)$	$\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n)$
f_1	$\cos 2\theta$	$+\infty$	0	$\Theta(n^{1/6})$	$\Theta(n^{-5/6})$
f_2	$\mathrm{sign}(\theta)$	1	0	$\Theta(n^{5/6})$	$\Theta(n^{-1/6})$
f_3	$\pi/2 - \lvert\theta\rvert$	2	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	> 0	$\Theta(n)$	$\Theta(1)$

Table 3.4: Target functions used in the experiments for the second order arc-cosine kernel without bias $k^{(2)}_{\text{w/o bias}}$, their values of β and μ_0, and theoretical rates for the normalized SC and the Bayesian generalization error from our theorems.

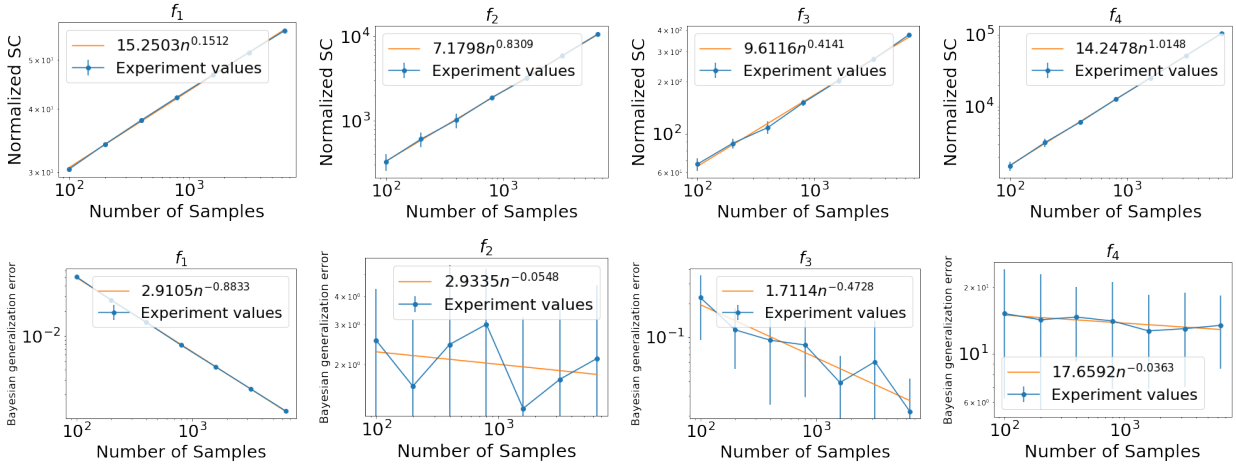

Figure 3.3: Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(2)}_{\text{w/o bias}}$ and the target functions in Table 3.4.

	function value	β	μ_0	$\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n)$	$\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n)$
f_1	$\cos 2\theta$	$+\infty$	0	$\Theta(n^{1/6})$	$\Theta(n^{-5/6})$
f_2	θ^2	2	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$
f_3	$(\lvert\theta\rvert - \pi/2)^2$	2	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	0	$\Theta(n^{5/6})$	$\Theta(n^{-1/6})$

Table 3.5: Target functions used in the experiments for the second order arc-cosine kernel with bias $k^{(2)}_{\text{w/ bias}}$, their values of β and μ_0, and theoretical rates for the normalized SC and the Bayesian generalization error from our theorems.

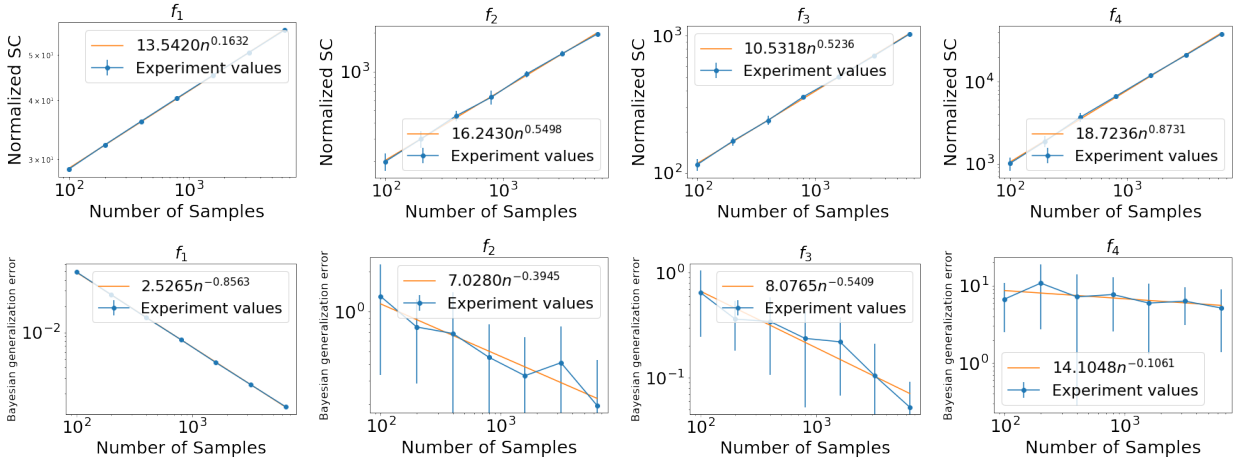

Figure 3.4: Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(2)}_{\text{w/ bias}}$ and the target functions in Table 3.5.

	function value	β	μ_0	$\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n)$	$\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n)$
f_1	$\cos 2\theta$	$+\infty$	> 0	$\Theta(n)$	$\Theta(1)$
f_2	$\mathrm{sign}(\theta)$	1	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$
f_3	$\pi/2 - \lvert\theta\rvert$	2	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	> 0	$\Theta(n)$	$\Theta(1)$

Table 3.6: Target functions used in the experiments for the zero order arc-cosine kernel without bias $k^{(0)}_{\text{w/o bias}}$, their values of β and μ_0, and theoretical rates for the normalized SC and the Bayesian generalization error from our theorems.

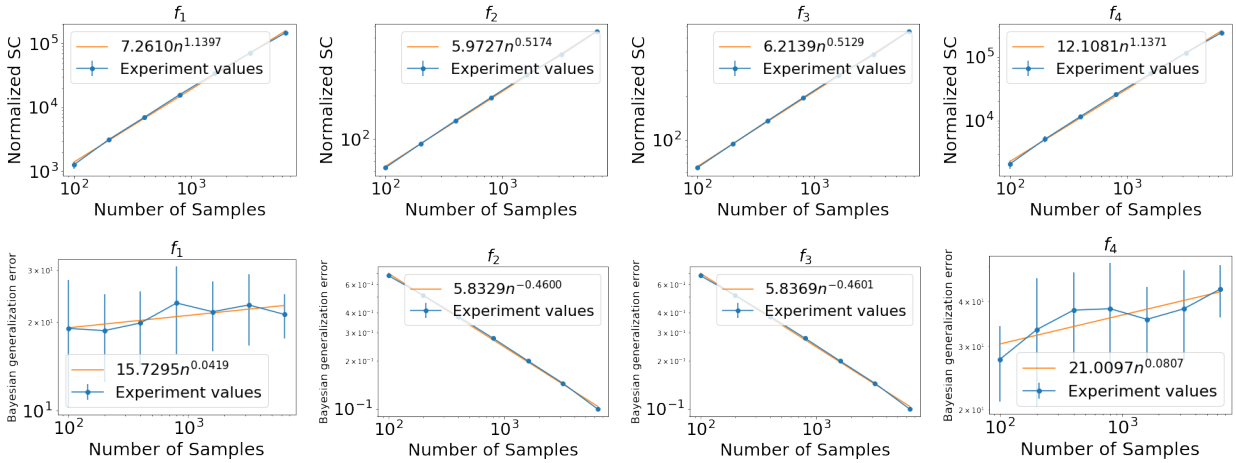

Figure 3.5: Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(0)}_{\text{w/o bias}}$ and the target functions in Table 3.6.

	function value	β	μ_0	$\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n)$	$\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n)$
f_1	$\cos 2\theta$	$+\infty$	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$
f_2	θ^2	2	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$
f_3	$(\lvert\theta\rvert - \pi/2)^2$	2	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$

Table 3.7: Target functions used in the experiments for the zero order arc-cosine kernel with bias $k^{(0)}_{\text{w/ bias}}$, their values of β and μ_0, and theoretical rates for the normalized SC and the Bayesian generalization error from our theorems.

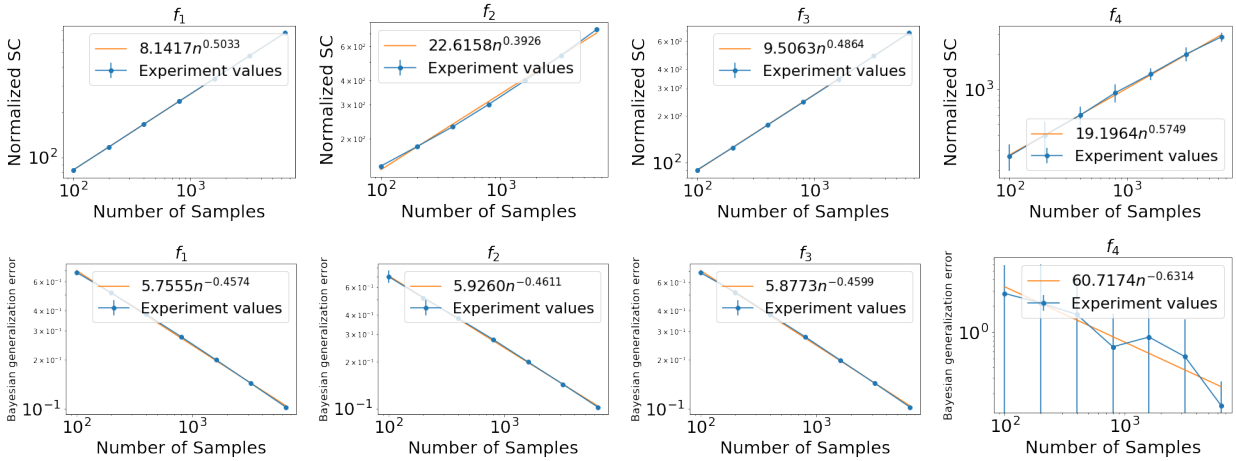

Figure 3.6: Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(0)}_{\text{w/ bias}}$ and the target functions in Table 3.7.

3.B Proofs Related to the Marginal Likelihood

Proof of Proposition 49. Let $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_n)^T$ be the outputs of the GP regression model on training inputs $\mathbf{x}$. Under the GP prior, the prior distribution of $\bar{\mathbf{y}}$ is $\mathcal{N}(0, K_n)$. Then the

evidence of the model is given as follows:

$$\begin{aligned} Z_n &= \int_{\mathbb{R}^n} \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\bar{y}_i - y_i)^2}{2\sigma^2}} \right) \frac{1}{(2\pi)^{n/2} \det(K_n)^{1/2}} e^{-\frac{1}{2}\bar{\mathbf{y}}^T K_n^{-1} \bar{\mathbf{y}}} \mathrm{d}\bar{\mathbf{y}} \\ &= \frac{1}{(2\pi)^n \sigma^n \det(K_n)^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{1}{2}\bar{\mathbf{y}}^T (K_n^{-1} + \frac{1}{\sigma^2} I)\bar{\mathbf{y}} + \frac{1}{\sigma^2}\bar{\mathbf{y}}^T \mathbf{y} - \frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y}} \mathrm{d}\bar{\mathbf{y}}. \end{aligned} \tag{3.26}$$

Letting $\tilde{K}_n^{-1} = K_n^{-1} + \frac{1}{\sigma^2} I$ and $\mu = \frac{1}{\sigma^2}\tilde{K}_n \mathbf{y}$, we have

$$\begin{aligned} Z_n &= \frac{1}{(2\pi)^n \sigma^n \det(K_n)^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{1}{2}(\bar{\mathbf{y}} - \mu)^T \tilde{K}_n^{-1}(\bar{\mathbf{y}} - \mu) - \frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} + \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu} \mathrm{d}\bar{\mathbf{y}} \\ &= \frac{1}{(2\pi)^n \sigma^n \det(K_n)^{1/2}} (2\pi)^{n/2} \det(\tilde{K}_n)^{1/2} e^{-\frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} + \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu} \\ &= \frac{\det(\tilde{K}_n)^{1/2}}{(2\pi)^{n/2} \sigma^n \det(K_n)^{1/2}} e^{-\frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} + \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu}. \end{aligned} \tag{3.27}$$

The normalized evidence is

$$\begin{aligned} Z_n^0 &= \frac{Z_n}{(2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T (\mathbf{y} - f(\mathbf{x}))}} \\ &= \frac{\det(\tilde{K}_n)^{1/2}}{\det(K_n)^{1/2}} e^{-\frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} + \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu + \frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T (\mathbf{y} - f(\mathbf{x}))}. \end{aligned} \tag{3.28}$$

So the normalized stochastic complexity is

$$
\begin{aligned}
F^0(D_n) &= -\log Z_n^0 \\
&= -\frac{1}{2}\log\det(\tilde{K}_n)^{1/2} + \frac{1}{2}\log\det(K_n)^{1/2} + \frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}\mu^T\tilde{K}_n^{-1}\mu \\
&\quad - \frac{1}{2\sigma^2}(\mathbf{y}-f(\mathbf{x}))^T(\mathbf{y}-f(\mathbf{x})) \\
&= -\frac{1}{2}\log\det(K_n^{-1}+\frac{1}{\sigma^2}I)^{-1} + \frac{1}{2}\log\det(K_n) + \frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{y} - \frac{1}{2\sigma^4}\mathbf{y}^T(K_n^{-1}+\frac{1}{\sigma^2}I)^{-1}\mathbf{y} \\
&\quad - \frac{1}{2\sigma^2}(\mathbf{y}-f(\mathbf{x}))^T(\mathbf{y}-f(\mathbf{x})) \\
&= \frac{1}{2}\log\det(I+\frac{K_n}{\sigma^2}) + \frac{1}{2\sigma^2}\mathbf{y}^T(I+\frac{K_n}{\sigma^2})^{-1}\mathbf{y} - \frac{1}{2\sigma^2}(\mathbf{y}-f(\mathbf{x}))^T(\mathbf{y}-f(\mathbf{x})). \\
&= \frac{1}{2}\log\det(I+\frac{K_n}{\sigma^2}) + \frac{1}{2\sigma^2}f(\mathbf{x})^T(I+\frac{K_n}{\sigma^2})^{-1}f(\mathbf{x}) + \frac{1}{2\sigma^2}\boldsymbol{\epsilon}^T(I+\frac{K_n}{\sigma^2})^{-1}\boldsymbol{\epsilon} - \frac{1}{2\sigma^2}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \\
&\quad + \frac{1}{2\sigma^2}\boldsymbol{\epsilon}^T(I+\frac{K_n}{\sigma^2})^{-1}f(\mathbf{x})
\end{aligned}
$$

.

(3.29)

After taking the expectation over noises $\boldsymbol{\epsilon}$, we get

$$
\mathbb{E}_{\boldsymbol{\epsilon}}F^0(D_n) = \frac{1}{2}\log\det(I+\frac{K_n}{\sigma^2}) + \frac{1}{2\sigma^2}f(\mathbf{x})^T(I+\frac{K_n}{\sigma^2})^{-1}f(\mathbf{x}) - \frac{1}{2}\mathrm{Tr}(I-(I+\frac{K_n}{\sigma^2})^{-1}).
\tag{3.30}
$$

This concludes the proof. □

3.C Helper Lemmas

Lemma 61. *Assume that $m \to \infty$ as $n \to \infty$. Given constants $a_1, a_2, s_1, s_2 > 0$, if $s_1 > 1$ and $s_2 s_3 > s_1 - 1$, we have that*

$$
\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} = \Theta(m^{\frac{1-s_1}{s_2}}).
\tag{3.31}
$$

If $s_1 > 1$ and $s_2 s_3 = s_1 - 1$, we have that

$$\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1 + a_2 m i^{-s_2})^{s_3}} = \Theta(m^{-s_3} \log m). \tag{3.32}$$

If $s_1 > 1$ and $s_2 s_3 < s_1 - 1$, we have that

$$\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1 + a_2 m i^{-s_2})^{s_3}} = \Theta(m^{-s_3}). \tag{3.33}$$

Overall, if $s_1 > 1$ and $m \to \infty$,

$$\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1 + a_2 m i^{-s_2})^{s_3}} = \begin{cases} \Theta(m^{\max\{-s_3, \frac{1-s_1}{s_2}\}}), & s_2 s_3 \neq s_1 - 1, \\ \Theta(m^{\frac{1-s_1}{s_2}} \log m), & s_2 s_3 = s_1 - 1. \end{cases} \tag{3.34}$$

Proof of Lemma 61. First, when $s_1 > 1$ and $s_2 s_3 > s_1 - 1$, we have that

$$\begin{aligned} \sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1 + a_2 m i^{-s_2})^{s_3}} &\leq \frac{a_1}{(1 + a_2 m)^{s_3}} + \int_{[1,+\infty]} \frac{a_1 x^{-s_1}}{(1 + a_2 m x^{-s_2})^{s_3}} \mathrm{d}x \\ &= \frac{a_1}{(1 + a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \int_{[1,+\infty]} \frac{a_1 (\frac{x}{m^{1/s_2}})^{-s_1}}{(1 + a_2 (\frac{x}{m^{1/s_2}})^{-s_2})^{s_3}} \mathrm{d}\frac{x}{m^{1/s_2}} \\ &= \frac{a_1}{(1 + a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \int_{[1/m^{1/s_2},+\infty]} \frac{a_1 x^{-s_1}}{(1 + a_2 x^{-s_2})^{s_3}} \mathrm{d}x \\ &= \Theta(m^{\frac{1-s_1}{s_2}}). \end{aligned}$$

On the other hand, we have

$$\begin{aligned} \sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1 + a_2 m i^{-s_2})^{s_3}} &\geq \int_{[1,R+1]} \frac{a_1 x^{-s_1}}{(1 + a_2 m x^{-s_2})^{s_3}} \mathrm{d}x \\ &= m^{\frac{1-s_1}{s_2}} \int_{[1,R+1]} \frac{a_1 (\frac{x}{m^{1/s_2}})^{-s_1}}{(1 + a_2 (\frac{x}{m^{1/s_2}})^{-s_2})^{s_3}} \mathrm{d}\frac{x}{m^{1/s_2}} \\ &= m^{\frac{1-s_1}{s_2}} \int_{[1/m^{1/s_2},(R+1)/m^{1/s_2}]} \frac{a_1 x^{-s_1}}{(1 + a_2 x^{-s_2})^{s_3}} \mathrm{d}x \\ &= \Theta(m^{\frac{1-s_1}{s_2}}). \end{aligned}$$

Second, when $s_1 > 1$ and $s_2 s_3 = s_1 - 1$, we have that

$$\begin{aligned}\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} &\le \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \int_{[1/m^{1/s_2},+\infty]} \frac{a_1 x^{-s_1}}{(1+a_2 x^{-s_2})^{s_3}} \mathrm{d}x \\ &\le \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} O(\log m^{(1/s_2)}) \\ &= \Theta(m^{\frac{1-s_1}{s_2}} \log n).\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} &\ge \int_{[1,R+1]} \frac{a_1 x^{-s_1}}{(1+a_2 m x^{-s_2})^{s_3}} \mathrm{d}x \\ &= m^{\frac{1-s_1}{s_2}} \int_{[1,R+1]} \frac{a_1 (\frac{x}{m^{1/s_2}})^{-s_1}}{(1+a_2 (\frac{x}{m^{1/s_2}})^{-s_2})^{s_3}} \mathrm{d}\frac{x}{m^{1/s_2}} \\ &= m^{\frac{1-s_1}{s_2}} \int_{[1/m^{1/s_2},(R+1)/m^{1/s_2}]} \frac{a_1 x^{-s_1}}{(1+a_2 x^{-s_2})^{s_3}} \mathrm{d}x \\ &= \Theta(m^{\frac{1-s_1}{s_2}} \log n).\end{aligned}$$

Third, when $s_1 > 1$ and $s_2 s_3 < s_1 - 1$, we have that

$$\begin{aligned}\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} &\le \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \int_{[1/m^{1/s_2},+\infty]} \frac{a_1 x^{-s_1}}{(1+a_2 x^{-s_2})^{s_3}} \mathrm{d}x \\ &\le \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \Theta(m^{(-1/s_2)(1-s_1+s_2 s_3)}) \\ &= \Theta(m^{-s_3}).\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} &\le \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \int_{[2/m^{1/s_2},(R+1)/m^{1/s_2}]} \frac{a_1 x^{-s_1}}{(1+a_2 x^{-s_2})^{s_3}} \mathrm{d}x \\ &\le \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \Theta(m^{(-1/s_2)(1-s_1+s_2 s_3)}) \\ &= \Theta(m^{-s_3}).\end{aligned}$$

Overall, if $s_1 > 1$,

$$\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} = \begin{cases} \Theta(m^{\max\{-s_3, \frac{1-s_1}{s_2}\}}), & s_2 s_3 \neq s_1 - 1, \\ \Theta(m^{-s_3} \log n), & s_2 s_3 = s_1 - 1. \end{cases} \tag{3.35}$$

□

Lemma 62. *Assume that $R = m^{\frac{1}{s_2}+\kappa}$ for $\kappa > 0$. Given constants $a_1, a_2, s_1, s_2 > 0$, if $s_1 \leq 1$, we have that*

$$\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} = \tilde{O}(\max\{m^{-s_3}, R^{1-s_1}\}). \tag{3.36}$$

Proof of Lemma 62. First, when $s_1 \leq 1$ and $s_2 s_3 > s_1 - 1$, we have that

$$\begin{aligned} \sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} &\leq \frac{a_1}{(1+a_2 m)^{s_3}} + \int_{[1,R]} \frac{a_1 x^{-s_1}}{(1+a_2 m x^{-s_2})^{s_3}} \mathrm{d}x \\ &= \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \int_{[1,R]} \frac{a_1 (\frac{x}{m^{1/s_2}})^{-s_1}}{(1+a_2 (\frac{x}{m^{1/s_2}})^{-s_2})^{s_3}} \mathrm{d}\frac{x}{m^{1/s_2}} \\ &= \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \int_{[1/m^{1/s_2}, R/m^{1/s_2}]} \frac{a_1 x^{-s_1}}{(1+a_2 x^{-s_2})^{s3}} \mathrm{d}x \\ &= \frac{a_1}{(1+a_2 m)^{s_3}} + \tilde{O}(m^{\frac{1-s_1}{s_2}} (\frac{R}{m^{1/s_2}})^{1-s_1}) \\ &= \tilde{O}(\max\{m^{-s_3}, R^{1-s_1}\}). \end{aligned}$$

Second, when $s_1 \leq 1$ and $s_2 s_3 \leq s_1 - 1$, we have that

$$\begin{aligned} \sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} &\leq \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \int_{[1/m^{1/s_2}, R/m^{1/s_2}]} \frac{a_1 x^{-s_1}}{(1+a_2 x^{-s_2})^{s_3}} \mathrm{d}x \\ &\leq \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \tilde{O}(m^{(-1/s_2)(1-s_1+s_2 s_3)} + (\frac{R}{m^{1/s_2}})^{1-s_1}) \\ &= \tilde{O}(\max\{m^{-s_3}, R^{1-s_1}\}). \end{aligned}$$

Overall, if $s_1 \leq 1$,

$$\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} = \tilde{O}(\max\{m^{-s_3}, R^{1-s_1}\}). \tag{3.37}$$

□

Lemma 63. *Assume that $f \in L^2(\Omega, \rho)$. Consider the random vector $f(\mathbf{x}) = (f(x_1), \ldots, f(x_n))^T$, where $x_1, \ldots, x_n$ are drawn i.i.d from ρ. Then with probability of at least $1 - \delta_1$, we have*

$$\|f(\mathbf{x})\|_2^2 = \sum_{i=1}^{n} f^2(x_i) = \tilde{O}\Big((\tfrac{1}{\delta_1} + 1)n\|f\|_2^2\Big),$$

where $\|f\|_2^2 = \int_{x\in\Omega} f^2(x)\mathrm{d}\rho(x)$.

Proof of Lemma 63. Given a positive number $C \geq \|f\|_2^2$, applying Markov's inequality we have

$$\mathbb{P}(f^2(X) > C) \leq \frac{1}{C}\|f\|_2^2.$$

Let A be the event that for all sample inputs $(x_i)_{i=1}^n$, $f^2(x_i) \leq C$. Then

$$\mathbb{P}(A) \geq 1 - n\mathbb{P}(f^2(X) > C) \geq 1 - \frac{1}{C}n\|f\|_2^2. \tag{3.38}$$

Define $\bar{f}^2(x) = \min\{f^2(x), C\}$. Then $\mathbb{E}\bar{f}^2(X) \leq \mathbb{E}f^2(X) = \|f\|_2^2$. So $|\bar{f}^2(X) - \mathbb{E}\bar{f}^2(X)| \leq \max\{C, \|f\|_2^2\} = C$ Since $0 \leq \bar{f}^2(x) \leq C$, we have

$$\mathbb{E}(\bar{f}^4(X)) \leq C\mathbb{E}(\bar{f}^2(X)) \leq C\|f\|_2^2. \tag{3.39}$$

So we have

$$\mathbb{E}|\bar{f}^2(X) - \mathbb{E}\bar{f}^2(X)|^2 \leq \mathbb{E}(\bar{f}^4(X)) \leq C\|f\|_2^2. \tag{3.40}$$

Applying Bernstein's inequality, we have

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n} \bar{f}^2(x_i) > t + n\mathbb{E}\bar{f}^2(X)) &\leq \exp\left(-\frac{t^2}{2(n\mathbb{E}|\bar{f}^2(X) - \mathbb{E}\bar{f}^2(X)|^2) + \frac{Ct}{3})}\right) \\
&\leq \exp\left(-\frac{t^2}{2(nC\|f\|_2^2 + \frac{Ct}{3})}\right) \\
&\leq \exp\left(-\frac{t^2}{4\max\{nC\|f\|_2^2, \frac{Ct}{3}\}}\right).
\end{aligned}$$

Hence, with probability of at least $1-\delta_1/2$ we have

$$\begin{aligned}\sum_{i=1}^{n}\bar{f}^2(x_i) \leq & \max\left\{\sqrt{4C\log\frac{2}{\delta_1}n\|f\|_2^2}, \frac{4C}{3}\log\frac{2}{\delta_1}\right\} + n\mathbb{E}\bar{f}^2(X) \\ \leq & \max\left\{\sqrt{4C\log\frac{2}{\delta_1}n\|f\|_2^2}, \frac{4C}{3}\log\frac{2}{\delta_1}\right\} + n\|f\|_2^2.\end{aligned} \tag{3.41}$$

When event A happens, $f^2(x_i) = \bar{f}^2(x_i)$ for all sample inputs. According to (3.38) and (3.41), with probability at least $1-\frac{1}{C}n\|f\|_2^2 - \delta_1/2$, we have

$$\sum_{i=1}^{n}f^2(x_i) = \sum_{i=1}^{n}\bar{f}^2(x_i) \leq \max\left\{\sqrt{4C\log\frac{2}{\delta_1}n\|f\|_2^2}, \frac{4C}{3}\log\frac{2}{\delta_1}\right\} + n\|f\|_2^2.$$

Choosing $C = \frac{2}{\delta_1}n\|f\|_2^2$, with probability of at least $1-\delta_1$ we have

$$\begin{aligned}\sum_{i=1}^{n}f^2(x_i) = \sum_{i=1}^{n}\bar{f}^2(x_i) &\leq \max\left\{\sqrt{\frac{8}{\delta_1}\log\frac{2}{\delta_1}n^2\|f\|_2^4}, \frac{8}{3\delta_1}n\|f\|_2^2\log\frac{2}{\delta_1}\right\} + n\|f\|_2^2 \\ &= \tilde{O}\Big((\tfrac{1}{\delta_1}+1)n\|f\|_2^2\Big).\end{aligned}$$

□

Lemma 64. *Assume that $f \in L^2(\Omega,\rho)$. Consider the random vector $f(\mathbf{x}) = (f(x_1),\ldots,f(x_n))^T$, where $x_1,\ldots,x_n$ are drawn i.i.d from ρ. Assume that $\|f\|_\infty = \sup_{x\in\Omega}f(x) \leq C$. With probability of at least $1-\delta_1$, we have*

$$\|f(\mathbf{x})\|_2^2 = \tilde{O}\left(\sqrt{C^2n\|f\|_2^2} + C^2\right) + n\|f\|_2^2,$$

where $\|f\|_2^2 = \int_{x\in\Omega}f^2(x)\mathrm{d}\rho(x)$.

Proof of Lemma 64. We have $|f^2(X) - \mathbb{E}f^2(X)| \leq \max\{C^2, \|f\|_2^2\} = C^2$ Since $0 \leq f^2(x) \leq C$, we have

$$\mathbb{E}(f^4(X)) \leq C^2\mathbb{E}(f^2(X)) \leq C^2\|f\|_2^2. \tag{3.42}$$

So we have

$$\mathbb{E}|f^2(X) - \mathbb{E}f^2(X)|^2 \leq \mathbb{E}(f^4(X)) \leq C^2\|f\|_2^2. \tag{3.43}$$

Applying Bernstein's inequality, we have

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n} f^2(x_i) > t + n\mathbb{E}f^2(X)) \leq & \exp\left(-\frac{t^2}{2(n\mathbb{E}|f^2(X) - \mathbb{E}f^2(X)|^2) + \frac{C^2 t}{3})}\right) \\
\leq & \exp\left(-\frac{t^2}{2(nC^2\|f\|_2^2 + \frac{C^2 t}{3})}\right) \\
\leq & \exp\left(-\frac{t^2}{4\max\{nC^2\|f\|_2^2, \frac{C^2 t}{3}\}}\right).
\end{aligned}$$

Hence, with probability of at least $1 - \delta_1$ we have

$$\begin{aligned}
\sum_{i=1}^{n} f^2(x_i) \leq & \max\left\{\sqrt{4C^2 \log\frac{1}{\delta_1} n\|f\|_2^2}, \frac{4C^2}{3}\log\frac{1}{\delta_1}\right\} + n\mathbb{E}f^2(X) \\
\leq & \tilde{O}\left(\max\left\{\sqrt{C^2 n\|f\|_2^2}, C^2\right\}\right) + n\|f\|_2^2 \\
\leq & \tilde{O}\left(\sqrt{C^2 n\|f\|_2^2} + C^2\right) + n\|f\|_2^2.
\end{aligned} \tag{3.44}$$

□

For the proofs in the reminder of this section, the definitions of the relevant quantities are given in Section 3.3.

Corollary 65. *With probability of at least* $1 - \delta_1$*, we have*

$$\|f_{>R}(\mathbf{x})\|_2^2 = \tilde{O}\left((\tfrac{1}{\delta_1} + 1)nR^{1-2\beta}\right).$$

Proof of Corollary 65. The L_2 norm of $f_{>R}(x)$ is given by $\|f_{>R}\|_2^2 = \sum_{p=R+1}^{\infty} \mu_p^2 \leq \frac{C_\mu}{2\beta-1} R^{1-2\beta}$. Applying Lemma 63 we get the result. □

Corollary 66. *For any $\nu \in \mathbb{R}^R$, with probability of at least $1-\delta_1$ we have*

$$\|\Phi_R \nu\|_2^2 = \tilde{O}\Big((\tfrac{1}{\delta_1}+1)n\|\nu\|_2^2\Big).$$

Proof of Corollary 66. Let $g(x) = \sum_{p=1}^R \nu_p \phi_p(x)$. Then $\Phi_R \nu = g(\mathbf{x})$. The L_2 norm of $g(x)$ is given by $\|g\|_2^2 = \sum_{p=1}^R \nu_p^2 = \|\nu\|_2^2$. Applying Lemma 63 we get the result. □

Next we consider the quantity, $\Phi_R^T \Phi_R - nI$. The key tool that we use is the matrix Bernstein inequality that describes the upper tail of a sum of independent zero-mean random matrices.

Lemma 67. *Let $D = \mathrm{diag}\{d_1, \ldots, d_R\}$, $d_1, \ldots, d_R > 0$ and $d_{\max} = \max\{d_1, \ldots, d_R\}$. Let $M = \max\{\sum_{p=0}^R d_p^2 \|\phi_p\|_\infty^2, d_{\max}^2\}$. Then with probability of at least $1-\delta$, we have*

$$\|D(\Phi_R^T \Phi_R - nI)D\|_2 \le \max\left\{\sqrt{n d_{\max}^2 M \log \tfrac{R}{\delta}}, M \log \tfrac{R}{\delta})\right\}. \tag{3.45}$$

Proof of Lemma 67. Let $Y_j = (\phi_1(x_j), \ldots, \phi_R(x_j))^T$ and $Z_j = DY_j$. It is easy to verify that $\mathbb{E}(Z_j Z_j^T) = D^2$. Then the left hand side of (3.45) is $\sum_{j=1}^n [Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)]$. We note that

$$\|Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)\|_2 \le \max\{\|Z_j Z_j^T\|_2, \|\mathbb{E}(Z_j Z_j^T)\|_2\} \le \max\{\|Z_j\|_2^2, d_{\max}^2\}.$$

For $\|Z_j\|_2^2$, we have

$$\|Z_j\|_2^2 = \sum_{p=0}^R d_p^2 \phi_p^2(x_j) \le \sum_{p=0}^R d_p^2 \|\phi_p\|_\infty^2, \tag{3.46}$$

we have

$$\|Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)\|_2 \le \max\{\textstyle\sum_{p=0}^R d_p^2 \|\phi_p\|_\infty^2, d_{\max}^2\}.$$

On the other hand,

$$\mathbb{E}[(Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T))^2] = \mathbb{E}[\|Z_j\|_2^2 Z_j Z_j^T] - (\mathbb{E}(Z_j Z_j^T))^2.$$

Since

$$\begin{aligned}\mathbb{E}[\|Z_j\|_2^2 Z_j Z_j^T] &\preccurlyeq \mathbb{E}[\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2 Z_j Z_j^T], \quad \text{(by (3.46))} \\ &= \sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2 \mathbb{E}[Z_j Z_j^T],\end{aligned}$$

we have

$$\begin{aligned}\|\mathbb{E}[(Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T))^2]\|_2 &\le \max\{\textstyle\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2 \|\mathbb{E}[Z_j Z_j^T]\|_2, d_{\max}^4\} \\ &\le \max\{\textstyle\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2 d_{\max}^2, d_{\max}^4\} \\ &\le d_{\max}^2 \max\{\textstyle\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2, d_{\max}^2\}.\end{aligned}$$

Using the matrix Bernstein inequality [Tro12, Theorem 6.1], we have

$$\begin{aligned}&\mathbb{P}(\|\sum_{j=1}^{n}[Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)]\|_2 > t) \\ \le& R\exp\left(\frac{-t^2}{2(n\|\mathbb{E}[(Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T))^2]\|_2 + \frac{t \max_j \|Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)\|_2}{3})}\right) \\ \le& R\exp\left(\frac{-t^2}{2(nd_{\max}^2 \max\{\sum_{p=0}^{R} d_p^2\|\phi_p\|_\infty^2, d_{\max}^2\} + \frac{t\max\{\sum_{p=0}^{R} d_p^2\|\phi_p\|_\infty^2, d_{\max}^2\}}{3})}\right) \\ =& R\exp\left(\frac{-t^2}{O(\max\{nd_{\max}^2 \max\{\sum_{p=0}^{R} d_p^2\|\phi_p\|_\infty^2, d_{\max}^2\}, t\max\{\sum_{p=0}^{R} d_p^2\|\phi_p\|_\infty^2, d_{\max}^2\}\})}\right).\end{aligned}$$

Then with probability of at least $1-\delta$, we have

$$\begin{aligned}&\|\sum_{j=1}^{n}[Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)]\|_2 \\ &\le \max\left\{\sqrt{nd_{\max}^2 \max\{\textstyle\sum_{p=0}^{R} d_p^2\|\phi_p\|_\infty^2,\ d_{\max}^2\}\log\frac{R}{\delta}}, \max\left\{\textstyle\sum_{p=0}^{R} d_p^2\|\phi_p\|_\infty^2,\ d_{\max}^2\right\}\log\frac{R}{\delta}\right\}.\end{aligned}$$

□

Corollary 68. *Suppose that the eigenvalues $(\lambda_p)_{p\ge1}$ satisfy Assumption 50, and the eigenfunctions satisfy Assumption 52. Assume $\sigma^2 = \Theta(n^t)$ where $1 - \frac{\alpha}{1+2\tau} < t < 1$ Let γ be a*

positive number such that $\frac{1+\alpha+2\tau-(1+2\tau+2\alpha)t}{2\alpha(1-t)} < \gamma \le 1$. *Then with probability of at least* $1-\delta$, *we have*

$$\begin{aligned}&\|\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R-nI)\Lambda_R^{\gamma/2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2\\&\quad\le O\bigg(n^{\frac{1+\alpha+2\tau-(1+2\tau+2\alpha)t}{2\alpha}-\gamma(1-t)}\sqrt{\log\tfrac{R}{\delta}}\bigg).\end{aligned}\tag{3.47}$$

Proof of Corollary 68. Use the same notation as in Lemma 67. Let $D=(I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}$. Then $d_{\max}^2 \le \frac{\sigma^{2\gamma}}{n^\gamma}$ and $\sum_{p=0}^R d_p^2\|\phi_p\|_\infty^2 \le \sum_{p=0}^R C_\phi^2\frac{\lambda_p^\gamma p^{2\tau}}{(1+\frac{n}{\sigma^2}\lambda_p)^\gamma} = O((\frac{n}{\sigma^2})^{\frac{1-\gamma\alpha+2\tau}{\alpha}})$, where the first inequality follows from Assumptions 50 and 52 and the last equality from Lemma 61. Then $M=\max\{\sum_{p=0}^R d_p^2\|\phi_p\|_\infty^2, d_{\max}^2\} = O((\frac{n}{\sigma^2})^{\frac{1-\gamma\alpha+2\tau}{\alpha}})$. Applying Lemma 67, we have

$$\begin{aligned}&\|\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R-nI)\Lambda_R^{\gamma/2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2\\&\le \tfrac{1}{\sigma^2}\max\left\{\sqrt{n\tfrac{\sigma^{2\gamma}}{n^\gamma}O((\tfrac{n}{\sigma^2})^{\frac{1-\gamma\alpha+2\tau}{\alpha}})\log\tfrac{R}{\delta}}, O((\tfrac{n}{\sigma^2})^{\frac{1-\gamma\alpha+2\tau}{\alpha}})\log\tfrac{R}{\delta}\right\}\\&= O(\tfrac{1}{\sigma^2}(\tfrac{n}{\sigma^2})^{\frac{1-2\gamma\alpha+2\tau}{2\alpha}}n^{\frac12}) = O(\sqrt{\log\tfrac{R}{\delta}}n^{\frac{(1-2\gamma\alpha+2\tau)(1-t)}{2\alpha}+\frac12-t})\\&= O\bigg(\sqrt{\log\tfrac{R}{\delta}}n^{\frac{1+\alpha+2\tau}{2\alpha}-\frac{(1+2\tau+2\alpha)t}{2\alpha}-\gamma(1-t)}\bigg).\end{aligned}\tag{3.48}$$

□

Corollary 69. *Suppose that the eigenvalues* $(\lambda_p)_{p\ge1}$ *satisfy Assumption 50, and the eigenfunctions satisfy Assumption 52. Let* $\tilde\Lambda_{1,R}=\mathrm{diag}\{1,\lambda_1,\dots,\lambda_R\}$. *Assume* $\sigma^2=\Theta(n^t)$ *where* $t<1$ *Let* γ *be a positive number such that* $\frac{1+2\tau}{\alpha}<\gamma\le1$. *Then with probability of at least* $1-\delta$, *we have*

$$\|(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\tilde\Lambda_{1,R}^{\gamma/2}(\Phi_R^T\Phi_R-nI)\tilde\Lambda_{1,R}^{\gamma/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2 \le O\bigg(\sqrt{\log\tfrac{R}{\delta}}n^{\frac12}\bigg).\tag{3.49}$$

Proof of Corollary 69. Use the same notation as in Lemma 67. Let $D=(I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\tilde\Lambda_{1,R}^{\gamma/2}$. Then $d_{\max}^2\le1$ and $\sum_{p=0}^R d_p^2\|\phi_p\|_\infty^2 \le C_\phi^2+\sum_{p=1}^R C_\phi^2\frac{\lambda_p^\gamma p^{2\tau}}{(1+\frac{n}{\sigma^2}\lambda_p)^\gamma} = C_\phi^2+O(n^{\frac{(1-\gamma\alpha+2\tau)(1-t)}{\alpha}}) = O(1)$ where the first inequality follows from Assumptions 50 and 52 and the second equality from

Lemma 61. Then $M = \max\{\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2, d_{\max}^2\} = O(1)$. Applying Lemma 67, we have

$$\begin{aligned}
&\|(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2 \\
&\leq \max\left\{\sqrt{\log\tfrac{R}{\delta} nO(1)}, \log\tfrac{R}{\delta} O(1)\right\} \\
&= O\left(\sqrt{\log\tfrac{R}{\delta}} n^{\frac{1}{2}}\right).
\end{aligned} \tag{3.50}$$

□

Corollary 70. *Suppose that the eigenvalues $(\lambda_p)_{p\geq 1}$ satisfy Assumption 50, and the eigenfunctions satisfy Assumption 52. Let $\Phi_{R+1:S} = (\phi_{R+1}(\mathbf{x}), \ldots, \phi_S(\mathbf{x}))$, and $\Lambda_{R+1:S} = (\lambda_{R+1}, \ldots, \lambda_S)$. Then with probability of at least $1-\delta$, we have*

$$\|\Lambda_{R+1:S}^{1/2}(\Phi_{R+1:S}^T\Phi_{R+1:S} - nI)\Lambda_{R+1:S}^{1/2}\|_2 \leq O\big(\log\tfrac{S-R}{\delta} \max\{n^{\frac{1}{2}} R^{\frac{1-2\alpha+2\tau}{2}}, R^{1-\alpha+2\tau}\}\big). \tag{3.51}$$

Proof of Corollary 70. Use the same notation as in Lemma 67. Let $D = \Lambda_{R+1:S}^{1/2}$. Then $d_{\max}^2 \leq \overline{C_\lambda} R^{-\alpha} = O(R^{-\alpha})$ and $\sum_{p=R+1}^{S} C_\phi^2 d_p^2 p^{2\tau} \leq \sum_{p=R+1}^{S} C_\phi^2 \overline{C_\lambda} p^{-\alpha} p^{2\tau} = O(R^{1-\alpha+2\tau})$, where the first inequality follows from Assumptions 50 and 52. Then $M = \max\{\sum_{p=R+1}^{S} C_\phi^2 d_p^2 p^{2\tau}, d_{\max}^2\} = O(R^{1-\alpha+2\tau})$. Applying Lemma 67, we have

$$\begin{aligned}
&\|(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2 \\
&\leq \max\left\{\sqrt{\log\tfrac{S-R}{\delta} nO(R^{-\alpha})O(R^{1-\alpha+2\tau})}, \log\tfrac{S-R}{\delta} O(R^{1-\alpha+2\tau}))\right\} \\
&= O\big(\log\tfrac{S-R}{\delta} \max\{n^{\frac{1}{2}} R^{\frac{1-2\alpha+2\tau}{2}}, R^{1-\alpha+2\tau}\}\big).
\end{aligned} \tag{3.52}$$

□

Lemma 71. *Under the assumptions of Corollary 70, with probability of at least $1-\delta$, we have*

$$\|\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\|_2 = \tilde{O}(\max\{nR^{-\alpha}, n^{\frac{1}{2}} R^{\frac{1-2\alpha+2\tau}{2}}, R^{1-\alpha+2\tau}\}).$$

Proof of Lemma 71. For $S \in \mathbb{N}$, we have

$$\begin{aligned}\|\Phi_{>S}\Lambda_{>S}\Phi_{>S}^T\|_2 \leq & \sum_{p=S+1}^{\infty} \|\Lambda_p \phi_p(\mathbf{x})\phi_p(\mathbf{x})^T\|_2 \\ = & \sum_{p=S+1}^{\infty} \lambda_p \|\phi_p(\mathbf{x})\|_2^2 \\ \leq & \sum_{p=S+1}^{\infty} \lambda_p n C_\phi^2 p^{2\tau} \\ = & O(nS^{1-\alpha+2\tau}).\end{aligned}$$

Let $S = R^{\frac{\alpha}{\alpha-1-2\tau}}$. Then we get $\|\Phi_{>S}\Lambda_{>S}\Phi_{>S}^T\|_2 = O(nR^{-\alpha})$.

Let $\Phi_{R+1:S} = (\phi_{R+1}(\mathbf{x}), \ldots, \phi_S(\mathbf{x}))$, $\Lambda_{R+1:S} = (\lambda_{R+1}, \ldots, \lambda_S)$. We then have

$$\begin{aligned}\|\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\|_2 \leq & \|\Phi_{>S}\Lambda_{>S}\Phi_{>S}^T\|_2 + \|\Phi_{R+1:S}\Lambda_{R+1:S}\Phi_{R+1:S}^T\|_2 \\ \leq & O(nR^{-\alpha}) + \|\Lambda_{R+1:S}^{1/2}\Phi_{R+1:S}^T\Phi_{R+1:S}\Lambda_{R+1:S}^{1/2}\|_2 \\ \leq & O(nR^{-\alpha}) + n\|\Lambda_{R+1:S}\|_2 + \|\Lambda_{R+1:S}^{1/2}(\Phi_{R+1:S}^T\Phi_{R+1:S} - nI)\Lambda_{R+1:S}^{1/2}\|_2 \\ \leq & O(nR^{-\alpha}) + O(nR^{-\alpha}) + O(\log\frac{R^{\frac{\alpha}{\alpha-1}} - R}{\delta}\max\{n^{\frac{1}{2}}R^{\frac{1-2\alpha+2\tau}{2}}, R^{1-\alpha+2\tau}\}) \\ = & \tilde{O}(\max\{nR^{-\alpha}, n^{\frac{1}{2}}R^{\frac{1-2\alpha+2\tau}{2}}, R^{1-\alpha+2\tau}\}),\end{aligned}$$

where in the fourth inequality we use Corollary 70. □

Corollary 72. *Assume that $\sigma^2 = \Theta(1)$. If $R = n^{\frac{1}{\alpha}+\kappa}$ where $0 < \kappa < \frac{\alpha-1-2\tau}{\alpha(1+2\tau)}$, then with probability of at least $1-\delta$, we have*

$$\|(I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2 \leq \|\tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2 = \tilde{O}(n^{-\kappa\alpha}) = o(1).$$

Proof of Corollary 72. By Lemma 71 and the assumption $R = n^{\frac{1}{\alpha}+\kappa}$, we have

$$\begin{aligned}\|(I + \tfrac{\Phi_R \Lambda_R \Phi_R^T}{\sigma^2})^{-1} \tfrac{\Phi_{>R} \Lambda_{>R} \Phi_{>R}^T}{\sigma^2}\|_2 \leq & \|\tfrac{\Phi_{>R} \Lambda_{>R} \Phi_{>R}^T}{\sigma^2}\|_2 \\ \leq & \tilde{O}(\max\{nR^{-\alpha}, n^{\frac{1}{2}} R^{\frac{1-2\alpha+2\tau}{2}}, R^{1-\alpha+2\tau}\}) \\ = & \tilde{O}(n^{-\kappa\alpha}).\end{aligned}$$

□

Lemma 73. *Assume that* $\|\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2 < 1$ *where* $\frac{1+2\tau}{\alpha} < \gamma \leq 1$. *We then have*

$$\begin{aligned}&(I + \tfrac{1}{\sigma^2}\Lambda_R \Phi_R^T \Phi_R)^{-1} \\ &= (I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1} + \sum_{j=1}^{\infty} (-1)^j \big(\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\big)^j (I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}.\end{aligned}$$

Proof of Lemma 73. First note that

$$\begin{aligned}&\|\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{1/2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_2 \\ &< \|\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2 < 1.\end{aligned}$$

Let $\tilde{\Lambda}_{\epsilon,R} = \mathrm{diag}\{\epsilon, \lambda_1, \ldots, \lambda_R\}$. Since $\Lambda_R = \mathrm{diag}\{0, \lambda_1, \ldots, \lambda_R\}$, we have that when ϵ is sufficiently small, $\|\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\tilde{\Lambda}_{\epsilon,R}^{1/2}(\Phi_R^T\Phi_R - nI)\tilde{\Lambda}_{\epsilon,R}^{1/2}(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\|_2 < 1$. Since all

diagonal entries of $\tilde{\Lambda}_{\epsilon,R}$ are positive, we have

$$
\begin{aligned}
&(I + \tfrac{1}{\sigma^2}\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1} \\
&\quad = (I + \tfrac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R} + \tfrac{1}{\sigma^2}\tilde{\Lambda}_{\epsilon,R}(\Phi_R^T\Phi_R - nI))^{-1} \\
&\quad = \tilde{\Lambda}_{\epsilon,R}^{1/2}(I + \tfrac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\Big[I + \tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\tilde{\Lambda}_{\epsilon,R}^{1/2}(\Phi_R^T\Phi_R - nI)\tilde{\Lambda}_{\epsilon,R}^{1/2}(I + \tfrac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\Big]^{-1} \\
&\qquad\qquad (I + \tfrac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\tilde{\Lambda}_{\epsilon,R}^{-1/2} \\
&\quad = (I + \tfrac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1} \\
&\quad + \sum_{j=1}^{\infty}\Bigg[(-1)^j\tilde{\Lambda}_{\epsilon,R}^{1/2}(I + \tfrac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\Big(\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\tilde{\Lambda}_{\epsilon,R}^{1/2}(\Phi_R^T\Phi_R - nI)\tilde{\Lambda}_{\epsilon,R}^{1/2} \\
&\qquad\qquad (I + \tfrac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\Big)^j(I + \tfrac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\tilde{\Lambda}_{\epsilon,R}^{-1/2}\Bigg] \\
&\quad = (I + \tfrac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1} + \sum_{j=1}^{\infty}(-1)^j\Big(\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\tilde{\Lambda}_{\epsilon,R}(\Phi_R^T\Phi_R - nI)\Big)^j(I + \tfrac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}.
\end{aligned}
$$

Letting $\epsilon \to 0$, we get

$$
\begin{aligned}
&(I + \tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1} \\
=&(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1} + \sum_{j=1}^{\infty}(-1)^j\left(\frac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\right)^j(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}.
\end{aligned}
$$

This concludes the proof. □

Lemma 74. *If* $\|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2 < 1$, *then we have*

$$
(I + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - (I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} = \sum_{j=1}^{\infty}(-1)^j\Big((I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\Big)^j(I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}. \tag{3.53}
$$

In particular, assume that $\sigma^2 = \Theta(1)$. *Let* $R = n^{\frac{1}{\alpha}+\kappa}$ *where* $0 < \kappa < \frac{\alpha-1-2\tau}{\alpha(1+2\tau)}$. *Then with probability of at least* $1-\delta$, *for sufficiently large* n, *we have* $\|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2 < 1$ *and* (3.53) *holds.*

Proof of Lemma 74. Define $\Phi_{>R} = (\phi_{R+1}(\mathbf{x}), \phi_{R+2}(\mathbf{x}), \ldots)$, $\Lambda_{>R} = \mathrm{diag}(\lambda_{R+1}, \lambda_{R+2}, \ldots)$. Then we have

$$\begin{aligned}
&(I + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - (I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} \\
=&(I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2} + \tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2})^{-1} - (I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} \\
=&\bigg(\Big(I + (I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\Big)^{-1} - I\bigg)(I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}.
\end{aligned}$$

By Corollary 72, for sufficiently large n, $\|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2 < 1$ with probability of at least $1 - \delta$. Hence

$$\begin{aligned}
&(I + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - (I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} \\
=&\bigg(\Big(I + (I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\Big)^{-1} - I\bigg)(I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} \\
=&\sum_{j=1}^{\infty}(-1)^j\Big((I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\Big)^j(I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}.
\end{aligned}$$

□

Lemma 75. *Assume that $\mu_0 = 0$ and $\sigma^2 = \Theta(n^t)$ where $1 - \frac{\alpha}{1+2\tau} < t < 1$. Let $R = n^{(\frac{1}{\alpha}+\kappa)(1-t)}$ where $0 < \kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{\alpha^2(1-t)}$. Then when n is sufficiently large, with probability of at least $1 - 2\delta$ we have*

$$\|(I + \tfrac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}f_R(\mathbf{x})\|_2 = \tilde{O}\bigg(\sqrt{(\tfrac{1}{\delta}+1)n}\cdot n^{\max\{-(1-t), \frac{(1-2\beta)(1-t)}{2\alpha}\}}\bigg). \tag{3.54}$$

Proof of Lemma 75. Let $\Lambda_{1:R} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_R\}$, $\Phi_{1:R} = (\phi_1(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_R(\mathbf{x}))$ and $\boldsymbol{\mu}_{1:R} = (\mu_1, \ldots, \mu_R)$. Since $\mu_0 = 0$, we have

$$(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}f_R(\mathbf{x}) = (I + \frac{1}{\sigma^2}\Phi_{1:R}\Lambda_{1:R}\Phi_{1:R}^T)^{-1}\Phi_{1:R}\boldsymbol{\mu}_{1:R}. \tag{3.55}$$

Using the Woodbury matrix identity, we have that

$$\begin{aligned}(I+\tfrac{1}{\sigma^2}\Phi_{1:R}\Lambda_{1:R}\Phi_{1:R}^T)^{-1}\Phi_{1:R}\boldsymbol{\mu}_{1:R} =&\big[I-\Phi_{1:R}(\sigma^2 I+\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\Lambda_{1:R}\Phi_{1:R}^T\big]\Phi_{1:R}\boldsymbol{\mu}_{1:R}\\ =&\Phi_{1:R}\boldsymbol{\mu}_{1:R}-\Phi_{1:R}(\sigma^2 I+\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R}\boldsymbol{\mu}_{1:R}\\ =&\Phi_{1:R}(I+\tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}.\end{aligned} \tag{3.56}$$

Let $A=(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1/2}\Lambda_{1:R}^{1/2}(\Phi_{1:R}^T\Phi_{1:R}-nI)\Lambda_{1:R}^{1/2}(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1/2}$.By Corollary 68, with probability of at least $1-\delta$, we have $\|\frac{1}{\sigma^2}A\|_2=\sqrt{\log\frac{R}{\delta}}n^{\frac{1-\alpha+2\tau}{2\alpha}-\frac{(1+2\tau)t}{2\alpha}}$. When n is sufficiently large, $\|\frac{1}{\sigma^2}A\|_2=o(1)$ is less than 1 because $1-\frac{\alpha}{1+2\tau}<t<1$. By Lemma 73, we have

$$\begin{aligned}&(I+\tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\\ =&(I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}+\sum_{j=1}^{\infty}(-1)^j\big(\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}(\Phi_{1:R}^T\Phi_{1:R}-nI)\big)^j(I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}.\end{aligned}$$

We then have

$$\begin{aligned}&\|(I+\frac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2\\ =&\left\|\left((I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}+\sum_{j=1}^{\infty}(-1)^j\big(\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}(\Phi_{1:R}^T\Phi_{1:R}-nI)\big)^j(I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\right)\boldsymbol{\mu}_{1:R}\right\|_2\\ \le&\|(I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2+\sum_{j=1}^{\infty}\left\|\big(\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}(\Phi_{1:R}^T\Phi_{1:R}-nI)\big)^j(I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\right\|_2.\end{aligned} \tag{3.57}$$

By Lemma 61 and Assumption 51, assuming that $\sup_{i\ge1}p_{i+1}-p_i=h$, we have

$$\begin{aligned}\|(I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2&\le\sqrt{\sum_{p=1}^{R}\frac{C_\mu^2p^{-2\beta}}{(1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2)^2}}=\Theta(n^{\max\{-(1-t),\frac{(1-2\beta)(1-t)}{2\alpha}\}}\log^{k/2}n),\\ \|(I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2&\ge\sqrt{\sum_{i=1}^{\lfloor\frac{R}{h}\rfloor}\frac{\underline{C_\mu^2}i^{-2\beta}}{(1+\frac{n}{\sigma^2}\overline{C_\lambda}(hi)^{-\alpha})^2}}=\Theta(n^{\max\{-(1-t),\frac{(1-2\beta)(1-t)}{2\alpha}\}}\log^{k/2}n)\end{aligned}$$

where $k = \begin{cases} 0, & 2\alpha \neq 2\beta - 1, \\ 1, & 2\alpha = 2\beta - 1. \end{cases}$. Overall we have

$$\|(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2 = \Theta(n^{(1-t)\max\{-1, \frac{1-2\beta}{2\alpha}\}} \log^{k/2} n). \tag{3.58}$$

Using the fact that $\|\frac{1}{\sigma^2}A\|_2 = \sqrt{\log \frac{R}{\delta}} n^{\frac{1-\alpha+2\tau}{2\alpha} - \frac{(1+2\tau)t}{2\alpha}}$ and $\|(I + \frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}\|_2 \leq n^{-1}$, we have

$$\begin{aligned} &\left\| \left(\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}(\Phi_{1:R}^T\Phi_{1:R} - nI)\right)^j (I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R} \right\|_2 \\ =& \left\| (I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-\frac{1}{2}}\Lambda_{1:R}^{\frac{1}{2}}(\tfrac{1}{\sigma^2}A)^j (I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-\frac{1}{2}}\Lambda_{1:R}^{\frac{1}{2}}\boldsymbol{\mu}_{1:R} \right\|_2 \\ \leq& \tilde{O}(n^{-\frac{1-t}{2}})\|\tfrac{1}{\sigma^2}A\|_2^j \|(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-\frac{1}{2}}\Lambda_{1:R}^{-\frac{1}{2}}\boldsymbol{\mu}_{1:R}\|_2 \end{aligned} \tag{3.59}$$

By Lemma 62 and the assumption $R = n^{(\frac{1}{\alpha}+\kappa)(1-t)}$,

$$\begin{aligned} \|(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-\frac{1}{2}}\Lambda_{1:R}^{-\frac{1}{2}}\boldsymbol{\mu}_{1:R}\|_2 \leq& \sqrt{\sum_{p=1}^{R} \frac{(\underline{C_\lambda} p^{-\alpha})^{-1} C_\mu^2 p^{-2\beta}}{(1 + n\underline{C_\lambda} p^{-\alpha}/\sigma^2)^1}} \\ =& \tilde{O}(\max\{n^{-(1-t)/2}, R^{1/2-\beta+\alpha/2}\}) \\ =& \tilde{O}(\max\{n^{-(1-t)/2}, n^{(\frac{1}{2}+\frac{1-2\beta}{2\alpha}+\kappa(1/2-\beta+\alpha/2))(1-t)}\}) \end{aligned} \tag{3.60}$$

We then have

$$\begin{aligned} &\left\| \left(\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}(\Phi_{1:R}^T\Phi_{1:R} - nI)\right)^j (I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R} \right\|_2 \\ =& \|\tfrac{1}{\sigma^2}A\|_2^j \tilde{O}(\max\{n^{-(1-t)}, n^{(\frac{1-2\beta}{2\alpha}+\kappa(1/2-\beta+\alpha/2))(1-t)}\}) \end{aligned} \tag{3.61}$$

By (3.57), (3.58) and (3.61), we have

$$\begin{aligned} &\|(I + \tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2 \\ =& \Theta(n^{(1-t)\max\{-1, \frac{1-2\beta}{2\alpha}\}} \log^{k/2} n) + \sum_{j=1}^{\infty} \|\frac{1}{\sigma^2}A\|_2^j \tilde{O}(\max\{n^{-(1-t)}, n^{(1-t)\frac{1-2\beta}{2\alpha}+\kappa(1-t)(1/2-\beta+\alpha/2)}\}) \\ =& \Theta(n^{(1-t)\max\{-1, \frac{1-2\beta}{2\alpha}\}} \log^{k/2} n) + \tilde{O}(n^{\frac{1-\alpha+2\tau}{2\alpha} - \frac{(1+2\tau)t}{2\alpha}}) \tilde{O}(\max\{n^{-(1-t)}, n^{(1-t)\frac{1-2\beta}{2\alpha}+\kappa(1-t)(1/2-\beta+\alpha/2)}\}). \end{aligned} \tag{3.62}$$

By assumption $\kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{\alpha^2(1-t)}$, we have that

$$\begin{aligned}&\kappa(1-t)(1/2-\beta+\alpha/2)+\frac{1-\alpha+2\tau}{2\alpha}-\frac{(1+2\tau)t}{2\alpha}\\&<\kappa\alpha(1-t)/2+\frac{1-\alpha+2\tau}{2\alpha}-\frac{(1+2\tau)t}{2\alpha}<0.\end{aligned}$$

Using (3.62), we then get

$$\begin{aligned}\|(I+\tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2&=\Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n)\\&=\frac{1+o(1)}{\sigma^2}\|(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2.\end{aligned}\tag{3.63}$$

By Corollary 66, with probability of at least $1-\delta$, we have

$$\begin{aligned}\|\Phi_{1:R}(I+\tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2=&\tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n}\|(I+\tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2)\\=&\tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n}\cdot n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}).\end{aligned}\tag{3.64}$$

From (3.56), we get $\|(I+\frac{1}{\sigma^2}\Phi_{1:R}\Lambda_{1:R}\Phi_{1:R}^T)^{-1}\Phi_{1:R}\boldsymbol{\mu}_{1:R}\|_2=\tilde{O}(\sqrt{(\frac{1}{\delta}+1)n}\cdot n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}})$. This concludes the proof. □

Lemma 76. *Assume that $\mu_0>0$ and $\sigma^2=\Theta(n^t)$ where $1-\frac{\alpha}{1+2\tau}<t<1$. Let $R=n^{\frac{1}{\alpha}+\kappa}$ where $0<\kappa<\frac{\alpha-1-2\tau+(1+2\tau)t}{\alpha^2}$. Then when n is sufficiently large, with probability of at least $1-2\delta$, we have*

$$\|(I+\tfrac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}f_R(\mathbf{x})\|_2=\tilde{O}\bigg(\sqrt{(\tfrac{1}{\delta}+1)n}\bigg).\tag{3.65}$$

Proof of Lemma 76. Using the Woodbury matrix identity, we have that

$$\begin{aligned}(I+\tfrac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}f_R(\mathbf{x})=&\big[I-\Phi_R(\sigma^2I+\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T\big]\Phi_R\boldsymbol{\mu}_R\\=&\Phi_R\boldsymbol{\mu}_R-\Phi_R(\sigma^2I+\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T\Phi_R\boldsymbol{\mu}_R\\=&\Phi_R(I+\tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R.\end{aligned}\tag{3.66}$$

Let $\boldsymbol{\mu}_{R,1} = (\mu_0, 0, \ldots, 0)$ and $\boldsymbol{\mu}_{R,2} = (0, \mu_1, \ldots, \mu_R)$. Then $\boldsymbol{\mu}_R = \boldsymbol{\mu}_{R,1} + \boldsymbol{\mu}_{R,2}$. Then we have

$$\|(I + \tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R\|_2 = \|(I + \tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_{R,1}\|_2 + \|(I + \tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_{R,2}\|_2. \tag{3.67}$$

According to (3.63) in the proof of Lemma 75, we have $\|(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_{R,2}\|_2 = \tilde{O}(n^{\max\{-(1-t), \frac{(1-t)(1-2\beta)}{2\alpha}\}})$. Next we estimate $\|(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_{R,1}\|_2$.

Let

$$A = (I + \frac{n}{\sigma^2}\Lambda_{1:R})^{-\gamma/2}\Lambda_{1:R}^{\gamma/2}(\Phi_{1:R}^T\Phi_{1:R} - nI)\Lambda_{1:R}^{\gamma/2}(I + \frac{n}{\sigma^2}\Lambda_{1:R})^{-\gamma/2}$$

where $\frac{1}{1-t}(\frac{1+\alpha+2\tau}{2\alpha} - \frac{(1+2\tau+2\alpha)t}{2\alpha}) < \gamma < 1$. Since $1 - \frac{\alpha}{1+2\tau} < t < 1$, $\frac{1}{1-t}(\frac{1+\alpha+2\tau}{2\alpha} - \frac{(1+2\tau+2\alpha)t}{2\alpha}) < 1$ so the range for γ is well-defined.By Corollary 68, with probability of at least $1-\delta$, we have $\|\frac{1}{\sigma^2}A\|_2 = \tilde{O}(\sqrt{\log\frac{R}{\delta}}n^{\frac{1+\alpha+2\tau}{2\alpha} - \frac{(1+2\tau+2\alpha)t}{2\alpha} - \gamma(1-t)}) = o(1)$. When n is sufficiently large, $\|\frac{1}{\sigma^2}A\|_2$ is less than 1 because $1 - \frac{\alpha}{1+2\tau} < t < 1$. By Lemma 73, we have

$$\begin{aligned}
&(I + \tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1} \\
=&(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1} + \sum_{j=1}^{\infty}(-1)^j\big(\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\big)^j(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}.
\end{aligned}$$

We then have

$$\begin{aligned}
&\|(I + \tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_{R,1}\|_2 \\
=&\left\|\left((I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1} + \sum_{j=1}^{\infty}(-1)^j\big(\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\big)^j(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\right)\boldsymbol{\mu}_{R,1}\right\|_2 \\
\leq&\left(\|(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_{R,1}\|_2 + \sum_{j=1}^{\infty}\left\|\big(\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\big)^j(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_{R,1}\right\|_2\right).
\end{aligned} \tag{3.68}$$

By Lemma 61,

$$\|(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_{R,1}\|_2 \leq \sqrt{\mu_0^2 + \sum_{p=1}^{R}\frac{C_\mu^2 p^{-2\beta}}{(1 + n\underline{C_\lambda}p^{-\alpha}/\sigma^2)^2}} = O(1). \tag{3.69}$$

Let $\tilde{\Lambda}_{1,R} = \text{diag}\{1, \lambda_1, \ldots, \lambda_R\}$ and $I_{0,R} = (0, 1, \ldots, 1)$. Then $\Lambda_R = \tilde{\Lambda}_{1,R} I_{0,R}$. Let $B = (I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\tilde{\Lambda}_{1,R}^{\gamma/2}(\Phi_R^T\Phi_R - nI)\tilde{\Lambda}_{1,R}^{\gamma/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}$. According to Corollary 69, we have $\|B\|_2 = O\big(\sqrt{\log \frac{R}{\delta}} n^{\frac{1}{2}}\big)$. Using the fact that $\|\frac{1}{\sigma^2}A\|_2 = \tilde{O}\big(\sqrt{\log \frac{R}{\delta}} n^{\frac{1+\alpha+2\tau}{2\alpha} - \frac{(1+2\tau+2\alpha)t}{2\alpha} - \gamma(1-t)}\big)$, we have

$$
\begin{aligned}
&\Big\| \big(\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\big)^j (I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_{R,1} \Big\|_2 \\
&= \frac{1}{\sigma^{2j}} \Big\| (I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}} \Lambda_R^{1-\frac{\gamma}{2}} \big(A(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1+\gamma}\Lambda_R^{1-\gamma}\big)^{j-1} B (I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}} \boldsymbol{\mu}_{R,1} \Big\|_2 \\
&\leq \frac{1}{\sigma^2}(n^{(-1+\frac{\gamma}{2}+(-1+\gamma)(j-1))(1-t)}\tilde{O}(\sqrt{\log \tfrac{R}{\delta}} n^{(j-1)(\frac{1+\alpha+2\tau}{2\alpha} - \frac{(1+2\tau+2\alpha)t}{2\alpha} - \gamma(1-t))}\sqrt{\log \tfrac{R}{\delta}} n^{\frac{1}{2}} \|\boldsymbol{\mu}_{R,1}\|_2 \\
&\leq n^{(-1+\frac{\gamma}{2})(1-t)+\frac{1}{2}-t}\tilde{O}(n^{\frac{[1-\alpha+2\tau-(1+2\tau)t](j-1)}{2\alpha}})\sqrt{\log \tfrac{R}{\delta}}\|\boldsymbol{\mu}_{R,1}\|_2 \\
&= \tilde{O}(n^{-\frac{1}{2}+\frac{\gamma}{2}(1-t)+\frac{[1-\alpha+2\tau-(1+2\tau)t](j-1)}{2\alpha}}).
\end{aligned}
\tag{3.70}
$$

Since $\frac{1}{1-t}(\frac{1+\alpha+2\tau}{2\alpha} - \frac{(1+2\tau+2\alpha)t}{2\alpha}) < \gamma < 1$ and $-\frac{1}{2} + \frac{1}{1-t}(\frac{1+\alpha+2\tau}{2\alpha} - \frac{(1+2\tau+2\alpha)t}{2\alpha})\frac{1-t}{2} < 0$, we can let γ be a little bit larger than $\frac{1}{1-t}(\frac{1+\alpha+2\tau}{2\alpha} - \frac{(1+2\tau+2\alpha)t}{2\alpha})$ and make $-\frac{1}{2} + \frac{\gamma}{2}(1-t) < 0$ holds. By (3.68), (3.69), (3.70), we have

$$
\begin{aligned}
&\|(I + \tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_{R,1}\|_2 \\
&\leq O(1) + \sum_{j=1}^{\infty} \tilde{O}(n^{-\frac{1}{2}+\frac{\gamma}{2}(1-t)+\frac{[1-\alpha+2\tau-(1+2\tau)t](j-1)}{2\alpha}}) \\
&\leq O(1) + o(1) = O(1).
\end{aligned}
\tag{3.71}
$$

According to (3.67), we have $\|(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R\|_2 = \tilde{O}(n^{\max\{-(1-t), \frac{(1-t)(1-2\beta)}{2\alpha}\}}) + O(1) = O(1)$. By Corollary 66, with probability of at least $1 - \delta$, we have

$$
\begin{aligned}
\|\Phi_R(I + \tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R\|_2 =& \tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n}\|(I + \tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R\|_2) \\
=& \tilde{O}\Big(\sqrt{(\tfrac{1}{\delta}+1)n}\Big).
\end{aligned}
$$

From (3.66), we get $\|(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}f_R(\mathbf{x})\|_2 = \tilde{O}\Big(\sqrt{(\frac{1}{\delta}+1)n}\Big)$. This concludes the proof. □

Lemma 77. *Assume that $\sigma^2 = \Theta(1)$. Let $R = n^{\frac{1}{\alpha}+\kappa}$ where $0 < \kappa < \frac{\alpha-1-2\tau}{\alpha^2}$. Assume that $\mu_0 = 0$. Then when n is sufficiently large, with probability of at least $1 - 3\delta$ we have*

$$\|(I + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} f_R(\mathbf{x})\|_2 = \tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n} \cdot n^{\max\{-1, \frac{1-2\beta}{2\alpha}\}}). \tag{3.72}$$

Assume that $\mu_0 > 0$. Then when n is sufficiently large, with probability of at least $1 - 3\delta$ we have

$$\|(I + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} f_R(\mathbf{x})\|_2 = \tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n}). \tag{3.73}$$

Proof of Lemma 77. We have

$$\begin{aligned} &(I + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} f_R(\mathbf{x}) \\ =&(I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} f_R(\mathbf{x}) + \Big((I + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - (I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\Big) f_R(\mathbf{x}). \end{aligned} \tag{3.74}$$

When $\mu_0 = 0$, by Lemma 75, with probability of at least $1 - 2\delta$, we have

$$\|(I + \tfrac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1} f_R(\mathbf{x})\|_2 = \tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n} \cdot n^{\max\{-1, \frac{1-2\beta}{2\alpha}\}}).$$

Since $\frac{\alpha-1-2\tau}{\alpha^2} < \frac{\alpha-1-2\tau}{\alpha(1+2\tau)}$, we apply Lemma 74 and Corollary 72 and get that with probability of at least $1 - \delta$, the second term in the right hand side of (3.74) is estimated as follows:

$$\begin{aligned} &\|\Big((I + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - (I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\Big) f_R(\mathbf{x})\|_2 \\ &= \|\sum_{j=1}^{\infty} (-1)^j \Big((I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} \tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\Big)^j (I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} f_R(\mathbf{x})\|_2 \\ &= \sum_{j=1}^{\infty} \Big\|\Big((I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} \tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\Big)\Big\|_2^j \|(I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} f_R(\mathbf{x})\|_2 \\ &= \sum_{j=1}^{\infty} \tilde{O}(n^{-j\kappa\alpha}) \tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n} \cdot n^{\max\{-1, \frac{1-2\beta}{2\alpha}\}}) \\ &= o(\sqrt{(\tfrac{1}{\delta}+1)n} \cdot n^{\max\{-1, \frac{1-2\beta}{2\alpha}\}}). \end{aligned}$$

Overall, from (3.74), we have that with probability $1-3\delta$,

$$\|(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2 = \tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n}\cdot n^{\max\{-1,\frac{1-2\beta}{2\alpha}\}}).$$

When $\mu_0 > 0$, using the same approach and Lemma 76, we can prove that $\|(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2 = \tilde{O}(\sqrt{(\frac{1}{\delta}+1)n})$. This concludes the proof. □

3.D Proof of the Main Results

3.D.1 Proofs Related to the Asymptotics of the Normalized Stochastic Complexity

Lemma 78. *Under Assumptions 50, 51 and 52, with probability of at least $1-2\delta$ we have, we have*

$$|T_{1,R}(D_n) - T_1(D_n)| = \tilde{O}\big(\tfrac{1}{\sigma^2}(nR^{1-\alpha} + n^{1/2}R^{1-\alpha+\tau} + R^{1-\alpha+2\tau})\big) \tag{3.75}$$

If $R = n^{\frac{1}{\alpha}+\kappa}$ where $\kappa > 0$, we have $|T_{1,R}(D_n) - T_1(D_n)| = o\Big(\frac{1}{\sigma^2}n^{\frac{1}{\alpha}}\Big)$. If we further assume that $0 < \kappa < \frac{\alpha-1-2\tau}{\alpha^2}$, $\mu_0 = 0$ and $\sigma^2 = \Theta(1)$, then for sufficiently large n with probability of at least $1-4\delta$ we have

$$|T_{2,R}(D_n) - T_2(D_n)| = \tilde{O}\Big((\tfrac{1}{\delta}+1)n^{\max\{(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2},1+\frac{1-2\beta}{\alpha}+\frac{(1-2\beta)\kappa}{2},-1-\kappa\alpha,1+\frac{1-2\beta}{\alpha}-\kappa\alpha\}}\Big). \tag{3.76}$$

Proof of Lemma 78. Define $\Phi_{>R} = (\phi_{R+1}(\mathbf{x}), \phi_{R+2}(\mathbf{x}), \ldots, \phi_p(\mathbf{x}), \ldots)$, and $\Lambda_{>R} = \mathrm{diag}(\lambda_{R+1}, \ldots, \lambda_p, \ldots)$. We then have

$$\begin{aligned}|T_1(D_n) - T_{1,R}(D_n)| = &\left|\frac{1}{2}\log\det(I+\frac{1}{\sigma^2}\Phi\Lambda\Phi^T) - \frac{1}{2}\log\det(I+\frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)\right| \\ &+ \frac{1}{2}\left|\mathrm{Tr}(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - \mathrm{Tr}(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\right|.\end{aligned} \tag{3.77}$$

As for the first term in the right hand side of (3.77), we have

$$\begin{aligned}
&\left|\frac{1}{2}\log\det(I+\frac{1}{\sigma^2}\Phi\Lambda\Phi^T)-\frac{1}{2}\log\det(I+\frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)\right| \\
=&\left|\frac{1}{2}\log\det\left((I+\frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}(I+\frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T+\frac{1}{\sigma^2}\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T)\right)\right| \\
=&\left|\frac{1}{2}\log\det\left(I+\frac{1}{\sigma^2}(I+\frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\right)\right| \\
=&\frac{1}{2}\left|\mathrm{Tr}\log\left(I+\frac{1}{\sigma^2}(I+\frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\right)\right|.
\end{aligned} \tag{3.78}$$

Given a concave function h and a matrix $B\in\mathbb{R}^{n\times n}$ whose eigenvalues $\zeta_1,\ldots,\zeta_n$ are all positive, we have that

$$\mathrm{Tr}\,h(B)=\textstyle\sum_{p=1}^n h(\zeta_i)\le nh(\frac{1}{n}\sum_{p=1}^n\zeta_i)\le nh(\frac{1}{n}\,\mathrm{Tr}\,B), \tag{3.79}$$

where we used Jensen's inequality. Using $h(x)=\log(1+x)$ in (3.79), with probability $1-\delta$, we have

$$\begin{aligned}
&\left|\tfrac{1}{2}\log\det(I+\tfrac{1}{\sigma^2}\Phi\Lambda\Phi^T)-\tfrac{1}{2}\log\det(I+\tfrac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)\right| \\
&\le\tfrac{n}{2}\log(1+\tfrac{1}{n}\,\mathrm{Tr}(\tfrac{1}{\sigma^2}(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T)) \\
&\le\tfrac{n}{2}\log(1+\tfrac{1}{n\sigma^2}\|(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\|_2\,\mathrm{Tr}(\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T)) \\
&\le\tfrac{n}{2}\log(1+\tfrac{1}{n\sigma^2}\textstyle\sum_{p=R+1}^\infty\lambda_p\|\phi_p(\mathbf{x})\|_2^2)\le\frac{1}{2\sigma^2}\sum_{p=R+1}^\infty\lambda_p\|\phi_p(\mathbf{x})\|_2^2 \\
&=\tfrac{1}{2\sigma^2}\textstyle\sum_{p=R+1}^\infty\lambda_p\Big(C_\phi^2\tilde{O}\Big(\sqrt{p^{2\tau}n\|\phi_p\|_2^2}+p^{2\tau}\Big)+n\|\phi_p\|_2^2\Big) \\
&=\tilde{O}(\tfrac{1}{\sigma^2}n\textstyle\sum_{p=R+1}^\infty\lambda_p+n^{1/2}\sum_{p=R+1}^\infty\lambda_p p^\tau+\sum_{p=R+1}^\infty\lambda_p p^{2\tau}) \\
&=\tilde{O}\big(\tfrac{1}{\sigma^2}(nR^{1-\alpha}+n^{1/2}R^{1-\alpha+\tau}+R^{1-\alpha+2\tau})\big)=o\Big(\tfrac{1}{\sigma^2}n^{\frac{1}{\alpha}}\Big),
\end{aligned} \tag{3.80}$$

where in the second inequality we use the fact that $\mathrm{Tr}\,AB\le\|A\|_2\,\mathrm{Tr}\,B$ when A and B are symmetric positive definite matrices, and in the last inequality we use Lemma 64.

As for the second term in the right hand side of (3.77), let $A=(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1/2}$. Then

we have

$$
\begin{aligned}
&\tfrac{1}{2}\Big|\mathrm{Tr}(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}-\mathrm{Tr}(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\Big| \\
&=\tfrac{1}{2}\bigg|\mathrm{Tr}A\bigg[I-(I+A(\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2})A)^{-1}\bigg]A\bigg| \\
&\le \tfrac{1}{2}\,\mathrm{Tr}\Big[I-(I+A(\tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2})A)^{-1}\Big] \\
&\le \tfrac{n}{2}(1-(1+\tfrac{1}{n}\,\mathrm{Tr}\,A(\tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2})A)^{-1}) \le \tfrac{n}{2}(1-(1+\tfrac{1}{n}\,\mathrm{Tr}(\tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}))^{-1}) \\
&\le \tfrac{n}{2}(1-(1+\tfrac{1}{n\sigma^2}\textstyle\sum_{p=R+1}^{\infty}\lambda_p\|\phi_p(\mathbf{x})\|_2^2))^{-1}) \le \tfrac{1}{2\sigma^2}\textstyle\sum_{p=R+1}^{\infty}\lambda_p\|\phi_p(\mathbf{x})\|_2^2 \\
&=\tilde{O}\big(\tfrac{1}{\sigma^2}(nR^{1-\alpha}+n^{1/2}R^{1-\alpha+\tau}+R^{1-\alpha+2\tau})\big)=o\bigg(\tfrac{1}{\sigma^2}n^{\frac{1}{\alpha}}\bigg),
\end{aligned}
$$

where in the first inequality we use the fact that $\|A\|_2 < 1$ and $\mathrm{Tr}\,ABA \le \|A\|_2^2\,\mathrm{Tr}\,B$ when A and B are symmetric positive definite matrices, in the second inequality we use $h(x) = 1 - 1/(1+x)$ in (3.79) and in the last equality we use the last few steps of (3.80). This concludes the proof of the first statement.

As for $|T_2(D_n) - T_{2,R}(D_n)|$, we have

$$
\begin{aligned}
|T_2(D_n)-T_{2,R}(D_n)| = &\Big|f(\mathbf{x})^T(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x})-f_R(\mathbf{x})^T(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\Big| \\
&+\Big|f_R(\mathbf{x})^T(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})-f_R(\mathbf{x})^T(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\Big|.
\end{aligned}
\tag{3.81}
$$

For the first term on the right-hand side of (3.81), we have

$$
\begin{aligned}
&\Big|f(\mathbf{x})^T(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x})-f_R(\mathbf{x})^T(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\Big| \\
&\le 2\Big|f_{>R}(\mathbf{x})^T(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\Big|+\Big|f_{>R}(\mathbf{x})^T(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_{>R}(\mathbf{x})\Big| \\
&\le 2\|f_{>R}(\mathbf{x})\|_2\|(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2+\|f_{>R}(\mathbf{x})\|_2\|(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}\|_2\|f_{>R}(\mathbf{x})\|_2 \\
&\le 2\|f_{>R}(\mathbf{x})\|_2\|(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2+\|f_{>R}(\mathbf{x})\|_2^2.
\end{aligned}
$$

Applying Corollary 65 and Lemma 77, with probability of at least $1-4\delta$, we have

$$\begin{aligned}
&\left|f(\mathbf{x})^T(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x}) - f_R(\mathbf{x})^T(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
&\leq 2\tilde{O}\left(\sqrt{(\tfrac{1}{\delta}+1)nR^{1-2\beta}}\right)\tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n}\cdot n^{\max\{-1,\frac{1-2\beta}{2\alpha}\}}) + \tilde{O}((\tfrac{1}{\delta}+1)nR^{1-2\beta}) \\
&= 2\tilde{O}\left((\tfrac{1}{\delta}+1)n^{1+(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}+\max\{-1,\frac{1-2\beta}{2\alpha}\}}\right) + \tilde{O}((\tfrac{1}{\delta}+1)n^{1+(\frac{1}{\alpha}+\kappa)(1-2\beta)}) \\
&= 2\tilde{O}\left((\tfrac{1}{\delta}+1)n^{1+(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}+\max\{-1,\frac{1-2\beta}{2\alpha}\}}\right),
\end{aligned}$$

where the last equality holds because $(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2} < \frac{1-2\beta}{2\alpha}$ when $\kappa > 0$.

As for the second term on the right-hand side of (3.81), according to Lemma 74, Corollary 72 and Lemma 75, we have

$$\begin{aligned}
&\left|f_R(\mathbf{x})^T(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x}) - f_R(\mathbf{x})^T(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
&= \left|\sum_{j=1}^{\infty}(-1)^j f_R(\mathbf{x})^T\left((I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\right)^j(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
&\leq \sum_{j=1}^{\infty}\|(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\|_2^{j-1}\cdot\|\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2^j\cdot\|(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2^2 \qquad (3.82) \\
&= \sum_{j=1}^{\infty}\tilde{O}(n^{-j\kappa\alpha})\tilde{O}((\tfrac{1}{\delta}+1)n^{1+\max\{-2,\frac{1-2\beta}{\alpha}\}}) \\
&= \tilde{O}((\tfrac{1}{\delta}+1)n^{1+\max\{-2,\frac{1-2\beta}{\alpha}\}-\kappa\alpha}).
\end{aligned}$$

By (3.81), we have

$$\begin{aligned}
|T_2(D_n) - T_{2,R}(D_n)| &= \tilde{O}\left((\tfrac{1}{\delta}+1)n^{1+(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}+\max\{-1,\frac{1-2\beta}{2\alpha}\}}\right) \\
&\quad + \tilde{O}\left((\tfrac{1}{\delta}+1)n^{1+\max\{-2,\frac{1-2\beta}{\alpha}\}-\kappa\alpha}\right) \\
&= \tilde{O}\left((\tfrac{1}{\delta}+1)n^{\max\{(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2},1+\frac{1-2\beta}{\alpha}+\frac{(1-2\beta)\kappa}{2},-1-\kappa\alpha,1+\frac{1-2\beta}{\alpha}-\kappa\alpha\}}\right).
\end{aligned}$$

This concludes the proof of the second statement. □

In Lemma 78, we gave a bound for $|T_{2,R}(D_n) - T_2(D_n)|$ when $n^{\frac{1}{\alpha}} < R < n^{\frac{1}{\alpha}+\frac{\alpha-1-2\tau}{\alpha^2}}$. For $R > n$, we note the following lemma:

Lemma 79. *Let $R = n^C$ and $\sigma^2 = n^t$. Assume that $C \geq 1$ and $C(1-\alpha+2\tau)-t < 0$. Under Assumptions 50, 51 and 52, for sufficiently large n and with probability of at least $1-3\delta$ we have*

$$|T_{2,R}(D_n) - T_2(D_n)| = \tilde{O}\big((\tfrac{1}{\delta}+1)\tfrac{1}{\sigma^2} n R^{\max\{1/2-\beta, 1-\alpha+2\tau\}}\big). \tag{3.83}$$

Proof of Lemma 79. Define $\Phi_{>R} = (\phi_{R+1}(\mathbf{x}), \phi_{R+2}(\mathbf{x}), \ldots, \phi_p(\mathbf{x}), \ldots)$, and $\Lambda_{>R} = \mathrm{diag}(\lambda_{R+1}, \ldots, \lambda_p, \ldots)$. Then we have

$$\begin{aligned}|T_2(D_n) - T_{2,R}(D_n)| = &\left| f(\mathbf{x})^T (I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} f(\mathbf{x}) - f_R(\mathbf{x})^T (I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} f_R(\mathbf{x}) \right| \\ &+ \left| f_R(\mathbf{x})^T (I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} f_R(\mathbf{x}) - f_R(\mathbf{x})^T (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} f_R(\mathbf{x}) \right|.\end{aligned} \tag{3.84}$$

For the first term on the right-hand side of (3.84), with probability $1-3\delta$ we have

$$\begin{aligned}&\left| f(\mathbf{x})^T (I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} f(\mathbf{x}) - f_R(\mathbf{x})^T (I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} f_R(\mathbf{x}) \right| \\ \leq &2\left| f_{>R}(\mathbf{x})^T (I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} f_R(\mathbf{x}) \right| + \left| f_{>R}(\mathbf{x})^T (I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} f_{>R}(\mathbf{x}) \right| \\ \leq &2\|f_{>R}(\mathbf{x})\|_2 \|(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}\|_2 \|f_R(\mathbf{x})\|_2 + \|f_{>R}(\mathbf{x})\|_2 \|(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}\|_2 \|f_{>R}(\mathbf{x})\|_2 \\ \leq &2\|f_{>R}(\mathbf{x})\|_2 \|f_R(\mathbf{x})\|_2 + \|f_{>R}(\mathbf{x})\|_2^2 \\ \leq &2\tilde{O}\left(\sqrt{(\frac{1}{\delta}+1) n R^{1-2\beta}}\right) \tilde{O}(\sqrt{(\frac{1}{\delta}+1)n} \cdot \|f\|_2) + \tilde{O}((\frac{1}{\delta}+1) n R^{1-2\beta}) \\ = &\tilde{O}\left((\frac{1}{\delta}+1) n R^{1/2-\beta}\right),\end{aligned}$$

where we used Corollary 65 and Lemma 63 for the last inequality.

The assumption $C(1-\alpha+2\tau)-t < 0$ means that $\frac{R^{1-\alpha+2\tau}}{\sigma^2} = o(1)$. For the second term

on the right-hand side of (3.84), by Lemmas 74 and 71, we have

$$
\begin{aligned}
&\left| f_R(\mathbf{x})^T (I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} f_R(\mathbf{x}) - f_R(\mathbf{x})^T (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} f_R(\mathbf{x}) \right| \\
=&\left| \sum_{j=1}^{\infty} (-1)^j f_R(\mathbf{x})^T \left((I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} \frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2} \right)^j (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} f_R(\mathbf{x}) \right| \\
\leq& \sum_{j=1}^{\infty} \|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\|_2^{j+1} \cdot \| \frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2} \|_2^j \cdot \|f_R(\mathbf{x})\|_2^2 \qquad (3.85) \\
=& \sum_{j=1}^{\infty} \tilde{O}(\frac{1}{\sigma^2} R^{j(1-\alpha+2\tau)}) \tilde{O}((\frac{1}{\delta} + 1) n \|f\|_2^2) \\
=& \tilde{O}((\frac{1}{\delta} + 1) \frac{1}{\sigma^2} n R^{1-\alpha+2\tau}).
\end{aligned}
$$

Using (3.84), we have

$$
\begin{aligned}
|T_2(D_n) - T_{2,R}(D_n)| &= \tilde{O}\left((\frac{1}{\delta} + 1) n R^{1/2-\beta} \right) + \tilde{O}((\frac{1}{\delta} + 1) n \frac{1}{\sigma^2} R^{1-\alpha+2\tau}) \\
&= \tilde{O}\left((\frac{1}{\delta} + 1) n \frac{1}{\sigma^2} R^{\max\{1/2-\beta, 1-\alpha+2\tau\}} \right).
\end{aligned}
$$

□

Next we consider the asympototics of $T_{1,R}(D_n)$ and $T_{2,R}(D_n)$.

Lemma 80. *Let* $A = (I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2} \Lambda_R^{\gamma/2} (\Phi_R^T \Phi_R - nI) \Lambda_R^{\gamma/2} (I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}$. *Assume that* $\|A\|_2 < 1$ *where* $\frac{1+2\tau}{\alpha} < \gamma \leq 1$. *Then we have*

$$
T_{2,R}(D_n) = \tfrac{n}{2\sigma^2} \boldsymbol{\mu}_R^T (I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1} \boldsymbol{\mu}_R + \tfrac{1}{2} \textstyle\sum_{j=1}^{\infty} (-1)^{j+1} E_j,
$$

where

$$
E_j = \boldsymbol{\mu}_R^T \tfrac{1}{\sigma^2} (I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1} (\Phi_R^T \Phi_R - nI) \big(\tfrac{1}{\sigma^2} (I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1} \Lambda_R (\Phi_R^T \Phi_R - nI) \big)^{j-1} (I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1} \boldsymbol{\mu}_R.
$$

Proof of Lemma 80. Let $\tilde{\Lambda}_{\epsilon,R} = \mathrm{diag}\{\epsilon, \lambda_1, \ldots, \lambda_R\}$. Since $\Lambda_R = \mathrm{diag}\{0, \lambda_1, \ldots, \lambda_R\}$, we have that when ϵ is sufficiently small, $\| \frac{1}{\sigma^2} (I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2} \tilde{\Lambda}_{\epsilon,R}^{1/2} (\Phi_R^T \Phi_R - nI) \tilde{\Lambda}_{\epsilon,R}^{1/2} (I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2} \|_2 < 1$.

Since all diagonal entries of $\tilde{\Lambda}_{\epsilon,R}$ are positive, we have

$$\begin{aligned}
&\frac{1}{2\sigma^2}\boldsymbol{\mu}_R^T\Phi_R^T(I+\frac{1}{\sigma^2}\Phi_R\tilde{\Lambda}_{\epsilon,R}\Phi_R^T)^{-1}\Phi_R\boldsymbol{\mu}_R \\
&=\frac{1}{2\sigma^2}\boldsymbol{\mu}_R^T\Phi_R^T\Big[I-\Phi_R(\sigma^2 I+\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1}\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Big]\Phi_R\boldsymbol{\mu}_R \\
&=\frac{1}{2\sigma^2}\boldsymbol{\mu}_R^T\Phi_R^T\Phi_R\boldsymbol{\mu}_R-\frac{1}{2\sigma^2}\boldsymbol{\mu}_R^T\Phi_R^T\Phi_R(\sigma^2 I+\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1}\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R\boldsymbol{\mu}_R \\
&=\frac{1}{2}\boldsymbol{\mu}_R^T\Phi_R^T\Phi_R(\sigma^2 I+\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R \\
&=\frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R(\sigma^2 I+\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R \\
&=\frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}\boldsymbol{\mu}_R-\frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}(I+\frac{1}{\sigma^2}\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R.
\end{aligned} \tag{3.86}$$

Using Lemma 73, we have

$$\begin{aligned}
&\frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}\boldsymbol{\mu}_R-\frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}(I+\frac{1}{\sigma^2}\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R \\
&=\frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}\boldsymbol{\mu}_R-\frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\boldsymbol{\mu}_R \\
&+\frac{1}{2}\sum_{j=1}^{\infty}(-1)^{j+1}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}\bigg(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\tilde{\Lambda}_{\epsilon,R}(\Phi_R^T\Phi_R-nI)\bigg)^j(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\boldsymbol{\mu}_R \\
&=\frac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\boldsymbol{\mu}_R \\
&+\frac{1}{2}\sum_{j=1}^{\infty}(-1)^{j+1}\boldsymbol{\mu}_R^T\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}(\Phi_R^T\Phi_R-nI)\bigg(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\tilde{\Lambda}_{\epsilon,R}(\Phi_R^T\Phi_R-nI)\bigg)^{j-1} \\
&\qquad(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\boldsymbol{\mu}_R
\end{aligned} \tag{3.87}$$

Letting $\epsilon \to 0$, we get

$$
\begin{aligned}
&T_{2,R}(D_n) \\
&= \frac{1}{2\sigma^2}\boldsymbol{\mu}_R^T\Phi_R^T(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}\Phi_R\boldsymbol{\mu}_R \\
&= \frac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R \\
&+ \frac{1}{2}\sum_{j=1}^{\infty}\Bigg[(-1)^{j+1}\boldsymbol{\mu}_R^T\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}(\Phi_R^T\Phi_R - nI)\bigg(\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\bigg)^{j-1} \\
&\qquad (I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\Bigg]
\end{aligned}
$$

This concludes the proof. □

Lemma 81. *Assume that $\sigma^2 = \Theta(1)$. Let $R = n^{\frac{1}{\alpha}+\kappa}$ where $0 < \kappa < \frac{\alpha-1-2\tau}{2\alpha^2}$. Under Assumptions 50, 51 and 52, with probability of at least $1-\delta$, we have*

$$T_{1,R}(D_n) = \big(\tfrac{1}{2}\log\det(I + \tfrac{n}{\sigma^2}\Lambda_R) - \tfrac{1}{2}\operatorname{Tr}\big(I - (I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\big)\big)(1 + o(1)) = \Theta(n^{\frac{1}{\alpha}}). \quad (3.88)$$

Furthermore, if we assume $\mu_0 = 0$, we have

$$T_{2,R}(D_n) = \big(\tfrac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\big)(1 + o(1)) = \begin{cases} \Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}), & \alpha \neq 2\beta - 1, \\ \Theta(\log n), & \alpha = 2\beta - 1. \end{cases} \quad (3.89)$$

Proof of Lemma 81. Let

$$A = (I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}, \quad (3.90)$$

where $\frac{1+\alpha+2\tau}{2\alpha} < \gamma \leq 1$. By Corollary 68, with probability of at least $1-\delta$, we have

$$\|A\|_2 = \tilde{O}(n^{\frac{1-2\gamma\alpha+\alpha+2\tau}{2\alpha}}). \quad (3.91)$$

When n is sufficiently large, $\|A\|_2$ is less than 1. Let $B = (I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{1/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}$. Then $\|B\|_2 = \frac{\sigma^{2(1-\gamma)}}{n^{1-\gamma}}\|A\|_2 = \tilde{O}(n^{\frac{1-\alpha+2\tau}{2\alpha}})$. Using the Woodbury matrix

identity, we compute $T_{1,R}(D_n)$ as follows:

$$
\begin{aligned}
&T_{1,R}(D_n)\\
&= \tfrac{1}{2}\log\det(I+\tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R) - \tfrac{1}{2}\mathrm{Tr}\Phi_R(\sigma^2 I+\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T\\
&= \tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R)\\
&\quad + \tfrac{1}{2}\log\det[I+\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}(\Phi_R^T\Phi_R-nI)\Lambda_R^{1/2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}]\\
&\quad - \tfrac{1}{2}\mathrm{Tr}(\sigma^2 I+\Lambda\Phi_R^T\Phi_R)^{-1}\Lambda\Phi_R^T\Phi_R\\
&= \tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R) + \tfrac{1}{2}\,\mathrm{Tr}\log[I+\tfrac{1}{\sigma^2}B] - \tfrac{1}{2}\mathrm{Tr}(I-\sigma^2(\sigma^2 I+\Lambda\Phi_R^T\Phi_R)^{-1}))\\
&= \tfrac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) + \tfrac{1}{2}\,\mathrm{Tr}\sum_{j=1}^{\infty}\tfrac{(-1)^{j-1}}{j}(\tfrac{1}{\sigma^2}B)^j\\
&\quad - \tfrac{1}{2}\mathrm{Tr}\left(I-(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}+\sum_{j=1}^{\infty}(-1)^j\big(\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\big)^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\right)\\
&= \big(\tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R) - \tfrac{1}{2}\mathrm{Tr}\big(I-(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\big)\big) + \tfrac{1}{2}\,\mathrm{Tr}\sum_{j=1}^{\infty}\tfrac{(-1)^{j-1}}{j}(\tfrac{1}{\sigma^2}B)^j\\
&\quad - \tfrac{1}{2}\mathrm{Tr}\left(\sum_{j=1}^{\infty}(-1)^j\tfrac{1}{\sigma^{2j}}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\right),
\end{aligned}
\tag{3.92}
$$

where in the last equality we apply Lemma 73.

Let $h(x)=\log(1+x)-(1-\frac{1}{1+x})$. It is easy to verify that $h(x)$ is increasing on $[0,+\infty)$.

As for the first term on the right hand side of (3.92), we have

$$\begin{aligned}
&\tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R)-\tfrac{1}{2}\operatorname{Tr}\big(I-(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\big)\\
=&\tfrac{1}{2}\sum_{p=1}^{R}\Big(\log(1+\tfrac{n}{\sigma^2}\lambda_p)-(1-\tfrac{1}{1+\frac{n}{\sigma^2}\lambda_p})\Big)\\
=&\tfrac{1}{2}\sum_{p=1}^{R}h(\tfrac{n}{\sigma^2}\lambda_p)\le\tfrac{1}{2}\sum_{p=1}^{R}h(\frac{n}{\sigma^2}\overline{C_\lambda}p^{-\alpha})\\
\le&\tfrac{1}{2}h(\tfrac{n}{\sigma^2}\overline{C_\lambda})+\tfrac{1}{2}\int_{[1,R]}h(\tfrac{n}{\sigma^2}\overline{C_\lambda}x^{-\alpha})\mathrm{d}x\\
=&\tfrac{1}{2}h(\frac{n}{\sigma^2}\overline{C_\lambda})+\tfrac{1}{2}n^{1/\alpha}\int_{[1/n^{1/\alpha},R/n^{1/\alpha}]}h(\tfrac{\overline{C_\lambda}}{\sigma^2}x^{-\alpha})\mathrm{d}x\\
=&\Theta(n^{1/\alpha}),
\end{aligned}$$

where in the last equality we use the fact that $\int_{[0,+\infty]}h(x^{-\alpha})\mathrm{d}x<\infty$. On the other hand, we have

$$\begin{aligned}
&\tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R)-\tfrac{1}{2}\operatorname{Tr}\big(I-(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\big)\\
=&\tfrac{1}{2}\sum_{p=1}^{R}h(\tfrac{n}{\sigma^2}\lambda_p)\ge\tfrac{1}{2}\sum_{p=1}^{R}h(\tfrac{n}{\sigma^2}\underline{C_\lambda}p^{-\alpha})\\
\ge&\tfrac{1}{2}\int_{[1,R+1]}h(\tfrac{n}{\sigma^2}\underline{C_\lambda}x^{-\alpha})\mathrm{d}x\\
=&\tfrac{1}{2}n^{1/\alpha}\int_{[1/n^{1/\alpha},(R+1)/n^{1/\alpha}]}h(\tfrac{1}{\sigma^2}\underline{C_\lambda}x^{-\alpha})\mathrm{d}x\\
=&\Theta(n^{1/\alpha}).
\end{aligned}$$

Overall, we have $\frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R)-\frac{1}{2}\operatorname{Tr}\big(I-(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\big)=\Theta(n^{1/\alpha})$.

As for the second term on the right hand side of (3.92), we have

$$\begin{aligned}
\left|\operatorname{Tr}\sum_{j=1}^{\infty}\tfrac{(-1)^{j-1}}{j}(\tfrac{1}{\sigma^2}B)^j\right|&\le R\sum_{j=1}^{\infty}\|\tfrac{1}{\sigma^2}B\|_2^j=R\sum_{j=1}^{\infty}\tfrac{1}{\sigma^{2j}}\tilde{O}(n^{\frac{j(1-\alpha+2\tau)}{2\alpha}})\\
&=R\tilde{O}(n^{\frac{1-\alpha+2\tau}{2\alpha}})=\tilde{O}(n^{\frac{1}{\alpha}+\kappa+\frac{1-\alpha+2\tau}{2\alpha}}).
\end{aligned}$$

As for the third term on the right hand side of (3.92), we have

$$\begin{aligned}
&\left|\mathrm{Tr}\left(\sum_{j=1}^{\infty}(-1)^j\tfrac{1}{\sigma^{2j}}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\right)\right| \\
\leq&\sum_{j=1}^{\infty}\left|\mathrm{Tr}\big(\tfrac{1}{\sigma^{2j}}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\big)\right| \\
\leq&R\sum_{j=1}^{\infty}\left\|\tfrac{1}{\sigma^{2j}}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\right\|_2 \\
\leq&R\sum_{j=1}^{\infty}\left\|\tfrac{1}{\sigma^{2j}}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\right\|_2 \\
\leq&R\sum_{j=1}^{\infty}\left\|\tfrac{1}{\sigma^{2j}}B^j\right\|_2=\tilde{O}(n^{\frac{1}{\alpha}+\kappa+\frac{1-\alpha+2\tau}{2\alpha}}).
\end{aligned}$$

Then the asymptotics of $T_{1,R}(D_n)$ is given by

$$\begin{aligned}
&T_{1,R}(D_n) \\
=&\tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R)-\tfrac{1}{2}\,\mathrm{Tr}\big(I-(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\big)+\tilde{O}(n^{\frac{1}{\alpha}+\kappa+\frac{1-\alpha+2\tau}{2\alpha}})+\tilde{O}(n^{\frac{1}{\alpha}+\kappa+\frac{1-\alpha+2\tau}{2\alpha}}) \\
=&\Theta(n^{1/\alpha})+\tilde{O}(n^{\frac{1}{\alpha}+\kappa+\frac{1-\alpha+2\tau}{2\alpha}}) \\
=&\Theta(n^{\frac{1}{\alpha}}),
\end{aligned}$$

where in the last inequality we use the assumption that $\kappa<\frac{\alpha-1-2\tau}{2\alpha}$. Since $\tilde{O}(n^{\frac{1}{\alpha}+\kappa+\frac{1-\alpha+2\tau}{2\alpha}})$ is lower order term compared to $\Theta(n^{\frac{1}{\alpha}})$, we further have

$$T_{1,R}(D_n)=\big(\tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R)-\tfrac{1}{2}\,\mathrm{Tr}\big(I-(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\big)\big)(1+o(1)).$$

This concludes the proof of the first statement.

Let $\Lambda_{1:R}=\mathrm{diag}\{\lambda_1,\ldots,\lambda_R\}$, $\Phi_{1:R}=(\phi_1(\mathbf{x}),\phi_1(\mathbf{x}),\ldots,\phi_R(\mathbf{x}))$ and $\boldsymbol{\mu}_{1:R}=(\mu_1,\ldots,\mu_R)$. Since $\mu_0=0$, we have $T_{2,R}(D_n)=\frac{1}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T\Phi_{1:R}^T(I+\frac{1}{\sigma^2}\Phi_{1:R}\Lambda_{1:R}\Phi_{1:R}^T)^{-1}\Phi_{1:R}\boldsymbol{\mu}_{1:R}$. According to

Lemma 80, we have

$$
\begin{aligned}
T_{2,R}(D_n) &= \frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R} \\
&+ \frac{1}{2}\sum_{j=1}^{\infty}\Bigg[(-1)^{j+1}\boldsymbol{\mu}_{1:R}^T\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}(\Phi_{1:R}^T\Phi_{1:R}-nI) \\
&\qquad \left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}(\Phi_{1:R}^T\Phi_{1:R}-nI)\right)^{j-1}\Bigg] \\
&= \frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R} \\
&+ \frac{1}{2}\sum_{j=1}^{\infty}\Bigg[(-1)^{j+1}\frac{1}{\sigma^{2j}}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2}\Lambda_{1:R}^{-\gamma/2}A\Big((I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma}\Lambda_{1:R}^{1-\gamma}A\Big)^{j-1} \\
&\qquad (I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2}\Lambda_{1:R}^{-\gamma/2}\boldsymbol{\mu}_{1:R}\Bigg]
\end{aligned}
\tag{3.93}
$$

where in the second to last equality we used the definition of A (3.90). As for the first term on the right hand side of (3.93), by Lemma 61, Assumption 50 and Assumption 51, we have

$$
\frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R} \le \frac{n}{2\sigma^2}\sum_{p=1}^{R}\frac{C_\mu^2 p^{-2\beta}}{1+\frac{n}{\sigma^2}\underline{C_\lambda}p^{-\alpha}} = \begin{cases}\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}), & \alpha\neq 2\beta-1,\\ \Theta(\log n), & \alpha = 2\beta-1.\end{cases}
$$

On the other hand, by Assumption 51, assuming that $\sup_{i\ge 1} p_{i+1}-p_i = h$, we have

$$
\begin{aligned}
\frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R} &\ge \frac{n}{2\sigma^2}\sum_{i=1}^{\lfloor\frac{R}{h}\rfloor}\frac{\underline{C_\mu^2}p_i^{-2\beta}}{1+\frac{n}{\sigma^2}\overline{C_\lambda}p_i^{-\alpha}} \\
&\ge \frac{n}{2\sigma^2}\sum_{i=1}^{\lfloor\frac{R}{h}\rfloor}\frac{\underline{C_\mu^2}i^{-2\beta}}{1+\frac{n}{\sigma^2}\overline{C_\lambda}(hi)^{-\alpha}} \\
&= \begin{cases}\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}), & \alpha\neq 2\beta-1,\\ \Theta(\log n), & \alpha = 2\beta-1.\end{cases}
\end{aligned}
$$

Overall, we have

$$\frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}=\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n),\ \text{where } k=\begin{cases}0, & \alpha\neq 2\beta-1,\\ 1, & \alpha=2\beta-1.\end{cases}$$

By Lemma 62, we have

$$\begin{aligned}\|(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2}\Lambda_{1:R}^{-\gamma/2}\boldsymbol{\mu}_{1:R}\|_2^2 \leq& \sum_{p=1}^{R}\frac{C_\mu^2 p^{-2\beta}(\underline{C_\lambda}p^{-\alpha})^{-\gamma}}{(1+\frac{n}{\sigma^2}\underline{C_\lambda}p^{-\alpha})^{2-\gamma}}\\ =&\tilde{O}(\max\{n^{-2+\gamma},R^{1-2\beta+\alpha\gamma}\})\\ =&\tilde{O}(n^{\max\{-2+\gamma,\frac{1-2\beta}{\alpha}+\gamma+\kappa(1-2\beta+\alpha\gamma)\}}).\end{aligned}\tag{3.94}$$

Using (3.91), the second term on the right hand side of (3.93) is computed as follows:

$$\begin{aligned}&\frac{1}{2}\sum_{j=1}^{\infty}\Bigg[(-1)^{j+1}\frac{1}{\sigma^{2j}}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2}\Lambda_{1:R}^{-\gamma/2}A\Big((I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma}\Lambda_{1:R}^{1-\gamma}A\Big)^{j-1}\\ &\qquad(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2}\Lambda_{1:R}^{-\gamma/2}\boldsymbol{\mu}_{1:R}\Bigg]\\ \leq&\frac{1}{2}\sum_{j=1}^{\infty}\frac{1}{\sigma^{2j}}\|A\|^j\Big(\frac{n}{\sigma^2}\Big)^{(-1+\gamma)(j-1)}\|(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2}\Lambda_{1:R}^{-\gamma/2}\boldsymbol{\mu}_{1:R}\|_2^2\\ \leq&\frac{1}{2}\sum_{j=1}^{\infty}\frac{1}{\sigma^{2j}}\tilde{O}(n^{\frac{j(1-2\gamma\alpha+\alpha+2\tau)}{2\alpha}})\Big(\frac{n}{\sigma^2}\Big)^{(-1+\gamma)(j-1)}\tilde{O}(n^{\max\{-2+\gamma,\frac{1-2\beta}{\alpha}+\gamma+\kappa(1-2\beta+\alpha\gamma)\}})\\ =&\tilde{O}(n^{\max\{-2+\gamma+\frac{1-2\gamma\alpha+\alpha+2\tau}{2\alpha},\frac{1-2\beta}{\alpha}+\gamma+\frac{1-2\gamma\alpha+\alpha+2\tau}{2\alpha}+\kappa(1-2\beta+\alpha\gamma)\}})\\ =&\tilde{O}(n^{\max\{-2+\frac{1+\alpha+2\tau}{2\alpha},\frac{1-2\beta}{\alpha}+\frac{1+\alpha+2\tau}{2\alpha}+\kappa(1-2\beta+\alpha\gamma)\}}).\end{aligned}\tag{3.95}$$

Since $\frac{1+\alpha+2\tau}{2\alpha} < \frac{1+\alpha+2\tau}{\alpha+1+2\tau} = 1$, we have $-2 + \frac{1+\alpha+2\tau}{2\alpha} < 0$.Also we have

$$\begin{aligned}
&\frac{1-2\beta}{\alpha} + \frac{1+\alpha+2\tau}{2\alpha} + \kappa(1-2\beta+\alpha\gamma) \\
=&\frac{1-2\beta}{\alpha} + 1 + \frac{1-\alpha+2\tau}{2\alpha} + \kappa(1-2\beta+\alpha\gamma) \\
\leq&\frac{1-2\beta}{\alpha} + 1 + \frac{1-\alpha+2\tau}{2\alpha} + \kappa\alpha\gamma \\
<&\frac{1-2\beta}{\alpha} + 1,
\end{aligned} \tag{3.96}$$

where the last inequality holds because $\kappa < \frac{\alpha-1-2\tau}{2\alpha^2}$ and $\gamma \leq 1$. Hence we have

$$\begin{aligned}
T_{2,R}(D_n) =&\frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I + \frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R} + \tilde{O}(n^{\max\{-2+\frac{1+\alpha+2\tau}{2\alpha}, \frac{1-2\beta}{\alpha}+\frac{1+\alpha+2\tau}{2\alpha}+\kappa(1-2\beta+\alpha\gamma)\}}) \\
=&\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n) + \tilde{O}(n^{\max\{-2+\frac{1+\alpha+2\tau}{2\alpha}, \frac{1-2\beta}{\alpha}+\frac{1+\alpha+2\tau}{2\alpha}+\kappa(1-2\beta+\alpha\gamma)\}}) \\
=&\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n).
\end{aligned}$$

where $k = \begin{cases} 0, & \alpha \neq 2\beta - 1, \\ 1, & \alpha = 2\beta - 1. \end{cases}$. Since $\tilde{O}(n^{\max\{-2+\frac{1+\alpha+2\tau}{2\alpha}, \frac{1-2\beta}{\alpha}+\frac{1+\alpha+2\tau}{2\alpha}+\kappa(1-2\beta+\alpha\gamma)\}})$ is lower order term compared to $\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n)$, we further have

$$T_{2,R}(D_n) = \Big(\frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I + \frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\Big)(1+o(1))$$

This concludes the proof of the second statement. □

Lemma 82. *Under Assumptions 50, 51 and 52, with probability of at least $1 - 5\delta$, we have*

$$T_1(D_n) = \left(\frac{1}{2}\log\det(I + \frac{n}{\sigma^2}\Lambda) - \frac{1}{2}\operatorname{Tr}\Big(I - (I + \frac{n}{\sigma^2}\Lambda)^{-1}\Big)\right)(1+o(1)) = \Theta(n^{\frac{1}{\alpha}}), \tag{3.97}$$

Furthermore, let $\delta = n^{-q}$ where $0 \leq q < \min\{\frac{(2\beta-1)(\alpha-1-2\tau)}{4\alpha^2}, \frac{\alpha-1-2\tau}{2\alpha}\}$. If we assume $\mu_0 = 0$,

we have

$$T_2(D_n) = \Big(\frac{n}{2\sigma^2}\boldsymbol{\mu}^T(I+\frac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu}\Big)(1+o(1)) = \begin{cases}\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}), & \alpha \neq 2\beta-1,\\ \Theta(\log n), & \alpha = 2\beta-1.\end{cases} \tag{3.98}$$

Proof of Lemma 82. Let $R = n^{\frac{1}{\alpha}+\kappa}$ where $0 \leq \kappa < \frac{\alpha-1-2\tau}{2\alpha^2}$. By Lemmas 78 and 81, with probability of at least $1-5\delta$ we have

$$|T_{1,R}(D_n) - T_1(D_n)| = \tilde{O}(n^{\frac{1}{\alpha}+\kappa(1-\alpha)}), \tag{3.99}$$

and

$$|T_{2,R}(D_n) - T_2(D_n)| = \tilde{O}\bigg((\frac{1}{\delta}+1)n^{\max\{(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}, 1+\frac{1-2\beta}{\alpha}+\frac{(1-2\beta)\kappa}{2}, -1-\kappa\alpha, 1+\frac{1-2\beta}{\alpha}-\kappa\alpha\}}\bigg) \tag{3.100}$$

as well as

$$T_{1,R}(D_n) = \bigg(\frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) - \frac{1}{2}\operatorname{Tr}\Big(I-(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Big)\bigg)(1+o(1)) = \Theta(n^{\frac{1}{\alpha}}), \tag{3.101}$$

and

$$T_{2,R}(D_n) = \Big(\frac{n}{2\sigma^2}\boldsymbol{\mu}^T(I+\frac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu}\Big)(1+o(1)) = \begin{cases}\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}), & \alpha \neq 2\beta-1,\\ \Theta(\log n), & \alpha = 2\beta-1.\end{cases} \tag{3.102}$$

We then have

$$T_1(D_n) = T_{1,R}(D_n) + T_{1,R}(D_n) - T_1(D_n) = \Theta(n^{\frac{1}{\alpha}}) + \tilde{O}(n^{\frac{1}{\alpha}+\kappa(1-\alpha)}) = \Theta(n^{\frac{1}{\alpha}}).$$

Since $\tilde{O}(n^{\frac{1}{\alpha}+\kappa(1-\alpha)})$ is lower order term compared to $\Theta(n^{\frac{1}{\alpha}})$, we further have

$$T_1(D_n) = \bigg(\frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) - \frac{1}{2}\operatorname{Tr}\Big(I-(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Big)\bigg)(1+o(1)) = \Theta(n^{\frac{1}{\alpha}})$$

Besides, we have

$$\begin{aligned}
&\log\det(I+\frac{n}{\sigma^2}\Lambda)-\log\det(I+\frac{n}{\sigma^2}\Lambda_R)\\
&=\sum_{p=R+1}^{\infty}\log(1+\frac{n}{\sigma^2}\lambda_p)\le\frac{n}{\sigma^2}\sum_{p=R+1}^{\infty}\lambda_p\le\frac{n}{\sigma^2}\sum_{p=R+1}^{\infty}C_\lambda p^{-\alpha}=\frac{n}{\sigma^2}O(R^{1-\alpha})\\
&=\frac{n}{\sigma^2}O(n^{(1-\alpha)(\frac{1}{\alpha}+\kappa)})\\
&=o(n^{\frac{1}{\alpha}}).
\end{aligned}$$

Then we have $\log\det(I+\frac{n}{\sigma^2}\Lambda_R)=\log\det(I+\frac{n}{\sigma^2}\Lambda)(1+o(1))$. Similarly we can prove $\mathrm{Tr}\big(I-(I+\frac{n}{\sigma^2}\Lambda)^{-1}\big)=\mathrm{Tr}\big(I-(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\big)(1+o(1))$. This concludes the proof of the first statement.

As for $T_2(D_n)$, we have

$$\begin{aligned}
&T_2(D_n)\\
&=T_{2,R}(D_n)+T_{2,R}(D_n)-T_2(D_n)\\
&=\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n)+\tilde{O}\bigg((\frac{1}{\delta}+1)n^{\max\{(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2},1+\frac{1-2\beta}{\alpha}+\frac{(1-2\beta)\kappa}{2},-1-\kappa\alpha,1+\frac{1-2\beta}{\alpha}-\kappa\alpha\}}\bigg)\\
&=\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n)+\tilde{O}\bigg(n^{q+\max\{(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2},1+\frac{1-2\beta}{\alpha}+\frac{(1-2\beta)\kappa}{2},-1-\kappa\alpha,1+\frac{1-2\beta}{\alpha}-\kappa\alpha\}}\bigg)
\end{aligned}$$

where we use $\delta=n^{-q}$, $k=\begin{cases}0, & \alpha\neq 2\beta-1,\\ 1, & \alpha=2\beta-1.\end{cases}$.

Since $0\le\kappa<\frac{\alpha-1-2\tau}{2\alpha^2}$ and $0\le q<\min\{\frac{(2\beta-1)(\alpha-1-2\tau)}{4\alpha^2},\frac{\alpha-1-2\tau}{2\alpha}\}$, we can choose $\kappa<\frac{\alpha-1-2\tau}{2\alpha^2}$ and κ is arbitrarily close to $\frac{\alpha-1-2\tau}{2\alpha^2}$ such that $0\le q<\min\{\frac{(2\beta-1)\kappa}{2},\kappa\alpha\}$. Then we have $(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}+q<0$, $-1-\kappa\alpha+q<0$, $\frac{(1-2\beta)\kappa}{2}+q<0$ and $-\kappa\alpha+q<0$. So we have

$$T_{2,R}(D_n)=\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n).$$

Since $\tilde{O}\bigg((\frac{1}{\delta}+1)n^{\max\{(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2},1+\frac{1-2\beta}{\alpha}+\frac{(1-2\beta)\kappa}{2},-1-\kappa\alpha,1+\frac{1-2\beta}{\alpha}-\kappa\alpha\}}\bigg)$ is lower order term compared

to $\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n)$, we further have

$$T_2(D_n) = T_{2,R}(D_n)(1+o(1)) = \Big(\frac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\Big)(1+o(1)).$$

Furthermore, we have

$$\begin{aligned}
&\boldsymbol{\mu}^T(I+\frac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}_{\boldsymbol{R}}^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_{\boldsymbol{R}} \\
&= \sum_{p=R+1}^{\infty}\frac{\mu_p^2}{(1+\frac{n}{\sigma^2}\lambda_p)} \le \sum_{p=R+1}^{\infty}\mu_p^2 \le \frac{n}{\sigma^2}\sum_{p=R+1}^{\infty}C_\mu^2 p^{-2\beta} = O(R^{1-2\beta}) \\
&= O(n^{(1-2\beta)(\frac{1}{\alpha}+\kappa)}) \\
&= o(n^{\frac{1-2\beta}{\alpha}}).
\end{aligned}$$

Then we have $\boldsymbol{\mu}^T(I+\frac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu} = \boldsymbol{\mu}_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R(1+o(1))$. This concludes the proof of the second statement. □

Proof of Theorem 53. Using Lemma 82 and noting that $\frac{1}{\alpha} > 0$, with probability of at least $1-5\tilde{\delta}$, we have

$$\begin{aligned}
\mathbb{E}_\epsilon F^0(D_n) &= T_1(D_n) + T_2(D_n) \\
&= \Bigg[\frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) - \frac{1}{2}\operatorname{Tr}\Big(I-(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Big) \\
&\quad + \frac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\Bigg](1+o(1)) \\
&= \Theta(n^{\max\{\frac{1}{\alpha},\frac{1-2\beta}{\alpha}+1\}})
\end{aligned}$$

Letting $\delta = 5\tilde{\delta}$, we get the result. □

In the case of $\mu_0 > 0$, we have the following lemma:

Lemma 83. *Assume that $\sigma^2 = \Theta(1)$. Let $R = n^{\frac{1}{\alpha}+\kappa}$ where $0 < \kappa < \frac{\alpha-1-2\tau}{\alpha^2}$. Assume that $\mu_0 > 0$. Under Assumptions 50, 51 and 52, for sufficiently large n with probability of at least*

$1-4\delta$ *we have*

$$|T_{2,R}(D_n) - T_2(D_n)| = \tilde{O}\bigg((\frac{1}{\delta}+1)n^{\max\{1+(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}, 1-\kappa\alpha\}}\bigg).. \tag{3.103}$$

Proof of Lemma 83. As for $|T_2(D_n) - T_{2,R}(D_n)|$, we have

$$\begin{aligned}|T_2(D_n) - T_{2,R}(D_n)| = &\bigg|f(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x}) - f_R(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\bigg| \\ &+\bigg|f_R(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x}) - f_R(\mathbf{x})^T(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\bigg|.\end{aligned} \tag{3.104}$$

For the first term on the right-hand side of (3.104), we have

$$\begin{aligned}&\bigg|f(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x}) - f_R(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\bigg| \\ \leq&2\bigg|f_{>R}(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\bigg| + \bigg|f_{>R}(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_{>R}(\mathbf{x})\bigg| \\ \leq&2\|f_{>R}(\mathbf{x})\|_2\|(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2 + \|f_{>R}(\mathbf{x})\|_2\|(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}\|_2\|f_{>R}(\mathbf{x})\|_2 \\ \leq&2\|f_{>R}(\mathbf{x})\|_2\|(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2 + \|f_{>R}(\mathbf{x})\|_2^2.\end{aligned}$$

Applying Corollary 65 and Lemma 77, with probability of at least $1-4\delta$, we have

$$\begin{aligned}&\bigg|f(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x}) - f_R(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\bigg| \\ \leq&2\tilde{O}\bigg(\sqrt{(\frac{1}{\delta}+1)nR^{1-2\beta}}\bigg)\tilde{O}(\sqrt{(\frac{1}{\delta}+1)n}) + \tilde{O}((\frac{1}{\delta}+1)nR^{1-2\beta}) \\ =&2\tilde{O}\bigg((\frac{1}{\delta}+1)n^{1+(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}}\bigg) + \tilde{O}((\frac{1}{\delta}+1)n^{1+(\frac{1}{\alpha}+\kappa)(1-2\beta)}) \\ =&2\tilde{O}\bigg((\frac{1}{\delta}+1)n^{1+(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}}\bigg).\end{aligned}$$

As for the second term on the right-hand side of (3.81), according to Lemma 74, Corollary

72 and Lemma 76, we have

$$\begin{aligned}
&\left|f_R(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})-f_R(\mathbf{x})^T(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
=&\left|\sum_{j=1}^{\infty}(-1)^j f_R(\mathbf{x})^T\left((I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\right)^j(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
\leq&\sum_{j=1}^{\infty}\|(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\|_2^{j-1}\cdot\|\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2^j\cdot\|(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2^2 \qquad (3.105)\\
=&\sum_{j=1}^{\infty}\tilde{O}(n^{-j\kappa\alpha})\tilde{O}((\frac{1}{\delta}+1)n) \\
=&\tilde{O}((\frac{1}{\delta}+1)n^{1-\kappa\alpha}).
\end{aligned}$$

By (3.81), we have

$$\begin{aligned}
|T_2(D_n)-T_{2,R}(D_n)|&=\tilde{O}\left((\frac{1}{\delta}+1)n^{1+(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}}\right)+\tilde{O}((\frac{1}{\delta}+1)n^{1-\kappa\alpha}) \\
&=\tilde{O}\left((\frac{1}{\delta}+1)n^{\max\{1+(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2},1-\kappa\alpha\}}\right).
\end{aligned}$$

□

Lemma 84. *Assume that $\sigma^2=\Theta(1)$. Let $R=n^{\frac{1}{\alpha}+\kappa}$ where $0<\kappa<\min\{\frac{\alpha-1-2\tau}{2\alpha^2},\frac{2\beta-1}{\alpha^2}\}$. Assume that $\mu_0>0$. Under Assumptions 50, 51 and 52, with probability of at least $1-\delta$, we have*

$$T_{2,R}(D_n)=\frac{n}{2\sigma^2}\mu_0^2+\tilde{O}(n^{\max\{\frac{1+7\alpha+2\tau}{8\alpha},1+\frac{1-2\beta}{\alpha}\}}). \qquad (3.106)$$

Proof of Lemma 84. Let

$$A=(I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R-nI)\Lambda_R^{\gamma/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}, \qquad (3.107)$$

where $\frac{1+\alpha+2\tau}{2\alpha}<\gamma\leq 1$. By Corollary 68, with probability of at least $1-\delta$, we have

$$\|A\|_2=\tilde{O}(n^{\frac{1-2\gamma\alpha+\alpha+2\tau}{2\alpha}}). \qquad (3.108)$$

When n is sufficiently large, $\|A\|_2$ is less than 1. Let $\boldsymbol{\mu}_{R,1} = (\mu_0, 0, \ldots, 0)$ and $\boldsymbol{\mu}_{R,2} = (0, \mu_1, \ldots, \mu_R)$. Then $\boldsymbol{\mu}_R = \boldsymbol{\mu}_{R,1} + \boldsymbol{\mu}_{R,2}$. Let $\tilde{\Lambda}_{1,R} = \mathrm{diag}\{1, \lambda_1, \ldots, \lambda_R\}$ and $I_{0,R} = (0, 1, \ldots, 1)$. Then $\Lambda_R = \tilde{\Lambda}_{1,R} I_{0,R}$. Let $B = (I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2} \tilde{\Lambda}_{1,R}^{1/2} (\Phi_R^T \Phi_R - nI) \tilde{\Lambda}_{1,R}^{1/2} (I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}$. By Corollary 69, we have $\|B\|_2 = O(\sqrt{\log \frac{R}{\delta}} n^{\frac{1}{2}})$. By Lemma 80, we have

$$
\begin{aligned}
&T_{2,R}(D_n) \\
&= \frac{n}{2\sigma^2} \boldsymbol{\mu}_R^T (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} \boldsymbol{\mu}_R \\
&+ \frac{1}{2} \sum_{j=1}^{\infty} \Bigg[(-1)^{j+1} \boldsymbol{\mu}_R^T \frac{1}{\sigma^2} (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} (\Phi_R^T \Phi_R - nI) \left(\frac{1}{\sigma^2} (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} \Lambda_R (\Phi_R^T \Phi_R - nI) \right)^{j-1} \\
&\qquad (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} \boldsymbol{\mu}_R \Bigg]
\end{aligned}
\tag{3.109}
$$

As for the first term on the right hand side of (3.109), by Lemma 61, we have

$$
\frac{n}{2\sigma^2} \mu^T (I + \frac{n}{\sigma^2}\Lambda)^{-1} \mu \leq \frac{n}{2\sigma^2} \left(\mu_0^2 + \sum_{p=1}^{R} \frac{C_\mu^2 p^{-2\beta}}{1 + \frac{n}{\sigma^2} \underline{C_\lambda} p^{-\alpha}} \right) = \frac{n}{2\sigma^2} \mu_0^2 + \tilde{O}(n^{\max\{0, 1 + \frac{1-2\beta}{\alpha}\}}).
$$

We define $Q_{1,j}$, $Q_{2,j}$ and $Q_{3,j}$ by

$$
\begin{aligned}
Q_{1,j} &= \boldsymbol{\mu}_{R,1}^T \frac{1}{\sigma^2} (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} (\Phi_R^T \Phi_R - nI) \left(\frac{1}{\sigma^2} (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} \Lambda_R (\Phi_R^T \Phi_R - nI) \right)^{j-1} \\
&\qquad (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} \boldsymbol{\mu}_{R,1} \\
Q_{2,j} &= \boldsymbol{\mu}_{R,1}^T \frac{1}{\sigma^2} (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} (\Phi_R^T \Phi_R - nI) \left(\frac{1}{\sigma^2} (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} \Lambda_R (\Phi_R^T \Phi_R - nI) \right)^{j-1} \\
&\qquad (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} \boldsymbol{\mu}_{R,2} \\
Q_{3,j} &= \boldsymbol{\mu}_{R,2}^T \frac{1}{\sigma^2} (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} (\Phi_R^T \Phi_R - nI) \left(\frac{1}{\sigma^2} (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} \Lambda_R (\Phi_R^T \Phi_R - nI) \right)^{j-1} \\
&\qquad (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} \boldsymbol{\mu}_{R,2}
\end{aligned}
\tag{3.110}
$$

The quantity $Q_{3,j}$ actually shows up in the case of $\mu_0 = 0$ in the proof of Lemma 81. By

(3.93), (3.95) and (3.96), we have that

$$|\sum_{j=1}^{\infty}(-1)^{j+1}Q_{3,j}| = |\sum_{j=1}^{\infty}(-1)^{j+1}\tilde{O}(n^{\frac{(j-1)(1-\alpha+2\tau)}{2\alpha}})o(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}})| = o(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}). \tag{3.111}$$

For $Q_{1,j}$, we have

$$\begin{aligned} Q_{1,1} &= \frac{1}{\sigma^{2j}}\boldsymbol{\mu}_{R,1}^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}B(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\boldsymbol{\mu}_{R,1} \\ &\leq \frac{1}{\sigma^{2j}}\|\boldsymbol{\mu}_{R,1}\|_2^2\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\|_2^2\|B\|_2 \\ &= O(\sqrt{\log\frac{R}{\delta}}n^{\frac{1}{2}}), \end{aligned}$$

where in the last equality we use $\|B\|_2 = O(\sqrt{\log\frac{R}{\delta}}n^{\frac{1}{2}})$. For $j \geq 2$, we have

$$\begin{aligned} Q_{1,j} &= \frac{1}{\sigma^{2j}}\boldsymbol{\mu}_{R,1}^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}B\Big((I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma}\Lambda_R^{1-\gamma}A\Big)^{j-2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma}\Lambda_R^{1-\gamma} \\ &\quad B(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\boldsymbol{\mu}_{R,1} \\ &\leq \frac{1}{\sigma^{2j}}\|\boldsymbol{\mu}_{R,1}\|_2^2\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\|_2^2\|B\|_2^2\|A\|_2^{j-2}\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma}\Lambda_R^{1-\gamma}\|_2^{j-1} \\ &= O(\log\frac{R}{\delta}n\cdot n^{\frac{(j-2)(1-2\gamma\alpha+\alpha+2\tau)}{2\alpha}}\cdot n^{-(1-\gamma)(j-1)}) \\ &= O(\log\frac{R}{\delta}n^{\gamma}\cdot n^{\frac{(j-2)(1-\alpha+2\tau)}{2\alpha}}). \end{aligned}$$

Then we have

$$|\sum_{j=1}^{\infty}(-1)^{j+1}Q_{1,j}| \leq O(\sqrt{\log\frac{R}{\delta}}n^{\frac{1}{2}}) + \sum_{j=2}^{\infty}O(\log\frac{R}{\delta}n^{\gamma}\cdot n^{\frac{(j-2)(1-\alpha+2\tau)}{2\alpha}}) = O(\log\frac{R}{\delta}n^{\gamma}) \tag{3.112}$$

For $Q_{2,j}$, we have

$$\begin{aligned} Q_{2,j} &= \frac{1}{\sigma^{2j}}\boldsymbol{\mu}_{R,1}^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}B\Big((I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma}\Lambda_R^{1-\gamma}A\Big)^{j-1}(I+\frac{n}{\sigma^2}\Lambda)^{-1+\frac{\gamma}{2}}\tilde{\Lambda}_{1,R}^{-\frac{\gamma}{2}}\boldsymbol{\mu}_{R,2} \\ &\leq \frac{1}{\sigma^{2j}}\|\boldsymbol{\mu}_{R,1}\|_2\|B\|_2\|A\|_2^{j-1}\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma}\Lambda_R^{1-\gamma}\|_2^{j-1}\|(I+\frac{n}{\sigma^2}\Lambda)^{-1+\frac{\gamma}{2}}\tilde{\Lambda}_{1,R}^{-\frac{\gamma}{2}}\boldsymbol{\mu}_{R,2}\|_2 \\ &= O(\sqrt{\log\frac{R}{\delta}}n^{\frac{1}{2}}\cdot n^{\frac{(j-1)(1-\alpha+2\tau)}{2\alpha}})\|(I+\frac{n}{\sigma^2}\Lambda)^{-1+\frac{\gamma}{2}}\tilde{\Lambda}_{1,R}^{-\frac{\gamma}{2}}\boldsymbol{\mu}_{R,2}\|_2. \end{aligned}$$

Since $\|(I+\frac{n}{\sigma^2}\Lambda)^{-1+\frac{\gamma}{2}}\tilde{\Lambda}_{1,R}^{-\frac{\gamma}{2}}\boldsymbol{\mu}_{R,2}\|_2$ is actually the case of $\mu_0=0$, we can use (3.94) in the proof of Lemma 81 and get

$$\begin{aligned}\|(I+\frac{n}{\sigma^2}\Lambda)^{-1+\frac{\gamma}{2}}\tilde{\Lambda}_{1,R}^{-\frac{\gamma}{2}}\boldsymbol{\mu}_{R,2}\|_2^2 =&\|(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2}\Lambda_{1:R}^{-\gamma/2}\boldsymbol{\mu}_{1:R}\|_2^2\\ =&\tilde{O}(n^{\max\{-2+\gamma,\frac{1-2\beta}{\alpha}+\gamma+\kappa(1-2\beta+\alpha\gamma)\}}\\ =&\tilde{O}(n^{\max\{-2+\gamma,\frac{1-2\beta}{\alpha}+\gamma+\kappa(1-2\beta+\alpha\gamma)\}})\\ =&o(n^{\gamma}),\end{aligned} \tag{3.113}$$

where in the last equality we use $\kappa<\frac{2\beta-1}{\alpha^2}$. Then we have

$$|\sum_{j=1}^{\infty}(-1)^{j+1}Q_{2,j}|\le\sum_{j=1}^{\infty}o(\sqrt{\log\frac{R}{\delta}}n^{\frac{1+\gamma}{2}}\cdot n^{\frac{(j-1)(1-\alpha+2\tau)}{2\alpha}})=o(\sqrt{\log\frac{R}{\delta}}n^{\frac{1+\gamma}{2}}) \tag{3.114}$$

Choosing $\gamma=\frac{1}{2}(1+\frac{1+\alpha+2\tau}{2\alpha})=\frac{1+3\alpha+2\tau}{4\alpha}<1$, we have

$$\begin{aligned}T_{2,R}(D_n)&=\frac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R+\sum_{j=1}^{\infty}(-1)^{j+1}(Q_{1,j}+Q_{2,j}+Q_{3,j})\\ &=\frac{n}{2\sigma^2}\mu_0^2+\tilde{O}(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}})+o(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}})+O(\log\frac{R}{\delta}n^{\gamma})+o(\sqrt{\log\frac{R}{\delta}}n^{\frac{1+\gamma}{2}})\\ &=\frac{n}{2\sigma^2}\mu_0^2+\tilde{O}(n^{\max\{\frac{1+\gamma}{2},1+\frac{1-2\beta}{\alpha}\}})\\ &=\frac{n}{2\sigma^2}\mu_0^2+\tilde{O}(n^{\max\{\frac{1+7\alpha+2\tau}{8\alpha},1+\frac{1-2\beta}{\alpha}\}}).\end{aligned}$$

□

Proof of Theorem 54. Let $R=n^{\frac{1}{\alpha}+\kappa}$ where $0<\kappa<\min\{\frac{\alpha-1-2\tau}{2\alpha^2},\frac{2\beta-1}{\alpha^2}\}$. Since $0\le q<\min\{\frac{2\beta-1}{2},\alpha\}\cdot\min\{\frac{\alpha-1-2\tau}{2\alpha^2},\frac{2\beta-1}{\alpha^2}\}$, we can choose $\kappa<\min\{\frac{\alpha-1-2\tau}{2\alpha^2},\frac{2\beta-1}{\alpha^2}\}$ and κ is arbitrarily close to $\kappa<\min\{\frac{\alpha-1-2\tau}{2\alpha^2},\frac{2\beta-1}{\alpha^2}\}$ such that $0\le q<\min\{\frac{(2\beta-1)\kappa}{2},\kappa\alpha\}$. Then we have $(\frac{1}{\alpha}+$

$\kappa)\frac{1-2\beta}{2} + q < 0$, and $-\kappa\alpha + q < 0$. As for $T_2(D_n)$, we have

$$\begin{aligned}
T_2(D_n) &\le T_{2,R}(D_n) + |T_{2,R}(D_n) - T_2(D_n)| \\
&= \frac{n}{2\sigma^2}\mu_0^2 + \tilde{O}(n^{\max\{\frac{1+7\alpha+2\tau}{8\alpha}, 1+\frac{1-2\beta}{\alpha}\}}) + \tilde{O}\Big((\tfrac{1}{\delta}+1)n^{\max\{1+(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}, 1-\kappa\alpha\}}\Big) \\
&= \frac{n}{2\sigma^2}\mu_0^2 + \tilde{O}(n^{\max\{\frac{1+7\alpha+2\tau}{8\alpha}, 1+\frac{1-2\beta}{\alpha}\}}) + \tilde{O}\Big(n^{q+\max\{1+(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}, 1-\kappa\alpha\}}\Big) \\
&= \frac{n}{2\sigma^2}\mu_0^2 + o(n).
\end{aligned}$$

By Lemma 82, we have $T_1(D_n) = O(n^{\frac{1}{\alpha}})$. Hence $\mathbb{E}_\epsilon F^0(D_n) = T_1(D_n) + T_2(D_n) = \frac{n}{2\sigma^2}\mu_0^2 + o(n)$. □

3.D.2 Proofs Related to the Asymptotics of the Generalization Error

Lemma 85. *Assume $\sigma^2 = \Theta(n^t)$ where $1 - \frac{\alpha}{1+2\tau} < t < 1$. Let $R = n^{(\frac{2\alpha-1}{\alpha(\alpha-1)}+1)(1-t)}$. Under Assumptions 50, 51 and 52, with probability of at least $1-\delta$ over sample inputs $(x_i)_{i=1}^n$, we have*

$$G_1(D_n) = \frac{1+o(1)}{2\sigma^2}\Big(\mathrm{Tr}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R - \|\Lambda_R^{1/2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\|_F^2\Big) = \frac{1}{\sigma^2}\Theta\Big(n^{\frac{(1-\alpha)(1-t)}{\alpha}}\Big). \tag{3.115}$$

Proof of Lemma 85. Let $G_{1,R}(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{1,R}(D_{n+1}) - T_{1,R}(D_n))$, where $R = n^C$ for some constant C. By Lemma 78, we have that

$$\begin{aligned}
|G_1(D_n) - G_{1,R}(D_n)| &= \big|\mathbb{E}_{(x_{n+1},y_{n+1})}[T_1(D_{n+1}) - T_{1,R}(D_{n+1})] - [T_1(D_n) - T_{1,R}(D_n)]\big| \\
&= \big|\mathbb{E}_{(x_{n+1},y_{n+1})}O((n+1)R^{1-\alpha})\big| + \big|O(nR^{1-\alpha})]\big| \\
&= O(\tfrac{1}{\sigma^2}nR^{1-\alpha}).
\end{aligned} \tag{3.116}$$

Define $\eta_R = (\phi_0(x_{n+1}), \phi_1(x_{n+1}), \ldots, \phi_R(x_{n+1}))^T$ and $\widetilde{\Phi}_R = (\Phi_R^T, \eta_R)^T$. As for $G_{1,R}(D_n)$, we

have

$$
\begin{aligned}
G_{1,R}(D_n) &= \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{1,R}(D_{n+1}) - T_{1,R}(D_n)) \\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2}\log\det(I + \frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2}) - \frac{1}{2}\mathrm{Tr}(I - (I + \frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})^{-1})\right) \\
&- \left(\frac{1}{2}\log\det(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2}) - \frac{1}{2}\mathrm{Tr}(I - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1})\right) \\
&= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det(I + \frac{\widetilde{\Phi_R}\Lambda_R\widetilde{\Phi_R}^T}{\sigma^2}) - \log\det(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})\right) \\
&- \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\mathrm{Tr}(I - (I + \frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})^{-1}) - \mathrm{Tr}(I - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1})\right).
\end{aligned}
\tag{3.117}
$$

As for the first term in the right hand side (3.117), we have

$$
\begin{aligned}
&\frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det(I + \frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2}) - \log\det(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})\right) \\
&= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det(I + \frac{\Lambda_R\widetilde{\Phi}_R^T\widetilde{\Phi}_R}{\sigma^2}) - \log\det(I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})\right) \\
&= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det(I + \frac{\Lambda_R\Phi_R^T\Phi_R + \eta_R\eta_R^T}{\sigma^2}) - \log\det(I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})\right) \\
&= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det\left((I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}(I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2} + \frac{\Lambda_R\eta_R\eta_R^T}{\sigma^2})\right)\right) \\
&= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det\left(I + (I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\frac{\Lambda_R\eta_R\eta_R^T}{\sigma^2}\right)\right) \\
&= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\left(1 + \frac{1}{\sigma^2}\eta_R^T(I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\Lambda_R\eta_R\right)\right)
\end{aligned}
$$

Let

$$
A = (I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{1/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}. \tag{3.118}
$$

According to Corollary 68, with probability of at least $1-\delta$, we have

$$
\|\frac{1}{\sigma^2}A\|_2 = O(\sqrt{\log\frac{R}{\delta}}n^{\frac{1-\alpha+2\tau}{2\alpha} - \frac{(1+2\tau)t}{2\alpha}}) = o(1).
$$

When n is sufficiently large, $\|\frac{1}{\sigma^2}A\|_2$ is less than 1. By Lemma 73, we have

$$
\begin{aligned}
&\eta_R^T(I+\frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\Lambda_R\eta_R\\
=&\eta_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R\eta_R+\sum_{j=1}^{\infty}(-1)^j\eta_R^T\bigg(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\bigg)^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R\eta_R\\
=&\eta_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R\eta_R+\sum_{j=1}^{\infty}(-1)^j\frac{1}{\sigma^{2j}}\eta_R^T(I+\frac{n}{\sigma^{2j}}\Lambda_R)^{-1/2}\Lambda_R^{1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}\eta_R\\
\leq&\eta_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R\eta_R+\sum_{j=1}^{\infty}\|\frac{1}{\sigma^2}A\|_2^j\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}\eta_R\|_2^2\\
\leq&\sum_{p=1}^{R}\phi_p^2(x_{n+1})\frac{\overline{C_\lambda}p^{-\alpha}}{1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2}+\sum_{j=1}^{\infty}\|\frac{1}{\sigma^2}A\|_2^j\sum_{p=1}^{R}\phi_p^2(x_{n+1})\frac{\overline{C_\lambda}p^{-\alpha}}{1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2}\\
\leq&\sum_{p=1}^{R}\frac{\overline{C_\lambda}p^{-\alpha}p^{2\tau}}{1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2}+\sum_{j=1}^{\infty}\|\frac{1}{\sigma^2}A\|_2^j\sum_{p=1}^{R}\frac{\overline{C_\lambda}p^{-\alpha}p^{2\tau}}{1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2}\\
\leq&O(n^{\frac{(1-\alpha+2\tau)(1-t)}{\alpha}})+\sum_{j=1}^{\infty}\|\frac{1}{\sigma^2}A\|_2^jO(n^{\frac{(1-\alpha+2\tau)(1-t)}{\alpha}})\\
=&O(n^{\frac{(1-\alpha+2\tau)(1-t)}{\alpha}})=o(1),
\end{aligned}
\tag{3.119}
$$

where we use Lemma 61 in the last inequality. Next we have

$$
\begin{aligned}
&\frac{1}{2}\bigg(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det(I+\frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})-\log\det(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})\bigg)\\
&=\frac{1}{2}\bigg(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\bigg(1+\frac{1}{\sigma^2}\eta_R^T(I+\frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\Lambda_R\eta_R\bigg)\bigg)\\
&=\frac{1}{2}\bigg(\mathbb{E}_{(x_{n+1},y_{n+1})}\bigg(\frac{1}{\sigma^2}\eta_R^T(I+\frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\Lambda_R\eta_R\bigg)(1+o(1))\bigg)\\
&=\frac{1}{2\sigma^2}\bigg(\mathrm{Tr}(I+\frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\Lambda_R\bigg)(1+o(1)),
\end{aligned}
$$

where in the last equality we use the fact that $\mathbb{E}_{(x_{n+1},y_{n+1})}\eta_R\eta_R^T = I$. By Lemma 73, we have

$$\begin{aligned}
&\operatorname{Tr}(I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\Lambda_R \\
=&\operatorname{Tr}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R + \sum_{j=1}^{\infty}(-1)^j \operatorname{Tr}\left(\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\right)^j (I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R \\
=&\operatorname{Tr}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R + \sum_{j=1}^{\infty}(-1)^j \operatorname{Tr}\frac{1}{\sigma^{2j}}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}A^j(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}.
\end{aligned}$$

By Lemma 61, we have

$$\begin{aligned}
\operatorname{Tr}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R &\le \sum_{p=1}^{R}\frac{\overline{C_\lambda}p^{-\alpha}}{1 + n\underline{C_\lambda}p^{-\alpha}/\sigma^2} = \Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}}) \\
\operatorname{Tr}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R &\ge \sum_{p=1}^{R}\frac{\underline{C_\lambda}p^{-\alpha}}{1 + n\overline{C_\lambda}p^{-\alpha}/\sigma^2} = \Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}}).
\end{aligned}$$

Overall,

$$\operatorname{Tr}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R = \Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}}). \tag{3.120}$$

Since $\|\frac{1}{\sigma^2}A\|_2^j = o(1)$, we have that the absolute values of diagonal entries of $\frac{1}{\sigma^{2j}}A^j$ are at most $o(1)$. Let $(A^j)_{p,p}$ denote the (p,p)-th entry of the matrix A^j. Then we have

$$\begin{aligned}
&\left|\operatorname{Tr}\frac{1}{\sigma^{2j}}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}A^j(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}\right| \\
=&\left|\sum_{p=1}^{R}\frac{\lambda_p\frac{1}{\sigma^{2j}}(A^j)_{p,p}}{1 + n\lambda_p/\sigma^2}\right| \le \sum_{p=1}^{R}\frac{\lambda_p\|\frac{1}{\sigma^{2j}}A\|_2^j}{1 + n\lambda_p/\sigma^2} = \Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})\tilde{O}(n^{\frac{j(1-\alpha+2\tau-(1+2\tau)t)}{2\alpha}}(\log R)^{j/2}),
\end{aligned} \tag{3.121}$$

where in the last step we used (3.120). According to (3.120) and (3.121), we have

$$
\begin{aligned}
&\frac{1}{2}\bigg(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det(I+\frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})-\log\det(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})\bigg)\\
&=\frac{1}{2\sigma^2}\bigg(\mathrm{Tr}(I+\frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\Lambda_R\bigg)(1+o(1))\\
&=\tfrac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})+\tfrac{1}{\sigma^2}\sum_{j=1}^{\infty}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})\tilde{O}(n^{\frac{j(1-\alpha+2\tau-(1+2\tau)t)}{2\alpha}}(\log R)^{j/2})\\
&=\tfrac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})+\tfrac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})o(1)=\tfrac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})\\
&=\frac{1}{2\sigma^2}\bigg(\mathrm{Tr}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R\bigg)(1+o(1)).
\end{aligned}
\tag{3.122}
$$

Using the Woodbury matrix identity, the second term in the right hand side (3.117) is given by

$$
\begin{aligned}
&\frac{1}{2}\bigg(\mathbb{E}_{(x_{n+1},y_{n+1})}\mathrm{Tr}(I-(I+\frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})^{-1}-\mathrm{Tr}(I-(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\bigg)\\
&=\frac{1}{2}\bigg(\mathbb{E}_{(x_{n+1},y_{n+1})}\,\mathrm{Tr}(\frac{1}{\sigma^2}\widetilde{\Phi}_R(I+\frac{1}{\sigma^2}\Lambda_R\widetilde{\Phi}_R^T\widetilde{\Phi}_R)^{-1}\Lambda_R\widetilde{\Phi}_R^T-\mathrm{Tr}(\frac{1}{\sigma^2}\Phi_R(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T\bigg)\\
&=\frac{1}{2}\bigg(\mathbb{E}_{(x_{n+1},y_{n+1})}\,\mathrm{Tr}(\frac{1}{\sigma^2}(I+\frac{1}{\sigma^2}\Lambda_R\widetilde{\Phi}_R^T\widetilde{\Phi}_R)^{-1}\Lambda_R\widetilde{\Phi}_R^T\widetilde{\Phi}_R-\mathrm{Tr}(\frac{1}{\sigma^2}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T\Phi_R\bigg)\\
&=-\frac{1}{2}\bigg(\mathbb{E}_{(x_{n+1},y_{n+1})}\,\mathrm{Tr}(I+\frac{1}{\sigma^2}\Lambda_R\widetilde{\Phi}_R^T\widetilde{\Phi}_R)^{-1}-\mathrm{Tr}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\bigg)\\
&=-\frac{1}{2}\bigg(\mathbb{E}_{(x_{n+1},y_{n+1})}\,\mathrm{Tr}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R+\frac{1}{\sigma^2}\Lambda_R\eta_R\eta_R^T)^{-1}-\mathrm{Tr}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\bigg)\\
&=\frac{1}{2\sigma^2}\bigg(\mathbb{E}_{(x_{n+1},y_{n+1})}\,\mathrm{Tr}\,\frac{(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\eta_R\eta_R^T(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}}{1+\frac{1}{\sigma^2}\eta_R^T(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\eta_R}\bigg),
\end{aligned}
$$

where the last equality uses the Sherman–Morrison formula. According to (3.119), we get

$$
\begin{aligned}
&\frac{1}{2\sigma^2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\operatorname{Tr}\frac{(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\eta_R\eta_R^T(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}}{1+\frac{1}{\sigma^2}\eta_R^T(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\eta_R}\right)\\
&=\frac{1}{2\sigma^2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\operatorname{Tr}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\eta_R\eta_R^T(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}(1+o(1))\right)\\
&=\frac{1+o(1)}{2\sigma^2}\operatorname{Tr}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\\
&=\frac{1+o(1)}{2\sigma^2}\operatorname{Tr}\Lambda_R^{1/2}(I+\frac{1}{\sigma^2}\Lambda_R^{1/2}\Phi_R^T\Phi_R\Lambda_R^{1/2})^{-1}\Lambda_R^{1/2}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\\
&=\frac{1+o(1)}{2\sigma^2}\operatorname{Tr}(I+\frac{1}{\sigma^2}\Lambda_R^{1/2}\Phi_R^T\Phi_R\Lambda_R^{1/2})^{-1}\Lambda_R^{1/2}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R^{1/2}\\
&=\frac{1+o(1)}{2\sigma^2}\operatorname{Tr}(I+\frac{1}{\sigma^2}\Lambda_R^{1/2}\Phi_R^T\Phi_R\Lambda_R^{1/2})^{-1}\Lambda_R(I+\frac{1}{\sigma^2}\Lambda_R^{1/2}\Phi_R^T\Phi_R\Lambda_R^{1/2})^{-1}\\
&=\frac{1+o(1)}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{1}{\sigma^2}\Lambda_R^{1/2}\Phi_R^T\Phi_R\Lambda_R^{1/2})^{-1}\|_F^2\\
&=\frac{1+o(1)}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}(I+\frac{1}{\sigma^2}A)^{-1}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2,
\end{aligned}
$$

where in the penultimate equality we use $\operatorname{Tr}(BB^T)=\|B\|_F^2$, $\|B\|_F$ is the Frobenius norm of A, and in the last equality we use the definition of A (3.118). Then we have

$$
\begin{aligned}
&\frac{1+o(1)}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}(I+\frac{1}{\sigma^2}A)^{-1}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2\\
&=\frac{1+o(1)}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}(I+\sum_{j=1}^{\infty}(-1)^j\frac{1}{\sigma^{2j}}A^j)(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2\\
&=\frac{1+o(1)}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}+\sum_{j=1}^{\infty}(-1)^j\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2.
\end{aligned}
\tag{3.123}
$$

By Lemma 61, we have

$$
\begin{aligned}
\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F\leq&\sqrt{\sum_{p=1}^{R}\frac{\overline{C_\lambda}p^{-\alpha}}{(1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2)^2}}=\Theta(n^{\frac{(1-\alpha)(1-t)}{2\alpha}})\\
\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F\geq&\sqrt{\sum_{p=1}^{R}\frac{\underline{C_\lambda}p^{-\alpha}}{(1+n\overline{C_\lambda}p^{-\alpha}/\sigma^2)^2}}=\Theta(n^{\frac{(1-\alpha)(1-t)}{2\alpha}}).
\end{aligned}
$$

Overall, we have

$$\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F = \Theta(n^{\frac{(1-\alpha)(1-t)}{2\alpha}}). \tag{3.124}$$

Since $\|\frac{1}{\sigma^2}A\|_2 = O(\sqrt{\log\frac{R}{\delta}}n^{\frac{1-\alpha+2\tau}{2\alpha}-\frac{(1+2\tau)t}{2\alpha}}) = o(1)$, we have

$$\begin{aligned}
&\|\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F \\
&\le \|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F\|\frac{1}{\sigma^2}A\|_2^j\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_2 \\
&= O(n^{\frac{(1-\alpha)(1-t)}{2\alpha}})\tilde{O}(n^{\frac{j(1-\alpha+2\tau-(1+2\tau)t)}{2\alpha}}(\log R)^{j/2}),
\end{aligned} \tag{3.125}$$

where in the first inequality we use the fact that $\|AB\|_F \le \|A\|_F\|B\|_2$ when B is symmetric. By Lemma 61, we have

$$\begin{aligned}
&\frac{1}{\sigma^{2j}}\left|\operatorname{Tr}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\right| \\
&= \left|\sum_{p=1}^R \frac{\lambda_p((\frac{1}{\sigma^2}A)^j)_{p,p}}{(1+n\lambda_p/\sigma^2)^2}\right| \le \sum_{p=1}^R \frac{\lambda_p\|\frac{1}{\sigma^2}A\|_2^j}{(1+n\lambda_p/\sigma^2)^2} = \Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})\tilde{O}(n^{\frac{j(1-\alpha+2\tau-(1+2\tau)t)}{2\alpha}}(\log R)^{j/2}),
\end{aligned} \tag{3.126}$$

According to (3.124), (3.125) and (3.126), we have

$$
\begin{aligned}
&\frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\mathrm{Tr}(I-(I+\frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})^{-1}-\mathrm{Tr}(I-(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\right)\\
=&\frac{1+o(1)}{2\sigma^2}\,\mathrm{Tr}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\\
=&\frac{1+o(1)}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}+\sum_{j=1}^{\infty}(-1)^j\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2\\
=&\frac{1+o(1)}{2\sigma^2}\Bigg(\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F^2+\sum_{j=1}^{\infty}\left\|\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\right\|_F^2\\
&\qquad+2\,\mathrm{Tr}\,\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\sum_{j=1}^{\infty}(-1)^j\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Bigg)\\
=&\frac{1+o(1)}{2\sigma^2}\Bigg(\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})+\sum_{j=1}^{\infty}\frac{1}{\sigma^{2j}}O(n^{\frac{(1-\alpha)(1-t)}{\alpha}})\tilde{O}(n^{\frac{j(1-\alpha+2\tau-(1+2\tau)t)}{2\alpha}}(\log R)^{j/2})\\
&\qquad+2\sum_{j=1}^{\infty}\frac{1}{\sigma^{2j}}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})\tilde{O}(n^{\frac{j(1-\alpha+2\tau-(1+2\tau)t)}{2\alpha}}(\log R)^{j/2})\Bigg)\\
=&\tfrac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})=\frac{1+o(1)}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F^2.
\end{aligned}
\tag{3.127}
$$

Combining (3.122) and (3.127) we get that $G_{1,R}(D_n)=\frac{1+o(1)}{2\sigma^2}(\mathrm{Tr}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R+\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F^2)=\frac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})$. From (3.116) we have that $G_1(D_n)\le G_{1,R}(D_n)+|G_1(D_n)-G_{1,R}(D_n)|=\frac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})+O(n\frac{1}{\sigma^2}R^{1-\alpha})$. Choosing $R=n^{(\frac{2\alpha-1}{\alpha(\alpha-1)}+1)(1-t)}$ we conclude the proof. □

Lemma 86. *Assume $\sigma^2=\Theta(n^t)$ where $1-\frac{\alpha}{1+2\tau}<t<1$. Let $S=n^D$. Assume that $\|\xi\|_2=1$. When n is sufficiently large, with probability of at least $1-2\delta$ we have*

$$
\|(I+\tfrac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\Lambda_S\xi\|_2=O(\sqrt{(\tfrac{1}{\delta}+1)n}\cdot n^{-(1-t)}).
\tag{3.128}
$$

Proof of Lemma 86. Using the Woodbury matrix identity, we have that

$$\begin{aligned}((I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\Lambda_S\xi =&\big[I-\Phi_S(\sigma^2 I+\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\Phi_S^T\big]\Phi_S\Lambda_S\xi\\ =&\Phi_S\Lambda_S\xi-\Phi_S(\sigma^2 I+\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\Phi_S^T\Phi_S\Lambda_S\xi\\ =&\Phi_S(I+\tfrac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\xi.\end{aligned} \tag{3.129}$$

Let $A=(I+\frac{n}{\sigma^2}\Lambda_S)^{-\gamma/2}\Lambda_S^{\gamma/2}(\Phi_S^T\Phi_S-nI)\Lambda_S^{\gamma/2}(I+\frac{n}{\sigma^2}\Lambda_S)^{-\gamma/2}$, where $\gamma>\frac{1+\alpha+2\tau-(1+2\tau+2\alpha)t}{2\alpha(1-t)}$. By Corollary 68, with probability of at least $1-\delta$, we have $\|\frac{1}{\sigma^2}A\|_2=\tilde{O}(n^{\frac{1+\alpha+2\tau-(1+2\tau+2\alpha)t}{2\alpha}-\gamma(1-t)})$. When n is sufficiently large, $\|\frac{1}{\sigma^2}A\|_2$ is less than 1. By Lemma 73, we have

$$\begin{aligned}&(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\\ =&(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}+\sum_{j=1}^{\infty}(-1)^j\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S(\Phi_S^T\Phi_S-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}.\end{aligned}$$

Then we have

$$\begin{aligned}&\|(I+\tfrac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\xi\|_2\\ =&\left\|\left((I+\frac{n}{\sigma^2}\Lambda_S)^{-1}+\sum_{j=1}^{\infty}(-1)^j\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S(\Phi_S^T\Phi_S-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\right)\Lambda_S\xi\right\|_2\\ \le&\left(\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\xi\|_2+\sum_{j=1}^{\infty}\left\|\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S(\Phi_S^T\Phi_S-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\xi\right\|_2\right).\end{aligned} \tag{3.130}$$

For the first term in the right hand side of the last equation, we have

$$\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\xi\|_2\le\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\|_2\|\xi\|_2\le\frac{\sigma^2}{n}=O(n^{-(1-t)}). \tag{3.131}$$

Using the fact that $\|\frac{1}{\sigma^2}A\|_2=\tilde{O}(n^{\frac{1+\alpha+2\tau-(1+2\tau+2\alpha)t}{2\alpha}-\gamma(1-t)})$ and $\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\|_2\le n^{-1}$, we

have

$$\begin{aligned}
&\left\|\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S(\Phi_S^T\Phi_S-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\xi\right\|_2\\
=&\frac{1}{\sigma^{2j}}\left\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\frac{\gamma}{2}}\Lambda_S^{1-\frac{\gamma}{2}}\left(A(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\gamma}\Lambda_S^{1-\gamma}\right)^{j-1}A(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\frac{\gamma}{2}}\Lambda_S^{-\frac{\gamma}{2}}\Lambda_S\xi\right\|_2\\
\leq&n^{(1-t)(-1+\frac{\gamma}{2}+(-1+\gamma)(j-1))}\tilde{O}(n^{\frac{j(1+\alpha+2\tau-(1+2\tau+2\alpha)t)}{2\alpha}-j\gamma(1-t)})\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\frac{\gamma}{2}}\Lambda_S^{1-\frac{\gamma}{2}}\xi\|_2\\
=&\tilde{O}(n^{-\frac{\gamma}{2}(1-t)+\frac{(1-\alpha+2\tau-(1+2\tau)t)j}{2\alpha}})\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\frac{\gamma}{2}}\Lambda_S^{1-\frac{\gamma}{2}}\|_2\|\xi\|_2\\
=&\tilde{O}(n^{-\frac{\gamma}{2}(1-t)+\frac{(1-\alpha+2\tau-(1+2\tau)t)j}{2\alpha}})O(n^{(-1+\gamma/2)(1-t)})\\
=&\tilde{O}(n^{-(1-t)+\frac{(1-\alpha+2\tau-(1+2\tau)t)j}{2\alpha}}).
\end{aligned}\tag{3.132}$$

Using (3.130), (3.131) and (3.132), we have

$$\begin{aligned}
&\|(I+\tfrac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\xi\|_2\\
=&\left(\tilde{O}(n^{-(1-t)})+\sum_{j=1}^{\infty}\tilde{O}(n^{-1+\frac{(1-\alpha+2\tau-(1+2\tau)t)j}{2\alpha}})\right)\\
=&\left(\tilde{O}(n^{-(1-t)})+\tilde{O}(n^{-1+\frac{1-\alpha+2\tau-(1+2\tau)t}{2\alpha}})\right)\\
=&\tilde{O}(n^{-(1-t)}).
\end{aligned}\tag{3.133}$$

By Corollary 66, with probability of at least $1-\delta$, we have

$$\begin{aligned}
\|\Phi_S(I+\tfrac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\xi\|_2=&\tilde{O}(\sqrt{(\frac{1}{\delta}+1)n}\|(I+\tfrac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\xi\|_2)\\
=&\tilde{O}(\sqrt{(\frac{1}{\delta}+1)n}\cdot n^{-(1-t)}).
\end{aligned}$$

From (3.129) we get $\|(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}f_S(\mathbf{x})\|_2=\tilde{O}(\sqrt{(\frac{1}{\delta}+1)n}\cdot n^{-(1-t)})$. This concludes the proof. □

Lemma 87. *Assume $\sigma^2=\Theta(n^t)$ where $1-\frac{\alpha}{1+2\tau}<t<1$. Let $\delta=n^{-q}$ where $0\leq q<\frac{[\alpha-(1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$. Under Assumptions 50, 51 and 52, assume that $\mu_0=0$. Let $R=n^{(\frac{1}{\alpha}+\kappa)(1-t)}$ where $0<\kappa<\frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$. Then with probability of at least $1-6\delta$ over*

sample inputs $(x_i)_{i=1}^n$*, we have*

$$G_2(D_n) = \frac{(1+o(1))}{2\sigma^2}\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\|_2^2 = \tfrac{1}{\sigma^2}\Theta(n^{\max\{-2(1-t),\frac{(1-2\beta)(1-t)}{\alpha}\}}\log^{k/2} n),$$

$$\textit{where } k = \begin{cases} 0, & 2\alpha \neq 2\beta - 1, \\ 1, & 2\alpha = 2\beta - 1. \end{cases}.$$

Proof of Lemma 87. Let $S = n^D$. Let $G_{2,S}(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{2,S}(D_{n+1}) - T_{2,S}(D_n))$. By Lemma 79, when $S > n^{\max\{1,\frac{-t}{(\alpha-1-2\tau)}\}}$ with probability of at least $1-3\delta$ we have that

$$\begin{aligned}
&|G_2(D_n) - G_{2,S}(D_n)| = |\mathbb{E}_{(x_{n+1},y_{n+1})}[T_2(D_{n+1}) - T_{2,S}(D_{n+1})] - [T_2(D_n) - T_{2,S}(D_n)]| \\
&= \left|\mathbb{E}_{(x_{n+1},y_{n+1})}\tilde{O}\big((\tfrac{1}{\delta}+1)\tfrac{1}{\sigma^2}(n+1)S^{\max\{1/2-\beta,1-\alpha+2\tau\}}\big) - \tilde{O}\big((\tfrac{1}{\delta}+1)\tfrac{1}{\sigma^2}nS^{\max\{1/2-\beta,1-\alpha+2\tau\}}\big)\right| \\
&= \tilde{O}\big((\tfrac{1}{\delta}+1)\tfrac{1}{\sigma^2}nS^{\max\{1/2-\beta,1-\alpha+2\tau\}}\big) \qquad (3.134) \\
& \qquad (3.135)
\end{aligned}$$

Let $\Lambda_{1:S} = \text{diag}\{\lambda_1,\ldots,\lambda_S\}$, $\Phi_{1:S} = (\phi_1(\mathbf{x}),\phi_1(\mathbf{x}),\ldots,\phi_S(\mathbf{x}))$ and $\boldsymbol{\mu}_{1:S} = (\mu_1,\ldots,\mu_S)$. Since $\mu_0 = 0$, we have $T_{2,S}(D_n) = \frac{1}{2\sigma^2}\boldsymbol{\mu}_{1:S}^T\Phi_{1:S}^T(I+\frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}\boldsymbol{\mu}_{1:S}$. Define $\eta_{1:S} = (\phi_1(x_{n+1}),\ldots,\phi_S(x_{n+1}))^T$ and $\widetilde{\Phi}_{1:S} = (\Phi_{1:S}^T,\eta_{1:S})^T$. In the proof of Lemma 80, we showed that

$$\begin{aligned}
T_{2,S}(D_n) &= \frac{1}{2\sigma^2}\boldsymbol{\mu}_{1:S}^T\Phi_{1:S}^T(I+\frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}\boldsymbol{\mu}_{1:S} \\
&= \frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}\boldsymbol{\mu}_{1:S} - \frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}.
\end{aligned}$$

We have

$$
\begin{aligned}
&G_{2,S}(D_n)\\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{2,S}(D_{n+1}) - T_{2,S}(D_n))\\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}\boldsymbol{\mu}_{1:S} - \frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}(I+\frac{1}{\sigma^2}\Lambda_{1:S}\widetilde{\Phi}_S^T\widetilde{\Phi}_S)^{-1}\boldsymbol{\mu}_{1:S}\right)\\
&- \left(\frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}\boldsymbol{\mu}_{1:S} - \frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S})\right)\\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S} - \frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}(I+\frac{1}{\sigma^2}\Lambda_{1:S}\widetilde{\Phi}_S^T\widetilde{\Phi}_S)^{-1}\boldsymbol{\mu}_{1:S}\right)\\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2\sigma^2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}\frac{(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\Lambda_{1:S}\eta_{1:S}\eta_{1:S}^T(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}}{1+\frac{1}{\sigma^2}\eta_{1:S}^T(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\Lambda_{1:S}\eta_{1:S}}\boldsymbol{\mu}_{1:S})\right)\\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2\sigma^2}\frac{\boldsymbol{\mu}_{1:S}^T(I+\frac{1}{\sigma^2}\Phi_{1:S}^T\Phi_{1:S}\Lambda_{1:S})^{-1}\eta_{1:S}\eta_{1:S}^T(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}}{1+\frac{1}{\sigma^2}\eta_{1:S}^T(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\Lambda_{1:S}\eta_{1:S}})\right)\\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1+o(1)}{2\sigma^2}\boldsymbol{\mu}_{1:S}^T(I+\frac{1}{\sigma^2}\Phi_{1:S}^T\Phi_{1:S}\Lambda_{1:S})^{-1}\eta_{1:S}\eta_{1:S}^T(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\right)\\
&= \frac{1+o(1)}{2\sigma^2}\boldsymbol{\mu}_{1:S}^T(I+\frac{1}{\sigma^2}\Phi_{1:S}^T\Phi_{1:S}\Lambda_{1:S})^{-1}(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\\
&= \frac{1+o(1)}{2\sigma^2}\|(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\|_2^2,
\end{aligned}
\tag{3.136}
$$

where in the fourth to last equality we used the Sherman–Morrison formula, in the third inequality we used (3.119) , and in the last equality we used the fact that $\mathbb{E}_{(x_{n+1},y_{n+1})}\eta_{1:S}\eta_{1:S}^T = I$.

Let $\hat{\boldsymbol{\mu}}_{1:R} = (\mu_1,\ldots,\mu_R,0,\ldots,0) \in \mathbb{R}^S$. Then we have

$$
\begin{aligned}
&\|(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\|_2\\
&\le \|(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2 + \|(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}(\boldsymbol{\mu}_{1:S}-\hat{\boldsymbol{\mu}}_{1:R})\|_2,\\
&\|(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\|_2\\
&\ge \|(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2 - \|(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}(\boldsymbol{\mu}_{1:S}-\hat{\boldsymbol{\mu}}_{1:R})\|_2.
\end{aligned}
\tag{3.137}
$$

Let $R = n^{(\frac{1}{\alpha}+\kappa)(1-t)}$ where $0 < \kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$. In Lemma 75, (3.63), we showed that

with probability of at least $1-\delta$,

$$\begin{aligned}\|(I+\tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2 &= \Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n)\\ &= (1+o(1))\|(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2,\end{aligned} \tag{3.138}$$

where $k=\begin{cases}0, & 2\alpha\neq 2\beta-1,\\ 1, & 2\alpha=2\beta-1.\end{cases}$. The same proof holds if we replace $\Phi_{1:R}$ with $\Phi_{1:S}$, $\Lambda_{1:R}$ with $\Lambda_{1:S}$, and $\boldsymbol{\mu}_{1:R}$ with $\hat{\boldsymbol{\mu}}_{1:R}$. We have

$$\begin{aligned}\|(I+\tfrac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2 &= \Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n)\\ &= (1+o(1))\|(I+\frac{n}{\sigma^2}\Lambda_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2.\end{aligned} \tag{3.139}$$

Next we bound $\|(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}(\boldsymbol{\mu}_{1:S}-\hat{\boldsymbol{\mu}}_{1:R})\|_2$. By Assumption 51, we have that $\|\boldsymbol{\mu}_{1:S}-\hat{\boldsymbol{\mu}}_{1:R}\|_2=O(R^{\frac{1-2\beta}{2}})$. For any $\xi\in\mathbb{R}^S$ and $\|\xi\|_2=1$, using the Woodbury matrix identity, with probability of at least $1-2\delta$ we have

$$\begin{aligned}&|\xi^T(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}(\boldsymbol{\mu}_{1:S}-\hat{\boldsymbol{\mu}}_{1:R})|\\ &=|\xi^T\bigg(I-\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T(I+\frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}\bigg)(\boldsymbol{\mu}_{1:S}-\hat{\boldsymbol{\mu}}_{1:R})|\\ &=|\xi^T(\boldsymbol{\mu}_{1:S}-\hat{\boldsymbol{\mu}}_{1:R})-\frac{1}{\sigma^2}\xi^T\Lambda_{1:S}\Phi_{1:S}^T(I+\frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}(\boldsymbol{\mu}_{1:S}-\hat{\boldsymbol{\mu}}_{1:R})|\\ &\leq\|\xi\|_2\|\boldsymbol{\mu}_{1:S}-\hat{\boldsymbol{\mu}}_{1:R}\|_2+\frac{1}{\sigma^2}|\xi^T\Lambda_{1:S}\Phi_{1:S}^T(I+\frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}(\boldsymbol{\mu}_{1:S}-\hat{\boldsymbol{\mu}}_{1:R})|\\ &\leq O(R^{\frac{1-2\beta}{2}})+\frac{1}{\sigma^2}\|(I+\frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}\Lambda_{1:S}\xi\|_2\|\Phi_{1:S}(\boldsymbol{\mu}_{1:S}-\hat{\boldsymbol{\mu}}_{1:R})\|_2\\ &=O(R^{\frac{1-2\beta}{2}})+\frac{1}{\sigma^2}O(\sqrt{(\frac{1}{\delta}+1)n}\cdot n^{-(1-t)})O(\sqrt{(\frac{1}{\delta}+1)n}R^{\frac{1-2\beta}{2}})\\ &=O((\frac{1}{\delta}+1)R^{\frac{1-2\beta}{2}}),\end{aligned}$$

where in the second to last step we used Corollary 66 to show $\|\Phi_{1:S}(\boldsymbol{\mu}_{1:S}-\hat{\boldsymbol{\mu}}_{1:R})\|_2=O(\sqrt{(\frac{1}{\delta}+1)n}R^{\frac{1-2\beta}{2}})$ with probability of at least $1-\delta$, and Lemma 86 to show that $\|(I+\frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}\Lambda_{1:S}\xi\|_2=O(\sqrt{(\frac{1}{\delta}+1)n}\cdot n^{-1})$ with probability of at least $1-\delta$. Since

$R = n^{(\frac{1}{\alpha}+\kappa)(1-t)}$, we have

$$|\xi^T(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}(\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R})| = O((\frac{1}{\delta}+1)n^{\frac{(1-2\beta)(1-t)}{2\alpha}+\frac{(1-2\beta)(1-t)\kappa}{2}}).$$

Since ξ is arbitrary, we have

$$\|(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}(\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R})\|_2 = O((\frac{1}{\delta}+1)n^{\frac{(1-2\beta)(1-t)}{2\alpha}+\frac{(1-2\beta)(1-t)\kappa}{2}}).$$

Since $0 \le q < \frac{[\alpha-(1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$ and $0 < \kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$, we can choose $\kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$ and κ is arbitrarily close to $\kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$ such that $0 \le q < \frac{(2\beta-1)(1-t)\kappa}{2}$. Then we have $\frac{(1-2\beta)(1-t)\kappa}{2} + q < 0$. From (3.137) and (3.139), we have

$$\begin{aligned}
&\|(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\|_2 \\
=&\Theta(n^{\max\{-(1-t),\frac{(1-2\beta)(1-t)}{2\alpha}\}}\log^{k/2} n) + O((\frac{1}{\delta}+1)n^{\frac{(1-2\beta)(1-t)}{2\alpha}+\frac{(1-2\beta)(1-t)\kappa}{2}}) \\
=&\Theta(n^{\max\{-(1-t),\frac{(1-2\beta)(1-t)}{2\alpha}\}}\log^{k/2} n) + O((n^{q+\frac{(1-2\beta)(1-t)}{2\alpha}+\frac{(1-2\beta)(1-t)\kappa}{2}}) \\
=&\Theta(n^{\max\{-(1-t),\frac{(1-2\beta)(1-t)}{2\alpha}\}}\log^{k/2} n) \\
=&(1+o(1))\|(I + \frac{n}{\sigma^2}\Lambda_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2 \\
=&(1+o(1))\|(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\|_2.
\end{aligned} \tag{3.140}$$

Hence $G_{2,S}(D_n) = \frac{1+o(1)}{2\sigma^2}\|(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\|_2^2 = \frac{1}{\sigma^2}\Theta(n^{(1-t)\max\{-2,\frac{1-2\beta}{\alpha}\}}\log^{k/2} n)$. Then by (3.134), we have

$$G_2(D_n) = \frac{1}{\sigma^2}\Theta(n^{\max\{-2(1-t),\frac{(1-2\beta)(1-t)}{\alpha}\}}\log^{k/2} n) + \tilde{O}\bigg((\frac{1}{\delta}+1)\frac{n}{\sigma^2}S^{\max\{1/2-\beta,1-\alpha+2\tau\}}\bigg).$$

Choosing $S = n^{\max\left\{1,\frac{-t}{(\alpha-1-2\tau)},\left(\frac{1+q+\min\{2,\frac{2\beta-1}{\alpha}\}}{\min\{\beta-1/2,\alpha-1-2\tau\}}+1\right)(1-t)\right\}}$, we get the result. □

Proof of Theorem 55. From Lemmas 85 and 87 and $\frac{1}{\alpha}-1 > -2$, we have that with probability

of at least $1-7\tilde{\delta}$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n) =& \frac{1+o(1)}{2\sigma^2}(\mathrm{Tr}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R - \|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F^2 + \|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\|_2^2)\\ =& \tfrac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}}) + \tfrac{1}{\sigma^2}\Theta(n^{\max\{-2(1-t),\frac{(1-2\beta)(1-t)}{\alpha}\}}\log^{k/2} n)\\ =& \tfrac{1}{\sigma^2}\Theta(n^{\max\{\frac{(1-\alpha)(1-t)}{\alpha},\frac{(1-2\beta)(1-t)}{\alpha}\}})\end{aligned} \tag{3.141}$$

where $k = \begin{cases} 0, & 2\alpha \neq 2\beta - 1 \\ 1, & 2\alpha = 2\beta - 1 \end{cases}$, and $R = n^{(\frac{1}{\alpha}+\kappa)(1-t)}$, $\kappa > 0$.

Furthermore, we have

$$\begin{aligned}&\mathrm{Tr}(I+\frac{n}{\sigma^2}\Lambda)^{-1}\Lambda - \mathrm{Tr}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R\\ &= \sum_{p=R+1}^{\infty}\frac{\lambda_p}{1+\frac{n}{\sigma^2}\lambda_p} \leq \sum_{p=R+1}^{\infty}\frac{C_\lambda p^{-\alpha}}{1+\frac{n}{\sigma^2}C_\lambda p^{-\alpha}} \leq \sum_{p=R+1}^{\infty} C_\lambda p^{-\alpha} = \frac{n}{\sigma^2}O(R^{1-\alpha})\\ &= O(n^{(1-\alpha)(1-t)(\frac{1}{\alpha}+\kappa)})\\ &= o(n^{\frac{(1-\alpha)(1-t)}{\alpha}}).\end{aligned}$$

Then we have

$$\mathrm{Tr}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R = \mathrm{Tr}(I+\frac{n}{\sigma^2}\Lambda)^{-1}\Lambda(1+o(1)). \tag{3.142}$$

Similarly we can prove

$$\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F^2 = \|\Lambda^{1/2}(I+\frac{n}{\sigma^2}\Lambda)^{-1}\|_F^2(1+o(1)) \tag{3.143}$$

$$\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\|_2^2 = \|(I+\frac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu}\|_2^2(1+o(1)) \tag{3.144}$$

Letting $\delta = 7\tilde{\delta}$, the proof is complete. □

In the case of $\mu_0 > 0$, we have the following lemma:

Lemma 88. *Let $\delta = n^{-q}$ where $0 \leq q < \frac{[\alpha-(1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$. Under Assumptions 50, 51 and 52, assume that $\mu_0 > 0$. Then with probability of at least $1-6\delta$ over sample inputs $(x_i)_{i=1}^n$,*

we have $G_2(D_n) = \frac{1}{2\sigma^2}\mu_0^2 + o(1)$.

Proof of Lemma 88. Let $S = n^D$. Let $G_{2,S}(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{2,S}(D_{n+1}) - T_{2,S}(D_n))$. By Lemma 79, when $S > n^{\max\{1,\frac{-t}{(\alpha-1-2\tau)}\}}$, with probability of at least $1-3\delta$ we have that

$$
\begin{aligned}
&|G_2(D_n) - G_{2,S}(D_n)| = |\mathbb{E}_{(x_{n+1},y_{n+1})}[T_2(D_{n+1}) - T_{2,S}(D_{n+1})] - [T_2(D_n) - T_{2,S}(D_n)]| \\
&= \left|\mathbb{E}_{(x_{n+1},y_{n+1})}\tilde{O}\big((\tfrac{1}{\delta}+1)\tfrac{1}{\sigma^2}(n+1)S^{\max\{1/2-\beta,1-\alpha+2\tau\}}\big) - \tilde{O}\big((\tfrac{1}{\delta}+1)\tfrac{1}{\sigma^2}nS^{\max\{1/2-\beta,1-\alpha+2\tau\}}\big)\right| \\
&= \tilde{O}\big((\tfrac{1}{\delta}+1)\tfrac{1}{\sigma^2}nS^{\max\{1/2-\beta,1-\alpha+2\tau\}}\big)
\end{aligned}
$$

Let $\Lambda_S = \mathrm{diag}\{\lambda_1,\ldots,\lambda_S\}$, $\Phi_S = (\phi_1(\mathbf{x}),\phi_1(\mathbf{x}),\ldots,\phi_S(\mathbf{x}))$ and $\boldsymbol{\mu}_S = (\mu_1,\ldots,\mu_S)$. Define $\eta_S = (\phi_0(x_{n+1}),\phi_1(x_{n+1}),\ldots,\phi_S(x_{n+1}))^T$ and $\widetilde{\Phi}_S = (\Phi_S^T,\eta_S)^T$. By the same technique as in the proof of Lemma 80, we replace Λ_R by $\tilde{\Lambda}_{\epsilon,R} = \mathrm{diag}\{\epsilon,\lambda_1,\ldots,\lambda_R\}$, let $\epsilon\to 0$ and show the counterpart of the result (3.136) in the proof of Lemma 87:

$$
\begin{aligned}
G_{2,S}(D_n) &= \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{2,S}(D_{n+1}) - T_{2,S}(D_n)) \\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2\sigma^2}\frac{\boldsymbol{\mu}_S^T(I+\frac{1}{\sigma^2}\Phi_S^T\Phi_S\Lambda_S)^{-1}\eta_S\eta_S^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S}{1+\frac{1}{\sigma^2}\eta_S^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\eta_S})\right) \\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1+o(1)}{2\sigma^2}\boldsymbol{\mu}_S^T(I+\frac{1}{\sigma^2}\Phi_S^T\Phi_S\Lambda_S)^{-1}\eta_S\eta_S^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S\right) \\
&= \frac{1+o(1)}{2\sigma^2}\boldsymbol{\mu}_S^T(I+\frac{1}{\sigma^2}\Phi_S^T\Phi_S\Lambda_S)^{-1}(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S \\
&= \frac{1+o(1)}{2\sigma^2}\|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S\|_2^2,
\end{aligned}
\tag{3.145}
$$

where in the fourth to last equality we used the Sherman–Morrison formula, in the third inequality we used (3.119) , and in the last equality we used the fact that $\mathbb{E}_{(x_{n+1},y_{n+1})}\eta_{1:S}\eta_{1:S}^T = I$.

Let $\hat{\boldsymbol{\mu}}_R = (\mu_0,\mu_1,\ldots,\mu_R,0,\ldots,0) \in \mathbb{R}^S$. Then we have

$$
\begin{aligned}
\|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S\|_2 &\le \|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\hat{\boldsymbol{\mu}}_R\|_2 + \|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)\|_2, \\
\|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S\|_2 &\ge \|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\hat{\boldsymbol{\mu}}_R\|_2 - \|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)\|_2.
\end{aligned}
\tag{3.146}
$$

Choose $R = n^{(\frac{1}{\alpha}+\kappa)(1-t)}$ where $0 < \kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{\alpha^2(1-t)}$. In Lemma 75, (3.63), we showed that with probability of at least $1-\delta$,

$$\begin{aligned}\|(I + \tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2 &= \Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2} n) \\ &= (1+o(1))\|(I + \frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2,\end{aligned} \tag{3.147}$$

where $k = \begin{cases} 0, & 2\alpha \neq 2\beta - 1, \\ 1, & 2\alpha = 2\beta - 1. \end{cases}$. The same proof holds if we replace $\Phi_{1:R}$ with $\Phi_{1:S}$, $\Lambda_{1:R}$ with $\Lambda_{1:S}$, and $\boldsymbol{\mu}_{1:R}$ with $\hat{\boldsymbol{\mu}}_{1:R}$. We have

$$\begin{aligned}\|(I + \tfrac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2 &= \Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2} n) \\ &= (1+o(1))\|(I + \frac{n}{\sigma^2}\Lambda_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2.\end{aligned} \tag{3.148}$$

So we have

$$\begin{aligned}\|(I + \tfrac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\hat{\boldsymbol{\mu}}_R\|_2 &= \mu_0 + \Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2} n) \\ &= \mu_0 + o(1).\end{aligned} \tag{3.149}$$

Next we bound $\|(I + \frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\boldsymbol{\mu}_S - \hat{\boldsymbol{\mu}}_R)\|_2$. By Assumption 51, we have that $\|\boldsymbol{\mu}_S - \hat{\boldsymbol{\mu}}_R\|_2 = O(R^{\frac{1-2\beta}{2}})$. For any $\xi \in \mathbb{R}^S$ and $\|\xi\|_2 = 1$, using the Woodbury matrix identity, with

probability of at least $1-2\delta$ we have

$$\begin{aligned}
&|\xi^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)|\\
&=|\xi^T\bigg(I-\frac{1}{\sigma^2}\Lambda_S\Phi_S^T(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\bigg)(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)|\\
&=|\xi^T(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)-\frac{1}{\sigma^2}\xi^T\Lambda_S\Phi_S^T(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)|\\
&\le\|\xi\|_2\|\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R\|_2+\frac{1}{\sigma^2}|\xi^T\Lambda_S\Phi_S^T(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)|\\
&\le O(R^{\frac{1-2\beta}{2}})+\frac{1}{\sigma^2}\|(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\Lambda_S\xi\|_2\|\Phi_S(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)\|_2\\
&=O(R^{\frac{1-2\beta}{2}})+\frac{1}{\sigma^2}O(\sqrt{(\frac{1}{\delta}+1)n}\cdot n^{-(1-t)})O(\sqrt{(\frac{1}{\delta}+1)nR^{\frac{1-2\beta}{2}}})\\
&=O((\frac{1}{\delta}+1)R^{\frac{1-2\beta}{2}}),
\end{aligned}$$

where in the second to last step we used Corollary 66 to show $\|\Phi_S(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)\|_2=O(\sqrt{(\frac{1}{\delta}+1)n}$ $R^{\frac{1-2\beta}{2}})$ with probability of at least $1-\delta$, and Lemma 86 to show that $\|(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\Lambda_S\xi\|_2$ $=O(\sqrt{(\frac{1}{\delta}+1)n}\cdot n^{-(1-t)})$ with probability of at least $1-\delta$. Since $R=n^{(\frac{1}{\alpha}+\kappa)(1-t)}$, we have

$$|\xi^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)|=O((\frac{1}{\delta}+1)n^{\frac{(1-2\beta)(1-t)}{2\alpha}+\frac{(1-2\beta)(1-t)\kappa}{2}}).$$

Since ξ is arbitrary, we have $\|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)\|_2=O((\frac{1}{\delta}+1)n^{\frac{(1-2\beta)(1-t)}{2\alpha}+\frac{(1-2\beta)(1-t)\kappa}{2}})$. Since $0\le q<\frac{[\alpha-(1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$ and $0<\kappa<\frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$, we can choose $\kappa<\frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$ and κ is arbitrarily close to $\kappa<\frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$ such that $0\le q<\frac{(2\beta-1)(1-t)\kappa}{2}$. Then we have $\frac{(1-2\beta)(1-t)\kappa}{2}+q<0$. From (3.146) and (3.149), we have

$$\begin{aligned}
&\|(I+\tfrac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S\|_2\\
=&\mu_0+\Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n)+O((\frac{1}{\delta}+1)n^{\frac{(1-2\beta)(1-t)}{2\alpha}+\frac{(1-2\beta)(1-t)\kappa}{2}})\\
=&\mu_0+\Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n)\\
=&\mu_0+o(1).
\end{aligned}\tag{3.150}$$

Hence $G_{2,S}(D_n)=\frac{1+o(1)}{2\sigma^2}\|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S\|_2^2=\frac{1}{2\sigma^2}\mu_0^2+o(1)$. Then by (3.145), $G_2(D_n)=$

$\frac{1}{2\sigma^2}\mu_0^2 + o(1) + \tilde{O}\big((\frac{1}{\delta}+1)nS^{\max\{1/2-\beta,1-\alpha\}}\big)$.

Choosing $S = n^{\max\left\{1, \frac{-t}{(\alpha-1-2\tau)}, \left(\frac{1+q+\min\{2,\frac{2\beta-1}{\alpha}\}}{\min\{\beta-1/2,\alpha-1-2\tau\}}+1\right)(1-t)\right\}}$, we get the result. □

Proof of Theorem 57. According to Lemma 88, $G_2(D_n) = \frac{1}{2\sigma^2}\mu_0^2 + o(1)$. By Lemma 85, we have $G_1(D_n) = \Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})$. Then $\mathbb{E}_\epsilon G(D_n) = G_1(D_n) + G_2(D_n) = \frac{1}{2\sigma^2}\mu_0^2 + o(1)$. □

3.D.3 Proofs Related to the Excess Mean Squared Generalization Error

Proof of Theorem 58. For $\mu_0 = 0$, we can show that

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}} M(D_n) &= \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{x_{n+1}}[\bar{m}(x_{n+1}) - f(x_{n+1})]^2 \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{x_{n+1}}[K_{x_{n+1}\mathbf{x}}(K_n + \sigma_{\text{model}}^2 I_n)^{-1}\mathbf{y} - f(x_{n+1})]^2 \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{x_{n+1}}[\eta^T\Lambda\Phi^T[\Phi\Lambda\Phi^T + \sigma_{\text{model}}^2 I_n)^{-1}(\Phi\mu + \boldsymbol{\epsilon}) - \eta^T\mu]^2 \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{x_{n+1}}[\eta^T\Lambda\Phi^T(\Phi\Lambda\Phi^T + \sigma_{\text{model}}^2 I_n)^{-1}\boldsymbol{\epsilon}]^2 \\
&+ \mathbb{E}_{x_{n+1}}\big[\eta^T\big(\Lambda\Phi^T(\Phi\Lambda\Phi^T + \sigma_{\text{model}}^2 I_n)^{-1}\Phi - I\big)\mu\big]^2 \\
&= \sigma_{\text{true}}^2 \operatorname{Tr}\Lambda\Phi^T(\Phi\Lambda\Phi^T + \sigma_{\text{model}}^2 I_n)^{-2}\Phi\Lambda \\
&+ \mu^T\Big(I + \tfrac{1}{\sigma_{\text{model}}^2}\Phi^T\Phi\Lambda\Big)^{-1}\Big(I + \tfrac{1}{\sigma_{\text{model}}^2}\Lambda\Phi^T\Phi\Big)^{-1}\mu \\
&= \tfrac{\sigma_{\text{true}}^2}{\sigma_{\text{model}}^2}\operatorname{Tr}(I + \tfrac{\Lambda\Phi^T\Phi}{\sigma_{\text{model}}^2})^{-1}\Lambda - \operatorname{Tr}(I + \tfrac{\Lambda\Phi^T\Phi}{\sigma_{\text{model}}^2})^{-2}\Lambda + \|(I + \tfrac{1}{\sigma_{\text{model}}^2}\Lambda\Phi^T\Phi)^{-1}\mu\|_2^2.
\end{aligned}
$$

According to (3.140) from the proof of Lemma 87, the truncation procedure (3.134) and (3.144), with probability of at least $1-\delta$ we have

$$
\|(I+\tfrac{1}{\sigma_{\text{model}}^2}\Lambda\Phi^T\Phi)^{-1}\mu\|_2^2 = \Theta(n^{\max\{-2(1-t),\frac{(1-2\beta)(1-t)}{\alpha}\}}\log^{k/2} n) = (1+o(1))\|(I+\tfrac{n}{\sigma_{\text{model}}^2}\Lambda)^{-1}\boldsymbol{\mu}\|_2^2,
$$

where $k = \begin{cases} 0, & 2\alpha \neq 2\beta - 1, \\ 1, & 2\alpha = 2\beta - 1. \end{cases}$.

According to (3.122) and (3.127) from the proof of Lemma 85, the truncation procedure

(3.116), (3.142) and (3.143), with probability of at least $1-\delta$ we have

$$\begin{aligned}
&\mathrm{Tr}(I+\tfrac{\Lambda\Phi^T\Phi}{\sigma^2_{\text{model}}})^{-1}\Lambda - \mathrm{Tr}(I+\tfrac{\Lambda\Phi^T\Phi}{\sigma^2_{\text{model}}})^{-2}\Lambda \\
=&\Big(\mathrm{Tr}(I+\tfrac{n}{\sigma^2_{\text{model}}}\Lambda)^{-1}\Lambda\Big)(1+o(1)) - \|\Lambda^{1/2}(I+\tfrac{n}{\sigma^2_{\text{model}}}\Lambda)^{-1}\|_F^2(1+o(1)) \\
=&\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}}).
\end{aligned}$$

Combining the above two equations we get

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}} M(D_n) \\
=&(1+o(1))\Big(\tfrac{\sigma^2_{\text{true}}}{\sigma^2_{\text{model}}}\Big(\mathrm{Tr}(I+\tfrac{n}{\sigma^2_{\text{model}}}\Lambda)^{-1}\Lambda - \|\Lambda^{1/2}(I+\tfrac{n}{\sigma^2_{\text{model}}}\Lambda)^{-1}\|_F^2\Big) + \|(I+\tfrac{n}{\sigma^2_{\text{model}}}\Lambda)^{-1}\boldsymbol{\mu}\|_2^2\Big) \\
=&\tfrac{\sigma^2_{\text{true}}}{\sigma^2_{\text{model}}}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}}) + \Theta(n^{\max\{-2(1-t),\frac{(1-2\beta)(1-t)}{\alpha}\}}\log^{k/2} n) \\
=&\sigma^2_{\text{true}}\Theta(n^{\frac{1-\alpha-t}{\alpha}}) + \Theta(n^{\max\{-2(1-t),\frac{(1-2\beta)(1-t)}{\alpha}\}}\log^{k/2} n) \\
=&\Theta\Big(\max\{\sigma^2_{\text{true}} n^{\frac{1-\alpha-t}{\alpha}}, n^{\frac{(1-2\beta)(1-t)}{\alpha}}\}\Big)
\end{aligned}$$

When $\mu_0 > 0$, according to (3.150) in the proof of Lemma 88 and the truncation procedure (3.134), with probability of at least $1-\delta$ we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}} M(D_n) =&\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}}) + \mu_0^2 + o(1) \\
=&\mu_0^2 + o(1).
\end{aligned}$$

□

CHAPTER 4

Asymptotic Spectrum of the NTK via a Power Series Expansion *

4.1 Introduction

The spectrum of the NTK is fundamental to both the optimization and generalization of wide networks. In chapter 3, we show that the asymptotic generalization error of kernel ridge regression is closely related to the asymptotic spectrum. Moreover, bounding the smallest eigenvalue of the NTK Gram matrix is a staple technique for establishing convergence guarantees for the optimization [DLL19, DZP19, OS20]. Furthermore, the full spectrum of the NTK Gram matrix governs the dynamics of the empirical risk [ADH19c], and the eigenvalues of the associated integral operator characterize the dynamics of the generalization error outside the training set [BM22b, BM22a].

The importance of the spectrum of the NTK has led to a variety of efforts to characterize its structure via random matrix theory and other tools [YS19b, FW20]. There is a broader body of work studying the closely related Conjugate Kernel, Fisher Information Matrix, and Hessian [PLR16b, PW17, PW18, LLC18, KAA20]. [VY21a] demonstrated that for ReLU networks the spectrum of the NTK integral operator asymptotically follows a power law, which is consistent with our results for the uniform data distribution. [BJK19] calculated the NTK spectrum for shallow ReLU networks under the uniform distribution, which was

*This chapter is adapted from [MJB23], with the permission from coauthors. Michael Murray proposed the idea of NTK power series and gave out the expression of power series coefficients. Benjamin Bowman studied the effective rank of the NTK by its power series. I studied and computed the asymptotic spectrum of the NTK by its power series.

then expanded to the nonuniform case by [BGG20]. [GYK20] and [BB21] analyzed the reproducing kernel Hilbert spaces of the NTK for ReLU networks and the Laplace kernel via the decay rate of the spectrum of the kernel. [MJB23] characterize a variety of attributes of the spectrum for fixed input dimension using the power series expansion of NTK.

In this chapter, we analyze the asymptotic spectrum of dot-product kernel using the power series. In Theorem 91 we derive coefficients for the power series expansion of the NTK under unit variance initialization, see Assumption 90. Consequently we are able to derive insights into the NTK spectrum, notably concerning the outlier eigenvalues as well as the asymptotic decay. In Theorem 94 we characterize the asymptotic behavior of the NTK spectrum for uniform data distributions on the sphere. Our result shows that faster decay in the NTK power series coefficients implies a faster decay in its spectrum. Moreover, for NTK of shallow ReLU networks, our result recover the result of [BJK19]. At the end, we comment on how the activation function of the shallow networks influences the RKHS of the NTK. In the remainder of this introductory section, we review some related work.

Analysis of NTK Spectrum: theoretical analysis of the NTK spectrum via random matrix theory was investigated by [YS19b, FW20] in the high dimensional limit. [VY21a] demonstrated that for ReLU networks the spectrum of the NTK integral operator asymptotically follows a power law, which is consistent with our results for the uniform data distribution. [BJK19] calculated the NTK spectrum for shallow ReLU networks under the uniform distribution, which was then expanded to the nonuniform case by [BGG20]. [GYK20] and [BB21] analyzed the reproducing kernel Hilbert spaces of the NTK for ReLU networks and the Laplace kernel via the decay rate of the spectrum of the kernel. In contrast to previous works, we are able to address the spectrum in the finite dimensional setting and characterize the impact of different activation functions on it.

Hermite Expansion: [DFS16b] used Hermite expansion to the study the expressivity of the Conjugate Kernel. [SAD22] used this technique to demonstrate that any dot product kernel can be realized by the NTK or Conjugate Kernel of a shallow, zero bias network. [OS20] use Hermite expansion to study the NTK and establish a quantitative bound on the

smallest eigenvalue for shallow networks. This approach was incorporated by [NM20] to handle convergence for deep networks, with sharp bounds on the smallest NTK eigenvalue for deep ReLU networks provided by [NMM21]. The Hermite approach was utilized by [PSG20] to analyze the smallest NTK eigenvalue of shallow networks under various activations. Finally, in a concurrent work [HZL22] use Hermite expansions to develop a principled and efficient polynomial based approximation algorithm for the NTK and CNTK. In contrast to the aforementioned works, here we employ the Hermite expansion to characterize both the outlier and asymptotic portions of the spectrum for both shallow and deep networks under general activations.

4.2 Notations and Preliminaries

For our notation, lower case letters, e.g., x, y, denote scalars, lower case bold characters, e.g., $\mathbf{x}, \mathbf{y}$ are for vectors, and upper case bold characters, e.g., $\mathbf{X}, \mathbf{Y}$, are for matrices. For natural numbers $k_1, k_2 \in \mathbb{N}$ we let $[k_1] = \{1, \ldots, k_1\}$ and $[k_2, k_1] = \{k_2, \ldots, k_1\}$. If $k_2 > k_1$ then $[k_2, k_1]$ is the empty set. We use $\|\cdot\|_p$ to denote the p-norm of the matrix or vector in question and as default use $\|\cdot\|$ as the operator or 2-norm respectively. We use $\mathbf{1}_{m\times n} \in \mathbb{R}^{m\times n}$ to denote the matrix with all entries equal to one. We define $\delta_{p=c}$ to take the value 1 if $p = c$ and be zero otherwise. We will frequently overload scalar functions $\phi : \mathbb{R} \to \mathbb{R}$ by applying them elementwise to vectors and matrices. The entry in the ith row and jth column of a matrix we access using the notation $[\mathbf{X}]_{ij}$. The Hadamard or entrywise product of two matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$ we denote $\mathbf{X} \odot \mathbf{Y}$ as is standard. The pth Hadamard power we denote $\mathbf{X}^{\odot p}$ and define it as the Hadamard product of $\mathbf{X}$ with itself p times,

$$\mathbf{X}^{\odot p} := \mathbf{X} \odot \mathbf{X} \odot \cdots \odot \mathbf{X}.$$

Given a Hermitian or symmetric matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$, we adopt the convention that $\lambda_i(\mathbf{X})$ denotes the ith largest eigenvalue,

$$\lambda_1(\mathbf{X}) \geq \lambda_2(\mathbf{X}) \geq \cdots \geq \lambda_n(\mathbf{X}).$$

Finally, for a square matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ we let $Tr(\mathbf{X}) = \sum_{i=1}^{n} [\mathbf{X}]_{ii}$ denote the trace.

4.2.1 Hermite Expansion

We say that a function $f \colon \mathbb{R} \to \mathbb{R}$ is square integrable with respect to the standard Gaussian measure $\gamma(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ if $\mathbb{E}_{X \sim \mathcal{N}(0,1)}[f(X)^2] < \infty$. We denote by $L^2(\mathbb{R}, \gamma)$ the space of all such functions. The normalized probabilist's Hermite polynomials are defined as

$$h_k(x) = \frac{(-1)^k e^{x^2/2}}{\sqrt{k!}} \frac{d^k}{dx^k} e^{-x^2/2}, \quad k = 0, 1, \ldots$$

and form a complete orthonormal basis in $L^2(\mathbb{R}, \gamma)$ [OD14, §11]. The Hermite expansion of a function $\phi \in L^2(\mathbb{R}, \gamma)$ is given by $\phi(x) = \sum_{k=0}^{\infty} \mu_k(\phi) h_k(x)$, where $\mu_k(\phi) = \mathbb{E}_{X \sim \mathcal{N}(0,1)}[\phi(X) h_k(X)]$ is the kth normalized probabilist's Hermite coefficient of ϕ.

4.2.2 NTK Parametrization

In what follows, for $n, d \in \mathbb{N}$ let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote a matrix which stores n points in $\mathbb{R}^d$ row-wise. Unless otherwise stated, we assume $d \leq n$ and denote the ith row of $\mathbf{X}_n$ as $\mathbf{x}_i$. In this chapter we consider fully-connected neural networks of the form $f^{(L+1)} \colon \mathbb{R}^d \to \mathbb{R}$ with $L \in \mathbb{N}$ hidden layers and a linear output layer. For a given input vector $\mathbf{x} \in \mathbb{R}^d$, the activation $f^{(l)}$ and preactivation $g^{(l)}$ at each layer $l \in [L+1]$ are defined via the following

recurrence relations,

$$
\begin{aligned}
&g^{(1)}(\mathbf{x}) = \gamma_w \mathbf{W}^{(1)}\mathbf{x} + \gamma_b \mathbf{b}^{(1)},\ f^{(1)}(\mathbf{x}) = \phi\big(g^{(1)}(\mathbf{x})\big),\\
&g^{(l)}(\mathbf{x}) = \frac{\sigma_w}{\sqrt{m_{l-1}}}\mathbf{W}^{(l)} f^{(l-1)}(\mathbf{x}) + \sigma_b \mathbf{b}^{(l)},\ f^{(l)}(\mathbf{x}) = \phi\big(g^{(l)}(\mathbf{x})\big),\ \forall l \in [2, L],\\
&g^{(L+1)}(\mathbf{x}) = \frac{\sigma_w}{\sqrt{m_L}}\mathbf{W}^{(L+1)} f^{(L)}(\mathbf{x}),\ f^{(L+1)}(\mathbf{x}) = g^{(L+1)}(\mathbf{x}).
\end{aligned}
\tag{4.1}
$$

The parameters $\mathbf{W}^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$ and $\mathbf{b}^{(l)} \in \mathbb{R}^{m_l}$ are the weight matrix and bias vector at the lth layer respectively, $m_0 = d$, $m_{L+1} = 1$, and $\phi\colon \mathbb{R} \to \mathbb{R}$ is the activation function applied elementwise. The variables $\gamma_w, \sigma_w \in \mathbb{R}_{>0}$ and $\gamma_b, \sigma_b \in \mathbb{R}_{\geq 0}$ correspond to weight and bias hyperparameters respectively. Let $\theta_l \in \mathbb{R}^p$ denote a vector storing the network parameters $(\mathbf{W}^{(h)}, \mathbf{b}^{(h)})_{h=1}^{l}$ up to and including the lth layer. The Neural Tangent Kernel [JGH18c] $\tilde{\Theta}^{(l)}\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ associated with $f^{(l)}$ at layer $l \in [L+1]$ is defined as

$$
\tilde{\Theta}^{(l)}(\mathbf{x}, \mathbf{y}) := \langle \nabla_{\theta_l} f^{(l)}(\mathbf{x}), \nabla_{\theta_l} f^{(l)}(\mathbf{y}) \rangle. \tag{4.2}
$$

We will mostly study the NTK under the following standard assumptions.

Assumption 89. *NTK initialization.*

1. *At initialization all network parameters are distributed as $\mathcal{N}(0,1)$ and are mutually independent.*

2. *The activation function satisfies $\phi \in L^2(\mathbb{R}, \gamma)$, is differentiable almost everywhere and its derivative, which we denote ϕ', also satisfies $\phi' \in L^2(\mathbb{R}, \gamma)$.*

3. *The widths are sent to infinity in sequence, $m_1 \to \infty, m_2 \to \infty, \ldots, m_L \to \infty$. We refer to this regime as the sequential infinite width limit.*

Under Assumption 89, for any $l \in [L+1]$, $\tilde{\Theta}^{(l)}(\mathbf{x}, \mathbf{y})$ converges in probability to a deterministic limit $\Theta^{(l)}\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ [JGH18c] and the network behaves like a kernelized linear predictor during training; see, e.g., [ADH19c, LXS19c, WGL20]. Given access to the

rows $(\mathbf{x}_i)_{i=1}^n$ of $\mathbf{X}$ the NTK matrix at layer $l \in [L+1]$, which we denote $\mathbf{K}_l$, is the $n \times n$ matrix with entries defined as

$$[\mathbf{K}_l]_{ij} = \frac{1}{n}\Theta^{(l)}(\mathbf{x}_i, \mathbf{x}_j), \ \forall (i,j) \in [n] \times [n]. \tag{4.3}$$

4.3 Expressing the NTK as a Power Series

We derive a power series for the NTK under the following assumptions on the network initialization hyperparameters.

Assumption 90. *The hyperparameters of the network satisfy* $\gamma_w^2 + \gamma_b^2 = 1$, $\sigma_w^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\phi(Z)^2] \leq 1$ *and* $\sigma_b^2 = 1 - \sigma_w^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\phi(Z)^2]$. *The data is normalized so that* $\|\mathbf{x}_i\| = 1$ *for all* $i \in [n]$.

Recall under Assumption 89 that the preactivations of the network are centered Gaussian processes [Nea96b, LBN18]. Assumption 90 ensures the preactivation of each neuron has unit variance and thus is reminiscent of the [LBO12], [GB10] and [HZR15] initializations, which are designed to avoid vanishing and exploding gradients. We refer the reader to Appendix 4.A.3 for a thorough discussion. Under Assumption 90 we will also show it is possible to write the NTK not only as a dot-product kernel, but also as an analytic power series on $[-1, 1]$. In order to state this result recall, given a function $f \in L^2(\mathbb{R}, \gamma)$, that we denote the pth normalized probabilist's Hermite coefficient of f as $\mu_p(f)$, we refer the reader to Appendix 4.A.4 for an overview of the Hermite polynomials and their properties. Furthermore, letting $\bar{a} = (a_j)_{j=0}^{\infty}$ denote a sequence of real numbers, then for any $p, k \in \mathbb{Z}_{\geq 0}$ we define

$$F(p, k, \bar{a}) = \begin{cases} 1, & k = 0 \text{ and } p = 0, \\ 0, & k = 0 \text{ and } p \geq 1, \\ \sum_{(j_i)\in\mathcal{J}(p,k)} \prod_{i=1}^k a_{j_i}, & k \geq 1 \text{ and } p \geq 0, \end{cases} \tag{4.4}$$

where

$$\mathcal{J}(p,k) := \left\{(j_i)_{i\in[k]} \;:\; j_i \geq 0 \;\forall i \in [k],\; \sum_{i=1}^{k} j_i = p\right\} \quad \text{for all } p \in \mathbb{Z}_{\geq 0},\; k \in \mathbb{N}.$$

Here $\mathcal{J}(p,k)$ is the set of all k-tuples of nonnegative integers which sum to p and $F(p,k,\bar{a})$ is therefore the sum of all ordered products of k elements of $\bar{a}$ whose indices sum to p. We are now ready to state the key result of this section, Theorem 91, whose proof is provided in Appendix 4.B.1.

Theorem 91. *Under Assumptions 89 and 90, for all $l \in [L+1]$*

$$n\mathbf{K}_l = \sum_{p=0}^{\infty} \kappa_{p,l}\left(\mathbf{X}\mathbf{X}^T\right)^{\odot p}. \tag{4.5}$$

The series for each entry $n[\mathbf{K}_l]_{ij}$ converges absolutely and the coefficients $\kappa_{p,l}$ are nonnegative and can be evaluated using the recurrence relationships

$$\kappa_{p,l} = \begin{cases} \delta_{p=0}\gamma_b^2 + \delta_{p=1}\gamma_w^2, & l = 1, \\ \alpha_{p,l} + \sum_{q=0}^{p} \kappa_{q,l-1}\upsilon_{p-q,l}, & l \in [2, L+1], \end{cases} \tag{4.6}$$

where

$$\alpha_{p,l} = \begin{cases} \sigma_w^2\mu_p^2(\phi) + \delta_{p=0}\sigma_b^2, & l = 2, \\ \sum_{k=0}^{\infty} \alpha_{k,2} F(p,k,\bar{\alpha}_{l-1}), & l \geq 3, \end{cases} \tag{4.7}$$

and

$$\upsilon_{p,l} = \begin{cases} \sigma_w^2\mu_p^2(\phi'), & l = 2, \\ \sum_{k=0}^{\infty} \upsilon_{k,2} F(p,k,\bar{\alpha}_{l-1}), & l \geq 3, \end{cases} \tag{4.8}$$

are likewise nonnegative for all $p \in \mathbb{Z}_{\geq 0}$ and $l \in [2, L+1]$.

To compute the coefficients of the NTK as per Theorem 91, the Hermite coefficients of both ϕ and ϕ' are required. Under Assumption 92 below, which has minimal impact on the generality of our results, this calculation can be simplified. In short, under Assumption 92

$\upsilon_{p,2} = (p+1)\alpha_{p+1,2}$ and therefore only the Hermite coefficients of ϕ are required. We refer the reader to Lemma 102 in Appendix 4.B.2 for further details.

Assumption 92. *The activation function $\phi\colon \mathbb{R} \to \mathbb{R}$ is absolutely continuous on $[-a, a]$ for all $a > 0$, differentiable almost everywhere, and is polynomially bounded, i.e., $|\phi(x)| = \mathcal{O}(|x|^\beta)$ for some $\beta > 0$. Further, the derivative $\phi'\colon \mathbb{R} \to \mathbb{R}$ satisfies $\phi' \in L^2(\mathbb{R}, \gamma)$.*

We remark that ReLU, Tanh, Sigmoid, Softplus and many other commonly used activation functions satisfy Assumption 92. In order to understand the relationship between the Hermite coefficients of the activation function and the coefficients of the NTK, we first consider the simple two-layer case with $L = 1$ hidden layers. From Theorem 91

$$\kappa_{p,2} = \sigma_w^2(1 + \gamma_w^2 p)\mu_p^2(\phi) + \sigma_w^2\gamma_b^2(1+p)\mu_{p+1}^2(\phi) + \delta_{p=0}\sigma_b^2. \tag{4.9}$$

As per Table 4.1, a general trend we observe across all activation functions is that the first few coefficients account for the large majority of the total NTK coefficient series.

Table 4.1: Percentage of $\sum_{p=0}^{\infty} \kappa_{p,2}$ accounted for by the first $T+1$ NTK coefficients assuming $\gamma_w^2 = 1$, $\gamma_b^2 = 0$, $\sigma_w^2 = 1$ and $\sigma_b^2 = 1 - \mathbb{E}[\phi(Z)^2]$.

$T =$	0	1	2	3	4	5
ReLU	43.944	77.277	93.192	93.192	95.403	95.403
Tanh	41.362	91.468	91.468	97.487	97.487	99.090
Sigmoid	91.557	99.729	99.729	99.977	99.977	99.997
Gaussian	95.834	95.834	98.729	98.729	99.634	99.634

However, the asymptotic rate of decay of the NTK coefficients varies significantly by activation function, due to the varying behavior of their tails. In Lemma 93 we choose ReLU, Tanh and Gaussian as prototypical examples of activations functions with growing, constant, and decaying tails respectively, and analyze the corresponding NTK coefficients in the two layer setting. For typographical ease we denote the zero mean Gaussian density function with variance σ^2 as $\omega_\sigma(z) := (1/\sqrt{2\pi\sigma^2})\exp(-z^2/(2\sigma^2))$.

Lemma 93. *Under Assumptions 89 and 90,*

1. *if* $\phi(z) = ReLU(z)$, *then* $\kappa_{p,2} = \delta_{(\gamma_b>0)\cup(p\ even)}\Theta(p^{-3/2})$,
2. *if* $\phi(z) = Tanh(z)$, *then* $\kappa_{p,2} = \mathcal{O}\Big(\exp\Big(-\frac{\pi\sqrt{p-1}}{2}\Big)\Big)$,
3. *if* $\phi(z) = \omega_\sigma(z)$, *then* $\kappa_{p,2} = \delta_{(\gamma_b>0)\cup(p\ even)}\Theta(p^{1/2}(\sigma^2+1)^{-p})$.

The trend we observe from Lemma 93 is that activation functions whose Hermite coefficients decay quickly, such as ω_σ, result in a faster decay of the NTK coefficients. We remark that analyzing the rates of decay in the deep setting is challenging due to the calculation of $F(p, k, \bar{\alpha}_{l-1})$ (4.4) and therefore leave this study to future work.

4.4 Analyzing the Asymptotic Spectrum of the NTK via its Power Series

We analyze the spectrum of kernel function K which is a dot-product kernel of the form $K(x_1, x_2) = \sum_{p=0}^{\infty} c_p \langle x_1, x_2\rangle^p$. Assuming the training data is uniformly distributed on a hypersphere it was shown by [BJK19, BM19] that the eigenfunctions of K are the spherical harmonics. The following theorem gives the eigenvalues of the kernel K in this setting.

Theorem 94. *Suppose that the training data are uniformly sampled from the unit hypersphere* $\mathbb{S}^d$, $d \geq 2$. *If the dot-product kernel function has the expansion* $K(x_1, x_2) = \sum_{p=0}^{\infty} c_p \langle x_1, x_2\rangle^p$ *where* $c_p \geq 0$, *then the eigenvalue of every spherical harmonic of frequency* k *is given by*

$$\overline{\lambda_k} = \frac{\pi^{d/2}}{2^{k-1}} \sum_{\substack{p\geq k \\ p-k\ is\ even}} c_p \frac{\Gamma(p+1)\Gamma(\frac{p-k+1}{2})}{\Gamma(p-k+1)\Gamma(\frac{p-k+1}{2}+k+d/2)},$$

where Γ *is the gamma function.*

A proof of Theorem 94 is provided in Appendix 4.C.2. This theorem connects the coefficients c_p of the kernel power series with the eigenvalues $\overline{\lambda_k}$ of the kernel. In particular,

given a specific decay rate for the coefficients c_p one may derive the decay rate of $\overline{\lambda_k}$ as illustrated in the following Corollary.

Corollary 95. *Under the same setting as in Theorem 94,*

1. *if* $c_p = \Theta(p^{-a})$ *where* $a \geq 1$, *then* $\overline{\lambda_k} = \Theta(k^{-d-2a+2})$,
2. *if* $c_p = \delta_{(p \text{ even})}\Theta(p^{-a})$, *then* $\overline{\lambda_k} = \delta_{(k \text{ even})}\Theta(k^{-d-2a+2})$,
3. *if* $c_p = \mathcal{O}\big(\exp\big(-a\sqrt{p}\big)\big)$, *then* $\overline{\lambda_k} = \mathcal{O}\Big(k^{-d+1/2}\exp\Big(-a\sqrt{k}\Big)\Big)$,
4. *if* $c_p = \Theta(p^{1/2}a^{-p})$, *then* $\overline{\lambda_k} = \mathcal{O}\big(k^{-d+1}a^{-k}\big)$ *and* $\overline{\lambda_k} = \Omega\big(k^{-d/2+1}2^{-k}a^{-k}\big)$.

A proof of Corollary 95 is provided in Appendix 4.C.2. For the NTK of a two-layer ReLU network with $\gamma_b > 0$, then according to Lemma 3.2 of [MJB23], we have $c_p = \kappa_{p,2} = \Theta(p^{-3/2})$. Therefore using Corollary 95 $\overline{\lambda_k} = \Theta(k^{-d-1})$. Notice here that k refers to the frequency, and the number of spherical harmonics of frequency at most k is $\Theta(k^d)$. Therefore, for the lth largest eigenvalue λ_l we have $\lambda_l = \Theta(l^{-(d+1)/d})$. This rate agrees with [BJK19] and [VY21a]. For the NTK of a two-layer ReLU network with $\gamma_b = 0$, the eigenvalues corresponding to the even frequencies are 0, which also agrees with [BJK19]. Corollary 95 and Lemma 3.2 of [MJB23] also shows the decay rates of eigenvalues for the NTK of two-layer networks with Tanh activation and Gaussian activation. We observe that when the coefficients of the kernel power series decay quickly then the eigenvalues of the kernel also decay quickly. As faster decay of the eigenvalues of the kernel implies a smaller RKHS [GYK20], Corollary 95 demonstrates that using ReLU results in a larger RKHS relative to using either Tanh or Gaussian activations. We numerically illustrate Corollary 95 in Figure 4.1, Appendix 4.C.1.

Appendix

The appendix is organized as follows.

- Appendix 4.A gives background material on Gaussan kernels, NTK, unit variance intitialization, and Hermite polynomial expansions.

- Appendix 4.B provides details for Section 4.3.
- Appendix 4.C provides details for Section 4.4.

4.A Background Material

4.A.1 Gaussian Kernel

Observe by construction that the flattened collection of preactivations at the first layer $(g^{(1)}(\mathbf{x}_i))_{i=1}^n$ form a centered Gaussian process, with the covariance between the αth and βth neuron being described by

$$\Sigma^{(1)}_{\alpha_\beta}(\mathbf{x}_i, \mathbf{x}_j) := \mathbb{E}[g^{(1)}_\alpha(\mathbf{x}_i) g^{(1)}_\beta(\mathbf{x}_j)] = \delta_{\alpha=\beta}\big(\gamma_w^2 \mathbf{x}_i^T \mathbf{x}_j + \gamma_b^2\big).$$

Under the Assumption 89, the preactivations at each layer $l \in [L+1]$ converge also in distribution to centered Gaussian processes [Nea96b, LBN18]. We remark that the sequential width limit condition of Assumption 89 is not necessary for this behavior, for example the same result can be derived in the setting where the widths of the network are sent to infinity simultaneously under certain conditions on the activation function [dHR18]. However, as our interests lie in analyzing the limit rather than the conditions for convergence to said limit, for simplicity we consider only the sequential width limit. As per [LBN18, Eq. 4], the covariance between the preactivations of the αth and βth neurons at layer $l \geq 2$ for any input pair $\mathbf{x}, \mathbf{y} \in \mathbb{R}$ are described by the following kernel,

$$\begin{aligned}\Sigma^{(l)}_{\alpha_\beta}(\mathbf{x}, \mathbf{y}) &:= \mathbb{E}[g^{(l)}_\alpha(\mathbf{x}) g^{(l)}_\beta(\mathbf{y})] \\ &= \delta_{\alpha=\beta}\Big(\sigma_w^2 \mathbb{E}_{g^{(l-1)} \sim \mathcal{GP}(0, \Sigma^{l-1})}[\phi(g^{(l-1)}_\alpha(\mathbf{x}))\phi(g^{(l-1)}_\beta(\mathbf{y}))] + \sigma_b^2\Big).\end{aligned}$$

We refer to this kernel as the Gaussian kernel. As each neuron is identically distributed and the covariance between pairs of neurons is 0 unless $\alpha = \beta$, moving forward we drop the subscript and discuss only the covariance between the preactivations of an arbitrary neuron

given two inputs. As per the discussion by [LBN18, Section 2.3], the expectations involved in the computation of these Gaussian kernels can be computed with respect to a bivariate Gaussian distribution, whose covariance matrix has three distinct entries: the variance of a preactivation of $\mathbf{x}$ at the previous layer, $\Sigma^{(l-1)}(\mathbf{x},\mathbf{x})$, the variance of a preactivation of $\mathbf{y}$ at the previous layer, $\Sigma^{(l)}(\mathbf{y},\mathbf{y})$, and the covariance between preactivations of $\mathbf{x}$ and $\mathbf{y}$, $\Sigma^{(l-1)}(\mathbf{x},\mathbf{y})$. Therefore the Gaussian kernel, or covariance function, and its derivative, which we will require later for our analysis of the NTK, can be computed via the the following recurrence relations, see for instance [LBN18, JGH18c, ADH19c, NMM21],

$$\begin{aligned}
\Sigma^{(1)}(\mathbf{x},\mathbf{y}) &= \gamma_w^2 \mathbf{x}^T\mathbf{x} + \gamma_b^2, \\
\mathbf{A}^{(l)}(\mathbf{x},\mathbf{y}) &= \begin{bmatrix} \Sigma^{(l-1)}(\mathbf{x},\mathbf{x}) & \Sigma^{(l-1)}(\mathbf{x},\mathbf{y}) \\ \Sigma^{(l-1)}(\mathbf{y},\mathbf{x}) & \Sigma^{(l-1)}(\mathbf{x},\mathbf{x}) \end{bmatrix} \\
\Sigma^{(l)}(\mathbf{x},\mathbf{y}) &= \sigma_w^2 \mathbb{E}_{(B_1,B_2)\sim\mathcal{N}(0,\mathbf{A}^{(l)}(\mathbf{x},\mathbf{y}))}[\phi(B_1)\phi(B_2)] + \sigma_b^2, \\
\dot{\Sigma}^{(l)}(\mathbf{x},\mathbf{y}) &= \sigma_w^2 \mathbb{E}_{(B_1,B_2)\sim\mathcal{N}(0,\mathbf{A}^{(l)}(\mathbf{x},\mathbf{y}))}[\phi'(B_1)\phi'(B_2)].
\end{aligned} \tag{4.10}$$

4.A.2 Neural Tangent Kernel (NTK)

Under Assumption 89 $\tilde{\Theta}^{(l)}$ converges in probability to a deterministic limit, which we denote $\Theta^{(l)}$. This deterministic limit kernel can be expressed in terms of the Gaussian kernels and their derivatives from Section 4.A.1 via the following recurrence relationships [JGH18c, Theorem 1],

$$\begin{aligned}
\Theta^{(1)}(\mathbf{x},\mathbf{y}) &= \Sigma^{(1)}(\mathbf{x},\mathbf{y}), \\
\Theta^{(l)}(\mathbf{x},\mathbf{y}) &= \Theta^{(l-1)}(\mathbf{x},\mathbf{y})\dot{\Sigma}^{(l)}(\mathbf{x},\mathbf{y}) + \Sigma^{(l)}(\mathbf{x},\mathbf{y}) \\
&= \Sigma^{(l)}(\mathbf{x},\mathbf{y}) + \sum_{h=1}^{l-1} \Sigma^{(h)}(\mathbf{x},\mathbf{y}) \left(\prod_{h'=h+1}^{l} \dot{\Sigma}^{(h')}(\mathbf{x},\mathbf{y}) \right) \forall l \in [2, L+1].
\end{aligned} \tag{4.11}$$

A useful expression for the NTK matrix, which is a straightforward extension and generalization of [NMM21, Lemma 3.1], is provided in Lemma 96 below.

Lemma 96. *(Based on Lemma 3.1 in [NMM21]) Under Assumption 89, a sequence of positive semidefinite matrices $(\mathbf{G}_l)_{l=1}^{L+1}$ in $\mathbb{R}^{n\times n}$, and the related sequence $(\dot{\mathbf{G}}_l)_{l=2}^{L+1}$ also in $\mathbb{R}^{n\times n}$, can be constructed via the following recurrence relationships,*

$$\begin{aligned}
\mathbf{G}_1 &= \gamma_w^2 \mathbf{X}\mathbf{X}^T + \gamma_b^2 \boldsymbol{1}_{n\times n},\\
\mathbf{G}_2 &= \sigma_w^2 \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_d)}[\phi(\mathbf{X}\mathbf{w})\phi(\mathbf{X}\mathbf{w})^T] + \sigma_b^2 \boldsymbol{1}_{n\times n},\\
\dot{\mathbf{G}}_2 &= \sigma_w^2 \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_n)}[\phi'(\mathbf{X}\mathbf{w})\phi'(\mathbf{X}\mathbf{w})^T],\\
\mathbf{G}_l &= \sigma_w^2 \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_n)}[\phi(\sqrt{\mathbf{G}_{l-1}}\mathbf{w})\phi(\sqrt{\mathbf{G}_{l-1}}\mathbf{w})^T] + \sigma_b^2 \boldsymbol{1}_{n\times n},\ l\in[3,L+1],\\
\dot{\mathbf{G}}_l &= \sigma_w^2 \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_n)}[\phi'(\sqrt{\mathbf{G}_{l-1}}\mathbf{w})\phi'(\sqrt{\mathbf{G}_{l-1}}\mathbf{w})^T],\ l\in[3,L+1].
\end{aligned} \tag{4.12}$$

The sequence of NTK matrices $(\mathbf{K}_l)_{l=1}^{L+1}$ can in turn be written using the following recurrence relationship,

$$\begin{aligned}
n\mathbf{K}_1 &= \mathbf{G}_1,\\
n\mathbf{K}_l &= \mathbf{G}_l + n\mathbf{K}_{l-1}\odot\dot{\mathbf{G}}_l\\
&= \mathbf{G}_l + \sum_{i=1}^{l-1}\Big(\mathbf{G}_i\odot\Big(\odot_{j=i+1}^{l}\dot{\mathbf{G}}_j\Big)\Big).
\end{aligned} \tag{4.13}$$

Proof. For the sequence $(\mathbf{G}_l)_{l=1}^{L+1}$ it suffices to prove for any $i,j\in[n]$ and $l\in[L+1]$ that

$$[\mathbf{G}_l]_{i,j} = \Sigma^{(l)}(\mathbf{x}_i,\mathbf{x}_j)$$

and $\mathbf{G}_l$ is positive semi-definite. We proceed by induction, considering the base case $l=1$ and comparing (4.12) with (4.10) then it is evident that

$$[\mathbf{G}_1]_{i,j} = \Sigma^{(1)}(\mathbf{x}_i,\mathbf{x}_j).$$

In addition, $\mathbf{G}_1$ is also clearly positive semi-definite as for any $\mathbf{u}\in\mathbb{R}^n$

$$\mathbf{u}^T\mathbf{G}_1\mathbf{u} = \gamma_w^2\|\mathbf{X}^T\mathbf{u}\|^2 + \gamma_b^2\|\mathbf{1}_n^T\mathbf{u}\|^2 \geq 0.$$

We now assume the induction hypothesis is true for $\mathbf{G}_{l-1}$. We will need to distinguish slightly between two cases, $l = 2$ and $l \in [3, L+1]$. The proof of the induction step in either case is identical. To this end, and for notational ease, let $\mathbf{V} = \mathbf{X}$, $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_d)$ when $l = 2$, and $\mathbf{V} = \sqrt{\mathbf{G}_{l-1}}$, $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_n)$ for $l \in [3, L+1]$. In either case we let $\mathbf{v}_i$ denote the ith row of $\mathbf{V}$. For any $i, j \in [n]$

$$[\mathbf{G}_l]_{ij} = \sigma_w^2 \mathbb{E}_{\mathbf{w}}[\phi(\mathbf{v}_i^T \mathbf{w})\phi(\mathbf{v}_j^T \mathbf{w})] + \sigma_b^2.$$

Now let $B_1 = \mathbf{v}_i^T \mathbf{w}$, $B_2 = \mathbf{v}_j^T \mathbf{w}$ and observe for any $\alpha_1, \alpha_2 \in \mathbb{R}$ that $\alpha_1 B_1 + \alpha_2 B_2 = \sum_k^n (\alpha_1 v_{ik} + \alpha_2 v_{jk}) w_k \sim \mathcal{N}(0, \|\alpha_1 \mathbf{v}_i + \alpha_2 \mathbf{v}_j\|^2)$. Therefore the joint distribution of (B_1, B_2) is a mean 0 bivariate normal distribution. Denoting the covariance matrix of this distribution as $\tilde{\mathbf{A}} \in \mathbb{R}^{2\times 2}$, then $[\mathbf{G}_l]_{ij}$ can be expressed as

$$[\mathbf{G}_l]_{ij} = \sigma_w^2 \mathbb{E}_{(B_1, B_2) \sim \tilde{\mathbf{A}}}[\phi(B_1)\phi(B_2)] + \sigma_b^2.$$

To prove $[\mathbf{G}_l]_{i,j} = \Sigma^{(l)}$ it therefore suffices to show that $\tilde{\mathbf{A}} = \mathbf{A}^{(l)}$ as per (4.10). This follows by the induction hypothesis as

$$\begin{aligned}
\mathbb{E}[B_1^2] &= \mathbf{v}_i^T \mathbf{v}_i = [\mathbf{G}_{l-1}]_{ii} = \Sigma^{(l-1)}(\mathbf{x}_i, \mathbf{x}_i), \\
\mathbb{E}[B_2^2] &= \mathbf{v}_j^T \mathbf{v}_j = [\mathbf{G}_{l-1}]_{jj} = \Sigma^{(l-1)}(\mathbf{x}_j, \mathbf{x}_j), \\
\mathbb{E}[B_1 B_2] &= \mathbf{v}_i^T \mathbf{v}_j = [\mathbf{G}_{l-1}]_{ij} = \Sigma^{(l-1)}(\mathbf{x}_i, \mathbf{x}_j).
\end{aligned}$$

Finally, $\mathbf{G}_l$ is positive semi-definite as long as $\mathbb{E}_{\mathbf{w}}[\phi(\mathbf{V}\mathbf{w})\phi(\mathbf{V}\mathbf{w})^T]$ is positive semi-definite. Let $M(\mathbf{w}) = \phi(\mathbf{V}\mathbf{w}) \in \mathbb{R}^{n\times n}$ and observe for any $\mathbf{w}$ that $M(\mathbf{w})M(\mathbf{w})^T$ is positive semi-definite. Therefore $\mathbb{E}_{\mathbf{w}}[M(\mathbf{w})M(\mathbf{w})^T]$ must also be positive semi-definite. Thus the inductive step is complete and we may conclude for $l \in [L+1]$ that

$$[\mathbf{G}_l]_{i,j} = \Sigma^{(l)}(\mathbf{x}_i, \mathbf{x}_j). \tag{4.14}$$

For the proof of the expression for the sequence $(\dot{\mathbf{G}}_l)_{l=2}^{L+1}$ it suffices to prove for any $i, j \in [n]$

and $l \in [L+1]$ that

$$[\dot{\mathbf{G}}_l]_{i,j} = \dot{\Sigma}^{(l)}(\mathbf{x}_i, \mathbf{x}_j).$$

By comparing (4.12) with (4.10) this follows immediately from (4.14). Therefore with (4.12) proven (4.13) follows from (4.11). □

4.A.3 Unit Variance Initialization

The initialization scheme for a neural network, particularly a deep neural network, needs to be designed with some care in order to avoid either vanishing or exploding gradients during training [GB10, HZR15, MM16, LBO12]. Some of the most popular initialization strategies used in practice today, in particular [LBO12] and [GB10] initialization, first model the preactivations of the network as Gaussian random variables and then select the network hyperparameters in order that the variance of these idealized preactivations is fixed at one. Under Assumption 89 this idealized model on the preactivations is actually realized and if we additionally assume the conditions of Assumption 90 hold then likewise the variance of the preactivations at every layer will be fixed at one. To this end, and as in [PLR16c, MAT22], consider the function $V \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ defined as

$$V(q) = \sigma_w^2 \mathbb{E}_{Z \sim \mathcal{N}(0,1)} \left[\phi(\sqrt{q} Z)^2 \right] + \sigma_b^2. \tag{4.15}$$

Noting that V is another expression for $\Sigma^{(l)}(\mathbf{x}, \mathbf{x})$, derived via a change of variables as per [PLR16c], the sequence of variances $(\Sigma^{(l)}(\mathbf{x}, \mathbf{x}))_{l=2}^{L}$ can therefore be generated as follows,

$$\Sigma^{(l)}(\mathbf{x}, \mathbf{x}) = V(\Sigma^{(l-1)}(\mathbf{x}, \mathbf{x})). \tag{4.16}$$

The linear correlation $\rho^{(l)} : \mathbb{R}^d \times \mathbb{R}^d \to [-1, 1]$ between the preactivations of two inputs $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we define as

$$\rho^{(l)}(\mathbf{x}, \mathbf{y}) = \frac{\Sigma^{(l)}(\mathbf{x}, \mathbf{y})}{\sqrt{\Sigma^{(l)}(\mathbf{x}, \mathbf{x}) \Sigma^{(l)}(\mathbf{y}, \mathbf{y})}}. \tag{4.17}$$

Assuming $\Sigma^{(l)}(\mathbf{x},\mathbf{x}) = \Sigma^{(l)}(\mathbf{y},\mathbf{y}) = 1$ for all $l \in [L+1]$, then $\rho^{(l)}(\mathbf{x},\mathbf{y}) = \Sigma^{(l)}(\mathbf{x},\mathbf{y})$. Again as in [MAT22] and analogous to (4.15), with $Z_1, Z_2 \sim \mathcal{N}(0,1)$ independent, $U_1 := Z_1$, $U_2(\rho) := (\rho Z_1 + \sqrt{1-\rho^2} Z_2)$ [1] we define the correlation function $R : [-1,1] \to [-1,1]$ as

$$R(\rho) = \sigma_w^2 \mathbb{E}[\phi(U_1)\phi(U_2(\rho))] + \sigma_b^2. \tag{4.18}$$

Noting under these assumptions that R is equivalent to $\Sigma^{(l)}(\mathbf{x},\mathbf{y})$, the sequence of correlations $(\rho^{(l)}(\mathbf{x},\mathbf{y}))_{l=2}^{L}$ can thus be generated as

$$\rho^{(l)}(\mathbf{x},\mathbf{y}) = R(\rho^{(l-1)}(\mathbf{x},\mathbf{y})).$$

As observed in [PLR16c, SGG17], $R(1) = V(1) = 1$, hence $\rho = 1$ is a fixed point of R. We remark that as all preactivations are distributed as $\mathcal{N}(0,1)$, then a correlation of one between preactivations implies they are equal. The stability of the fixed point $\rho = 1$ is of particular significance in the context of initializing deep neural networks successfully. Under mild conditions on the activation function one can compute the derivative of R, see e.g., [PLR16c, SGG17, MAT22], as follows,

$$R'(\rho) = \sigma_w^2 \mathbb{E}[\phi'(U_1)\phi'(U_2(\rho))]. \tag{4.19}$$

Observe that the expression for $\dot{\Sigma}^{(l)}$ and R' are equivalent via a change of variables [PLR16c], and therefore the sequence of correlation derivatives may be computed as

$$\dot{\Sigma}^{(l)}(\mathbf{x},\mathbf{y}) = R'(\rho^{(l)}(\mathbf{x},\mathbf{y})).$$

With the relevant background material now in place we are in a position to prove Lemma 97.

Lemma 97. *Under Assumptions 89 and 90 and defining* $\chi = \sigma_w^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\phi'(Z)^2] \in \mathbb{R}_{>0}$,

[1] We remark that U_1, U_2 are dependent and identically distributed as $U_1, U_2 \sim \mathcal{N}(0,1)$.

then for all $i, j \in [n]$, $l \in [L+1]$

- $[\mathbf{G}_{n,l}]_{ij} \in [-1, 1]$ *and* $[\mathbf{G}_{n,l}]_{ii} = 1$,
- $[\dot{\mathbf{G}}_{n,l}]_{ij} \in [-\chi, \chi]$ *and* $[\dot{\mathbf{G}}_{n,l}]_{ii} = \chi$.

Furthermore, the NTK is a dot product kernel, meaning $\Theta(\mathbf{x}_i, \mathbf{x}_j)$ *can be written as a function of the inner product between the two inputs,* $\Theta(\mathbf{x}_i^T \mathbf{x}_j)$.

Proof. Recall from Lemma 96 and its proof that for any $l \in [L+1]$, $i, j \in [n]$ $[\mathbf{G}_{n,l}]_{ij} = \Sigma^{(l)}(\mathbf{x}_i, \mathbf{x}_j)$ and $[\dot{\mathbf{G}}_{n,l}]_{ij} = \dot{\Sigma}^{(l)}(\mathbf{x}_i, \mathbf{x}_j)$. We first prove by induction $\Sigma^{(l)}(\mathbf{x}_i, \mathbf{x}_i) = 1$ for all $l \in [L+1]$. The base case $l = 1$ follows as

$$\Sigma^{(1)}(\mathbf{x}, \mathbf{x}) = \gamma_w^2 \mathbf{x}^T \mathbf{x} + \gamma_b^2 = \gamma_w^2 + \gamma_b^2 = 1.$$

Assume the induction hypothesis is true for layer $l-1$. With $Z \sim \mathcal{N}(0, 1)$, then from (4.15) and (4.16)

$$\begin{aligned}
\Sigma^{(l)}(\mathbf{x}, \mathbf{x}) &= V(\Sigma^{(l-1)}(\mathbf{x}, \mathbf{x})) \\
&= \sigma_w^2 \mathbb{E}\left[\phi^2\left(\sqrt{\Sigma^{(l-1)}(\mathbf{x}, \mathbf{x})} Z\right)\right] + \sigma_b^2 \\
&= \sigma_w^2 \mathbb{E}\big[\phi^2(Z)\big] + \sigma_b^2 \\
&= 1,
\end{aligned}$$

thus the inductive step is complete. As an immediate consequence it follows that $[\mathbf{G}_l]_{ii} = 1$. Also, for any $i, j \in [n]$ and $l \in [L+1]$,

$$\Sigma^{(l)}(\mathbf{x}_i, \mathbf{x}_j) = \rho^{(l)}(\mathbf{x}_i, \mathbf{x}_j) = R(\rho^{(l-1)}(\mathbf{x}_i, \mathbf{x}_j)) = R(...R(R(\mathbf{x}_i^T \mathbf{x}_j))).$$

Thus we can consider $\Sigma^{(l)}$ as a univariate function of the input correlation $\Sigma : [-1, 1] \to [-1, 1]$ and also conclude that $[\mathbf{G}_l]_{ij} \in [-1, 1]$. Furthermore,

$$\dot{\Sigma}^{(l)}(\mathbf{x}_i, \mathbf{x}_j) = R'(\rho^{(l)}(\mathbf{x}_i, \mathbf{x}_j)) = R'(R(...R(R(\mathbf{x}_i^T \mathbf{x}_j)))),$$

which likewise implies $\dot{\Sigma}$ is a dot product kernel. Recall now the random variables introduced to define R: $Z_1, Z_2 \sim \mathcal{N}(0,1)$ are independent and $U_1 = Z_1$, $U_2 = (\rho Z_1 + \sqrt{1-\rho^2} Z_2)$. Observe U_1, U_2 are dependent but identically distributed as $U_1, U_2 \sim \mathcal{N}(0,1)$. For any $\rho \in [-1,1]$ then applying the Cauchy-Schwarz inequality gives

$$|R'(\rho)|^2 = \sigma_w^4 |\mathbb{E}[\phi'(U_1)\phi'(U_2)]|^2 \leq \sigma_w^4 \mathbb{E}[\phi'(U_1)^2]\mathbb{E}[\phi'(U_2)^2] = \sigma_w^4 \mathbb{E}[\phi'(U_1)^2]^2 = |R'(1)|^2.$$

As a result, under the assumptions of the lemma $\dot{\Sigma}^{(l)} : [-1,1] \to [-\chi, \chi]$ and $\dot{\Sigma}^{(l)}(\mathbf{x}_i, \mathbf{x}_i) = \chi$. From this it immediately follows that $[\dot{\mathbf{G}}_l]_{ij} \in [-\chi, \chi]$ and $[\dot{\mathbf{G}}_l]_{ii} = \chi$ as claimed. Finally, as $\Sigma : [-1,1] \to [-1,1]$ and $\dot{\Sigma} : [-1,1] \to [-\chi, \chi]$ are dot product kernels, then from (4.11) the NTK must also be a dot product kernel and furthermore a univariate function of the pairwise correlation of its input arguments. □

The following corollary, which follows immediately from Lemma 97 and (4.13), characterizes the trace of the NTK matrix in terms of the trace of the input gram.

Corollary 98. *Under the same conditions as Lemma 97, suppose ϕ and σ_w^2 are chosen such that $\chi = 1$. Then*

$$Tr(\mathbf{K}_{n,l}) = l. \tag{4.20}$$

4.A.4 Hermite Expansions

We say that a function $f : \mathbb{R} \to \mathbb{R}$ is square integrable w.r.t. the standard Gaussian measure $\gamma = e^{-x^2/2}/\sqrt{2\pi}$ if $\mathbb{E}_{x \sim \mathcal{N}(0,1)}[f(x)^2] < \infty$. We denote by $L^2(\mathbb{R}, \gamma)$ the space of all such functions. The probabilist's Hermite polynomials are given by

$$H_k(x) = (-1)^k e^{x^2/2} \frac{d^k}{dx^k} e^{-x^2/2}, \quad k = 0, 1, \ldots.$$

The first three Hermite polynomials are $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = (x^2 - 1)$. Let $h_k(x) = \frac{H_k(x)}{\sqrt{k!}}$ denote the normalized probabilist's Hermite polynomials. The normalized Hermite polynomials form a complete orthonormal basis in $L^2(\mathbb{R}, \gamma)$ [OD14, §11]: in all that

follows, whenever we reference the Hermite polynomials, we will be referring to the normalized Hermite polynomials. The Hermite expansion of a function $\phi \in L^2(\mathbb{R}, \gamma)$ is given by

$$\phi(x) = \sum_{k=0}^{\infty} \mu_k(\phi) h_k(x), \tag{4.21}$$

where

$$\mu_k(\phi) = \mathbb{E}_{X \sim \mathcal{N}(0,1)}[\phi(X) h_k(X)] \tag{4.22}$$

is the kth normalized probabilist's Hermite coefficient of ϕ. In what follows we shall make use of the following identities.

$$\forall k \geq 1,\ h_k'(x) = \sqrt{k} h_{k-1}(x), \tag{4.23}$$

$$\forall k \geq 1,\ x h_k(x) = \sqrt{k+1} h_{k+1}(x) + \sqrt{k} h_{k-1}(x). \tag{4.24}$$

$$\begin{aligned} h_k(0) &= \begin{cases} 0, & \text{if } k \text{ is odd} \\ \frac{1}{\sqrt{k!}} (-1)^{\frac{k}{2}} (k-1)!! & \text{if } k \text{ is even} \end{cases}, \\ \text{where } k!! &= \begin{cases} 1, & k \leq 0 \\ k \cdot (k-2) \cdots 5 \cdot 3 \cdot 1, & k > 0 \text{ odd} \\ k \cdot (k-2) \cdots 6 \cdot 4 \cdot 2, & k > 0 \text{ even} \end{cases}. \end{aligned} \tag{4.25}$$

We also remark that the more commonly encountered physicist's Hermite polynomials, which we denote $\tilde{H}_k$, are related to the normalized probablist's polynomials as follows,

$$h_k(z) = \frac{2^{-k/2} \tilde{H}_k(z/\sqrt{2})}{\sqrt{k!}}.$$

The Hermite expansion of the activation function deployed will play a key role in determining the coefficients of the NTK power series. In particular, the Hermite coefficients of ReLU are as follows.

Lemma 99. *[DFS16b] For $\phi(z) = \max\{0, z\}$ the Hermite coefficients are given by*

$$\mu_k(\phi) = \begin{cases} 1/\sqrt{2\pi}, & k = 0, \\ 1/2, & k = 1, \\ (k-3)!!/\sqrt{2\pi k!}, & k \text{ even and } k \geq 2, \\ 0, & k \text{ odd and } k > 3. \end{cases} \tag{4.26}$$

4.B Expressing the NTK as a Power Series

4.B.1 Deriving a Power Series for the NTK

We will require the following minor adaptation of [NM20, Lemma D.2]. We remark this result was first stated for ReLU and Softplus activations in the work of [OS20, Lemma H.2].

Lemma 100. *For arbitrary $n, d \in \mathbb{N}$, let $\mathbf{A} \in \mathbb{R}^{n \times d}$. For $i \in [n]$, we denote the ith row of $\mathbf{A}$ as $\mathbf{a}_i$, and further assume that $\|\mathbf{a}_i\| = 1$. Let $\phi : \mathbb{R} \to \mathbb{R}$ satisfy $\phi \in L^2(\mathbb{R}, \gamma)$ and define*

$$\mathbf{M} = \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, \boldsymbol{I}_n)}[\phi(\mathbf{A}\mathbf{w})\phi(\mathbf{A}\mathbf{w})^T] \in \mathbb{R}^{n \times n}.$$

Then the matrix series

$$\mathbf{S}_K = \sum_{k=0}^{K} \mu_k^2(\phi) \left(\mathbf{A}\mathbf{A}^T\right)^{\odot k}$$

converges uniformly to $\mathbf{M}$ as $K \to \infty$.

The proof of Lemma 100 follows exactly as in [NM20, Lemma D.2], and is in fact slightly simpler due to the fact we assume the rows of $\mathbf{A}$ are unit length and $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_d)$ instead of $\sqrt{d}$ and $\mathbf{w} \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ respectively. For the ease of the reader, we now recall the following definitions, which are also stated in Section 4.3. Letting $\bar{\alpha}_l := (\alpha_{p,l})_{p=0}^{\infty}$ denote a sequence of

real coefficients, then

$$F(p,k,\bar{\alpha}_l) := \begin{cases} 1 & k=0 \text{ and } p=0, \\ 0 & k=0 \text{ and } p \geq 1, \\ \sum_{(j_i)\in\mathcal{J}(p,k)} \prod_{i=1}^{k} \alpha_{j_i,l} & k \geq 1 \text{ and } p \geq 0, \end{cases} \tag{4.27}$$

where

$$\mathcal{J}(p,k) := \{(j_i)_{i\in[k]} \ : \ j_i \geq 0 \ \forall i \in [k], \ \sum_{i=1}^{k} j_i = p\}$$

for all $p \in \mathbb{Z}_{\geq 0}$, $k \in \mathbb{Z}_{\geq 1}$.

We are now ready to derive power series for elements of $(\mathbf{G}_l))_{l=1}^{L+1}$ and $(\dot{\mathbf{G}}_l))_{l=2}^{L+1}$.

Lemma 101. *Under Assumptions 89 and 90, for all $l \in [2, L+1]$*

$$\mathbf{G}_l = \sum_{k=0}^{\infty} \alpha_{k,l} (\mathbf{X}\mathbf{X}^T)^{\odot k}, \tag{4.28}$$

where the series for each element $[\mathbf{G}_l]_{ij}$ converges absolutely and the coefficients $\alpha_{p,l}$ are nonnegative. The coefficients of the series (4.28) for all $p \in \mathbb{Z}_{\geq 0}$ can be expressed via the following recurrence relationship,

$$\alpha_{p,l} = \begin{cases} \sigma_w^2 \mu_p^2(\phi) + \delta_{p=0}\sigma_b^2, & l = 2, \\ \sum_{k=0}^{\infty} \alpha_{k,2} F(p,k,\bar{\alpha}_{l-1}), & l \geq 3. \end{cases} \tag{4.29}$$

Furthermore,

$$\dot{\mathbf{G}}_l = \sum_{k=0}^{\infty} \upsilon_{k,l} (\mathbf{X}\mathbf{X}^T)^{\odot k}, \tag{4.30}$$

where likewise the series for each entry $[\dot{\mathbf{G}}_l]_{ij}$ converges absolutely and the coefficients $\upsilon_{p,l}$ for

all $p \in \mathbb{Z}_{\geq 0}$ are nonnegative and can be expressed via the following recurrence relationship,

$$\upsilon_{p,l} = \begin{cases} \sigma_w^2 \mu_p^2(\phi'), & l = 2, \\ \sum_{k=0}^{\infty} \upsilon_{k,2} F(p, k, \bar{\alpha}_{l-1}), & l \geq 3. \end{cases} \tag{4.31}$$

Proof. We start by proving (4.28) and (4.29). Proceeding by induction, consider the base case $l = 2$. From Lemma 96

$$\mathbf{G}_2 = \sigma_w^2 \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)}[\phi(\mathbf{X}\mathbf{w})\phi(\mathbf{X}\mathbf{w})^T] + \sigma_b^2 \mathbf{1}_{n \times n}.$$

By the assumptions of the lemma, the conditions of Lemma 100 are satisfied and therefore

$$\begin{aligned} \mathbf{G}_2 &= \sigma_w^2 \sum_{k=0}^{\infty} \mu_k^2(\phi) \big(\mathbf{X}\mathbf{X}^T\big)^{\odot k} + \sigma_b^2 \mathbf{1}_{n \times n} \\ &= \alpha_{0,2} \mathbf{1}_{n \times n} + \sum_{k=1}^{\infty} \alpha_{k,2} \big(\mathbf{X}\mathbf{X}^T\big)^{\odot k}. \end{aligned}$$

Observe the coefficients $(\alpha_{k,2})_{k \in \mathbb{Z}_{\geq 0}}$ are nonnegative. Therefore, for any $i, j \in [n]$ using Lemma 97 the series for $[\mathbf{G}_l]_{ij}$ satisfies

$$\sum_{k=0}^{\infty} |\alpha_{k,2}| \big|\langle \mathbf{x}_i, \mathbf{x}_j \rangle^k\big| \leq \sum_{k=0}^{\infty} \alpha_{k,2} \langle \mathbf{x}_i, \mathbf{x}_i \rangle^k \quad = [\mathbf{G}_l]_{ii} = 1 \tag{4.32}$$

and so must be absolutely convergent. With the base case proved we proceed to assume the inductive hypothesis holds for arbitrary $\mathbf{G}_l$ with $l \in [2, L]$. Observe

$$\mathbf{G}_{l+1} = \sigma_w^2 \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)}[\phi(\mathbf{A}\mathbf{w})\phi(\mathbf{A}\mathbf{w})^T] + \sigma_b^2 \mathbf{1}_{n \times n},$$

where $\mathbf{A}$ is a matrix square root of $\mathbf{G}_l$, meaning $\mathbf{G}_l = \mathbf{A}\mathbf{A}$. Recall from Lemma 96 that $\mathbf{G}_l$ is also symmetric and positive semi-definite, therefore we may additionally assume, without loss of generality, that $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric, which conveniently implies $\mathbf{G}_{n,l} = \mathbf{A}\mathbf{A}^T$. Under the assumptions of the lemma the conditions for Lemma 97 are satisfied and as a result

$[\mathbf{G}_{n,l}]_{ii} = \|\mathbf{a}_i\| = 1$ for all $i \in [n]$, where we recall $\mathbf{a}_i$ denotes the ith row of $\mathbf{A}$. Therefore we may again apply Lemma 96,

$$\begin{aligned}
\mathbf{G}_{l+1} &= \sigma_w^2 \sum_{k=0}^{\infty} \mu_k^2(\phi) \left(\mathbf{A}\mathbf{A}^T\right)^{\odot k} + \sigma_b^2 \mathbf{1}_{n\times n} \\
&= (\sigma_w^2 \mu_0^2(\phi) + \sigma_b^2)\mathbf{1}_{n\times n} + \sigma_w^2 \sum_{k=1}^{\infty} \mu_k^2(\phi) (\mathbf{G}_{n,l})^{\odot k} \\
&= (\sigma_w^2 \mu_0^2(\phi) + \sigma_b^2)\mathbf{1}_{n\times n} + \sigma_w^2 \sum_{k=1}^{\infty} \mu_k^2(\phi) \left(\sum_{m=0}^{\infty} \alpha_{m,l} (\mathbf{X}\mathbf{X}^T)^{\odot m}\right)^{\odot k},
\end{aligned}$$

where the final equality follows from the inductive hypothesis. For any pair of indices $i, j \in [n]$

$$[\mathbf{G}_{l+1}]_{ij} = (\sigma_w^2 \mu_0^2(\phi) + \sigma_b^2) + \sigma_w^2 \sum_{k=1}^{\infty} \mu_k^2(\phi) \left(\sum_{m=0}^{\infty} \alpha_{m,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^m\right)^k.$$

By the induction hypothesis, for any $i, j \in [n]$ the series $\sum_{m=0}^{\infty} \alpha_{m,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^m$ is absolutely convergent. Therefore, from the Cauchy product of power series and for any $k \in \mathbb{Z}_{\geq 0}$ we have

$$\left(\sum_{m=0}^{\infty} \alpha_{m,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^m\right)^k = \sum_{p=0}^{\infty} F(p, k, \bar{\alpha}_l) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p, \tag{4.33}$$

where $F(p, k, \bar{\alpha}_l)$ is defined in (4.4). By definition, $F(p, k, \bar{\alpha}_l)$ is a sum of products of positive coefficients, and therefore $|F(p, k, \bar{\alpha}_l)| = F(p, k, \bar{\alpha}_l)$. In addition, recall again by Assumption 90 and Lemma 97 that $[\mathbf{G}_l]_{ii} = 1$. As a result, for any $k \in \mathbb{Z}_{\geq 0}$, as $|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| \leq 1$

$$\sum_{p=0}^{\infty} |F(p, k, \bar{\alpha}_l) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p| \leq \left(\sum_{m=0}^{\infty} \alpha_{m,l}\right)^k = [\mathbf{G}_{n,l}]_{ii} = 1 \tag{4.34}$$

and therefore the series $\sum_{p=0}^{\infty} F(p, k, \bar{\alpha}_l) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p$ converges absolutely. Recalling from the proof of the base case that the series $\sum_{p=1}^{\infty} \alpha_{p,2}$ is absolutely convergent and has only

nonnegative elements, we may therefore interchange the order of summation in the following,

$$\begin{aligned}
[\mathbf{G}_{l+1}]_{ij} &= (\sigma_w^2\mu_0^2(\phi) + \sigma_b^2) + \sigma_w^2 \sum_{k=1}^{\infty} \mu_k^2(\phi) \left(\sum_{p=0}^{\infty} F(p,k,\bar{\alpha}_l) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \right) \\
&= \alpha_{0,2} + \sum_{k=1}^{\infty} \alpha_{k,2} \left(\sum_{p=0}^{\infty} F(p,k,\bar{\alpha}_l) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \right) \\
&= \alpha_{0,2} + \sum_{p=0}^{\infty} \left(\sum_{k=1}^{\infty} \alpha_{k,2} F(p,k,\bar{\alpha}_l) \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p.
\end{aligned}$$

Recalling the definition of $F(p,k,l)$ in (4.4), in particular $F(0,0,\bar{\alpha}_l) = 1$ and $F(p,0,\bar{\alpha}_l) = 0$ for $p \in \mathbb{Z}_{\geq 1}$, then

$$\begin{aligned}
[\mathbf{G}_{l+1}]_{ij} &= \left(\alpha_{0,2} + \sum_{k=1}^{\infty} \alpha_{k,2} F(0,k,\bar{\alpha}_l) \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^0 + \sum_{p=1}^{\infty} \left(\sum_{k=1}^{\infty} \alpha_{k,2} F(p,k,\bar{\alpha}_l) \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \\
&= \left(\sum_{k=0}^{\infty} \alpha_{k,2} F(0,k,\bar{\alpha}_l) \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^0 + \sum_{p=1}^{\infty} \left(\sum_{k=0}^{\infty} \alpha_{k,2} F(p,k,\bar{\alpha}_l) \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \\
&= \sum_{p=0}^{\infty} \left(\sum_{k=0}^{\infty} \alpha_{k,2} F(p,k,\bar{\alpha}_l) \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \\
&= \sum_{p=0}^{\infty} \alpha_{p,l+1} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p.
\end{aligned}$$

As the indices $i, j \in [n]$ were arbitrary we conclude that

$$\mathbf{G}_{l+1} = \sum_{p=0}^{\infty} \alpha_{p,l+1} \big(\mathbf{X}\mathbf{X}^T \big)^{\odot p}$$

as claimed. In addition, by inspection and using the induction hypothesis it is clear that the coefficients $(\alpha_{p,l+1})_{p=0}^{\infty}$ are nonnegative. Therefore, by an argument identical to (4.32), the series for each entry of $[\mathbf{G}_{l+1}]_{ij}$ is absolutely convergent. This concludes the proof of (4.28) and (4.29).

We now turn our attention to proving the (4.30) and (4.31). Under the assumptions of the lemma the conditions for Lemmas 96 and 100 are satisfied and therefore for the base case

$l = 2$

$$\begin{aligned}\dot{\mathbf{G}}_2 &= \sigma_w^2 \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_n)}[\phi'(\mathbf{X}\mathbf{w})\phi'(\mathbf{X}\mathbf{w})^T] \\ &= \sigma_w^2 \sum_{k=0}^{\infty} \mu_k^2(\phi')\big(\mathbf{X}\mathbf{X}^T\big)^{\odot k} \\ &= \sum_{k=0}^{\infty} \upsilon_{k,2}\big(\mathbf{X}\mathbf{X}^T\big)^{\odot k}.\end{aligned}$$

By inspection the coefficients $(\upsilon_{p,2})_{p=0}^{\infty}$ are nonnegative and as a result by an argument again identical to (4.32) the series for each entry of $[\dot{\mathbf{G}}_2]_{ij}$ is absolutely convergent. For $l \in [2, L]$, from (4.28) and its proof there is a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ such that $\mathbf{G}_l = \mathbf{A}\mathbf{A}^T$. Again applying Lemma 100

$$\begin{aligned}\dot{\mathbf{G}}_{n,l+1} &= \sigma_w^2 \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_n)}[\phi'(\mathbf{A}\mathbf{w})\phi'(\mathbf{A}\mathbf{w})^T] \\ &= \sigma_w^2 \sum_{k=0}^{\infty} \mu_k^2(\phi')\big(\mathbf{A}\mathbf{A}^T\big)^{\odot k} \\ &= \sum_{k=0}^{\infty} \upsilon_{k,2}(\mathbf{G}_{n,l})^{\odot k} \\ &= \sum_{k=0}^{\infty} \upsilon_{k,2}\left(\sum_{p=0}^{\infty} \alpha_{p,l}\big(\mathbf{X}\mathbf{X}^T\big)^{\odot p}\right)^{\odot k}\end{aligned}$$

Analyzing now an arbitrary entry $[\dot{\mathbf{G}}_{l+1}]_{ij}$, by substituting in the power series expression for $\mathbf{G}_l$ from (4.28) and using (4.33) we have

$$\begin{aligned}[\dot{\mathbf{G}}_{l+1}]_{ij} &= \sum_{k=0}^{\infty} \upsilon_{k,2}\left(\sum_{p=0}^{\infty} \alpha_{p,l}\langle \mathbf{x}_i, \mathbf{x}_j\rangle^p\right)^k \\ &= \sum_{k=0}^{\infty} \upsilon_{k,2}\left(\sum_{p=0}^{\infty} F(p,k,\bar{\alpha}_l)\langle \mathbf{x}_i, \mathbf{x}_j\rangle^p\right) \\ &= \sum_{p=0}^{\infty}\left(\sum_{k=0}^{\infty} \upsilon_{k,2} F(p,k,\bar{\alpha}_l)\right)\langle \mathbf{x}_i, \mathbf{x}_j\rangle^p \\ &= \sum_{p=0}^{\infty} \upsilon_{p,l+1}\langle \mathbf{x}_i, \mathbf{x}_j\rangle^p.\end{aligned}$$

Note that exchanging the order of summation in the third equality above is justified as for

any $k \in \mathbb{Z}_{\geq 0}$ by (4.34) we have $\sum_{p=0}^{\infty} F(p,k,\bar{\alpha}_l)|\langle \mathbf{x}_i, \mathbf{x}_j\rangle|^p \leq 1$ and therefore

$$\sum_{k=0}^{\infty}\sum_{p=0}^{\infty} \upsilon_{k,2} F(p,k,\bar{\alpha}_l)\langle \mathbf{x}_i, \mathbf{x}_j\rangle^p$$

converges absolutely. As the indices $i,j \in [n]$ were arbitrary we conclude that

$$\dot{\mathbf{G}}_{l+1} = \sum_{p=0}^{\infty} \upsilon_{p,l+1}\big(\mathbf{X}\mathbf{X}^T\big)^{\odot p}$$

as claimed. Finally, by inspection the coefficients $(\upsilon_{p,l+1})_{p=0}^{\infty}$ are nonnegative, therefore, and again by an argument identical to (4.32), the series for each entry of $[\dot{\mathbf{G}}_{n,l+1}]_{ij}$ is absolutely convergent. This concludes the proof. □

We are now prove the key result of Section 4.3.

Theorem 91. *Under Assumptions 89 and 90, for all $l \in [L+1]$*

$$n\mathbf{K}_l = \sum_{p=0}^{\infty} \kappa_{p,l}\big(\mathbf{X}\mathbf{X}^T\big)^{\odot p}. \tag{4.5}$$

The series for each entry $n[\mathbf{K}_l]_{ij}$ converges absolutely and the coefficients $\kappa_{p,l}$ are nonnegative and can be evaluated using the recurrence relationships

$$\kappa_{p,l} = \begin{cases} \delta_{p=0}\gamma_b^2 + \delta_{p=1}\gamma_w^2, & l = 1, \\ \alpha_{p,l} + \sum_{q=0}^{p} \kappa_{q,l-1}\upsilon_{p-q,l}, & l \in [2, L+1], \end{cases} \tag{4.6}$$

where

$$\alpha_{p,l} = \begin{cases} \sigma_w^2\mu_p^2(\phi) + \delta_{p=0}\sigma_b^2, & l = 2, \\ \sum_{k=0}^{\infty} \alpha_{k,2} F(p,k,\bar{\alpha}_{l-1}), & l \geq 3, \end{cases} \tag{4.7}$$

and

$$\upsilon_{p,l} = \begin{cases} \sigma_w^2\mu_p^2(\phi'), & l = 2, \\ \sum_{k=0}^{\infty} \upsilon_{k,2} F(p,k,\bar{\alpha}_{l-1}), & l \geq 3, \end{cases} \tag{4.8}$$

are likewise nonnegative for all $p \in \mathbb{Z}_{\geq 0}$ *and* $l \in [2, L+1]$.

Proof. We proceed by induction. The base case $l = 1$ follows trivially from Lemma 96. We therefore assume the induction hypothesis holds for an arbitrary $l - 1 \in [1, L]$. From (4.13) and Lemma 101

$$\begin{aligned} n\mathbf{K}_l &= \mathbf{G}_l + n\mathbf{K}_{l-1} \odot \dot{\mathbf{G}}_l \\ &= \left(\sum_{p=0}^{\infty} \alpha_{p,l} \big(\mathbf{X}\mathbf{X}^T\big)^{\odot p}\right) + \left(n \sum_{q=0}^{\infty} \kappa_{q,l-1} \big(\mathbf{X}\mathbf{X}^T\big)^{\odot q}\right) \odot \left(\sum_{w=0}^{\infty} \upsilon_{w,l} \big(\mathbf{X}\mathbf{X}^T\big)^{\odot w}\right). \end{aligned}$$

Therefore, for arbitrary $i, j \in [n]$

$$[n\mathbf{K}_l]_{ij} = \sum_{p=0}^{\infty} \alpha_{p,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p + \left(n \sum_{q=0}^{\infty} \kappa_{q,l-1} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^q\right) \left(\sum_{w=0}^{\infty} \upsilon_{w,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^w\right).$$

Observe $n \sum_{q=0}^{\infty} \kappa_{q,l-1} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^q = \Theta^{(l-1)}(\mathbf{x}_i, \mathbf{x}_j)$ and therefore the series must converge due to the convergence of the NTK. Furthermore, $\sum_{w=0}^{\infty} \upsilon_{w,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^w = [\dot{\mathbf{G}}_{n,l}]_{ij}$ and therefore is absolutely convergent by Lemma 101. As a result, by Merten's Theorem the product of these two series is equal to their Cauchy product. Therefore

$$\begin{aligned} [n\mathbf{K}_l]_{ij} &= \sum_{p=0}^{\infty} \alpha_{p,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p + \sum_{p=0}^{\infty} \left(\sum_{q=0}^{p} \kappa_{q,l-1} \upsilon_{p-q,l}\right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \\ &= \sum_{p=0}^{\infty} \left(\alpha_{p,l} + \sum_{q=0}^{p} \kappa_{q,l-1} \upsilon_{p-q,l}\right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \\ &= \sum_{p=0}^{\infty} \kappa_{p,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p, \end{aligned}$$

from which the (4.5) immediately follows. □

4.B.2 Analyzing the Coefficients of the NTK Power Series

In this section we study the coefficients of the NTK power series stated in Theorem 91. Our first observation is that, under additional assumptions on the activation function ϕ,

the recurrence relationship (4.6) can be simplified in order to depend only on the Hermite expansion of ϕ.

Lemma 102. *Under Assumption 92 the Hermite coefficients of ϕ' satisfy*

$$\mu_k(\phi') = \sqrt{k+1}\mu_{k+1}(\phi)$$

for all $k \in \mathbb{Z}_{\geq 0}$.

Proof. Note for each $n \in \mathbb{N}$ as ϕ is absolutely continuous on $[-n, n]$ it is differentiable a.e. on $[-n, n]$. It follows by the countable additivity of the Lebesgue measure that ϕ is differentiable a.e. on $\mathbb{R}$. Furthermore, as ϕ is polynomially bounded we have $\phi \in L^2(\mathbb{R}, e^{-x^2/2}/\sqrt{2\pi})$. Fix $a > 0$. Since ϕ is absolutely continuous on $[-a, a]$ it is of bounded variation on $[-a, a]$. Also note that $h_k(x)e^{-x^2/2}$ is of bounded variation on $[-a, a]$ due to having a bounded derivative. Thus we have by Lebesgue-Stieltjes integration-by-parts (see e.g. Chapter 3 of [Fol99a])

$$\begin{aligned}
&\int_{-a}^{a} \phi'(x)h_k(x)e^{-x^2/2}dx \\
&= \phi(a)h_k(a)e^{-a^2/2} - \phi(-a)h_k(-a)e^{-a^2/2} + \int_{-a}^{a} \phi(x)[xh_k(x) - h_k'(x)]e^{-x^2/2}dx \\
&= \phi(a)h_k(a)e^{-a^2/2} - \phi(-a)h_k(-a)e^{-a^2/2} + \int_{-a}^{a} \phi(x)\sqrt{k+1}h_{k+1}(x)e^{-x^2/2}dx,
\end{aligned}$$

where in the last line above we have used the fact that (4.23) and (4.24) imply that $xh_k(x) - h_k'(x) = \sqrt{k+1}h_{k+1}(x)$. Thus we have shown

$$\begin{aligned}
&\int_{-a}^{a} \phi'(x)h_k(x)e^{-x^2/2}dx \\
&= \phi(a)h_k(a)e^{-a^2/2} - \phi(-a)h_k(-a)e^{-a^2/2} + \int_{-a}^{a} \phi(x)\sqrt{k+1}h_{k+1}(x)e^{-x^2/2}dx.
\end{aligned}$$

We note that since $|\phi(x)h_k(x)| = \mathcal{O}(|x|^{\beta+k})$ we have that as $a \to \infty$ the first two terms above

vanish. Thus by sending $a \to \infty$ we have

$$\int_{-\infty}^{\infty} \phi'(x) h_k(x) e^{-x^2/2} dx = \int_{-\infty}^{\infty} \sqrt{k+1} \phi(x) h_{k+1}(x) e^{-x^2/2} dx.$$

After dividing by $\sqrt{2\pi}$ we get the desired result. □

In particular, under Assumption 92, and as highlighted by Corollary 103, which follows directly from Lemmas 101 and 102, the NTK coefficients can be computed only using the Hermite coefficients of ϕ.

Corollary 103. *Under Assumptions 89, 90 and 92, for all $p \in \mathbb{Z}_{\geq 0}$*

$$\upsilon_{p,l} = \begin{cases} (p+1)\alpha_{p+1,2}, & l = 2, \\ \sum_{k=0}^{\infty} \upsilon_{k,2} F(p, k, \bar{\alpha}_{l-1}), & l \geq 3. \end{cases} \tag{4.35}$$

With these results in place we proceed to analyze the decay of the coefficients of the NTK for depth two networks. As stated in the main text, the decay of the NTK coefficients depends on the decay of the Hermite coefficients of the activation function deployed. This in turn is strongly influenced by the behavior of the tails of the activation function. To this end we roughly group activation functions into three categories: growing tails, flat or constant tails and finally decaying tails. Analyzing each of these groups in full generality is beyond the scope of this chapter, we therefore instead study the behavior of ReLU, Tanh and Gaussian activation functions, being prototypical and practically used examples of each of these three groups respectively. We remark that these three activation functions satisfy Assumption 92. For typographical ease we let $\omega_\sigma(z) := (1/\sqrt{2\pi\sigma^2}) \exp(-z^2/(2\sigma^2))$ denote the Gaussian activation function with variance σ^2.

Lemma 93. *Under Assumptions 89 and 90,*

1. *if $\phi(z) = ReLU(z)$, then $\kappa_{p,2} = \delta_{(\gamma_b > 0) \cup (p \text{ even})} \Theta(p^{-3/2})$,*
2. *if $\phi(z) = Tanh(z)$, then $\kappa_{p,2} = \mathcal{O}\Big(\exp\Big(-\frac{\pi\sqrt{p-1}}{2}\Big)\Big)$,*

3. if $\phi(z) = \omega_\sigma(z)$, then $\kappa_{p,2} = \delta_{(\gamma_b>0)\cup(p\ even)}\Theta(p^{1/2}(\sigma^2+1)^{-p})$.

Proof. Recall (4.9),

$$\kappa_{p,2} = \sigma_w^2(1+\gamma_w^2 p)\mu_p^2(\phi) + \sigma_w^2\gamma_b^2(1+p)\mu_{p+1}^2(\phi) + \delta_{p=0}\sigma_b^2.$$

In order to bound $\kappa_{p,2}$ we proceed by using Lemma 99 to bound the square of the Hermite coefficients. We start with ReLU. Note Lemma 99 actually provides precise expressions for the Hermite coefficients of ReLU, however, these are not immediately easy to interpret. Observe from Lemma 99 that above index $p = 2$ all odd indexed Hermite coefficients are 0. It therefore suffices to bound the even indexed terms, given by

$$\mu_p(ReLU) = \frac{1}{\sqrt{2\pi}}\frac{(p-3)!!}{\sqrt{p!}}.$$

Observe from (4.25) that for p even

$$h_p(0) = (-1)^{p/2}\frac{(p-1)!!}{\sqrt{p!}},$$

therefore

$$\mu_p(ReLU) = \frac{1}{\sqrt{2\pi}}\frac{(p-3)!!}{\sqrt{p!}} = \frac{1}{\sqrt{2\pi}}\frac{|h_p(0)|}{p-1}.$$

Analyzing now $|h_p(0)|$,

$$\frac{(p-1)!!}{\sqrt{p!}} = \frac{\prod_{i=1}^{p/2}(2i-1)}{\sqrt{\prod_{i=1}^{p/2}(2i-1)2i}} = \sqrt{\frac{\prod_{i=1}^{p/2}(2i-1)}{\prod_{i=1}^{p/2}2i}} = \sqrt{\frac{(p-1)!!}{p!!}}.$$

Here, the expression inside the square root is referred to in the literature as the Wallis ratio, for which the following lower and upper bounds are available [Kaz56],

$$\sqrt{\frac{1}{\pi(p+0.5)}} < \frac{(p-1)!!}{p!!} < \sqrt{\frac{1}{\pi(p+0.25)}}. \tag{4.36}$$

As a result

$$|h_p(0)| = \Theta(p^{-1/4})$$

and therefore

$$\mu_p(ReLU) = \begin{cases} \Theta(p^{-5/4}), & p \text{ even}, \\ 0, & p \text{ odd}. \end{cases}$$

As $(p+1)^{-3/2} = \Theta(p^{-3/2})$, then from (4.9)

$$\begin{aligned} \kappa_{p,2} &= \Theta((p\mu_p^2(ReLU) + \delta_{\gamma_b>0}(p+1)\mu_{p+1}^2(ReLU))) \\ &= \Theta((\delta_{p \text{ even}} p^{-3/2} + \delta_{(p \text{ odd})\cap(\gamma_b>0)}(p+1)^{-3/2})) \\ &= \Theta\big(\delta_{(p \text{ even})\cup((p \text{ odd})\cap(\gamma_b>0))} p^{-3/2}\big) \\ &= \delta_{(p \text{ even})\cup(\gamma_b>0)} \Theta\big(p^{-3/2}\big) \end{aligned}$$

as claimed in item *1.*

We now proceed to analyze $\phi(z) = Tanh(z)$. From [PSG20, Corollary F.7.1]

$$\mu_p(Tanh') = \mathcal{O}\bigg(\exp\bigg(-\frac{\pi\sqrt{p}}{4}\bigg)\bigg).$$

As Tanh satisfies the conditions of Lemma 102

$$\mu_p(Tanh) = p^{-1/2}\mu_{p-1}(Tanh') = \mathcal{O}\bigg(p^{-1/2}\exp\bigg(-\frac{\pi\sqrt{p-1}}{4}\bigg)\bigg).$$

Therefore the result claimed in item *2.* follows as

$$\begin{aligned} \kappa_{p,2} &= \mathcal{O}((p\mu_p^2(Tanh) + (p+1)\mu_{p+1}^2(Tanh))) \\ &= \mathcal{O}\bigg(\exp\bigg(-\frac{\pi\sqrt{p-1}}{2}\bigg) + \exp\bigg(-\frac{\pi\sqrt{p}}{2}\bigg)\bigg) \\ &= \mathcal{O}\bigg(\exp\bigg(-\frac{\pi\sqrt{p-1}}{2}\bigg)\bigg). \end{aligned}$$

Finally, we now consider $\phi(z) = \omega_\sigma(z)$ where $\omega_\sigma(z)$ is the density function of $\mathcal{N}(0, \sigma^2)$.

Similar to ReLU, analytic expressions for the Hermite coefficients of $\omega_\sigma(z)$ are known (see e.g., Theorem 2.9 in [Dav21]),

$$\mu_p^2(\omega_\sigma) = \begin{cases} \frac{p!}{((p/2)!)^2 2^p 2\pi(\sigma^2+1)^{p+1}}, & \text{p even,} \\ 0, & \text{p odd.} \end{cases}$$

For p even

$$(p/2)! = p!!2^{-p/2}.$$

Therefore

$$\frac{p!}{(p/2)!(p/2)!} = 2^p \frac{p!}{p!!p!!} = 2^p \frac{(p-1)!!}{p!!}.$$

As a result, for p even and using (4.36), it follows that

$$\mu_p^2(\omega_\sigma) = \frac{(\sigma^2+1)^{-(p+1)}}{2\pi} \frac{(p-1)!!}{p!!} \quad = \Theta(p^{-1/2}(\sigma^2+1)^{-p}).$$

Finally, since $(p+1)^{1/2}(\sigma^2+1)^{-p-1} = \Theta(p^{1/2}(\sigma^2+1)^{-p})$, then from (4.9)

$$\begin{aligned} \kappa_{p,2} &= \Theta((p\mu_p^2(\omega_\sigma) + \delta_{\gamma_b>0}(p+1)\mu_{p+1}^2(\omega_\sigma))) \\ &= \Theta\big(\delta_{(p \text{ even})\cup((p \text{ odd})\cap(\gamma_b>0))} p^{1/2}(\sigma^2+1)^{-p}\big) \\ &= \delta_{(p \text{ even})\cup(\gamma_b>0)} \Theta\big(p^{1/2}(\sigma^2+1)^{-p}\big) \end{aligned}$$

as claimed in item *3.* □

4.C Analyzing the Spectrum of the NTK via its power series

4.C.1 Experimental validation of results on the NTK spectrum

To test our theory in Section 4.4, we numerically plot the spectrum of NTK of two-layer feedforward networks with ReLU, Tanh, and Gaussian activations in Figure 4.1. The input data are uniformly drawn from $\mathbb{S}^2$. Notice that when $d = 2$, $k = \Theta(\ell^{1/2})$. Then

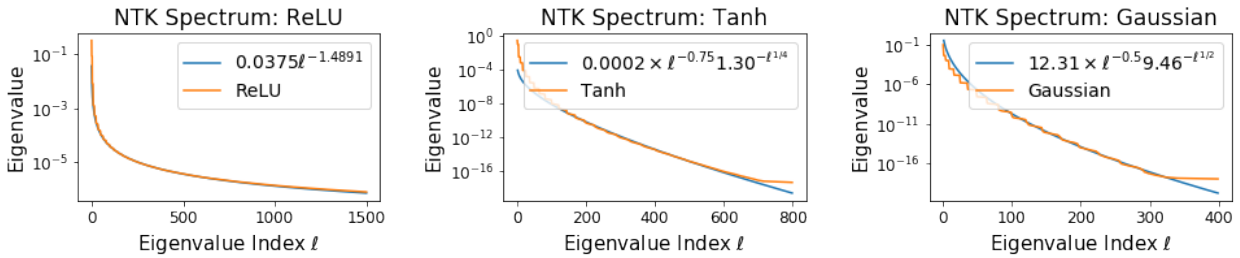

Figure 4.1: NTK spectrum of two-layer fully connected networks with ReLU, Tanh and Gaussian activations under the NTK parameterization. The orange curve is the experimental eigenvalue. The blue curves in the left shows the regression fit for the experimental eigenvalues as a function of eigenvalue index ℓ in the form of $\lambda_\ell = a\ell^{-b}$ where a and b are unknown parameters determined by regression. The blue curves in the middle shows the regression fit for the experimental eigenvalues in the form of $\lambda_\ell = a\ell^{-0.75}b^{-l^{1/4}}$. The blue curves in the right shows the regression fit for the experimental eigenvalues in the form of $\lambda_\ell = a\ell^{-0.5}b^{-l^{1/2}}$.

Corollary 95 shows that for the ReLU activation $\lambda_\ell = \Theta(\ell^{-3/2})$, for the Tanh activation $\lambda_\ell = O\big(\ell^{-3/4}\exp(-\frac{\pi}{2}\ell^{1/4})\big)$, and for the Gaussian activation $\lambda_\ell = O(\ell^{-1/2}2^{-\ell^{1/2}})$. These theoretical decay rates for the NTK spectrum are verified by the experimental results in Figure 4.1.

4.C.2 Analysis of the Asymptotic Spectrum: Uniform Data

Theorem 94. *Suppose that the training data are uniformly sampled from the unit hypersphere $\mathbb{S}^d$, $d \geq 2$. If the dot-product kernel function has the expansion $K(x_1, x_2) = \sum_{p=0}^{\infty} c_p \langle x_1, x_2\rangle^p$ where $c_p \geq 0$, then the eigenvalue of every spherical harmonic of frequency k is given by*

$$\overline{\lambda_k} = \frac{\pi^{d/2}}{2^{k-1}} \sum_{\substack{p \geq k \\ p-k \text{ is even}}} c_p \frac{\Gamma(p+1)\Gamma(\frac{p-k+1}{2})}{\Gamma(p-k+1)\Gamma(\frac{p-k+1}{2}+k+d/2)},$$

where Γ is the gamma function.

Proof. Let $\theta(t) = \sum_{p=0}^{\infty} c_p t^p$, then $K(x_1, x_2) = \theta(\langle x_1, x_2\rangle)$ According to Funk Hecke theorem

[BJK19, Section 4.2], we have

$$\overline{\lambda_k} = \mathrm{Vol}(\mathbb{S}^{d-1}) \int_{-1}^{1} \theta(t) P_{k,d}(t) (1-t^2)^{\frac{d-2}{2}} \mathrm{d}t, \tag{4.37}$$

where $\mathrm{Vol}(\mathbb{S}^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ is the volume of the hypersphere $\mathbb{S}^{d-1}$, and $P_{k,d}(t)$ is the Gegenbauer polynomial, given by

$$P_{k,d}(t) = \frac{(-1)^k}{2^k} \frac{\Gamma(d/2)}{\Gamma(k+d/2)} \frac{1}{(1-t^2)^{(d-2)/2}} \frac{d^k}{dt^k} (1-t^2)^{k+(d-2)/2},$$

and Γ is the gamma function.

From (4.37) we have

$$\begin{aligned}
\overline{\lambda_k} &= \mathrm{Vol}(\mathbb{S}^{d-1}) \int_{-1}^{1} \theta(t) P_{k,d}(t) (1-t^2)^{\frac{d-2}{2}} \mathrm{d}t \\
&= \frac{2\pi^{d/2}}{\Gamma(d/2)} \int_{-1}^{1} \theta(t) \frac{(-1)^k}{2^k} \frac{\Gamma(d/2)}{\Gamma(k+d/2)} \frac{d^k}{dt^k} (1-t^2)^{k+(d-2)/2} \mathrm{d}t \\
&= \frac{2\pi^{d/2}}{\Gamma(d/2)} \frac{(-1)^k}{2^k} \frac{\Gamma(d/2)}{\Gamma(k+d/2)} \sum_{p=0}^{\infty} c_p \int_{-1}^{1} t^p \frac{d^k}{dt^k} (1-t^2)^{k+(d-2)/2} \mathrm{d}t.
\end{aligned} \tag{4.38}$$

Using integration by parts, we have

$$\begin{aligned}
&\int_{-1}^{1} t^p \frac{d^k}{dt^k} (1-t^2)^{k+(d-2)/2} \mathrm{d}t \\
&= t^p \frac{d^{k-1}}{dt^{k-1}} (1-t^2)^{k+(d-2)/2} \Big|_{-1}^{1} - p \int_{-1}^{1} t^{p-1} \frac{d^{k-1}}{dt^{k-1}} (1-t^2)^{k+(d-2)/2} \mathrm{d}t \\
&= -p \int_{-1}^{1} t^{p-1} \frac{d^{k-1}}{dt^{k-1}} (1-t^2)^{k+(d-2)/2} \mathrm{d}t,
\end{aligned} \tag{4.39}$$

where the last line in (4.39) holds because $\frac{d^{k-1}}{dt^{k-1}} (1-t^2)^{k+(d-2)/2} = 0$ when $t = 1$ or $t = -1$.

When $p < k$, repeat the above procedure (4.39) p times, we get

$$\begin{aligned}\int_{-1}^{1} t^p \frac{d^k}{dt^k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t &= (-1)^p p! \int_{-1}^{1} \frac{d^{k-p}}{dt^{k-p}}(1-t^2)^{k+(d-2)/2}\mathrm{d}t \\ &= (-1)^p p! \frac{d^{k-p-1}}{dt^{k-p-1}}(1-t^2)^{k+(d-2)/2}\Big|_{-1}^{1} \\ &= 0. \end{aligned} \tag{4.40}$$

When $p \geq k$, repeat the above procedure (4.39) k times, we get

$$\int_{-1}^{1} t^p \frac{d^k}{dt^k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t = (-1)^k p(p-1)\cdots(p-k+1)\int_{-1}^{1} t^{p-k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t. \tag{4.41}$$

When $p-k$ is odd, $t^{p-k}(1-t^2)^{k+(d-2)/2}$ is an odd function, then

$$\int_{-1}^{1} t^{p-k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t = 0. \tag{4.42}$$

When $p-k$ is even,

$$\begin{aligned}\int_{-1}^{1} t^{p-k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t &= 2\int_{0}^{1} t^{p-k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t \\ &= \int_{0}^{1} (t^2)^{(p-k-1)/2}(1-t^2)^{k+(d-2)/2}\mathrm{d}t^2 \\ &= B\left(\frac{p-k+1}{2}, k+d/2\right) \\ &= \frac{\Gamma(\frac{p-k+1}{2})\Gamma(k+d/2)}{\Gamma(\frac{p-k+1}{2}+k+d/2)}, \end{aligned} \tag{4.43}$$

where B is the beta function.

Plugging (4.43) , (4.40) and (4.42) into (4.41), we get

$$\begin{aligned}
&\int_{-1}^{1} t^p \frac{d^k}{dt^k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t \\
&= \begin{cases} (-1)^k p(p-1)\dots(p-k+1)\frac{\Gamma(\frac{p-k+1}{2})\Gamma(k+d/2)}{\Gamma(\frac{p-k+1}{2}+k+d/2)}, & p-k \text{ is even and } p \geq k, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{4.44}$$

Plugging (4.44) into (4.38), we get

$$\begin{aligned}
\overline{\lambda_k} &= \frac{2\pi^{d/2}}{\Gamma(d/2)}\frac{(-1)^k}{2^k}\frac{\Gamma(d/2)}{\Gamma(k+d/2)} \sum_{\substack{p\geq k \\ p-k \text{ is even}}} c_p (-1)^k p(p-1)\dots(p-k+1)\frac{\Gamma(\frac{p-k+1}{2})\Gamma(k+d/2)}{\Gamma(\frac{p-k+1}{2}+k+d/2)} \\
&= \frac{\pi^{d/2}}{2^{k-1}} \sum_{\substack{p\geq k \\ p-k \text{ is even}}} c_p \frac{p(p-1)\dots(p-k+1)\Gamma(\frac{p-k+1}{2})}{\Gamma(\frac{p-k+1}{2}+k+d/2)} \\
&= \frac{\pi^{d/2}}{2^{k-1}} \sum_{\substack{p\geq k \\ p-k \text{ is even}}} c_p \frac{\Gamma(p+1)\Gamma(\frac{p-k+1}{2})}{\Gamma(p-k+1)\Gamma(\frac{p-k+1}{2}+k+d/2)}.
\end{aligned}$$

□

Corollary 95. *Under the same setting as in Theorem 94,*

1. *if $c_p = \Theta(p^{-a})$ where $a \geq 1$, then $\overline{\lambda_k} = \Theta(k^{-d-2a+2})$,*
2. *if $c_p = \delta_{(p \text{ even})}\Theta(p^{-a})$, then $\overline{\lambda_k} = \delta_{(k \text{ even})}\Theta(k^{-d-2a+2})$,*
3. *if $c_p = \mathcal{O}\big(\exp\big(-a\sqrt{p}\big)\big)$, then $\overline{\lambda_k} = \mathcal{O}\Big(k^{-d+1/2}\exp\Big(-a\sqrt{k}\Big)\Big)$,*
4. *if $c_p = \Theta(p^{1/2}a^{-p})$, then $\overline{\lambda_k} = \mathcal{O}\big(k^{-d+1}a^{-k}\big)$ and $\overline{\lambda_k} = \Omega\big(k^{-d/2+1}2^{-k}a^{-k}\big)$.*

Proof of Corollary 4.C.2, part 1. We first prove $\overline{\lambda_k} = O(k^{-d-2a+2})$. Suppose that $c_p \leq Cp^{-a}$ for some constant C, then according to Theorem 94 we have

$$\overline{\lambda_k} \leq \frac{\pi^{d/2}}{2^{k-1}} \sum_{\substack{p\geq k \\ p-k \text{ is even}}} Cp^{-a} \frac{\Gamma(p+1)\Gamma(\frac{p-k+1}{2})}{\Gamma(p-k+1)\Gamma(\frac{p-k+1}{2}+k+d/2)}.$$

According to Stirling's formula, we have

$$\Gamma(z) = \sqrt{\frac{2\pi}{z}}\Big(\frac{z}{e}\Big)^z\bigg(1 + O\Big(\frac{1}{z}\Big)\bigg). \tag{4.45}$$

Then for any $z \geq \frac{1}{2}$, we can find constants C_1 and C_2 such that

$$C_1\sqrt{\frac{2\pi}{z}}\Big(\frac{z}{e}\Big)^z \leq \Gamma(z) \leq C_2\sqrt{\frac{2\pi}{z}}\Big(\frac{z}{e}\Big)^z. \tag{4.46}$$

Then

$$\begin{aligned}
\overline{\lambda_k} &\leq \frac{\pi^{d/2}}{2^{k-1}}\frac{C_2^2}{C_1^2}\sum_{\substack{p\geq k\\ p-k \text{ is even}}} Cp^{-a}\frac{\sqrt{\frac{2\pi}{p+1}}\big(\frac{p+1}{e}\big)^{p+1}\sqrt{\frac{2\pi}{\frac{p-k+1}{2}}}\Big(\frac{\frac{p-k+1}{2}}{e}\Big)^{\frac{p-k+1}{2}}}{\sqrt{\frac{2\pi}{p-k+1}}\big(\frac{p-k+1}{e}\big)^{p-k+1}\sqrt{\frac{2\pi}{\frac{p-k+1}{2}+k+d/2}}\Big(\frac{\frac{p-k+1}{2}+k+d/2}{e}\Big)^{\frac{p-k+1}{2}+k+d/2}} \\
&= \frac{\pi^{d/2}}{2^{k-1}}\frac{C_2^2C}{C_1^2}\sum_{\substack{p\geq k\\ p-k \text{ is even}}} p^{-a}\frac{e^{\frac{d}{2}}\sqrt{\frac{2}{p+1}}(p+1)^{p+1}\big(\frac{p-k+1}{2}\big)^{\frac{p-k+1}{2}}}{(p-k+1)^{p-k+1}\sqrt{\frac{1}{\frac{p-k+1}{2}+k+d/2}}\big(\frac{p-k+1}{2}+k+d/2\big)^{\frac{p-k+1}{2}+k+d/2}} \\
&= \frac{\pi^{d/2}}{2^{k-1}}\frac{C_2^2C}{C_1^2}\sum_{\substack{p\geq k\\ p-k \text{ is even}}} p^{-a}\frac{e^{\frac{d}{2}}2^{\frac{-p+k}{2}}(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}\big(\frac{p-k+1}{2}+k+d/2\big)^{\frac{p-k}{2}+k+d/2}} \\
&= 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2C}{C_1^2}\sum_{\substack{p\geq k\\ p-k \text{ is even}}} \frac{p^{-a}(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}(p+k+1+d)^{\frac{p+k+d}{2}}}.
\end{aligned} \tag{4.47}$$

We define

$$f_a(p) = \frac{p^{-a}(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}(p+k+1+d)^{\frac{p+k+d}{2}}}. \tag{4.48}$$

By applying the chain rule to $e^{\log f_a(p)}$, we have that the derivative of f_a is

$$\begin{aligned}
f_a'(p) = &\frac{(p+1)^{p+\frac{1}{2}}p^{-a}}{2(p-k+1)^{\frac{p-k+1}{2}}(p+k+d+1)^{\frac{p+k+d}{2}}} \\
&\cdot\bigg(-\frac{2a}{p} - \frac{k+d}{(p+1)(p+k+d+1)} + \log(1+\frac{k^2-d(p-k+1)}{(p-k+1)(p+k+d+1)})\bigg).
\end{aligned} \tag{4.49}$$

Let $g_a(p) = -\frac{2a}{p} - \frac{k+d}{(p+1)(p+k+d+1)} + \log(1+\frac{k^2-d(p-k+1)}{(p-k+1)(p+k+d+1)})$. Then $g_a(p)$ and $f_a'(p)$ have the

same sign. Next we will show that $g_a(p) \geq 0$ for $k \leq p \leq \frac{k^2}{d+24a}$ when k is large enough.

First when $p \geq k$ and $\frac{k^2-d(p-k+1)}{(p-k+1)(p+k+d+1)} \geq 1$, we have

$$g_a(p) \geq -\frac{2a}{k} - \frac{k+d}{(k+1)(k+k+d+1)} + \log(2) \geq 0, \tag{4.50}$$

when k is sufficiently large.

Second when $p \geq k$ and $0 \leq \frac{k^2-d(p-k+1)}{(p-k+1)(p+k+d+1)} \leq 1$, since $\log(1+x) \geq \frac{x}{2}$ for $0 \leq x \leq 1$, we have

$$\begin{aligned} g_a(p) &\geq -\frac{2a}{p} - \frac{k+d}{(p+1)(p+k+d+1)} + \frac{k^2-d(p-k+1)}{2(p-k+1)(p+k+d+1)} \\ &\geq -\frac{2a}{p} - \frac{k+d}{(p+1)(p+k+d+1)} + \frac{k^2-dp}{2p(p+k+d+1)}. \end{aligned}$$

When $p \leq \frac{k^2}{d+24a}$, we have $k^2 - dp \geq 24ap$. Then

$$\frac{k^2-dp}{4p(p+k+d+1)} \geq \frac{24ap}{4p(p+k+d+1)} \geq \frac{6ap}{(p+1)(p+k+d+1)} \geq \frac{k+d}{(p+1)(p+k+d+1)}$$

when k is sufficiently large. Also we have

$$\frac{k^2-dp}{4r(p+k+d+1)} \geq \frac{24ap}{4r(p+k+d+1)} \geq \frac{6a}{p+k+d+1} \geq \frac{2a}{p}$$

when k is sufficiently large.

Combining all the arguments above, we conclude that $g_a(p) \geq 0$ and $f_a'(p) \geq 0$ when $k \leq p \leq \frac{k^2}{d+24a}$. Then when $k \leq p \leq \frac{k^2}{d+24a}$, we have

$$f_a(p) \leq f_a\left(\frac{k^2}{d+24a}\right). \tag{4.51}$$

When $p \geq \frac{k^2}{d+24a}$, we have

$$\begin{aligned} f_a(p) &= \frac{p^{-a}(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}(p+k+1+d)^{\frac{p+k+d}{2}}} \\ &= \frac{p^{-a}(p+1)^{p+\frac{1}{2}}}{((p+1)^2-k^2+d(p-k+1))^{\frac{p-k+1}{2}}(p+k+1+d)^{\frac{2k+d-1}{2}}} \\ &= \frac{p^{-a}(p+1)^{-\frac{d}{2}}}{\left(1-\frac{k^2-d(p-k+1)}{(p+1)^2}\right)^{\frac{p-k+1}{2}}\left(1+\frac{k+d}{p+1}\right)^{\frac{2k+d-1}{2}}} \\ &\leq \frac{p^{-a-\frac{d}{2}}}{\left(1-\frac{k^2-d(p-k+1)}{(p+1)^2}\right)^{\frac{p-k+1}{2}}}. \end{aligned}$$

If $k^2-d(p-k+1)<0$, $\left(1-\frac{k^2-d(p-k+1)}{(p+1)^2}\right)^{\frac{p-k+1}{2}} \geq 1$. If $k^2-d(p-k+1)\geq 0$, i.e., $p \leq \frac{k^2+dk-d}{d}$, for sufficiently large k, we have

$$\begin{aligned} \left(1-\frac{k^2-d(p-k+1)}{(p+1)^2}\right)^{\frac{p-k+1}{2}} &\geq \left(1-\frac{k^2-d(\frac{k^2}{d+24a}-k+1)}{(\frac{k^2}{d+24a}+1)^2}\right)^{\frac{\frac{k^2+dk-d}{d}-k+1}{2}} \\ &\geq \left(1-\frac{48a(d+24a)}{k^2}\right)^{\frac{k^2}{2d}} \\ &\geq e^{-\frac{k^2}{2d}\frac{48a(d+24a)}{k^2}} = e^{-\frac{48a(d+24a)}{2d}}, \end{aligned}$$

which is a constant independent of k. Then for $p \geq \frac{k^2}{d+24a}$, we have

$$f_a(p) \leq e^{\frac{48a(d+24a)}{2d}} p^{-a-\frac{d}{2}}. \tag{4.52}$$

Finally we have

$$\begin{aligned}
\overline{\lambda_k} &= 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2C}{C_1^2}\sum_{\substack{p\geq k\\ p-k \text{ is even}}} f_a(p)\\
&\leq O\left(\sum_{\substack{k\leq p\leq \frac{k^2}{d+24a}\\ p-k \text{ is even}}} f_a(p) + \sum_{\substack{p\geq \frac{k^2}{d+24a}\\ p-k \text{ is even}}} f_a(p)\right)\\
&\leq O\left(\left(\frac{k^2}{d+24a}-k+1\right)f_a\left(\frac{k^2}{d+24a}\right) + \sum_{\substack{p\geq \frac{k^2}{d+24a}\\ p-k \text{ is even}}} e^{\frac{48a(d+24a)}{2d}}p^{-a-\frac{d}{2}}\right)\\
&\leq O\Bigg(\left(\frac{k^2}{d+24a}-k+1\right)e^{\frac{48a(d+24a)}{2d}}\left(\frac{k^2}{d+24a}\right)^{-a-\frac{d}{2}}\\
&\qquad + e^{\frac{48a(d+24a)}{2d}}\frac{1}{a+\frac{d}{2}-1}\left(\frac{k^2}{d+24a}-1\right)^{1-a-\frac{d}{2}}\Bigg)\\
&= O(k^{-d-2a+2}).
\end{aligned}$$

Next we prove $\overline{\lambda_k} = \Omega(k^{-d-2a+2})$. Since c_p are nonnegative and $c_p = \Theta(p^{-a})$, we have that $c_p \geq C'p^{-a}$ for some constant C'. Then we have

$$\overline{\lambda_k} \geq \frac{\pi^{d/2}}{2^{k-1}}\sum_{\substack{p\geq k\\ p-k \text{ is even}}} C'p^{-a}\frac{\Gamma(p+1)\Gamma(\frac{p-k+1}{2})}{\Gamma(p-k+1)\Gamma(\frac{p-k+1}{2}+k+d/2)}. \tag{4.53}$$

According to Stirling's formula (4.45) and (4.46), using the similar argument as (4.47) we

have

$$\overline{\lambda_k} \geq \frac{\pi^{d/2}}{2^{k-1}} \frac{C_1^2}{C_2^2} \sum_{\substack{p \geq k \\ p-k \text{ is even}}} C' p^{-a} \frac{\sqrt{\frac{2\pi}{p+1}} \left(\frac{p+1}{e}\right)^{p+1} \sqrt{\frac{2\pi}{\frac{p-k+1}{2}}} \left(\frac{\frac{p-k+1}{2}}{e}\right)^{\frac{p-k+1}{2}}}{\sqrt{\frac{2\pi}{p-k+1}} \left(\frac{p-k+1}{e}\right)^{p-k+1} \sqrt{\frac{2\pi}{\frac{p-k+1}{2}+k+d/2}} \left(\frac{\frac{p-k+1}{2}+k+d/2}{e}\right)^{\frac{p-k+1}{2}+k+d/2}} \tag{4.54}$$

$$= 2\pi^{d/2} \frac{2^{\frac{d}{2}} e^{\frac{d}{2}} C_1^2 C'}{C_2^2} \sum_{\substack{p \geq k \\ p-k \text{ is even}}} \frac{p^{-a}(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}} (p+k+1+d)^{\frac{p+k+d}{2}}} \tag{4.55}$$

$$\geq 2\pi^{d/2} \frac{2^{\frac{d}{2}} e^{\frac{d}{2}} C_1^2 C'}{C_2^2} \sum_{\substack{p \geq k^2 \\ p-k \text{ is even}}} f_a(p), \tag{4.56}$$

where $f_a(p)$ is defined in (4.48). When $p \geq k^2$, we have

$$\begin{aligned} f_a(p) &= \frac{p^{-a}(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}} (p+k+1+d)^{\frac{p+k+d}{2}}} \\ &= \frac{p^{-a}(p+1)^{p+\frac{1}{2}}}{((p+1)^2 - k^2 + d(p-k+1))^{\frac{p-k+1}{2}} (p+k+1+d)^{\frac{2k+d-1}{2}}} \\ &\geq \frac{(p+1)^{-a-\frac{d}{2}}}{\left(1 - \frac{k^2 - d(p-k+1)}{(p+1)^2}\right)^{\frac{p-k+1}{2}} \left(1 + \frac{k+d}{p+1}\right)^{\frac{2k+d-1}{2}}} \end{aligned}$$

For sufficiently large k, $k^2 - d(p-k+1) < 0$. Then we have

$$\begin{aligned} \left(1 - \frac{k^2 - d(p-k+1)}{(p+1)^2}\right)^{\frac{p-k+1}{2}} &= \left(1 - \frac{k^2 - d(p-k+1)}{(p+1)^2}\right)^{\frac{-(p+1)^2}{k^2 - d(p-k+1)} \cdot \frac{-k^2 + d(p-k+1)}{(p+1)^2} \cdot \frac{p-k+1}{2}} \\ &\leq e^{\frac{-k^2 + d(p-k+1)}{(p+1)^2} \cdot \frac{p-k+1}{2}} \\ &\leq e^{\frac{dp^2}{2p^2}} = e^{\frac{d}{2}} \end{aligned}$$

which is a constant independent of k. Also, for sufficiently large k, we have

$$\begin{aligned}\left(1+\frac{k+d}{p+1}\right)^{\frac{2k+d-1}{2}} &= \left(1+\frac{k+d}{p+1}\right)^{\frac{p+1}{k+d}\frac{k+d}{p+1}\frac{2k+d-1}{2}} \\ &\le e^{\frac{k+d}{p+1}\frac{2k+d-1}{2}} \\ &\le e^{\frac{3k^2}{2r}} = e^{\frac{3}{2}}\end{aligned}$$

Then for $p \ge k^2$, we have $f_a(p) \ge e^{-\frac{d}{2}-\frac{3}{2}}(p+1)^{-a-\frac{d}{2}}$.

Finally we have

$$\overline{\lambda_k} \ge 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_1^2C'}{C_2^2} \sum_{\substack{p\ge k^2 \\ p-k \text{ is even}}} f_a(p) \tag{4.57}$$

$$\ge 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_1^2C'}{C_2^2} \sum_{\substack{p\ge k^2 \\ p-k \text{ is even}}} e^{-\frac{d}{2}-\frac{3}{2}}(p+1)^{-a-\frac{d}{2}} \tag{4.58}$$

$$\ge 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_1^2C'}{C_2^2}e^{-\frac{d}{2}-\frac{3}{2}}\frac{1}{2(a+\frac{d}{2}-1)}(k^2+2)^{1-a-\frac{d}{2}} \tag{4.59}$$

$$= \Omega(k^{-d-2a+2}). \tag{4.60}$$

Overall, we have $\overline{\lambda_k} = \Theta(k^{-d-2a+2})$. □

Proof of Corollary 4.C.2, part 2. It is easy to verify that $\overline{\lambda_k} = 0$ when k is even because $c_p = 0$ when $p \ge k$ and $p-k$ is even. When k is odd, the proof of Theorem 94 still applies. □

Proof of Corollary 4.C.2, part 3. Since $c_p = \mathcal{O}\big(\exp\big(-a\sqrt{p}\big)\big)$, we have that $c_p \le Ce^{-a\sqrt{p}}$ for some constant C. Similar to (4.47), we have

$$\overline{\lambda_k} \le 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2C}{C_1^2} \sum_{\substack{p\ge k \\ p-k \text{ is even}}} \frac{e^{-a\sqrt{p}}(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}(p+k+1+d)^{\frac{p+k+d}{2}}}. \tag{4.61}$$

Use the definition in (4.48) and let $a = 0$, we have

$$f_0(p) = \frac{(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}(p+k+1+d)^{\frac{p+k+d}{2}}}. \tag{4.62}$$

Then according to (4.51) and (4.52), for sufficiently large k, we have $f_0(p) \le f_0\left(\frac{k^2}{d}\right)$ when $k \le p \le \frac{k^2}{d}$ and $f_0(p) \le C_3 p^{-\frac{d}{2}}$ for some constant C_3 when $p \ge \frac{k^2}{d}$. Then when $k \le p \le \frac{k^2}{d}$, we have $f_0(p) \le f_0\left(\frac{k^2}{d}\right) \le C_3\left(\frac{k^2}{d}\right)^{-\frac{d}{2}}$. When $p \ge \frac{k^2}{d}$, we have $f_0(p) \le C_3 p^{-\frac{d}{2}} \le C_3\left(\frac{k^2}{d}\right)^{-\frac{d}{2}}$. Overall, for all $p \ge k$, we have

$$f_0(p) \le C_3\left(\frac{k^2}{d}\right)^{-\frac{d}{2}}. \tag{4.63}$$

Then we have

$$\begin{aligned}
\overline{\lambda_k} &\le 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2 C}{C_1^2} \sum_{\substack{p\ge k \\ p-k \text{ is even}}} e^{-a\sqrt{p}} f_0(p) && (4.64)\\
&\le 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2 C_3 C}{C_1^2}\left(\frac{k^2}{d}\right)^{-\frac{d}{2}} \sum_{\substack{p\ge k \\ p-k \text{ is even}}} e^{-a\sqrt{p}} && (4.65)\\
&\le 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2 C_3 C}{C_1^2}\left(\frac{k^2}{d}\right)^{-\frac{d}{2}} \frac{2e^{-a\sqrt{k-1}}(a\sqrt{k-1}+1)}{a^2} && (4.66)\\
&= \mathcal{O}\left(k^{-d+1/2}\exp\left(-a\sqrt{k}\right)\right) && (4.67)
\end{aligned}$$

□

Proof of Corollary 4.C.2, part 4. Since $c_p = \Theta(p^{1/2}a^{-p})$, we have that $c_p \le Cp^{1/2}a^{-p}$ for some constant C. Similar to (4.47), we have

$$\overline{\lambda_k} \le 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2 C}{C_1^2} \sum_{\substack{p\ge k \\ p-k \text{ is even}}} \frac{p^{1/2}a^{-p}(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}(p+k+1+d)^{\frac{p+k+d}{2}}}. \tag{4.68}$$

Use the definition in (4.48) and let $a = 0$, we have

$$f_0(p) = \frac{(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}(p+k+1+d)^{\frac{p+k+d}{2}}}. \tag{4.69}$$

Then according to (4.51) and (4.52), for sufficiently large k, we have $f_0(p) \leq f_0\Big(\frac{k^2}{d}\Big)$ when $k \leq p \leq \frac{k^2}{d}$ and $f_0(p) \leq C_3 p^{-\frac{d}{2}}$ for some constant C_3 when $p \geq \frac{k^2}{d}$. Then when $k \leq p \leq \frac{k^2}{d}$, we have $p^{1/2} f_0(p) \leq p^{1/2} f_0\Big(\frac{k^2}{d}\Big) \leq C_3 \Big(\frac{k^2}{d}\Big)^{1/2} \Big(\frac{k^2}{d}\Big)^{-\frac{d}{2}}$. When $p \geq \frac{k^2}{d}$, we have $p^{1/2} f_0(p) \leq C_3 p^{1/2} p^{-\frac{d}{2}} \leq C_3 \Big(\frac{k^2}{d}\Big)^{-\frac{d}{2}+\frac{1}{2}}$. Overall, for all $p \geq k$, we have

$$p^{1/2} f_0(p) \leq C_3 \left(\frac{k^2}{d}\right)^{-\frac{d}{2}+\frac{1}{2}}. \tag{4.70}$$

Then we have

$$\begin{aligned}
\overline{\lambda_k} &\leq 2\pi^{d/2} \frac{2^{\frac{d}{2}} e^{\frac{d}{2}} C_2^2 C}{C_1^2} \sum_{\substack{p \geq k \\ p-k \text{ is even}}} p^{1/2} a^{-p} f_0(p) && (4.71)\\
&\leq 2\pi^{d/2} \frac{2^{\frac{d}{2}} e^{\frac{d}{2}} C_2^2 C_3 C}{C_1^2} \left(\frac{k^2}{d}\right)^{-\frac{d}{2}+\frac{1}{2}} \sum_{\substack{p \geq k \\ p-k \text{ is even}}} a^{-p} && (4.72)\\
&\leq 2\pi^{d/2} \frac{2^{\frac{d}{2}} e^{\frac{d}{2}} C_2^2 C_3 C}{C_1^2} \left(\frac{k^2}{d}\right)^{-\frac{d}{2}+\frac{1}{2}} \frac{1}{\log a} a^{-(k-1)} && (4.73)\\
&= \mathcal{O}\big(k^{-d+1} a^{-k}\big). && (4.74)
\end{aligned}$$

On the other hand, since $c_p = \Theta(p^{1/2} a^{-p})$, we have that $c_p \geq C' p^{1/2} a^{-p}$ for some constant C'.

Similar to (4.55), we have

$$\overline{\lambda_k} \geq 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_1^2C'}{C_2^2}\sum_{\substack{p\geq k\\ p-k \text{ is even}}}\frac{p^{1/2}a^{-p}(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}(p+k+1+d)^{\frac{p+k+d}{2}}} \tag{4.75}$$

$$\geq 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_1^2C'}{C_2^2}\frac{k^{1/2}a^{-k}(k+1)^{k+\frac{1}{2}}}{(k-k+1)^{\frac{k-k+1}{2}}(k+k+1+d)^{\frac{k+k+d}{2}}} \tag{4.76}$$

$$= \Omega\left(\frac{k^{-d/2+1}a^{-k}(k+1)^k}{(k+k+1+d)^k}\right). \tag{4.77}$$

Since $(k+1)^k = k^k(1+1/k)^k = \Theta(k^k)$. Similarly, $(k+k+1+d)^k = \Theta((2k)^k)$. Then we have

$$\overline{\lambda_k} = \Omega\left(\frac{k^{-d/2+1}a^{-k}(k+1)^k}{(k+k+1+d)^k}\right) \tag{4.78}$$

$$= \Omega\left(\frac{k^{-d/2+1}a^{-k}k^k}{(2k)^k}\right) \tag{4.79}$$

$$= \Omega\big(k^{-d/2+1}2^{-k}a^{-k}\big). \tag{4.80}$$

□

CHAPTER 5

Conclusion

In previous chapters we analyze the generalization of wide neural networks through linearization and kernel learning. This approach overcomes the drawback of the traditional statistical learning theory and can be applied to overparametrized networks. Also the linearization and kernel learning approach explains the phenomenon observed by [ZBH21] that deep networks can fit randoms labels while still have good generalization performance. In Chapter 2 we show that under gradient descent the wide neural networks would fit the training data by a smooth function, thus the networks can fit random labels while also generalize well if the target function is smooth. In Chapter 3 we show that the target function is learnable if it lies in the span of the eigenfunctions with positive eigenvalues. Thus we answer the question why wide neural networks learns a function and when a function is learnable.

Our results also have both theoretical and practical potential applications. In Chapter 2 we show that training a wide neural network by gradient descent is equivalent to fitting the training data by some kind of splines. In reverse, if we want to fit some splines, for example, fit a surface from a point cloud, we can use wide neural networks to do this task. This method has already been explored in [WTB21]. In Chapter 3 we show that the decay rate of the generalization error for kernel learning using the NTK can be characterized by the decay rate of the NTK spectrum. This result could explain the spectral bias [RBA19b] to a certain extent.

However, there is still a long way to go in understanding the generalization of deep learning. We raise several issues about the liberalization and kernel learning approach and discuss possible future works in the following.

First, whether kernel learning can explain the performance of deep networks is still unclear. In many practical tasks, the state-of-the-art kernel method cannot achieve the same performance as the state-of-the-art deep learning method. [ADH19b] showed that the performance of kernel learning with Convolutional NTK (CNTK) is 6% lower than the performance of the corresponding finite deep net architecture. Many people believes that kernel learning performs worse that deep learning because deep networks has the ability to learn good feature representations while kernel learning uses fixed features. There are some current works showing when and why the neural networks outperform the kernel method [GMM20, AL19]. Nevertheless, the kernel learning with NTK have comparable performance to the deep networks and understanding why the NTK is better than traditional kernels is still an important approach to understand deep learning.

Second, our method in Chapter 2 only applies to shallow feedforward networks. Whether we can generalize the result to more complicated architectures such as convolutional networks remains an open question and requires future work. Our method in Chapter 3 applies to any kernel, thus we can study convolutional networks by studying CNTK. However, the spectrum of CNTK is not well studied and it would be interesting to show the spectrum of CNTK on the natural image dataset in the future.

Third, it is well-known that the kernel spectrum is highly related to the data distribution, but what is the exact relation remains an open question. In Chapter 4 the NTK power series gives a bit of hints of the relation between the data and the NTK, but a better understanding of that requires more future works. [PZA21] shows that natural image data has a low-dimensional structure despite the high ambient dimension. Neural networks are able to use this low-dimensional structure to overcome the curse of dimensionality [Bac17]. It would be interesting to explore how to use the low-dimensional structure to analyze the spectrum of NTK and CNTK, which has a direct impact to generalization of deep networks according to Chapter 3.

References

[ADH19a] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. "Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks." In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 322–332. PMLR, 2019.

[ADH19b] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. "On exact computation with an infinitely wide neural net." In *Advances in Neural Information Processing Systems*, volume 32, pp. 8139–8148, 2019.

[ADH19c] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. "On Exact Computation with an Infinitely Wide Neural Net." In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[AFS92] Shun'ichi Amari, Naotake Fujita, and Shigeru Shinomoto. "Four types of learning curves." *Neural Computation*, **4**(4):605–618, 1992.

[AL19] Zeyuan Allen-Zhu and Yuanzhi Li. "What can resnet learn efficiently, going beyond kernels?" *Advances in Neural Information Processing Systems*, **32**, 2019.

[ALS19a] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. "A Convergence Theory for Deep Learning via Over-Parameterization." In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR, 2019.

[ALS19b] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. "A Convergence Theory for Deep Learning via Over-Parameterization." In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[AM93] Shun'ichi Amari and Noboru Murata. "Statistical theory of learning curves under entropic loss criterion." *Neural Computation*, **5**(1):140–153, 1993.

[ANW67] J. H. Ahlberg, Edwin N. Nilson, and J. L. Walsh. *The Theory of Splines and Their Applications*. ISSN. Elsevier Science, 1967.

[AS96a] Felix Abramovich and David M. Steinberg. "Improved inference in nonparametric regression using L_k-smoothing splines." *Journal of Statistical Planning and Inference*, **49**(3):327 – 341, 1996.

[AS96b] Felix Abramovich and David M Steinberg. "Improved inference in nonparametric regression using Lk-smoothing splines." *Journal of Statistical Planning and Inference*, **49**(3):327–341, 1996.

[Bac17] Francis Bach. "Breaking the curse of dimensionality with convex neural networks." *The Journal of Machine Learning Research*, **18**(1):629–681, 2017.

[Bar93] Andrew R Barron. "Universal approximation bounds for superpositions of a sigmoidal function." *IEEE Transactions on Information theory*, **39**(3):930–945, 1993.

[Bar98] Andrew R. Barron. "Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems." In Dawid A. Bernardo J., Berger J. and Smith A., editors, *Bayesian statistics*, volume 6, pp. 27–52. Oxford University Press, 1998.

[BB21] Alberto Bietti and Francis Bach. "Deep Equals Shallow for ReLU Networks in Kernel Regimes." In *International Conference on Learning Representations*, 2021.

[BCP20] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. "Spectrum dependent learning curves in kernel regression and wide neural networks." In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1024–1034, 2020.

[BDK21] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. "Explaining neural scaling laws." *arXiv preprint arXiv:2102.06701*, 2021.

[BGG20] Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. "Frequency Bias in Neural Networks for Input of Non-Uniform Density." In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 685–694. PMLR, 2020.

[BGL21] Aristide Baratin, Thomas George, César Laurent, R Devon Hjelm, Guillaume Lajoie, Pascal Vincent, and Simon Lacoste-Julien. "Implicit Regularization via Neural Feature Alignment." In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2269–2277. PMLR, 13–15 Apr 2021.

[BHM21] Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon van Handel, and Amir Yehudayoff. "A theory of universal learning." In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 532–541, 2021.

[Bis95] Christopher Bishop. "Regularization and Complexity Control in Feed-forward Networks." In *Proceedings International Conference on Artificial Neural Networks ICANN'95*, volume 1, pp. 141–148. EC2 et Cie, January 1995.

[BJK19] Ronen Basri, David W. Jacobs, Yoni Kasten, and Shira Kritchman. "The Convergence Rate of Neural Networks for Learned Functions of Different Frequencies." In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32*, pp. 4763–4772, 2019.

[BM18] Gilles Blanchard and Nicole Mücke. "Optimal rates for regularization of statistical inverse learning problems." *Foundations of Computational Mathematics*, **18**(4):971–1013, 2018.

[BM19] Alberto Bietti and Julien Mairal. "On the Inductive Bias of Neural Tangent Kernels." In *Advances in Neural Information Processing Systems*, volume 32, pp. 12873–12884, 2019.

[BM22a] Benjamin Bowman and Guido Montúfar. "Implicit Bias of MSE Gradient Optimization in Underparameterized Neural Networks." In *International Conference on Learning Representations*, 2022.

[BM22b] Benjamin Bowman and Guido Montúfar. "Spectral Bias Outside the Training Set for Deep Networks in the Kernel Regime." *CoRR*, **abs/2206.02927**, 2022.

[BMM18] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. "To understand deep learning we need to understand kernel learning." In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 541–549, 2018.

[Bra06] Mikio L. Braun. "Accurate error bounds for the eigenvalues of the kernel matrix." *The Journal of Machine Learning Research*, **7**:2303–2328, 2006.

[BVB21] Alberto Bietti, Luca Venturi, and Joan Bruna. "On the Sample Complexity of Learning with Geometric Stability." *arXiv preprint arXiv:2106.07148*, 2021.

[CB20] Lénaïc Chizat and Francis Bach. "Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss." In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1305–1338. PMLR, 09–12 Jul 2020.

[CBP21] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. "Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks." *Nature communications*, **12**(1):1–12, 2021.

[CD07] Andrea Caponnetto and Ernesto De Vito. "Optimal rates for the regularized least-squares algorithm." *Foundations of Computational Mathematics*, **7**(3):331–368, 2007.

[CFW21] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. "Towards Understanding the Spectral Bias of Deep Learning." In Zhi-Hua Zhou, editor,

Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pp. 2205–2211. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[CG19] Yuan Cao and Quanquan Gu. "Generalization bounds of stochastic gradient descent for wide and deep neural networks." *Advances in neural information processing systems*, **32**, 2019.

[CLK21] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. "Generalization Error Rates in Kernel Regression: The Crossover from the Noiseless to Noisy Regime." *arXiv preprint arXiv:2105.15004*, 2021.

[COB19] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. "On lazy training in differentiable programming." In *Advances in Neural Information Processing Systems*, pp. 2933–2943, 2019.

[CPB19] Niladri Chatterji, Aldo Pacchiano, and Peter Bartlett. "Online learning with kernel losses." In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 971–980, 2019.

[CS09] Youngmin Cho and Lawrence K Saul. "Kernel methods for deep learning." In *Advances in Neural Information Processing Systems*, volume 22, pp. 342–350, 2009.

[Dan17] Amit Daniely. "SGD Learns the Conjugate Kernel Class of the Network." In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Dav21] Tom Davis. "A GENERAL EXPRESSION FOR HERMITE EXPANSIONS WITH APPLICATIONS." 2021.

[DFS16a] Amit Daniely, Roy Frostig, and Yoram Singer. "Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity." In *Advances In Neural Information Processing Systems*, volume 29, pp. 2253–2261, 2016.

[DFS16b] Amit Daniely, Roy Frostig, and Yoram Singer. "Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity." In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[dHR18] Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. "Gaussian Process Behaviour in Wide Deep Neural Networks." In *International Conference on Learning Representations*, 2018.

[DLL19] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. "Gradient Descent Finds Global Minima of Deep Neural Networks." In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685. PMLR, 2019.

[Dom20] Pedro Domingos. "Every model learned by gradient descent is approximately a kernel machine." *arXiv preprint arXiv:2012.00152*, 2020.

[DPB17] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. "Sharp Minima Can Generalize For Deep Nets." In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[Du 08] Wilna Du Toit. *Radial basis function interpolation*. PhD thesis, Stellenbosch: Stellenbosch University, 2008.

[DZP19] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. "Gradient Descent Provably Optimizes Over-parameterized Neural Networks." In *International Conference on Learning Representations*, 2019.

[EL06] PPB Eggermont and VN LaRiccia. "Uniform error bounds for smoothing splines." *Lecture Notes-Monograph Series*, pp. 220–237, 2006.

[Fol99a] G. B. Folland. *Real analysis: Modern techniques and their applications*. Wiley, New York, 1999.

[Fol99b] Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

[FS20] Simon Fischer and Ingo Steinwart. "Sobolev Norm Learning Rates for Regularized Least-Squares Algorithms." *Journal of Machine Learning Research*, **21**:1–38, 2020.

[FW20] Zhou Fan and Zhichao Wang. "Spectra of the Conjugate Kernel and Neural Tangent Kernel for linear-width neural networks." In *Advances in Neural Information Processing Systems*, volume 33, pp. 7710–7721. Curran Associates, Inc., 2020.

[GB10] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256. PMLR, 2010.

[Ger01] G German. "Smoothing and non-parametric regression." *International Journal of Systems Science*, 2001.

[GHR18] Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. "Gaussian Process Behaviour in Wide Deep Neural Networks." In *International Conference on Learning Representations*, 2018.

[GLS18a] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. "Characterizing Implicit Bias in Terms of Optimization Geometry." In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[GLS18b] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. "Implicit Bias of Gradient Descent on Linear Convolutional Networks." In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pp. 9461–9471. Curran Associates, Inc., 2018.

[GMM20] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "When do neural networks outperform kernel methods?" *Advances in Neural Information Processing Systems*, **33**:14820–14830, 2020.

[GRA19] Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. "Deep Convolutional Networks as shallow Gaussian Processes." In *International Conference on Learning Representations*, 2019.

[GYK20] Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Basri Ronen. "On the Similarity between the Laplace and Neural Tangent Kernels." In *Advances in Neural Information Processing Systems*, volume 33, pp. 1451–1461. Curran Associates, Inc., 2020.

[HKS96] David Haussler, Michael Kearns, H Sebastian Seung, and Naftali Tishby. "Rigorous learning curve bounds from statistical mechanics." *Machine Learning*, **25**(2-3):195–236, 1996.

[HM76] Charles A Hall and W Weston Meyer. "Optimal error bounds for cubic spline interpolation." *Journal of Approximation Theory*, **16**(2):105–122, 1976.

[HNA17] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. "Deep learning scaling is predictable, empirically." *arXiv preprint arXiv:1712.00409*, 2017.

[HO97] David Haussler and Manfred Opper. "Mutual information, metric entropy and cumulative relative entropy risk." *The Annals of Statistics*, **25**(6):2451–2492, 1997.

[HTW19] Jakob Heiss, Josef Teichmann, and Hanna Wutte. "How implicit regularization of Neural Networks affects the learned function - Part I." *arXiv preprint arXiv:1911.02903*, 2019.

[HZL22] Insu Han, Amir Zandieh, Jaehoon Lee, Roman Novak, Lechao Xiao, and Amin Karbasi. "Fast Neural Kernel Embeddings for General Activations." In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[HZR15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015.

[JBM22] Hui Jin, Pradeep Kr. Banerjee, and Guido Montufar. "Learning Curves for Gaussian Process Regression with Power-Law Priors and Targets." In *International Conference on Learning Representations*, 2022.

[JCO19] Kwang-Sung Jun, Ashok Cutkosky, and Francesco Orabona. "Kernel truncated randomized ridge regression: Optimal rates and low noise acceleration." *Advances in Neural Information Processing Systems*, **32**:15358–15367, 2019.

[JGH18a] Arthur Jacot, Franck Gabriel, and Clement Hongler. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks." In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pp. 8571–8580. Curran Associates, Inc., 2018.

[JGH18b] Arthur Jacot, Franck Gabriel, and Clement Hongler. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks." In *Advances in Neural Information Processing Systems*, volume 31, pp. 8571–8580, 2018.

[JGH18c] Arthur Jacot, Franck Gabriel, and Clement Hongler. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks." In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[JM20] Hui Jin and Guido Montúfar. "Implicit bias of gradient descent for mean squared error regression with wide neural networks." *arXiv preprint arXiv:2006.07356*, 2020.

[JT19] Ziwei Ji and Matus Telgarsky. "The implicit bias of gradient descent on non-separable data." In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1772–1798, Phoenix, USA, 25–28 Jun 2019. PMLR.

[KAA20] Ryo Karakida, Shotaro Akaho, and Shun ichi Amari. "Universal statistics of Fisher information in deep neural networks: mean field approach." *Journal of Statistical Mechanics: Theory and Experiment*, **2020**(12):124005, 2020.

[Kaz56] Donat K. Kazarinoff. "On Wallis' formula." *Edinburgh Mathematical Notes*, **40**:19–21, 1956.

[KHK19] Kenji Kawaguchi, Jiaoyang Huang, and Leslie Pack Kaelbling. "Effect of depth and width on local minima in deep learning." *Neural computation*, **31**(7):1462–1498, 2019.

[KHS18] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. "Gaussian processes and kernel methods: A review on connections and equivalences." *arXiv preprint arXiv:1807.02582*, 2018.

[KMN17] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. "On large-batch training for deep learning: Generalization gap and sharp minima." In *International Conference on Learning Representations*, 2017.

[Kwa17] Mateusz Kwaśnicki. "Ten equivalent definitions of the fractional Laplace operator." *Fractional Calculus and Applied Analysis*, **20**(1):7–51, 2017.

[LBN18] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. "Deep Neural Networks as Gaussian Processes." In *International Conference on Learning Representations*, 2018.

[LBO12] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. *Efficient BackProp*, pp. 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[LG10] Ziyue Liu and Wensheng Guo. "Data driven adaptive spline smoothing." *Statistica Sinica*, pp. 1143–1163, 2010.

[LG15] Loic Le Gratiet and Josselin Garnier. "Asymptotic analysis of the learning curve for Gaussian process regression." *Machine Learning*, **98**(3):407–433, 2015.

[LLC18] Cosme Louart, Zhenyu Liao, and Romain Couillet. "A RANDOM MATRIX APPROACH TO NEURAL NETWORKS." *The Annals of Applied Probability*, **28**(2):1190–1248, 2018.

[LSP18] Jaehoon Lee, Jascha Sohl-Dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. "Deep Neural Networks as Gaussian Processes." In *International Conference on Learning Representations*, 2018.

[LSP20] Jaehoon Lee, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. "Finite Versus Infinite Neural Networks: an Empirical Study." In *Advances in Neural Information Processing Systems*, volume 33, pp. 15156–15172, 2020.

[LVM19] Marco Loog, Tom Viering, and Alexander Mey. "Minimizers of the empirical risk and risk monotonicity." In *Advances in Neural Information Processing Systems*, volume 32, pp. 7478–7487, 2019.

[LXS19a] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent." In *Advances in Neural Information Processing Systems*, volume 32, pp. 8572–8583, 2019.

[LXS19b] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent." In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pp. 8572–8583. Curran Associates, Inc., 2019.

[LXS19c] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent." In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[MAT22] M. Murray, V. Abrol, and J. Tanner. "Activation function design for deep networks: linearity and effective initialisation." *Applied and Computational Harmonic Analysis*, **59**:117–154, 2022. Special Issue on Harmonic Analysis and Machine Learning.

[MBG18] Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. "Gradient descent quantizes ReLU network features." *arXiv preprint arXiv:1803.08367*, 2018.

[MJB23] Michael Murray, Hui Jin, Benjamin Bowman, and Guido Montufar. "Characterizing the spectrum of the NTK via a power series expansion." In *Submitted to The Eleventh International Conference on Learning Representations*, 2023. under review.

[MM16] Dmytro Mishkin and Jiri Matas. "All you need is a good init." In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, Conference Track Proceedings*, 2016.

[MO01a] Dörthe Malzahn and Manfred Opper. "Learning curves for Gaussian processes models: Fluctuations and universality." In *International Conference on Artificial Neural Networks*, pp. 271–276, 2001.

[MO01b] Dörthe Malzahn and Manfred Opper. "Learning curves for Gaussian processes regression: A framework for good approximations." In *Advances in Neural Information Processing Systems*, volume 13, pp. 273–279, 2001.

[MPC14] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. "On the number of linear regions of deep neural networks." *Advances in neural information processing systems*, **27**, 2014.

[MU08] Richard B Melrose and Gunther Uhlmann. *An introduction to microlocal analysis*. Department of Mathematics, Massachusetts Institute of Technology, 2008.

[Nas73] C. Nasim. "The Solution of an Integral Equation." *Proceedings of the American Mathematical Society*, **40**(1):95–101, 1973.

[Nea96a] Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, Berlin, Heidelberg, 1996.

[Nea96b] Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, Berlin, Heidelberg, 1996.

[Nea96c] Radford M Neal. "Priors for infinite networks." In *Bayesian Learning for Neural Networks*, pp. 29–53. Springer, 1996.

[NH17] Quynh Nguyen and Matthias Hein. "The loss surface of deep and wide neural networks." In *International conference on machine learning*, pp. 2603–2612. PMLR, 2017.

[NM20] Quynh Nguyen and Marco Mondelli. "Global Convergence of Deep Networks with One Wide Layer Followed by Pyramidal Topology." In *Advances in Neural Information Processing Systems*, volume 33, pp. 11961–11972. Curran Associates, Inc., 2020.

[NMM21] Quynh Nguyen, Marco Mondelli, and Guido Montúfar. "Tight Bounds on the Smallest Eigenvalue of the Neural Tangent Kernel for Deep ReLU Networks." In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8119–8129. PMLR, 2021.

[NS21] Atsushi Nitanda and Taiji Suzuki. "Optimal Rates for Averaged Stochastic Gradient Descent under Neural Tangent Kernel Regime." In *International Conference on Learning Representations*, 2021.

[NTS15] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. "In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning." In *ICLR (Workshop)*, 2015.

[NTS17] Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. "Geometry of optimization and implicit regularization in deep learning." *arXiv preprint arXiv:1705.03071*, 2017.

[NXB19] Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. "Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes." In *International Conference on Learning Representations*, 2019.

[OD14] Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, 2014.

[OM02] Manfred Opper and D. Malzahn. "A variational approach to learning curves." In *Advances in Neural Information Processing Systems*, volume 14, pp. 463–469, 2002.

[OS19] Samet Oymak and Mahdi Soltanolkotabi. "Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path?" In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4951–4960, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[OS20] Samet Oymak and Mahdi Soltanolkotabi. "Toward Moderate Overparameterization: Global Convergence Guarantees for Training Shallow Neural Networks." *IEEE Journal on Selected Areas in Information Theory*, **1**(1), 2020.

[OT10] Peter Orbanz and Yee Whye Teh. "Bayesian nonparametric models." In *Encyclopedia of Machine Learning*, pp. 81–89. Springer, 2010.

[OV99] Manfred Opper and Francesco Vivarelli. "General bounds on Bayes errors for regression with Gaussian processes." In *Advances in Neural Information Processing Systems*, volume 11, pp. 302–308, 1999.

[OWS20] Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. "A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case." In *International Conference on Learning Representations*, 2020.

[PGM19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

[PLR16a] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. "Exponential expressivity in deep neural networks through transient chaos." *Advances in neural information processing systems*, **29**, 2016.

[PLR16b] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. "Exponential expressivity in deep neural networks through transient chaos." In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[PLR16c] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. "Exponential expressivity in deep neural networks through transient chaos." In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[PN19] Rahul Parhi and Robert D. Nowak. "Minimum "Norm" Neural Networks are Splines." *arXiv preprint arXiv:1910.02333*, 2019.

[PN21] Rahul Parhi and Robert D Nowak. "What Kinds of Functions do Deep Neural Networks Learn? Insights from Variational Spline Theory." *arXiv preprint arXiv:2105.03361*, 2021.

[Pot81] Evelyn Dianne Hatton Potter. *Multivariate polyharmonic spline interpolation.* Iowa State University, 1981.

[PSG20] Abhishek Panigrahi, Abhishek Shetty, and Navin Goyal. "Effect of Activation Functions on the Training of Overparametrized Neural Nets." In *International Conference on Learning Representations*, 2020.

[PSH06] Alexandre Pintore, Paul Speckman, and Chris C. Holmes. "Spatially adaptive smoothing splines." *Biometrika*, **93**(1):113–125, 03 2006.

[PW17] Jeffrey Pennington and Pratik Worah. "Nonlinear random matrix theory for deep learning." In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[PW18] Jeffrey Pennington and Pratik Worah. "The Spectrum of the Fisher Information Matrix of a Single-Hidden-Layer Neural Network." In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[PZA21] Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. "The Intrinsic Dimension of Images and Its Impact on Learning." In *International Conference on Learning Representations*, 2021.

[Rag83] David L Ragozin. "Error bounds for derivative estimates based on spline smoothing of exact or noisy data." *Journal of approximation theory*, **37**(4):335–355, 1983.

[RBA19a] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. "On the spectral bias of neural networks." In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.

[RBA19b] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. "On the Spectral Bias of Neural Networks." In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[RJK19] Basri Ronen, David Jacobs, Yoni Kasten, and Shira Kritchman. "The convergence rate of neural networks for learned functions of different frequencies." *Advances in Neural Information Processing Systems*, **32**:4761–4771, 2019.

[RWW95] Klaus Ritter, Grzegorz W Wasilkowski, and Henryk Woźniakowski. "Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions." *The Annals of Applied Probability*, pp. 518–540, 1995.

[SAD22] James Benjamin Simon, Sajant Anand, and Mike Deweese. "Reverse Engineering the Neural Tangent Kernel." In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20215–20231. PMLR, 17–23 Jul 2022.

[SBR10] Curtis B Storlie, Howard D Bondell, and Brian J Reich. "A locally adaptive penalty for estimation of functions with varying roughness." *Journal of Computational and Graphical Statistics*, **19**(3):569–589, 2010.

[SDP20] Justin Sahs, Aneel Damaraju, Ryan Pyle, Onur Tavaslioglu, Josue Ortega Caro, Hao Yang Lu, and Ankit Patel. "A Functional Characterization of Randomly Initialized Gradient Descent in Deep ReLU Networks.", 2020.

[Seg19] Karel Segeth. "Multivariate smooth interpolation that employs polyharmonic functions." *Programs and Algorithms of Numerical Mathematics*, pp. 140–148, 2019.

[SES19] Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. "How do infinite width bounded norm networks look in function space?" In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2667–2690, Phoenix, USA, 25–28 Jun 2019. PMLR.

[SGG17] Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. "Deep Information Propagation." In *International Conference on Learning Representations (ICLR)*, 2017.

[SGW20] Stefano Spigler, Mario Geiger, and Matthieu Wyart. "Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm." *Journal of Statistical Mechanics: Theory and Experiment*, **2020**(12):124001, 2020.

[SH02] Peter Sollich and Anason Halees. "Learning curves for Gaussian process regression: Approximations and bounds." *Neural Computation*, **14**(6):1393–1428, 2002.

[SHN18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. "The implicit bias of gradient descent on separable data." *The Journal of Machine Learning Research*, **19**(1):2822–2878, 2018.

[SHS09] Ingo Steinwart, Don R Hush, Clint Scovel, et al. "Optimal Rates for Regularized Least Squares Regression." In *Conference on Learning Theory*, pp. 79–93, 2009.

[SKF08] Matthias W. Seeger, Sham M. Kakade, and Dean P. Foster. "Information consistency of nonparametric Gaussian process methods." *IEEE Transactions on Information Theory*, **54**(5):2376–2382, 2008.

[Sol87] Donald C Solmon. "Asymptotic formulas for the dual Radon transform and applications." *Mathematische Zeitschrift*, **195**(3):321–343, 1987.

[Sol99] Peter Sollich. "Learning curves for Gaussian processes." In *Advances in Neural Information Processing Systems*, volume 11, pp. 344–350, 1999.

[Sol01] Peter Sollich. "Gaussian Process Regression with Mismatched Models." In *Advances in Neural Information Processing Systems*, volume 13, pp. 519–526, 2001.

[SPD20] Justin Sahs, Ryan Pyle, Aneel Damaraju, Josue Ortega Caro, Onur Tavaslioglu, Andy Lu, and Ankit Patel. "Shallow Univariate ReLU Networks as Splines: Initialization, Loss Surface, Hessian, & Gradient Flow Dynamics.", 2020.

[SS12] Ingo Steinwart and Clint Scovel. "Mercer's Theorem on General Domains: On the Interaction between Measures, Kernels, and RKHSs." *Constructive Approximation*, **35**:363–417, 2012.

[Ste12] Michael L Stein. *Interpolation of spatial data: Some theory for kriging*. Springer Science & Business Media, 2012.

[Tro12] Joel A Tropp. "User-friendly tail bounds for sums of random matrices." *Foundations of computational mathematics*, **12**(4):389–434, 2012.

[Val84] Leslie G Valiant. "A theory of the learnable." *Communications of the ACM*, **27**(11):1134–1142, 1984.

[Val18] Ernesto Araya Valdivia. "Relative concentration bounds for the spectrum of kernel matrices." *arXiv preprint arXiv:1812.02108*, 2018.

[Ver18] Roman Vershynin. "Four lectures on probabilistic methods for data science." In M.W. Mahoney, J.C. Duchi, and A.C. Gilbert, editors, *The Mathematics of Data*, IAS/Park City Mathematics Series, pp. 231–271. American Mathematical Society, 2018.

[VKP21] Sattar Vakili, Kia Khezeli, and Victor Picheny. "On information gain and regret bounds in Gaussian process bandits." In *International Conference on Artificial Intelligence and Statistics*, pp. 82–90, 2021.

[VL21] Tom Viering and Marco Loog. "The Shape of Learning Curves: A Review." *arXiv preprint arXiv:2103.10948*, 2021.

[VML19] Tom Viering, Alexander Mey, and Marco Loog. "Open problem: Monotonicity of learning." In *Conference on Learning Theory*, pp. 3198–3201, 2019.

[VV11] Aad Van Der Vaart and Harry Van Zanten. "Information Rates of Nonparametric Gaussian Process Methods." *Journal of Machine Learning Research*, **12**(6), 2011.

[VY21a] Maksim Velikanov and Dmitry Yarotsky. "Explicit loss asymptotics in the gradient descent training of neural networks." In *Advances in Neural Information Processing Systems*, volume 34, pp. 2570–2582. Curran Associates, Inc., 2021.

[VY21b] Maksim Velikanov and Dmitry Yarotsky. "Universal scaling laws in the gradient descent training of neural networks." *arXiv preprint arXiv:2105.00507*, 2021.

[Wat09] Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory.* Cambridge University Press, 2009.

[WDS13] Xiao Wang, Pang Du, and Jinglai Shen. "Smoothing splines with varying smoothing parameter." *Biometrika*, **100**(4):955–970, 2013.

[Wen04] Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[WGL20] Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. "Kernel and Rich Regimes in Overparametrized Models." In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673. PMLR, 2020.

[Wid63] Harold Widom. "Asymptotic behavior of the eigenvalues of certain integral equations." *Transactions of the American Mathematical Society*, **109**(2):278–295, 1963.

[Wil97] Christopher K.I. Williams. "Computing with Infinite Networks." In *Advances in Neural Information Processing Systems*, volume 9, pp. 295–301, 1997.

[WR06] Christopher K Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning.* MIT press, 2006.

[WTB21] Francis Williams, Matthew Trager, Joan Bruna, and Denis Zorin. "Neural splines: Fitting 3d surfaces with infinitely-wide neural networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9949–9958, 2021.

[WTP19] Francis Williams, Matthew Trager, Daniele Panozzo, Claudio Silva, Denis Zorin, and Joan Bruna. "Gradient dynamics of shallow univariate relu networks." In *Advances in Neural Information Processing Systems*, pp. 8378–8387, 2019.

[WV00] Christopher K.I. Williams and Francesco Vivarelli. "Upper and lower bounds on the learning curve for Gaussian processes." *Machine Learning*, **40**(1):77–102, 2000.

[WZW17] Lei Wu, Zhanxing Zhu, and E Weinan. "Towards Understanding Generalization of Deep Learning: Perspective of Loss Landscapes." *arXiv preprint arXiv:1706.10239*, 2017.

[Yan19] Greg Yang. "Wide Feedforward or Recurrent Neural Networks of Any Architecture are Gaussian Processes." In *Advances in Neural Information Processing Systems*, volume 32, pp. 9951–9960, 2019.

[YS19a] Greg Yang and Hadi Salman. "A fine-grained spectral perspective on neural networks." *arXiv preprint arXiv:1907.10599*, 2019.

[YS19b] Greg Yang and Hadi Salman. "A Fine-Grained Spectral Perspective on Neural Networks.", 2019.

[ZBH17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization." In *International Conference on Learning Representations, ICLR 2017*, 2017.

[ZBH21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning (still) requires rethinking generalization." *Communications of the ACM*, **64**(3):107–115, 2021.

[ZCZ20] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. "Gradient descent optimizes over-parameterized deep ReLU networks." *Machine learning*, **109**(3):467–492, 2020.

[ZXL20] Yaoyu Zhang, Zhi-Qin John Xu, Tao Luo, and Zheng Ma. "A type of generalization error induced by initialization in deep neural networks." In Jianfeng Lu and Rachel Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pp. 144–164, Princeton University, Princeton, NJ, USA, 20–24 Jul 2020. PMLR.